

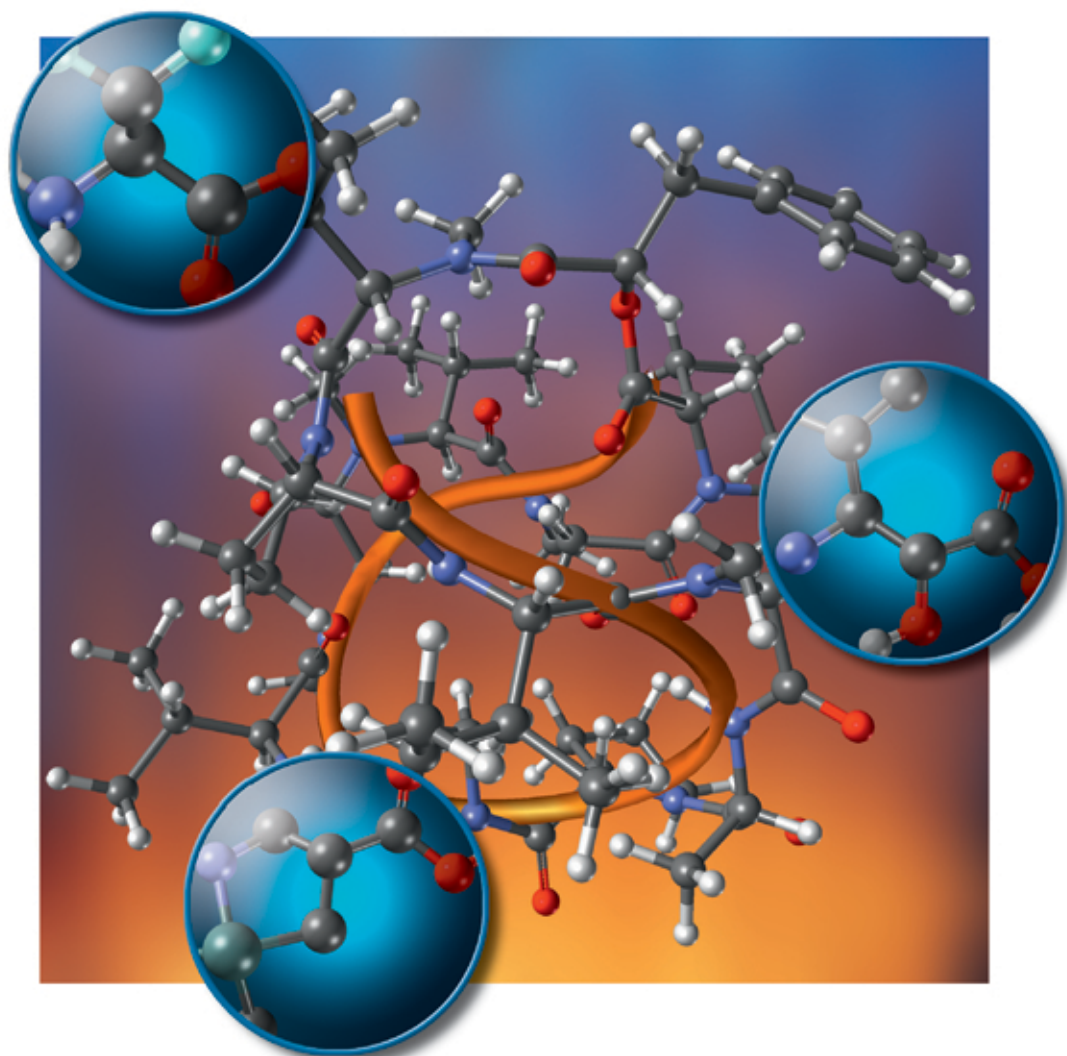
Edited by Andrew B. Hughes

 WILEY-VCH

# Amino Acids, Peptides and Proteins in Organic Chemistry

Volume 5

Analysis and Function of Amino Acids  
and Peptides





*Edited by*  
*Andrew B. Hughes*

**Amino Acids, Peptides  
and Proteins in  
Organic Chemistry**

## ***Further Reading***

Pignataro, B. (ed.)

### **Ideas in Chemistry and Molecular Sciences Advances in Synthetic Chemistry**

2010

ISBN: 978-3-527-32539-9

Tulla-Puche, Judit / Albericio,  
Fernando (eds.)

### **The Power of Functional Resins in Organic Synthesis**

2008

ISBN: 978-3-527-31936-7

Eicher, T., Hauptmann, S., Speicher, A.

### **The Chemistry of Heterocycles Structure, Reactions, Synthesis, and Applications**

2011

ISBN: 978-3-527-32868-0 (Hardcover)

ISBN: 978-3-527-32747-8 (Softcover)

Royer, J. (ed.)

### **Asymmetric Synthesis of Nitrogen Heterocycles**

2009

ISBN: 978-3-527-32036-3

Drauz, K., Gröger, H., May, O. (eds.)

### **Enzyme Catalysis in Organic Synthesis Third, Completely Revised and Enlarged Edition**

3 Volumes

2011

ISBN: 978-3-527-32547-4

Fessner, W.-D., Anthonsen, T.

### **Modern Biocatalysis Stereoselective and Environmentally Friendly Reactions**

2009

ISBN: 978-3-527-32071-4

Lutz, S., Bornscheuer, U. T. (eds.)

### **Protein Engineering Handbook**

2 Volume Set

2009

ISBN: 978-3-527-31850-6

Castanho, Miguel / Santos, Nuno (eds.)

### **Peptide Drug Discovery and Development**

Translational Research in Academia  
and Industry

2011

ISBN: 978-3-527-32891-8

Sewald, N., Jakubke, H.-D.

### **Peptides: Chemistry and Biology**

2009

ISBN: 978-3-527-31867-4

JNicolau, K. C., Chen, J. S.

### **Classics in Total Synthesis III New Targets, Strategies, Methods**

2011

ISBN: 978-3-527-32958-8 (Hardcover)

ISBN: 978-3-527-32957-1 (Softcover)

*Edited by*  
*Andrew B. Hughes*

# **Amino Acids, Peptides and Proteins in Organic Chemistry**

Volume 5 - Analysis and Function of  
Amino Acids and Peptides



WILEY-VCH Verlag GmbH & Co. KGaA

#### The Editor

**Andrew B. Hughes**

La Trobe University  
Department of Chemistry  
Victoria 3086  
Australia

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

**Library of Congress Card No.:** applied for

#### **British Library Cataloguing-in-Publication Data**

A catalogue record for this book is available from the British Library.

#### **Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

© 2012 Wiley-VCH Verlag & Co. KGaA,  
Boschstr. 12, 69469 Weinheim, Germany

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

**Composition** Thomson Digital, Noida, India

**Printing and Binding** betz-druck GmbH, Darmstadt

**Cover Design** Schulz Grafik Design, Fußgönheim

Printed in the Federal Republic of Germany

Printed on acid-free paper

**Print ISBN:** 978-3-527-32104-9

**ePDF ISBN:** 978-3-527-63185-8

**oBook ISBN:** 978-3-527-63184-1

## Contents

### List of Contributors XV

<b>1</b>	<b>Mass Spectrometry of Amino Acids and Proteins</b>	<b>1</b>
	<i>Simin D. Maleknia and Richard Johnson</i>	
1.1	Introduction	1
1.1.1	Mass Terminology	1
1.1.2	Components of a Mass Spectrometer	4
1.1.3	Resolution and Mass Accuracy	6
1.1.4	Accurate Analysis of ESI Multiply Charged Ions	10
1.1.5	Fragment Ions	11
1.2	Basic Protein Chemistry and How it Relates to MS	21
1.2.1	Mass Properties of the Polypeptide Chain	21
1.2.2	<i>In Vivo</i> Protein Modifications	21
1.2.3	<i>Ex Vivo</i> Protein Modifications	26
1.3	Sample Preparation and Data Acquisition	28
1.3.1	Top-Down Versus Bottom-Up Proteomics	28
1.3.2	Shotgun Versus Targeted Proteomics	28
1.3.3	Enzymatic Digestion for Bottom-Up Proteomics	29
1.3.4	Liquid Chromatography and Capillary Electrophoresis for Mixtures in Bottom-Up	30
1.4	Data Analysis of LC-MS/MS (or CE-MS/MS) of Mixtures	32
1.4.1	Identification of Proteins from MS/MS Spectra of Peptides	32
1.4.2	<i>De Novo</i> Sequencing	35
1.5	MS of Protein Structure, Folding, and Interactions	36
1.5.1	Methods to Mass-Tag Structural Features	37
1.6	Conclusions and Perspectives	40
	References	40
<b>2</b>	<b>X-Ray Structure Determination of Proteins and Peptides</b>	<b>51</b>
	<i>Andrew J. Fisher</i>	
2.1	Introduction	51
2.1.1	Light Microscopy	51
2.1.2	X-Rays and Crystallography at the Start	52

2.1.3	X-Ray Crystallography Today	53
2.1.4	Limitations of X-Ray Crystallography	54
2.2	Growing Crystals	55
2.2.1	Why Crystals?	55
2.2.2	Basic Methods of Growing Protein Crystals	55
2.2.3	Protein Sample	59
2.2.4	Preliminary Crystal Analysis	59
2.2.5	Mounting Crystals for X-Ray Analysis	61
2.3	Symmetry and Space Groups	62
2.3.1	Crystals and the Unit Cell	62
2.3.2	Point Groups	65
2.3.3	Space Groups	66
2.3.4	Asymmetric Unit	67
2.4	X-Ray Scattering and Diffraction	67
2.4.1	X-Rays and Mathematical Representation of Waves	67
2.4.2	Interaction of X-Rays with Matter	70
2.4.3	Crystal Lattice, Miller Indices, and the Reciprocal Space	73
2.4.4	X-Ray Diffraction from a Crystal: Bragg's Law	75
2.4.5	Bragg's Law in Reciprocal Space	77
2.4.6	Fourier Transform Equation from a Lattice	79
2.4.7	Friedel's Law and the Electron Density Equation	80
2.5	Collecting and Processing Diffraction Data	82
2.5.1	Data Collection Strategy	82
2.5.2	Symmetry and Scaling Data	83
2.6	Solving the Structure (Determining Phases)	83
2.6.1	Molecular Replacement	83
2.6.2	Isomorphous Replacement	85
2.6.3	MAD	88
2.7	Analyzing and Refining the Structure	90
2.7.1	Electron Density Interpretation and Model Building	90
2.7.2	Protein Structure Refinement	91
2.7.3	Protein Structure Validation	93
	References	94

### **3 Nuclear Magnetic Resonance of Amino Acids, Peptides, and Proteins** 97

*Andrea Bernini and Pierandrea Temussi*

3.1	Introduction	97
3.1.1	Active Nuclei in NMR	98
3.1.2	Energy Levels and Spin States	98
3.1.3	Main NMR Parameters (Glossary)	99
3.1.3.1	Chemical Shift	99
3.1.3.2	Scalar Coupling Constants	100
3.1.3.3	NOE	100
3.1.3.4	RDC	101



3.2	Amino Acids	101
3.2.1	Historical Significance	101
3.2.2	Amino Acids Structure	101
3.2.3	Random Coil Chemical Shift	102
3.2.4	Spin Systems	105
3.2.5	Labile Protons	110
3.2.6	Contemporary Relevance: Metabolomics	112
3.3	Peptides	113
3.3.1	Historical Significance	113
3.3.2	Oligopeptides as Models for Conformational Transitions in Proteins	114
3.3.3	Bioactive Peptides	116
3.3.4	Choice of the Solvent	117
3.3.4.1	Transport Fluids	118
3.3.4.2	Membranes	120
3.3.4.3	Receptor Cavities	122
3.3.5	Ensemble Calculations	125
3.3.6	Selected Examples from the Major Fields of Bioactive Peptides	125
3.3.6.1	Aspartame	125
3.3.6.2	Opioids	126
3.3.6.3	Transmembrane Helices	127
3.3.6.4	Cyclopeptides	128
3.4	Proteins	129
3.4.1	An Alternative to or a Validation of Diffractometric Methods?	129
3.4.2	Protein Spectra	129
3.4.3	Wüthrich's Protocol	130
3.4.3.1	Sample Preparation	131
3.4.3.2	Recording NMR Spectra	131
3.4.3.3	Sequential Assignment	131
3.4.3.4	Conformational Constraints	132
3.4.3.5	Model Building	134
3.4.4	Recent Developments	134
3.4.5	Selected Structures	136
3.4.5.1	Superoxide Dismutases	137
3.4.5.2	Malate Synthase G	137
3.4.5.3	Interactions	138
3.5	Conclusions	145
	References	146
<b>4</b>	<b>Structure and Activity of N-Methylated Peptides</b>	<b>155</b>
	<i>Raymond S. Norton</i>	
4.1	Introduction	155
4.2	Conformational Effects of N-Methylation	157
4.3	Effects of N-Methylation on Bioactive Peptides	159

4.3.1	Thyrotropin-Releasing Hormone	159
4.3.2	Cyclic Peptides	159
4.3.3	Somatostatin Analogs	160
4.3.4	Antimalarial Peptide	161
4.4	Concluding Remarks	162
	References	163
<b>5</b>	<b>High-Performance Liquid Chromatography of Peptides and Proteins</b>	<b>167</b>
	<i>Reinhard I. Boysen and Milton T.W. Hearn</i>	
5.1	Introduction	167
5.2	Basic Terms and Concepts in Chromatography	169
5.3	Chemical Structure of Peptides and Proteins	173
5.3.1	Biophysical Properties of Peptides and Proteins	173
5.3.2	Conformational Properties of Peptides and Proteins	176
5.3.3	Optical Properties of Peptides and Proteins	176
5.4	HPLC Separation Modes in Peptide and Protein Analysis	177
5.4.1	SEC	178
5.4.2	RPC	179
5.4.3	NPC	181
5.4.4	HILIC	181
5.4.5	ANPC	183
5.4.6	HIC	184
5.4.7	IEX	187
5.4.8	AC	188
5.5	Method Development from Analytical to Preparative Scale Illustrated for HP-RPC	189
5.5.1	Development of an Analytical Method	190
5.5.2	Scaling Up to Preparative Chromatography	196
5.5.3	Fractionation	198
5.5.4	Analysis of the Quality of the Fractionation	198
5.6	Multidimensional HPLC	198
5.6.1	Purification of Peptides and Proteins by MD-HPLC Methods	200
5.6.2	Fractionation of Complex Peptide and Protein Mixtures by MD-HPLC	202
5.6.3	Operational Strategies for MD-HPLC Methods	202
5.6.3.1	Off-line Coupling Mode for MD-HPLC Methods	202
5.6.3.2	On-Line Coupling Mode for MD-HPLC Methods	203
5.6.4	Design of an Effective MD-HPLC Scheme	203
5.6.4.1	Orthogonality of Chromatographic Modes	203
5.6.4.2	Compatibility Matrix of Chromatographic Modes	205
5.7	Conclusions	206
	References	207

<b>6</b>	<b>Local Surface Plasmon Resonance and Electrochemical Biosensing Systems for Analyzing Functional Peptides</b>	<b>211</b>
	<i>Masato Saito and Eiichi Tamiya</i>	
6.1	Localized Surface Plasmon Resonance (LSPR)-Based Microfluidics Biosensor for the Detection of Insulin Peptide Hormone	211
6.1.1	LSPR and Micro Total Analysis Systems	211
6.1.2	Microfluidic LSPR Chip Fabrication and LSPR Measurement	212
6.1.3	Detection of the Insulin–Anti-Insulin Antibody Reaction on a Chip	213
6.2	Electrochemical LSPR-Based Label-Free Detection of Melittin	215
6.2.1	Melittin and E-LSPR	215
6.2.2	Fabrication of E-LSPR Substrate and Formation of the Hybrid Bilayer Membrane	215
6.2.3	Measurements of Membrane-Based Sensors for Peptide Toxin	217
6.3	Label-Free Electrochemical Monitoring of $\beta$ -Amyloid ( $A\beta$ ) Peptide Aggregation	218
6.3.1	Alzheimer’s $A\beta$ Aggregation and Electrochemical Detection Method	218
6.3.2	Label-Free Electrochemical Detection of $A\beta$ Aggregation	219
	References	221
<b>7</b>	<b>Surface Plasmon Resonance Spectroscopy in the Biosciences</b>	<b>225</b>
	<i>Jing Yuan, Yinqiu Wu, and Marie-Isabel Aguilar</i>	
7.1	Introduction	225
7.2	SPR-Based Optical Biosensors	225
7.3	Principle of Operation of SPR Biosensors	226
7.4	Description of a SPR Instrument	228
7.4.1	Sensor Surface	228
7.4.2	Flow System	229
7.4.3	Detection System	230
7.5	Application of SPR in Immunosensor Design	230
7.5.1	Assay Development	232
7.5.1.1	Immobilization of the Analyte to a Specific Chip Surface	232
7.5.1.2	Assay Design	233
7.6	Application of SPR in Membrane Interactions	234
7.6.1	General Protocols for Membrane Interaction Studies by SPR	236
7.6.1.1	Liposome Preparation	236
7.6.1.2	Formation of Bilayer Systems	236
7.6.1.3	Analyte Binding to the Membrane System	237
7.6.1.4	Membrane Binding of Antimicrobial Peptides by SPR	238
7.7	Data Analysis	240
7.7.1	Linearization Analysis	240
7.7.2	Numerical Integration Analysis	241
7.7.3	Steady-State Approximations	242

7.8	Conclusions	243
	References	244
<b>8</b>	<b>Atomic Force Microscopy of Proteins</b>	<b>249</b>
	<i>Adam Mechler</i>	
8.1	Foreword	249
8.1.1	Importance of Asking the Right Question	250
8.2	AFM	250
8.2.1	Principle and Basic Modes of Operation	250
8.2.2	How Does a Tip Tap?	251
8.3	Bioimaging Highlights	253
8.3.1	Protein Oligomerization, Aggregation, and Fibers	253
8.3.2	Membrane Binding and Lysis	255
8.3.3	Ion Channel Activity	257
8.3.4	Protein–DNA-Specific Binding	261
8.4	Issues	261
8.4.1	Resolution	262
8.4.2	Imaging Force	263
8.4.3	Repetitive Stress	264
8.4.4	Artifacts Related to too Low Free Amplitude	265
8.4.5	Transient Force and Bandwidth	266
8.4.6	Accuracy of Surface Tracking	266
8.4.7	Step Artifacts	268
8.5	Force Measurements	269
8.6	Liquid Imaging	269
8.7	Sample Preparation for Bioimaging	272
8.7.1	Adhesion	272
8.7.2	Physical Entrapment	273
8.7.3	Chemical Binding	274
8.8	Outlook	274
	References	275
<b>9</b>	<b>Solvent Interactions with Proteins and Other Macromolecules</b>	<b>277</b>
	<i>Satoshi Ohtake, Yoshiko Kita, Kouhei Tsumoto, and Tsutomu Arakawa</i>	
9.1	Introduction	277
9.2	Solvent Applications	280
9.2.1	Research	280
9.2.2	Precipitation	287
9.2.3	Chromatography	288
9.2.4	Protein Refolding	296
9.2.5	Formulation	297
9.3	Solvent Application for Viruses	300
9.3.1	Isolation and Purification of Viruses	301
9.3.2	Stabilization and Formulation of Viruses	302
9.3.3	Inactivation of Viruses	309

9.4	Solvent Application for DNA	310
9.4.1	Isolation and Purification of DNA	310
9.4.2	Stability of DNA in a Cosolvent System	312
9.5	Mechanism	314
9.5.1	Physical Mechanism	315
9.5.1.1	Hydration	315
9.5.1.2	Excluded Volume	318
9.5.2	Thermodynamic Interaction	322
9.5.2.1	Group Interaction: Model Compound Solubility	322
9.5.3	Preferential Interaction	328
9.6	Protein–Solvent Interactions in Frozen and Freeze-Dried Systems	342
9.6.1	Frozen Systems	342
9.6.2	Freeze-Dried System	345
9.7	Conclusions	348
	References	349
<b>10</b>	<b>Role of Cysteine</b>	<b>361</b>
	<i>Lalla A. Ba, Torsten Burkholz, Thomas Schneider, and Claus Jacob</i>	
10.1	Sulfur: A Redox Chameleon with Many Faces	361
10.2	Three Faces of Thiols: Nucleophilicity, Redox Activity, and Metal Binding	365
10.3	Towards a Dynamic Picture of Disulfide Bonds	371
10.4	Chemical Protection and Regulation via S-Thiolation	374
10.5	“Dormant” Catalytic Sites	378
10.6	Peroxiredoxin/Sulfiredoxin Catalysis and Control Pathway	379
10.7	Higher Sulfur Oxidation States: From the Shadows to the Heart of Biological Sulfur Chemistry	384
10.8	Cysteine as a Target for Oxidants, Metal Ions, and Drug Molecules	388
10.9	Conclusions and Outlook	390
	References	391
<b>11</b>	<b>Role of Disulfide Bonds in Peptide and Protein Conformation</b>	<b>395</b>
	<i>Keith K. Khoo and Raymond S. Norton</i>	
11.1	Introduction	395
11.2	Probing the Role of Disulfide Bonds	396
11.3	Contribution of Disulfide Bonds to Protein Stability	396
11.4	Role of Disulfide Bonds in Protein Folding	397
11.5	Role of Individual Disulfide Bonds in Protein Structure	399
11.6	Disulfide Bonds in Protein Dynamics	401
11.7	Disulfide Bonding Patterns and Protein Topology	403
11.7.1	Conservation and Evolution of Disulfide Bonding Patterns	403
11.7.2	Conservation of Disulfide Bonds	404
11.7.3	Cysteine Framework and Disulfide Connectivity	404
11.7.4	Non-Native Disulfide Connectivities	407

11.8	Applications	408
11.9	Conclusions	409
	References	410
<b>12</b>	<b>Quantitative Mass Spectrometry-Based Proteomics</b>	<b>419</b>
	<i>Shao-En Ong</i>	
12.1	Introduction	419
12.2	Quantification in Biological MS	420
12.2.1	Label-Free Approaches in Quantitative MS Proteomics	423
12.2.2	SIL in Quantitative Proteomics	425
12.3	Identifying Proteins Interacting with Small Molecules with Quantitative Proteomics	430
12.4	Conclusions	433
	References	434
<b>13</b>	<b>Two-Dimensional Gel Electrophoresis and Protein/Polypeptide Assignment</b>	<b>439</b>
	<i>Takashi Manabe and Ya Jin</i>	
13.1	Introduction	439
13.2	Aim of Protein Analysis and Development of 2-DE Techniques	439
13.3	Current Status of 2-DE Techniques	441
13.3.1	Denaturing 2-DE for the Separation of Polypeptides	442
13.3.1.1	Principle	442
13.3.1.2	Procedures	444
13.3.1.3	Specific Features	445
13.3.2	Nondenaturing 2-DE for the Separation of Biologically Active Proteins and Protein Complexes	445
13.3.2.1	Principle	445
13.3.2.2	Procedures	446
13.3.2.3	Specific Features	447
13.3.3	Blue-Native 2-DE for the Detection of Protein–Protein Interactions	448
13.3.3.1	Principle	448
13.3.3.2	Procedures	448
13.3.3.3	Specific Features	449
13.3.4	Visualization of Proteins Separated on 2-DE Gels	449
13.3.4.1	Fixing Before CBB, Silver, or Fluorescent Dye Staining	450
13.3.4.2	CBB Staining	450
13.3.4.3	Silver Staining	450
13.3.4.4	Reverse Staining with Zinc-Imidazole	451
13.3.4.5	Fluorescent Dye Staining	451
13.3.4.6	Quantitation	451
13.4	Development of Protein Assignment Techniques on 2-DE Gels and Current Status of Mass Spectrometric Techniques	452
13.4.1	Development of Protein Assignment Techniques	452

13.4.2	MS-Based Assignment Techniques Utilizing Amino Acid Sequence Databases	454
13.4.2.1	Sample Preparation for MS Analysis	455
13.4.2.2	MALDI-TOF-MS and PMF	456
13.4.2.3	MS/MS and Peptide Sequence Search	459
13.5	Conclusions	460
	References	460
<b>14</b>	<b>Bioinformatics Tools for Detecting Post-Translational Modifications in Mass Spectrometry Data</b>	<b>463</b>
	<i>Patricia M. Palagi, Erik Arhné, Markus Müller, and Frédérique Lisacek</i>	
14.1	Introduction	463
14.2	PTM Discovery with MS	465
14.2.1	Detecting PTMs in MS and MS/MS Data	466
14.2.2	Discovering PTMs in MS or MS/MS Data	468
14.2.3	PTM Prediction Tools	469
14.2.3.1	From MS Data	469
14.2.3.2	From Sequence Data	469
14.3	Database Resources for PTM Analysis	470
14.4	Conclusions	473
	References	473
	<b>Index</b>	<b>477</b>





## List of Contributors

**Marie-Isabel Aguilar**

Monash University  
Department of Biochemistry and  
Molecular Biology  
Wellington Road  
Clayton, Victoria 3800  
Australia

**Tsutomu Arakawa**

Alliance Protein Laboratories  
3957 Corte Cancion  
Thousand Oaks, CA 91360  
USA

**Erik Arhne**

Swiss Institute of Bioinformatics  
Proteome Informatics Group  
1 rue Michel Servet  
1211 Geneva 4  
Switzerland

**Lalla A. Ba**

University of Saarland  
School of Pharmacy  
Division of Bioorganic Chemistry  
Campus B 2.1  
66123 Saarbrücken  
Germany

**Andrea Bernini**

University of Siena  
Department of Molecular Biology  
via Fiorentina 1  
53100 Siena  
Italy

**Reinhard I. Boysen**

Monash University  
ARC Special Research  
Centre for Green Chemistry  
Building 75, Wellington Road  
Clayton, Victoria 3800  
Australia

**Torsten Burkholz**

University of Saarland  
School of Pharmacy  
Division of Bioorganic Chemistry  
Campus B 2.1  
66123 Saarbrücken  
Germany

**Andrew J. Fisher**

University of California  
Departments of Chemistry and  
Molecular & Cell Biology  
One Shields Avenue  
Davis, CA 95616  
USA

**Milton T.W. Hearn**

Monash University  
ARC Special Research  
Centre for Green Chemistry  
Building 75, Wellington Road  
Clayton, Victoria 3800  
Australia

**Patricia Hernandez**

Swiss Institute of Bioinformatics  
Proteome Informatics Group  
1 rue Michel Servet  
1211 Geneva 4  
Switzerland

**Claus Jacob**

University of Saarland  
School of Pharmacy  
Division of Bioorganic Chemistry  
Campus B 2.1  
66123 Saarbrücken  
Germany

**Ya Jin**

Ehime University  
Faculty of Science  
Department of Chemistry  
2-5 Bunkyo-cho  
Matsuyama City 790-8577  
Japan

**Richard Johnson**

Institute for Systems Biology  
1441 N. 34th Street  
Seattle, WA 98103-1299  
USA

**Keith K. Khoo**

The Walter and Eliza Hall Institute of  
Medical Research  
Structural Biology Division  
1G Royal Parade  
Parkville, Victoria 3052  
Australia

and

University of Melbourne  
Department of Medical Biology  
Parkville, Victoria 3010  
Australia

**Yoshiko Kita**

Keio University  
School of Medicine  
Department of Pharmacology  
35 Shinanomachi, Shinjuku-ku  
Tokyo 160-8582  
Japan

**Frederique Lisacek**

Swiss Institute of Bioinformatics  
Proteome Informatics Group  
1 rue Michel Servet  
1211 Geneva 4  
Switzerland

**Simin D. Maleknia**

University of New South Wales  
School of Biological, Earth and  
Environmental Sciences  
Sydney, NSW 2052  
Australia

**Takashi Manabe**

Ehime University  
Faculty of Science  
Department of Chemistry  
2-5 Bunkyo-cho  
Matsuyama City 790-8577  
Japan

**Adam Mechler**

La Trobe University  
Department of Chemistry  
Physical Sciences 3, Bundoora Campus  
Bundoora  
Victoria 3086  
Australia

**Markus Mueller**

Swiss Institute of Bioinformatics  
 Proteome Informatics Group  
 1 rue Michel Servet  
 1211 Geneva 4  
 Switzerland

**Raymond S. Norton**

Monash University  
 Monash Institute of Pharmaceutical  
 Sciences  
 Medicinal Chemistry and Drug Action  
 381 Royal Parade  
 Parkville, Victoria 3052  
 Australia

**Satoshi Ohtake**

Aridis Pharmaceuticals  
 Research and Development  
 5941 Optical Court  
 San Jose, CA 95138  
 USA

**Shao-En Ong**

University of Washington  
 Department of Pharmacology  
 Health Sciences Center, Room E-401  
 Seattle, WA 98195-7280  
 USA

**Patricia M. Palagi**

Swiss Institute of Bioinformatics  
 Proteome Informatics Group  
 1 rue Michel Servet  
 1211 Geneva 4  
 Switzerland

**Masato Saito**

Osaka University  
 Graduate School of Engineering  
 Department of Applied Physics  
 2-1 Yamadaoka, Suita  
 Osaka 565-0871  
 Japan

**Thomas Schneider**

University of Saarland  
 School of Pharmacy  
 Division of Bioorganic Chemistry  
 Campus B 2.1  
 66123 Saarbrücken  
 Germany

**Eiichi Tamiya**

Osaka University  
 Graduate School of Engineering  
 Department of Applied Physics  
 2-1 Yamadaoka, Suita  
 Osaka 565-0871  
 Japan

**Pierandrea Temussi**

Università di Napoli Federico II  
 Dipartimento di Chimica  
 Via Cinthia  
 Complesso Monte S. Angelo  
 80126 Napoli  
 Italy

and

National Institute for Medical Research  
 Division of Molecular Structure  
 The Ridgeway  
 London NW7 1AA  
 UK

**Kouhei Tsumoto**

University of Tokyo  
 Medical Proteomics Laboratory  
 4-6-1 Shirokanedai, Minato-ku  
 Tokyo 106-8639  
 Japan

***Yinqiu Wu***

New Zealand Institute for Plant and  
Food Research  
120 Mt Albert Rd, Sandringham  
Auckland 1025  
New Zealand

***Jing Yuan***

New Zealand Institute for Plant and  
Food Research  
East Street 3214, Hamilton  
New Zealand

# 1

## Mass Spectrometry of Amino Acids and Proteins

*Simin D. Maleknia and Richard Johnson*

### 1.1

#### Introduction

##### 1.1.1

#### Mass Terminology

Like most matter (with the exception of, say, neutron stars), proteins and peptides are mostly made of nothing – an ephemeral cloud of electrons with very little mass surrounding tiny and very dense atomic nuclei that contain nearly all of the mass (i.e., peptides and proteins are made of atoms). Atoms have mass and the unit of mass that is most convenient to use is called the atomic mass unit (abbreviated amu or u) or in biological circles a Dalton (Da). Over the years, physicists and chemists have argued about what standard to use to define an atomic mass unit, but the issue seems to have been settled in 1959 when the General Assembly of the International Union of Pure and Applied Chemistry defined an atomic mass unit as being exactly 1/12 of the mass of the most abundant carbon isotope ( $^{12}\text{C}$ ) in its unbound lowest energy state. Therefore, one atom of  $^{12}\text{C}$  has a mass of 12.0000 u. Using this as the standard, one proton has a measured mass of 1.00728 u and one neutron is slightly heavier at 1.00866 u. One  $^{12}\text{C}$  atom contains six protons and six neutrons, the sum of which is clearly more than the mass of 12.0000 u. A carbon atom is less than the sum of its parts, and the reason is that the protons and neutrons in a carbon nucleus are in a lower energy state than free protons and neutrons. Energy and mass are interchangeable via Einstein's famous equation ( $E = mc^2$ ), and so this "mass defect" is a result of the nuclear forces that hold neutrons and protons together within an atom. This mass defect also serves as a reminder of why people like A. Q. Khan are so dangerous [1].

Each element is defined by the number of protons per nucleus (e.g., carbon atoms always have six protons), but each element can have variable numbers of neutrons. Elements with differing numbers of neutrons are called isotopes and each isotope possesses a different mass. In some cases, the additional neutrons result in stable isotopes, which are particularly useful in mass spectrometry (MS) in a method called

isotope dilution. Examples in the proteomic field that employ isotope dilution methodology include the use of the stable isotopes  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ , and  $^{18}\text{O}$ , as applied in methods such as ICAT (isotope-coded affinity tags) [2], SILAC (stable isotope labeling with amino acids in cell culture) [3], or enzymatic incorporation of  $^{18}\text{O}$  water [4]. Whereas some isotopes are stable, others are not and will undergo radioactive decay. For example, hydrogen with one neutron is stable (deuterium), but if there are two additional neutrons (a tritium atom) the atoms will decay to helium (two protons and one neutron) plus a negatively charged  $\beta$ -particle and a neutrino. Generally, if there are sufficient amounts of a radioactive isotope to produce an abundant mass spectral signal, the sample is likely to be exceedingly radioactive, the instrumentation would have become contaminated, and the operator would likely come to regret having performed the analysis. Therefore, mass spectrometrists will typically concern themselves with stable isotopes. Each element has a different propensity to take on different numbers of neutrons. For example, fluorine has nine protons and always 10 neutrons; however, bromine with 35 protons is evenly split between possessing either 44 or 46 neutrons. There are most likely interesting reasons for this, but they are not particularly relevant to a description of the use of MS in the analysis of proteins.

What is relevant is the notion of “monoisotopic” versus “average” versus “nominal” mass. The monoisotopic mass of a molecule is calculated using the masses of the most abundant isotope of each element present in the molecule. For peptides, this means using the specific masses for the isotopes of each element that possess the highest natural abundance (e.g.,  $^1\text{H}$ ,  $^{12}\text{C}$ ,  $^{14}\text{N}$ ,  $^{16}\text{O}$ ,  $^{31}\text{P}$ , and  $^{32}\text{S}$  as shown in Table 1.1). The “average” or “chemical” mass is calculated using an average of the isotopes for each element, weighted for natural abundance. For elements found in most biological molecules, the most abundant isotope contains the fewest neutrons

**Table 1.1** Mass and abundance values for some biochemically relevant elements.

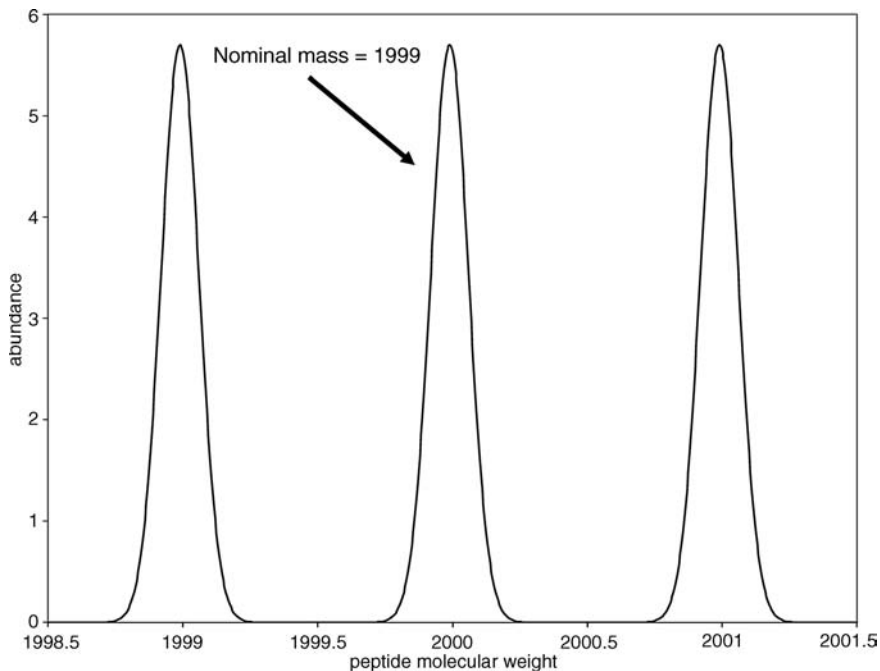
Element	Average mass	Isotope	Monoisotopic mass	Abundance (%)
Hydrogen	1.008	$^1\text{H}$	1.00783	99.985
		$^2\text{H}$	2.01410	0.015
Carbon	12.011	$^{12}\text{C}$	12	98.90
		$^{13}\text{C}$	13.00335	1.10
Nitrogen	14.007	$^{14}\text{N}$	14.00307	99.63
		$^{15}\text{N}$	15.00011	0.37
Oxygen	15.999	$^{16}\text{O}$	15.99491	99.76
		$^{17}\text{O}$	16.99913	0.04
		$^{18}\text{O}$	17.99916	0.200
Phosphorus	30.974	$^{31}\text{P}$	30.97376	100
Sodium	22.990	$^{23}\text{Na}$	22.98977	100
Sulfur	32.064	$^{32}\text{S}$	31.97207	95.02
		$^{33}\text{S}$	32.97146	0.75
		$^{34}\text{S}$	33.96787	4.21
		$^{36}\text{S}$	35.96708	0.02

and the less abundant isotopes are of greater mass. Therefore, the monoisotopic masses calculated for peptides are less than what are calculated using average elemental masses. The term “nominal mass” refers to the integer value of the most abundant isotope for each element. For example, the nominal masses of H, C, N, and O are 1, 12, 14, and 16, respectively. A rough conversion between nominal and monoisotopic peptide masses is shown as [5]:

$$M_c = 1.000495 \cdot M_n \quad (1.1)$$

$$D_m = 0.03 + 0.02 \cdot M_n/1000 \quad (1.2)$$

where  $M_c$  is the estimated monoisotopic peptide mass calculated from a nominal mass,  $M_n$ .  $D_m$  is the estimated standard deviation at a given nominal mass. For example, peptides with a nominal mass of 1999 would be expected, on average, to have a monoisotopic mass of around 1999.99 with a standard deviation of 0.07 u. Therefore, 99.7% of all peptides (3 standard deviations) at a nominal mass 1999 would be found at monoisotopic masses between 1999.78 and 2000.20 (Figure 1.1).



**Figure 1.1** Predicting monoisotopic from nominal molecular weights. Using the equations from Wool and Smilansky [5], peptides with nominal molecular weights of 1998, 1999, and 2000 would on average be expected to have monoisotopic molecular

weights of 1998.99, 1999.99, and 2000.99 with standard deviations of 0.07. The difference between monoisotopic and nominal masses is called the mass defect and this value scales with mass.

As can be seen, at around mass 2000, the mass defect in a peptide molecule is just about one whole mass unit. Most of this mass defect is due to the large number of hydrogen atoms present in a peptide of this size. The mass defect associated with nitrogen and oxygen tends to cancel out, and carbon by definition has no mass defect. The other important observation that can be made from this example is that 99.7% of peptides with a nominal mass of 1999 will be found between 1999.78 and 2000.20. Therefore, a molecule that is accurately measured to be 2000.45 cannot be a standard peptide and must either not be a peptide at all or is a peptide that has been modified with elements not typically found in peptides.

### 1.1.2

#### Components of a Mass Spectrometer

At minimum, a mass spectrometer has an ionization source, a mass analyzer, an ion detector, and some means of reporting the data. For the purposes here, there is no need to go into any detail at all regarding the ion detection and although there are many historically interesting methods of recording and reporting data (photographic plates, UV-sensitive paper, etc.), nowadays one simply uses a computer. The ionization source and the mass analyzer are the two components that need to be well understood.

Historically, ionization was limited to volatile molecules that were amenable to gas phase ionization methods such as electron impact. Over time, other techniques were developed that allowed for ionization of larger polar molecules – techniques such as fast atom bombardment (FAB) or field desorption ionization. However, these had relatively poor sensitivity requiring 0.1–1 nmol of peptide and, with the exception of plasma desorption ionization – a technique that used toxic radioactive californium, were generally not capable of ionizing larger molecules like proteins. Remarkably, two different ionization methods were developed in the late 1980s that did allow for sensitive ionization of larger molecules – electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). Posters presented at the 1988 American Society for Mass Spectrometry conference by John Fenn's group showed mass spectra of several proteins [6, 7], which revealed the general nature of ESI of peptides and proteins. Namely, a series of heterogeneous multiply protonated ions are observed, where the maximum number of charges is roughly dependent on the number of basic sites in the protein or peptide. Conveniently, this puts the ions at mass-to-charge ( $m/z$ ) ratios typically below 4000, which is a range suitable for just about all mass analyzers (see below). In a series of papers between 1985 and 1988, Hillenkamp and Karas described the essentials of MALDI [8–10]. Also, Tanaka presented a poster at a Joint Japan–China Symposium on Mass Spectrometry in 1987 showing a pentamer of lysozyme using laser desorption from a glycerol matrix containing metal shavings [11]. These early results showed the general nature of MALDI – singly charged ions predominate and therefore the mass analyzer must be capable of measuring ions with very high  $m/z$  ratios.



It is desirable for users to have some basic understanding of the different types of mass analyzers that are available. At one time multiselector analyzers [12] were well-liked (back when FAB ionization was popular), but quickly became dinosaurs for protein work after the discovery of ESI. It was too difficult to deal with the electrical arcs that tended to arise when trying to couple kiloelectronvolt source voltages with a wet acidic atmospheric spray. ESI was initially most readily coupled to quadrupole mass filters, which operated at much lower voltages. Quadrupole mass filters [13], as the name implies, are made from four parallel rods where at appropriate frequency and voltages, ions at specific masses can oscillate without running into a rod or escaping from between the rods. Given a little push (a few electronvolts potential) the oscillating ions will pass through the length of the parallel rods and be detected at the other end. Both quadrupole mass filters and multiselector instruments suffer from slow scan rates and poor sensitivity due to their low duty cycle. Instrument vendors have therefore been busy developing more sensitive analyzers. The ion traps [14, 15] are largely governed by the same equations for ion motion as quadrupole mass filters, but possess a greater duty cycle (and sensitivity). For those unafraid of powerful super cooled magnets, and who possess sufficiently deep pockets to pay for the initial outlay and subsequent liquid helium consumption, Fourier transform ion cyclotron resonance (FT-ICR) provides a high-mass-accuracy and high-resolution mass analyzer [16]. In this case, the ions circle within a very high vacuum cell under the influence of a strong magnetic field. The oscillating ions induce a current in a pair of detecting electrodes, where the frequency of oscillation is related to the  $m/z$  ratio. Detection of an oscillating current is also performed in Orbitrap instruments [17, 18], except in this case the ions circle around a spindle-shaped electrode rather than magnetic field lines. The time-of-flight hybrid (TOF) mass analyzer [19, 20] is, at least in principle, the simplest analyzer of all – it is an empty tube. Ions are accelerated down the empty tube and, as the name implies, the TOF is measured and is related to the  $m/z$  ratio (big ions move slowly and little ones move fast).

Tandem MS is a concept that is independent of the specific type of mass analyzer, but should be understood when discussing mass analyzers. As the name implies, tandem MS employs two stages of mass analysis, where the two analyzers can be scanned in various ways depending on the experiment. In the most common type of experiment, the first analyzer is statically passing an ion of a specific mass into a fragmentation region, where the selected ions are fragmented somehow (see below) and the resulting fragment ions are mass analyzed by the second mass analyzer. These so-called daughter, or product, ion scans are usually what are meant when referring to an “MS/MS spectra.” However, there are other types of tandem MS experiments that are occasionally performed. One is where the first mass analyzer is statically passing a precursor ion (as in the aforementioned product ion scan) and the second analyzer is also statically monitoring one, or a few, specific fragment ions. This so-called selected reaction monitoring (SRM) experiment is particularly useful in the quantitation of known molecules. There are other less frequently used tandem MS scans (e.g., neutral loss scans) and it should be noted

that only certain combinations of specific analyzers are capable of performing certain kinds of scans.

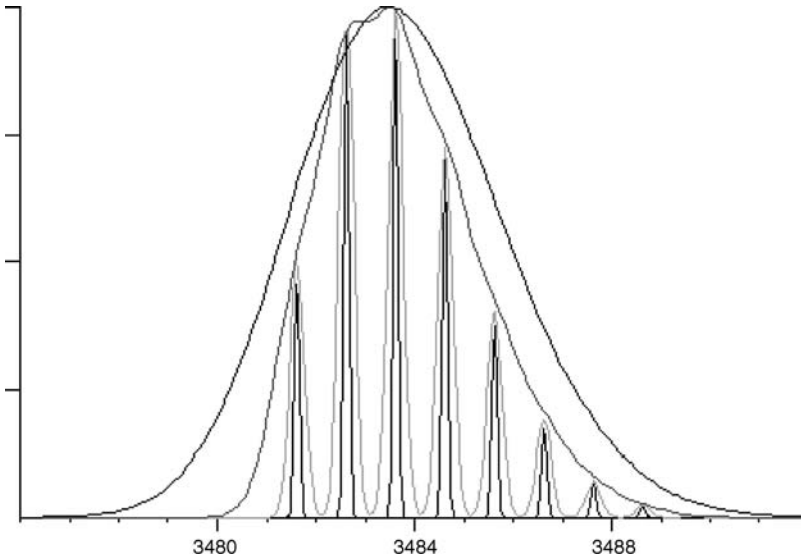
There are various combinations of mass analyzers used in different mass spectrometers. One of the more popular has been the quadrupole/TOF hybrid (quadrupole/time-of-flight hybrid Q-TOF) [21], which uses the quadrupole as a mass filter for precursor selection and the TOF is used to analyze the resulting fragment ions. Ion trap/time-of-flight hybrids are also sold and provide additional stages of tandem MS compared to the quadrupole/linear ion trap hybrid (Q-trap). The Q-TOF hybrid [22] is a unique instrument in that it can be thought of as a triple-quadrupole instrument where the third quadrupole can alternatively be used as a linear ion trap. There is consequently a great deal of flexibility in the types of experiments that can be done on such a mass spectrometer. The tandem TOF (TOF-TOF) [20] is an instrument that allows acquisition of tandem mass spectra or single-stage mass spectra of MALDI-generated ions. A timed electrode is used for precursor selection, which sweeps away all ions except those passing at a certain time (i.e.,  $m/z$ ) when the electrode is turned off momentarily. The selected packet of ions is then slowed down, possibly subjected to collision-induced dissociation (CID), and reaccelerated for the final TOF mass analysis of the fragments. The Orbitrap analyzer is purchased as a linear ion trap/Orbitrap hybrid and the same vendor sells their ion cyclotron resonance ICR instrument as a linear ion trap/ICR hybrid. It is beyond the scope of this chapter to go into any further details regarding the operation of the mass analyzers. Furthermore, it seems likely that the field will continue to change in the coming years, where instrument vendors will make further changes.

### 1.1.3

#### **Resolution and Mass Accuracy**

Regardless of the mass spectrometer, the user needs to understand their capabilities and limitations. Sensitivity has been a driving force for the development of many of the newer mass spectrometers. It is also a difficult parameter to evaluate, and one has to be careful not to simply evaluate the ability and tenacity of each vendor's application chemist when sending test samples out. Dynamic range is a parameter that is useful in the context of quantitative measurements and for most instruments it is around  $10^4$ . Some instruments can perform unique scan types (e.g., the Q-trap), or are more sensitive at performing SRM quantitative experiments (triple-quadrupole and Q-trap instruments). The scan speed or rate of MS/MS spectra acquisition is an instrument parameter that is relevant when attempting a deeper analysis of a complex mixture in a given amount of time. This latter issue is particularly important when analyzing complex proteomic samples.

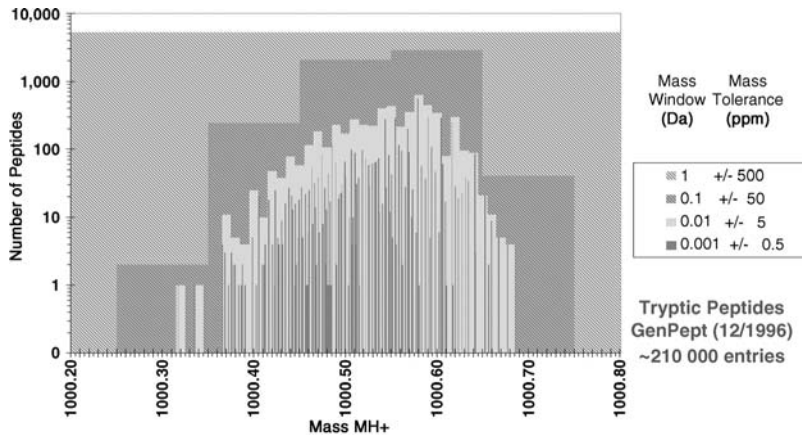
Two analyzer-dependent parameters are particularly important – mass accuracy and resolution. Resolution is defined as a unit-less ratio of mass divided by the peak width and is typically measured halfway up the peak. Figure 1.2 shows the peak shapes calculated for the peptide glucagon at various resolution values. At this mass, a resolution of 10 000 is sufficient to provide baseline separation of



**Figure 1.2** Effect of mass spectrometric resolution on peak shape. Shown are the calculated peak shapes for the  $(M + H)^+$  ion of porcine glucagon (monoisotopic mass of 3481.62 Da and average mass of 3483.8 Da) at various resolution values: 30 000 (inner most narrow peaks), 10 000 (outer most broad peak), 3000 (outer most broad peak), and 1000 (outer most broad peak).

each isotope peak and the higher resolution of 30 000 results in the narrowing of each isotope peak. As the resolution drops below 10 000 the valley between each isotope becomes higher until at 3000 the isotope cluster becomes a single broad unresolved peak. As the resolution drops further (blue), the single broad peak gets even fatter. Resolution is important to the extent that one needs to know if it is sufficient to separate the isotope peaks of a particular sample. If not, then a centroid of a broad unresolved peak (e.g., 1000 or 3000 for glucagon) is going to be closest to the peptide mass calculated using average elemental mass. Alternatively, if the resolution is sufficient to resolve the isotope peaks, and it is possible for the data system to accurately and consistently identify the monoisotopic  $^{12}\text{C}$  peak, then this observed peptide mass will be closest to that calculated using monoisotopic elemental masses.

Why do high resolution and high mass accuracy go hand in hand? One does not hear of low-resolution, high-mass-accuracy instruments, for instance. There are at least two reasons. First, it is not possible to determine a very accurate average elemental mass, which is weighted for isotope abundance. Chemical and physical fractionation processes occurring in nature result in variable amounts of each isotope in different samples. For example, the different photosynthetic processes (e.g., C3 and C4) will fractionate  $^{13}\text{C}$  slightly differently. Hence, corn will tend to have a slightly higher percentage of  $^{13}\text{C}$  than a tree. Therefore, in contrast to monoisotopic masses, average elemental masses come with fairly substantial error bars. The second reason



**Figure 1.3** Role of high mass accuracy in reducing false-positives from database searches. This histogram (from [23]) shows the number of tryptic peptides at different mass accuracies for a 1996 GenPept database. For a nominal molecular weight of 1000, there are around 5000 tryptic peptides if the measurements are accurate to 0.5 Da

(500 ppm). If the mass measurements are accurate to 0.01 Da (5 ppm), which are routinely available for Orbitrap and certain Q-TOF instruments, the number of possible tryptic peptides in the database drops to one to a couple hundred, depending on the specific mass window.

why higher resolution usually results in higher mass accuracy is that as a mono-isotopically resolved peak becomes narrower, any slight variation in the peak position is also reduced. Due to factors such as overlapping peaks and ion statistics, it is not possible to consistently and accurately measure a much wider unresolved isotope cluster at low resolution. Hence, the type of mass analyzer will determine the resolution and mass accuracy.

There are three types of resolution (and mass accuracy) for tandem MS that are associated with the precursor ion, precursor selection, and fragment ions. The precursor and fragment ion resolution and accuracy may be identical (e.g., for Q-TOF or ion traps) or different (e.g., for ion trap-FT-MS hybrids or TOF-TOF). The importance of being able to more accurately determine peptide masses was clearly demonstrated by Clauser *et al.* [23] as shown in Figure 1.3, which depicts a histogram of the number of tryptic peptides at different mass accuracies. For a 1996 GenPept database, there are around 5000 tryptic peptides at a nominal mass of 1000 with a tolerance of  $\pm 0.5$  Da (500 ppm). However, if the tolerance is tightened up to  $\pm 0.05$  Da, then the number of tryptic peptides drops by an amount that is dependent on the mass. There are fewer peptides at either the low- or high-mass end of the histogram, such that there are only two tryptic peptides in the database at a measurement of  $1000.3 \pm 0.05$  Da. Likewise, there are only 30–40 peptides with a mass of  $1000.7 \pm 0.05$  Da. Most of the tryptic peptides at a nominal mass of 1000 are in the range of 1000.45–1000.65, so a tolerance of  $\pm 0.05$  Da in the middle of this histogram will reduce the number of possible tryptic peptides from 5000 to 2000–3000. When using a database search program that identifies peptides from

their MS/MS spectra, a tighter precursor mass tolerance will result in fewer candidate sequences, which has the desirable effect of reducing the chances of an incorrect identification.

Database search programs (e.g., Mascot [24] or SEQUEST [25]) assume that there is only a single precursor and that all of the fragment ions are derived from that one precursor ion. For more complicated samples it is quite possible that more than one precursor is selected at a time and the likelihood of this happening is dependent on the precursor selection resolution. Typical ion traps select the precursor using a window that is three or four  $m/z$  units wide, Q-TOFs are similar, and TOF-TOFs have a precursor resolution of around 400 (e.g., at  $m/z$  1000, any peak at 997.5 will have its transmission reduced by half). The shape of this precursor selection window is also important – a sharp cutoff to zero transmission is good and a slow taper is not. Sometimes an extraneous low-intensity precursor is not a problem, as long as most of the fragment ion intensity is associated with the major precursor and the precursor mass that is associated with the resulting MS/MS spectrum is from the correct precursor ion. Search programs will still identify the major peptide, since there will only be a few low-intensity fragment ions left over. However, one can readily imagine several scenarios where mass selection of multiple precursors would be a problem. For example, suppose a minor precursor fragments really well, but the major precursor does not. In this case, the MS/MS spectrum contains fragment ions from the minor precursor, but the precursor mass that is used in the database search is derived from the major one. Or, a low-intensity precursor triggers a data-dependent MS/MS acquisition, but another very intense ion that is a few  $m/z$  units away contributes much of the fragment ion intensity. In such instances, where the fragment ions are derived from more than one precursor, search programs may get the wrong answer because the wrong precursor mass was used or there are too many leftover fragment ions and the scoring algorithm penalizes one of the correct sequences. Tighter selection windows with abrupt cutoffs (high precursor selection resolution) reduce the likelihood of this occurring. Improved database search algorithms would also help.

One of the major challenges in proteomics is high-throughput analysis. The high resolving power of FT-ICR instruments offers less than 1 ppm mass measurement accuracy and the peptide identification protocol of accurate mass tags (AMTs) now affords protein identification without the need for tandem MS/MS. Combining the AMT information with high-performance liquid chromatography (HPLC) elution times and MS/MS is referred to as peptide potential mass and time tags (PMTs) [26]. This approach expedites the analysis of samples from the same proteome through shotgun proteomics – a method of identifying proteins in complex mixtures by combining HPLC and MS/MS [27]. Once a peptide has been correctly identified through AMT and MS/MS with an assigned PMT, the information is stored in a database. This strategy greatly increases analysis throughput by eliminating the need for time-consuming MS/MS analyses. Accurate mass measurements are now routinely practiced in applications involving organisms with limited proteomes, including proteotyping the influenza virus [28], and the rapid differentiation of seasonal and pandemic stains [29].

## 1.1.4

**Accurate Analysis of ESI Multiply Charged Ions**

It is important to briefly describe the deconvolution algorithms used to translate  $m/z$  ratios of multiply charged ions generated during ESI to zero-charge molecular mass values. The accurate assignment of multiply charged ions is significant in proteomics applications, both in the analysis of the intact proteins and for the identification of fragment ions by MS/MS. For low-resolution mass spectra, algorithms were originally developed by assuming the nature of charge-carrying species or considering only a limited set of charge carrying species (i.e., proton, sodium) [30, 31]. For two ions ( $m_a/z_a$  and  $m_b/z_b$ ) that differ by one charge unit and both contain the same charge-carrying species, the charge on ion a ( $z_a$ ) is given by Eq. (1.3), where  $m_p$  is the mass of a proton, and the molecular weight is derived from Eq. (1.4):

$$z_a = (m_b/z_b - m_p) / (m_b/z_b - m_a/z_a) \quad (1.3)$$

$$\text{molecular weight} = z_a(m_a/z_a) - z_a m_p \quad (1.4)$$

The advantage of high-resolution electrospray mass spectra is that the ion charge can be derived directly from the reciprocal of the mass-to-charge separation between adjacent isotopic peaks ( $1/\Delta m/z$ ) for any multiply charged ion – referred to as the isotope spacing method [32]. Although the isotope spacing method is direct, complexities arising from spectral noise and overlapping peaks may result in inaccurate ion charge determination; furthermore, distinguishing  $1/z$  and  $1/(z + 1)$  for high charge state ions ( $z > 10$ ), would require mass accuracies of a few parts per million, which is not achieved routinely. To overcome some of these limitations, algorithms of Zscore [33] and THRASH [34] combined pattern recognition techniques to the isotope spacing method. For example, the THRASH algorithm matches the experimental abundances with theoretical isotopic distributions based on the model amino acid “averagine” ( $C_{4.938} H_{7.7583} N_{1.3577} O_{1.4773} S_{0.0417}$ ) [35]; however, this requirement restricts its application to a specific group of compounds and elemental compositions (i.e., proteins). The AID-MS [36] and PTFT [37] algorithms further advanced the latter algorithms by incorporating peak-finding routines to locate possible isotopic clusters and to overcome the problems associated with overlapping peaks.

A unique algorithm, CRAM (charge ratio analysis method) [38–40], deconvolutes electrospray mass spectra solely from the  $m/z$  values of multiply charged ions. The algorithm first determines the ion charge by correlating the ratio of  $m/z$  values for any two (i.e., consecutive or nonconsecutive) multiply charged ions to the unique ratios of two integers. The mass, and subsequently the identity of the charge carrying species, is then determined from  $m/z$  values and charge states of any two ions. For the analysis of high-resolution electrospray mass spectra, CRAM correlates isotopic peaks that share the same isotopic compositions. This process is also performed through the CRAM process after correcting the multiply charged ions to their lowest common ion charge. CRAM does not require prior knowledge of the elemental composition of a

molecule and as such does not rely at all on correlating experimental isotopic patterns with the theoretical patterns (i.e., known compositions), and therefore CRAM could be applied to mass spectral data for a range of compounds (i.e., including unspecified compositions).

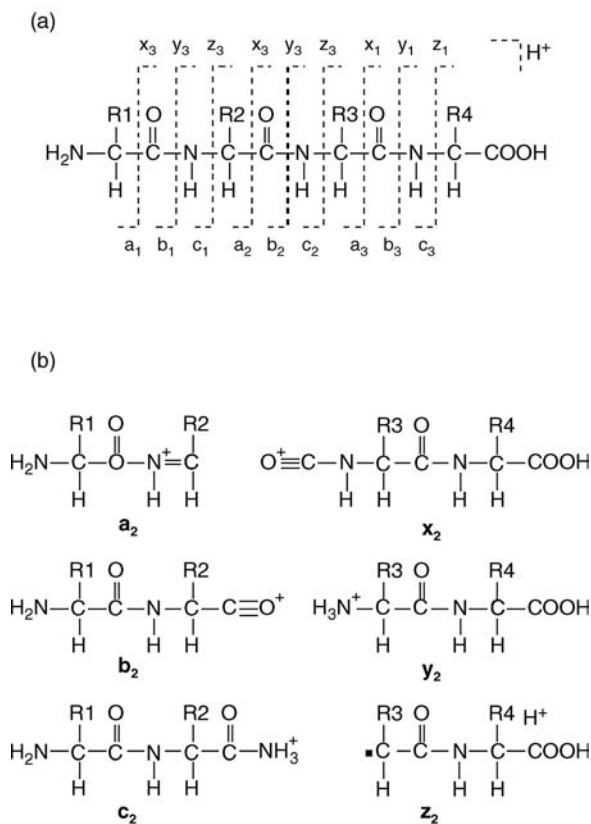
### 1.1.5

#### Fragment Ions

Although a considerable amount of work has been done in order to understand fragmentations of negatively charged peptide ions [41], the majority of protein identification work has employed positively charged peptide ions [42]. This is partially due to a general fear and ignorance of negatively charged peptides, but mostly because peptide signals are typically more abundant in the positive ion mode and the fragment ions are more likely to delineate a large portion of the peptide sequence. The following discussion is centered on fragmentation of peptide cations.

Depending on the type of mass spectrometer used, one can expect to generate fragment ions from three different processes – low-energy CID, high-energy CID, and electron capture (or transfer) dissociation (electron capture dissociation ECD or electron transfer dissociation ETD). Low-energy CID is the most common means of fragmenting peptide ions and occurs when the precursor ions collide with neutral collision gas with kinetic energies less than 500–1000 eV. This is the situation for any instrument with a quadrupole collision cell (triple-quadrupole, Q-trap, or Q-TOF), or any ion trap, including ion trap hybrids. A different process known as postsource decay (PSD) occurs in MALDI-TOF and MALDI-TOF-TOF instruments (when operated without collision gas). In PSD, precursor ions resulting from the MALDI process are sufficiently stable to stay intact during the initial acceleration into the flight tube, but they then fall apart in transit through the flight tube after full acceleration. These PSD-derived ions are largely identical to what is produced by low-energy CID. Figure 1.4(a) shows the peptide fragmentation nomenclature originally devised by Roepstorff and Fohlman [43], where the three possible bonds in a residue of a peptide are cleaved and the resulting fragment ion designated as X, Y, or Z (charge retained on the C-terminal fragments), or A, B, or C (N-terminal fragments). In addition to cleavage of the bond, different fragment ions also have variable numbers of hydrogen atoms and protons transferred to them. For a time there was considerable discussion as to whether the hydrogen transfer should be designated by tick marks (e.g., Y'' for two hydrogen atoms transferred to a Y cleavage ion) or by “+2” (e.g., Y + 2) designations. Biemann [44] subsequently proposed a similar designation whereby the letters went to lower case and the proper number of hydrogen atom transfers was assumed, without ticks or anything else. These are high stake issues, since adopting a specific nomenclature could dramatically increase one’s citation index.

At the most simplistic level, low-energy CID and PSD produce *b* and *y* ions. The structures shown in Figure 1.4(b) are not strictly accurate, but they illustrate how to go about calculating the masses of any fragment ion. The concept of a “residue mass” is



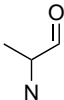
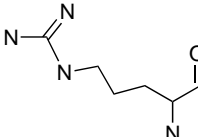
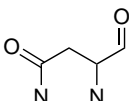
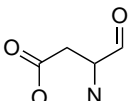
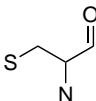
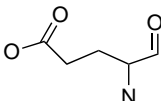
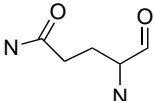
**Figure 1.4** Nomenclature for positive ion peptide fragments. Roepstorff nomenclature [43] is shown in (a). X, Y, and Z denote C-terminal fragments and A, B, and C denote N-terminal fragments. Fragment ions also have variable numbers of hydrogen atoms and protons transferred to them, as shown in

(b), which uses the Biemann nomenclature [44]. Low-energy CID of peptides in positive mode generally produces *b*-type and *y*-type ions. ETD and ECD generally produce *c*-type and *z*-type ions. The *z*-type ions are odd-electron radical cations, whereas the others are all even-electron cations.

that this is the mass of an amino acid within a peptide (i.e., it is the mass of an amino acid minus the mass of water, which is lost when amino acids polymerize to form peptides). Table 1.2 gives the average and monoisotopic residue masses for the common amino acids. It can be seen from Figure 1.4 that a *b* ion would be calculated by summing the residue masses and adding the mass of a single hydrogen atom (assuming that the peptide has an unmodified N-terminus). Likewise, a *y* ion would be calculated by summing the appropriate residue masses and then adding the mass of water plus a proton. The formulae for calculating the various peptide fragment ions are summarized in Table 1.3. It is believed that the actual structure of a *y* ion is the same as a protonated peptide and what is shown in Figure 1.4(b) is probably an



Table 1.2 Amino acid residue masses.

Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Alanine C <sub>3</sub> H <sub>5</sub> NO	Ala	A	71.03712	71.08	
Arginine C <sub>6</sub> H <sub>12</sub> N <sub>4</sub> O	Arg	R	156.10112	156.19	
Asparagine C <sub>4</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub>	Asn	N	114.04293	114.10	
Aspartic acid C <sub>4</sub> H <sub>5</sub> NO <sub>3</sub>	Asp	D	115.02695	115.09	
Asn or Asp	Asx	B			
Cysteine C <sub>3</sub> H <sub>5</sub> NOS	Cys	C	103.00919	103.14	
Glutamic acid C <sub>5</sub> H <sub>7</sub> NO <sub>3</sub>	Glu	E	129.04260	129.12	
Glutamine C <sub>5</sub> H <sub>8</sub> N <sub>2</sub> O <sub>2</sub>	Gln	Q	128.05858	128.13	
Glu or Gln	Glx	Z			

(Continued)

Table 1.2 (Continued)

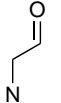
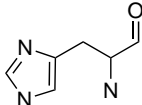
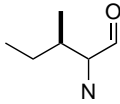
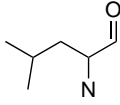
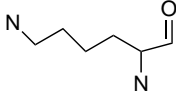
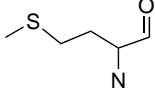
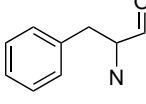

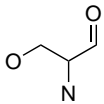
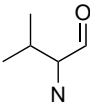
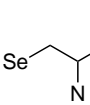
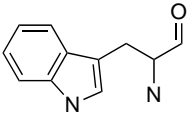
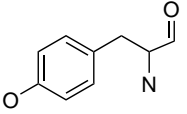
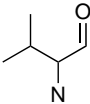
Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Glycine C <sub>2</sub> H <sub>3</sub> NO	Gly	G	57.02147	57.05	
Histidine C <sub>6</sub> H <sub>7</sub> N <sub>3</sub> O	His	H	137.05891	137.14	
Isoleucine C <sub>6</sub> H <sub>11</sub> NO	Ile	I	113.08407	113.16	
Leucine C <sub>6</sub> H <sub>11</sub> NO	Leu	L	113.08407	113.16	
Lysine C <sub>6</sub> H <sub>12</sub> N <sub>2</sub> O	Lys	K	128.09497	128.17	
Methionine C <sub>5</sub> H <sub>9</sub> NOS	Met	M	131.04049	131.19	
Phenylalanine C <sub>9</sub> H <sub>9</sub> NO	Phe	F	147.06842	147.18	
Proline C <sub>5</sub> H <sub>7</sub> NO	Pro	P	97.05277	97.12	
Serine C <sub>3</sub> H <sub>5</sub> NO <sub>2</sub>	Ser	S	87.03203	87.08	

Table 1.2 (Continued)

Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Threonine $C_4H_7NO_2$	Thr	T	101.04768	101.10	
Selenocysteine $C_3H_5NOSe$	SeC	U	150.95364	150.03	
Tryptophan $C_{11}H_{10}N_2O$	Trp	W	186.07932	186.21	
Tyrosine $C_9H_9NO_2$	Tyr	Y	163.06333	163.18	
Unknown	Xaa	X			
Valine $C_5H_9NO$	Val	V	99.06842	99.13	

accurate depiction of that type of fragment ion, although the site of protonation will vary. In contrast, the *b* ion structure in Figure 1.4(b) is almost certainly incorrect and instead is probably a five-membered ring structure [45]. The mechanism of formation of *b*-type ions most likely involves the carbonyl oxygen of the residue N-terminal to the cleavage site, which explains why one never observes  $b_1$  ions in peptides with free N-termini. Acylated peptides will produce  $b_1$  ions, since there is an N-terminal carbonyl available to induce the cleavage reaction.

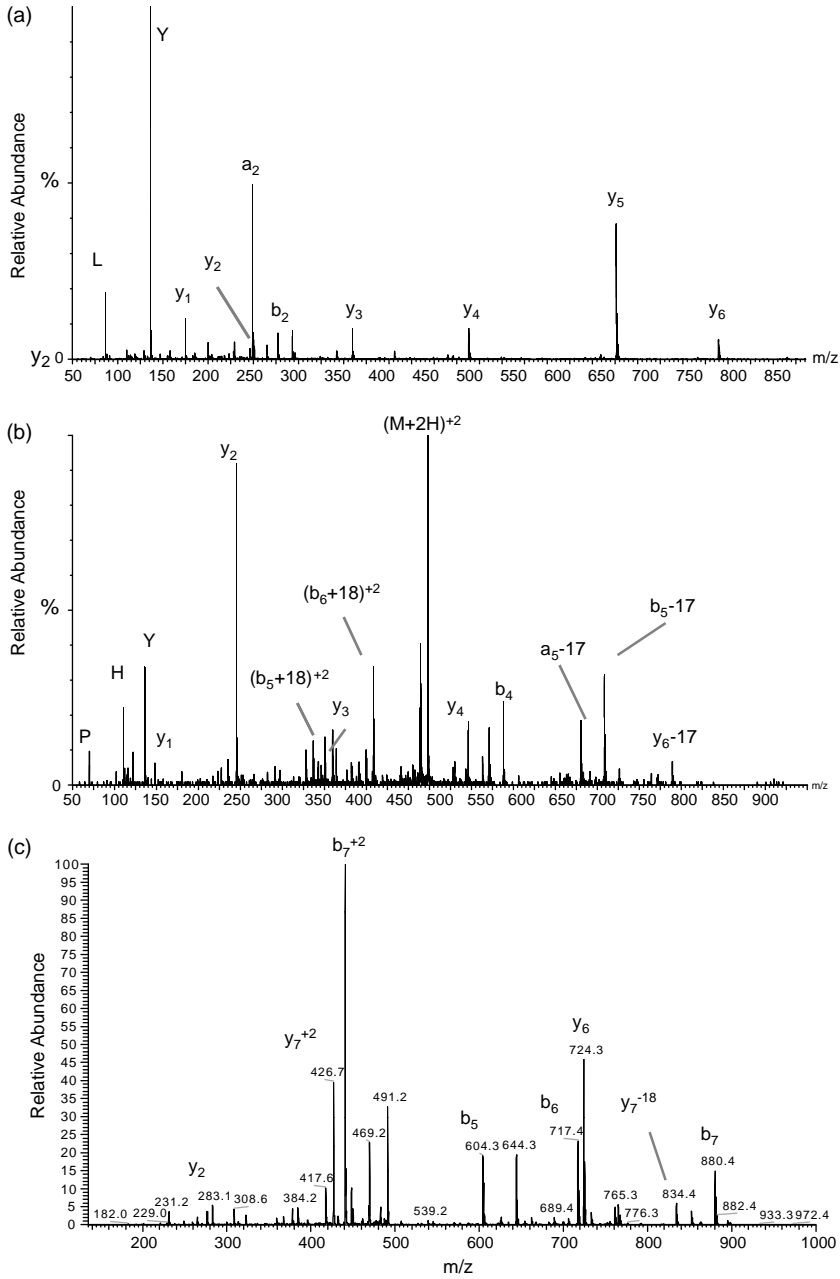
The concept of a “mobile proton” provides a useful framework for understanding the low-energy CID peptide fragmentation process [46]. In solution, the sites of peptide protonation are likely to be the N-terminal amino group, the lysine amino group, the histidine imidazole side-chain, or the guanidino group on arginine. In the gas phase, however, the peptide backbone amides are of comparable basicity to all but

**Table 1.3** Calculating the masses of positively charged fragment ions.

Ion type	Neutral molecular weight of the fragment
<i>a</i>	$[N] + [M] - CO - H$
<i>a</i> -H <sub>2</sub> O	$a - 18.0106$
<i>a</i> -NH <sub>3</sub>	$a - 17.0266$
<i>b</i>	$[N] + [M] - H$
<i>b</i> -H <sub>2</sub> O	$b - 18.0106$
<i>b</i> -NH <sub>3</sub>	$b - 17.0266$
<i>c</i>	$[N] + [M] + NH_2$
<i>d</i>	<i>a</i> – partial side-chain
<i>x</i>	$[C] + [M] + CO - H$
$\gamma$	$[C] + [M] + H$
$\gamma$ -H <sub>2</sub> O	$\gamma - 18.0106$
$\gamma$ -NH <sub>3</sub>	$\gamma - 17.0266$
<i>z</i>	$[C] + [M] - NH$
<i>v</i>	$\gamma$ – complete side-chain
<i>w</i>	<i>z</i> – partial side-chain

[N] is the mass of the N-terminus (e.g., 1.0078 Da for unmodified peptides and 43.0184 Da for acetylated N-terminus). [C] is the mass of the C-terminus (e.g., 17.0027 Da for unmodified peptides and 16.0187 Da for amidated C-terminus). [M] is the sum of the amino acid residue masses (see Table 1.1) that are contained within the fragment ion. CO is the combined mass of oxygen plus carbon atoms (27.9949 Da) and H is the mass of a proton (1.0078 Da). To calculate the *m/z* value of a fragment ion, add the mass of the protons to the neutral mass calculated from the table and divide by the number of protons added.

the arginine guanidino group. Therefore, in the absence of arginine, it takes only a little bit of collisional energy to scramble the site of protonation such that the ionized peptide is actually a population of ions that differ in the site of protonation (e.g., protonation occurring at any of the backbone amides or the side-chains). Protonation of the backbone amide is required for the production of *b*- or  $\gamma$ -type fragment ions and such cleavages that require protonation are called “charge promoted” fragmentations. Hence, as long as there is a mobile proton that can be sprinkled across the peptide backbone, one can expect to see a fairly contiguous series of *b*- and/or  $\gamma$ -type ions (e.g., Figure 1.5a). A major snag in this simplified view of low-energy CID of peptides is that the arginine guanidino group has such high gas-phase basicity that it essentially immobilizes a single proton. If there are at least as many arginine residues as protons, then to create *b*- or  $\gamma$ -type fragments, additional energy is required to “mobilize” one of the protons that would otherwise prefer to be stuck to the guanidino group. This additional energy will also result in the production of new and undesirable fragment ion types, such that the resulting spectra no longer possess the anticipated contiguous *b*- and  $\gamma$ -type fragment ion series (Figure 1.5b). One can see why low-energy CID of electrospray ionized tryptic peptides has been so successful, since most tryptic peptides will have no more than one arginine at the C-terminus, yet



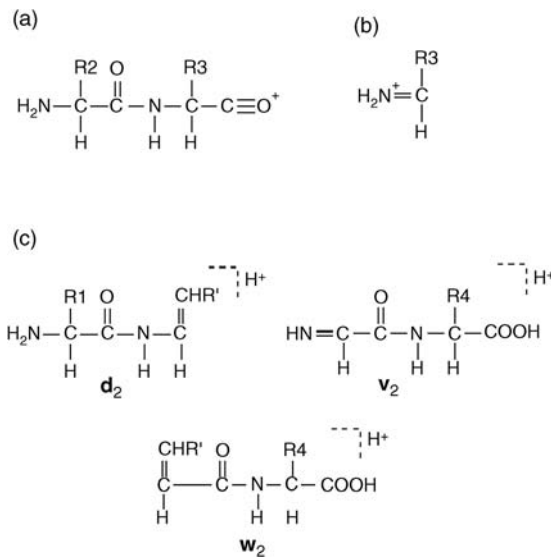
**Figure 1.5** Effect of arginine on fragment ion formation. (a) CID of  $(M+H)^{2+}$  precursor ion of the tryptic peptide YLYEAIAR, where one of the protons is “mobile” and induces a contiguous series of  $\gamma$ -type ions plus some  $b$ -type ions. (b) CID of  $(M+H)^{2+}$  precursor ion of the peptide YSRRHPE, which has two arginine residues and therefore no “mobile” proton.

Atypical fragmentations are seen and the sequence is impossible to determine. (c) CID of  $(M+H)^{2+}$  precursor ion of the peptide FKGRDIYT, which has a mobile proton that induces  $b$ - and  $\gamma$ -type fragmentations. However, the arginine in the middle of the peptide prevents formation of a contiguous series of ions.

be able to take on two protons – one for the arginine side-chain and one “mobile” proton to produce the  $b/\gamma$  fragment ions. Even for cases where there is a mobile proton, the presence of arginine in the middle of a peptide sequence can have adverse consequences as illustrated in Figure 1.5(c). Here, the mobile proton allows the production of  $b$ - and  $\gamma$ -type fragments; however, cleavages near the arginine are of reduced intensity and overall sequence coverage is sparse.

Low-energy CID produces a few additional fragment ion types and the resulting spectra possess certain characteristics that are useful to note. Under “mobile proton” conditions, the presence of proline in a peptide typically results in intense  $\gamma$ -type (and sometimes the corresponding  $b$ -type) ions resulting from cleavage on the N-terminal side of proline. Concomitantly, cleavage on the C-terminal side of proline is nonexistent or very much reduced. These effects are due to a combination of increased gas-phase basicity of the proline nitrogen and the unusual ring structure of the proline side-chain that inhibits the attack of the carbonyl on the N-terminal side of the proline. Under “mobile proton” conditions, histidine promotes fragmentation at its C-terminal side, resulting in enhanced abundance of the corresponding  $b/\gamma$  fragment ions. Sometimes a  $b/\gamma$  cleavage will occur twice in the same molecule, resulting in a fragment ion that contains neither the peptide’s original C- or N-terminus (Figure 1.6a). These “internal fragment ions” usually only contain a few residues and are often present if one of the two required  $b/\gamma$  fragmentations is particularly abundant. For example, cleavage at the N-terminal side of proline is sometimes so facile that this fragment will often fragment again, resulting in “internal fragment ions” that have the proline at the N-terminal side of the internal fragment ion. The  $b$ - and  $\gamma$ -type fragment ions often undergo an additional neutral loss of a molecule of water or ammonia. These ions are often designated as  $b - 17$  or  $b - 18$ , and so on. Under mobile proton conditions, these ions are usually less abundant than their corresponding  $b$ - or  $\gamma$ -type ion. The exceptions are when the N-terminal amino acid is glutamine or carbamidomethylated cysteine, in which case cyclization of the N-terminal amino acid and loss of ammonia occurs quite readily, resulting in abundant  $b - 17$  ions. Likewise, an N-terminal glutamic acid can cyclize and lose water, and the  $b - 18$  ions can be more abundant than the corresponding  $b$  fragment ions. In some cases, a  $b$ -type fragment ion can lose a molecule of carbon monoxide to form an  $a$ -type ion (28 Da less than the  $b$ -type fragment ion), although these seem to be more prominent for the lower mass fragments (e.g., it is not uncommon to find  $a_2$  ions that are of comparable intensity to the  $b_2$  ion in low-energy CID). Single amino acid immonium ions (Figure 1.6b) are often seen when MS/MS spectra acquisition includes this low mass region. Certain immonium ions are particularly diagnostic for the presence of their corresponding amino acid – leucine and isoleucine ( $m/z$  86), methionine ( $m/z$  104), histidine ( $m/z$  110), phenylalanine ( $m/z$  120), tyrosine ( $m/z$  136), and tryptophan ( $m/z$  159).

For peptide ions undergoing low-energy CID that lack a mobile proton, there are some additional fragment ions that become more prominent. Abundant ions resulting from cleavage at the C-terminal side of aspartic acid were first noticed in



**Figure 1.6** Additional ion types. (a) Internal ions are formed when a *b*/ $\gamma$ -type fragmentation occurs twice in the same molecule. R2 and R3 denote side-chains of the second and third amino acids in the original peptide sequence.

(b) Single amino acid immonium ions are observed if data acquisition includes lower mass regions. (c) Additional ions have been observed in high-energy CID (above 1 keV), but not at low energy.

MALDI-PSD spectra [47]. It later became clear that in the absence of a mobile proton, the side-chain carboxylic protons from aspartic acid (and to a lesser extent glutamic acid) can provide the necessary proton to catalyze a *b*/ $\gamma$  fragmentation [46]. This was first observed in the MALDI-PSD spectra, since the MALDI-derived singly charged ions need only a single arginine residue to lose the mobility of its one proton. Low-energy CID of peptide ions lacking a mobile proton also seem to be subject to the formation of a fragment ion that is sometimes called “*b* + 18” [45]. This is a rearrangement that occurs where the C-terminal residue is lost, but the C-terminal -OH group, plus a proton, are transferred to the ion. The designation “*b* + 18” refers to the fact that these have the mass of a *b*-type fragment ion plus the mass of water; however, the mechanism that gives rise to them is not related to the *b*-type fragmentation mechanism. Finally, it should be mentioned that low-energy CID of “nonmobile” peptide ions will often give more abundant neutral losses of water and ammonia (e.g., example, one might observe a  $\gamma$  - 17 ion in the absence of the corresponding  $\gamma$ -type fragment ion). For low-energy CID, MS/MS spectra from peptides with a mobile proton will exhibit the standard *b*- and  $\gamma$ -type fragment ions, and are most readily identified using database search programs. Likewise, spectra from peptides containing aspartic or glutamic acid in the absence of a mobile proton are also fairly readily interpreted. However, a “nonmobile proton” MS/MS spectrum of a peptide lacking aspartic or glutamic acid can be the most difficult type of peptide

to identify in a database search. This is especially true when the arginine is in the middle of the peptide.

The old multisection instruments were capable of subjecting peptide ions to much higher collision energy than the currently popular quadrupole collision cell and ion-trap instruments. At collision energies above 1 keV peptide ions can undergo alternative fragmentation pathways. In addition to the *b*/ $\gamma$  fragments seen for low-energy CID, high-energy CID can induce some additional “charge remote” fragmentations (Figure 1.6c), including the *d*- and *w*-type fragment ions that allows for the distinction between leucine and isoleucine [48, 49]. In general, these high-energy CID fragmentations seemed not to be influenced by the presence or absence of a mobile proton, which made it easier to derive sequences *de novo* directly from the spectra without recourse to searching a sequence database [50]. As already mentioned, these instruments are not used much anymore, but high-energy collisions are still relevant for one of the more modern instruments. If collision gas is used in a MALDI-TOF-TOF instrument [20], the collision energies can be as high as a couple of kiloelectronvolts, and the resulting MS/MS spectra will contain the *d*-,  $\nu$ -, and *w*-type fragment ions.

ECD is a process whereby an isolated multiply charged peptide ion captures a low-energy thermal electron, and the resulting radical cation becomes sufficiently unstable and fragments to produce *c*- and *z*-type fragment ions (Figure 1.4) [51]. Of key importance is that ECD induces fragmentation in a manner that does not result in intramolecular vibrational energy redistribution. In contrast, the additional energy acquired in CID is redistributed across the many vibrational modes of the entire molecule with the end result being that the weakest bonds break first, which often leaves insufficient energy for further peptide backbone cleavages. For example, low-energy CID of peptides containing phospho-serine or phospho-threonine usually results in a facile neutral loss of phosphoric acid. Sometimes the phosphate group stays attached, but usually not. The problem with this is that low-energy CID spectra of phosphopeptides typically exhibit a very abundant phosphoric acid neutral loss, but have tiny *b*/ $\gamma$ -type fragment ions that may not rise above the noise. Hence, the user is left knowing that they have a phosphopeptide, but not which one. Glycopeptides behave similarly. In contrast, the ECD fragmentation process leaves the phosphate or carbohydrate attached to the *c*- and *z*-type fragment ions, which allows for one to identify the protein and pinpoint the site of phosphorylation or glycosylation [52].

The trapping of thermal electrons for use in ECD has only been possible in FT-ICR instruments, which happen to be the most expensive type of mass spectrometer. Avoiding this expense provided some of the impetus in the development of ETD [53], where anionic molecules are trapped in a linear ion trap (using radiofrequency electrical fields) and are mixed with multiply charged cationic peptide analyte ions. Given the appropriate anion (one with low electron affinity), an electron is transferred to the peptide cation in an exothermic process that induces the production of the same *c*- and *z*-type fragment ions observed in ECD (Figure 1.4). ETD is sufficiently rapid that it can be used in conjunction with LC-MS/MS, and is sometimes used along with CID (i.e., data-dependent analysis might trigger the acquisition of both an



ETD and a CID spectrum from the same precursor ion). Similar to ECD, ETD seems to be particularly useful for the analysis of post-translational modifications (PTMs) that are otherwise labile under CID conditions (e.g., phosphorylation) [54]. For shotgun protein identifications, CID and ETD appear to be complementary in that ETD tends to be more successful at identifying peptide precursor ions with higher charge density, whereas CID is better at precursors with one to three protons [55].

## 1.2

### Basic Protein Chemistry and How it Relates to MS

#### 1.2.1

##### Mass Properties of the Polypeptide Chain

Proteins are linear chains of monomers made up of 20 standard amino acids (Table 1.2) that can be as massive as a few mega-Daltons (e.g., titin), but are typically in the range of 10–100 kDa. Proteins can exhibit a fairly wide range of physical properties such as solubility and hydrophobicity, which can make it difficult to find a universal means of separating and isolating them. Although most proteins are soluble in the buffers used for sodium dodecylsulfate–polyacrylamide gel electrophoresis (SDS–PAGE), the resulting separation leaves the proteins within a polyacrylamide gel matrix from which it is difficult and inefficient to extract the intact proteins. Proteolytic digestion and release of peptides derived from proteins entrained in gel slices is relatively efficient (in-gel digestion) [56]. In contrast to intact proteins, peptides derived from proteins via proteolysis tend to have more uniform distributions of physical properties that make them amenable to standard peptide separation techniques such as HPLC. There will often be a subset of proteolytic peptides for each protein that exhibit favorable properties with respect to chromatography and ionization. Therefore, most proteomics involves the so-called bottom-up approach of first ravaging proteins with one protease or another, analyzing the resulting peptide bits, and then trying to deduce which proteins were present in the first place.

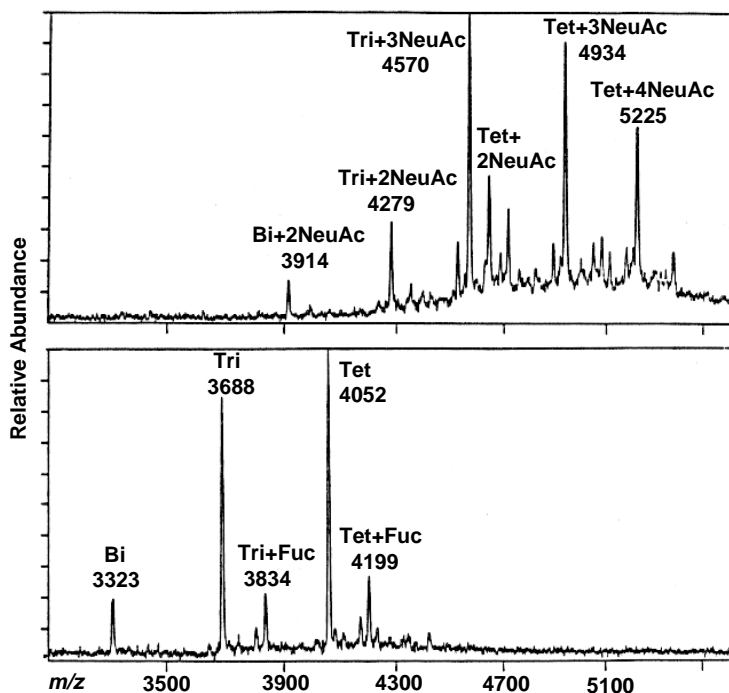
#### 1.2.2

##### *In Vivo* Protein Modifications

The standard amino acids can be decorated with a variety of biologically significant modifications. There are a couple of Web resources that list the various modifications that have been observed and these should be used whenever unexpected mass shifts are observed ([www.unimod.org](http://www.unimod.org) and <http://www.abrf.org/index.cfm/dm.home>). Some of these modifications will alter more than just the residue mass, possibly making the modified peptide more or less readily ionized, hydrophilic, or soluble. In some cases, the modification introduces a chemical bond that is particularly labile to mass spectrometric fragmentation. Therefore, interpreting spectra from modified

peptides is often a tricky business that involves more than just adding the right masses together. What follows is a brief description of just a few of the more common protein modifications, focusing on chemical and physical properties that are relevant to their analysis by MS.

Glycosylation is one of the more common modifications, which can be subdivided into at least four categories – *N*-linked, *O*-linked, *C*-mannosylation, and cytosolic *O*-GlcNAc modifications (GlcNAc = acetylglucosamine). *C*-Mannosylation is a modification of tryptophan in a WXXW motif, where the first tryptophan is modified by mannose via a carbon–carbon bond [57]. This bond is stable to low-energy CID, and can therefore be readily pinpointed using standard tandem MS methods. The *O*-GlcNAc modification is a very interesting modification involved in signal transduction pathways that occurs on nuclear and cytoplasmic proteins in eukaryotic cells. Specific serine and threonine residues are modified by *N*-acetylglucosylaminyl-transferase and *O*-GlcNAcase, which are the two enzymes that dynamically attach and remove this single monosaccharide [58, 59]. Low-energy CID of *O*-GlcNAc-peptides tends to produce abundant fragment ions resulting from loss of the monosaccharide leaving the modified serine intact, whereas ETD preferentially cleaves peptide bonds thereby leaving the *O*-GlcNAc attached to the modified residue [60]. In contrast to *O*-GlcNAc modification, the more standard extracellular *N*- and *O*-linked glycosylation are polymeric in nature, and the carbohydrate structures are typically large (above 2000 Da for *N*-linked) and heterogeneous at any given site of modification. Sites of *N*-linked glycosylation are determined relatively easily by use of *N*-glycosidase F, which removes the entire carbohydrate from the side-chain of asparagine and in the process converts it to aspartic acid [61, 62]. This modification only occurs at a specific sequence motif consisting of asparagine, followed by any residue except proline, which is then followed by serine, threonine, or cysteine. Thus, identification of aspartic acid in place of asparagine in such a motif after treatment by *N*-glycosidase F is sometimes considered to be sufficient for identifying a site of *N*-linked glycosylation. Given that asparagine is capable of chemically deamidating [63], absolute proof can be obtained by performing the enzymatic deglycosylation reaction in the presence of  $^{18}\text{O}$  water, which is incorporated into the deamidated aspartic acid side-chain. Extracellular *O*-linked carbohydrates are smaller than *N*-linked (only a few carbohydrate monomers) and are attached to serine or threonine, but within no clear sequence motif. Unlike *N*-glycosylation, there is no robust means of removing the glycan prior to mass spectrometric analysis, and determinations of *O*-linked glycopeptides and proteins can be quite challenging [64]. Enrichment of glycopeptides and glycoproteins is typically accomplished using lectin-affinity chromatography [65–68]. The analysis of carbohydrate heterogeneity for a specific glycopeptide site is best performed in a stepwise manner by treating the proteolytic sample with appropriate enzymes [69]. For example, *N*-acetyl neuraminic acid residues are easily removed with neuraminidase (Figure 1.7). This stepwise enzymatic treatment is beneficial for revealing fine structural details of complex glycopeptides mixtures. CID of glycopeptides results in fragmentation of glycosidic bonds allowing for characterization of the carbohydrate portion, but fragmentation of peptide bonds is usually absent. For determination of the site of glycosylation, ETD can provide peptide fragmentation. For isolation of



**Figure 1.7** Negative-ion MALDI-TOF mass spectra displaying carbohydrate heterogeneity (bi-, tri-, and tetra-antennary) for glycopeptide site III (IQATFFYFTPN<sup>+</sup>KTE) obtained from Staph V8 digest of human  $\alpha_1$ -acid glycoprotein; (top) before and (bottom) after treatment with

neuraminidase that removes *N*-acetyl neuraminic acid (NeuAc) residues while retaining other sugar moieties including Fucose (Fuc). The matrix solution was a 2 : 1 mixture of 2-aminobenzoic acid: nicotinic acid and a nitrogen laser at 337 nm was used.

*N*-linked peptides, peptide identification, and determining sites of *N*-linked modification, the “glycocapture” technique is particularly useful [70, 71].

Phosphorylation is a dynamic modification that most often occurs on serine, threonine, and tyrosine. Low-energy CID of peptides containing phospho-Ser/Thr tends to produce abundant fragment ions resulting from the neutral loss of phosphoric acid (loss of 98 Da), where the sequence-specific ions (*b*- and *y*-type) are much less intense. In contrast, ETD tends to leave the phosphate intact while still promoting sequence-specific fragment ions (*c*- and *z*-type), which makes pinpointing the site of phosphorylation more reliable [54]. Phospho-tyrosine is more stable, and sequence specific ions that still possess the phosphate are prominent in these CID spectra. There is evidence that during CID, phosphate groups can migrate from one site to another within a peptide molecule [72] and this is particularly pronounced in the absence of a mobile proton. Multiple phosphorylations of individual proteins seem to be quite common, which adds to the difficulty of analysis. Moreover, not all protein phosphorylation appears to be functionally relevant and, more importantly, different phosphorylation sites on the same protein may regulate different processes.

Thus, the challenge for understanding how phosphorylation modulates a given biological pathway is to discover which phosphorylation sites on a given protein are the relevant ones and how phosphorylation at those sites changes in response to various stimuli. The ability to derive quantitative information on specific phosphorylation sites is imperative to this goal. A confounding effect of phosphorylation analysis is that phosphorylated peptides can behave differently from their nonphosphorylated counterparts (e.g., changing solubility, ability to be ionized, chromatographic behavior, or tendency to adsorb to surfaces). The presence of phosphate near a predicted proteolytic site may inhibit proteolysis, which can make it difficult to make direct quantitative comparisons between phosphorylated and nonphosphorylated peptides encompassing the same site of modification. Even if the physical properties were identical, the oftentimes low stoichiometry of phosphorylation means that much larger amounts of sample needs to be analyzed in order to detect the phosphorylated peptides. For this reason, phosphopeptides and phosphoproteins are often enriched prior to mass spectral analysis (e.g., [73–77]). Of course, by enriching a phosphopeptide and removing its unphosphorylated counterpart, it becomes impossible to determine stoichiometry.

There is a class of proteins called ubiquitin-like modifiers (UBLs; proteins including ubiquitin, NEDD8, ISG15, SUMO1, etc.) that are all used by cells to tag other protein substrates on specific lysine residues [78]. This tagging of protein substrates by UBLs serves a variety of purposes ranging from targeting the substrate for degradation to signaling functions. In some cases, a single UBL will modify a particular lysine residue; in other cases, long chains of polymerized UBLs are attached to a substrate lysine. Although structurally more complicated than phosphorylation, this PTM is similar in that it can be dynamic. There are enzymes that put UBLs on substrates and others that take them off. Like phosphorylation, the stoichiometry of the modification can be quite low and care must be taken that the deubiquitinating enzymes do not remove the UBLs during sample preparation. From the standpoint of MS, it is important to note that all of these UBLs are attached to the substrate lysine  $\epsilon$ -amino group via an amide linkage to the UBL's C-terminus that always ends with Gly–Gly. In the case of ubiquitin itself, the Gly–Gly sequence is preceded by an Arg residue, which upon tryptic digestion of the substrate leaves the formerly ubiquitinated lysine tagged with a Gly–Gly [79, 80]. This often forms the basis for the identification of ubiquitination sites, and it should be pointed out that the amide linkage of Gly–Gly to the  $\epsilon$ -amino group of lysine is quite stable to CID (in contrast to Ser/Thr phosphorylation or *N*- and *O*-glycosylation). Other UBLs may not have this arginine residue. For example, SUMO1-modified proteins have a 19-amino-acid peptide appended to the lysine  $\epsilon$ -amino group, which makes the identification via CID a bit of a challenge [81].

Acylation of protein N-terminal amino groups and lysine  $\epsilon$ -amino groups is a common PTM. Acetylation of the protein N-terminus occurs on over half of eukaryotic cytosolic proteins; myristoylation of N-terminal glycine is also found in a small number of cytoplasmic proteins. Acylation produces an amide bond that is stable to low-energy CID, which makes MS/MS analysis considerably easier than the more labile modifications described above (e.g., glycosylation and phosphorylation).

Acetylation of lysine side-chains is a reversible modification that appears to be involved in a variety of cellular processes [82]. Acetylation of the lysine side-chain prevents proteolytic cleavage by trypsin, which therefore makes it difficult to make quantitative comparisons with the unmodified form if trypsin is used. Acetylation also reduces both the solution and gas-phase basicity of the lysine side-chain, which would likely influence peptide ionization and charge state, as well as retention time in cation-exchange chromatography. Enrichment of acetylated peptides can be accomplished using anti-acetyl lysine antibodies [82, 83].

Disulfide bonds are one of a few protein modifications that result in a loss of mass (two hydrogen atoms per cysteine pair). In principal, the determination of disulfide bonds is a simple matter of measuring the mass of proteolytic peptides before and after reduction, where one looks for ions that disappear after reduction as well as the appearance of the corresponding peptide ions containing reduced cysteine. In practice, disulfide determination is rarely this simple. In order to unambiguously assign disulfide linkages, proteolytic cleavage sites must be located between every cysteine, which they often are not. Moreover, proteins with intact disulfide bonds are often refractory to proteolytic degradation, so the intended cleavages often do not occur. A typical outcome for a disulfide experiment is to identify some of the reduced peptides, but not the original disulfide-linked peptide (or vice versa). Or sometimes one of the reduced ions is observed, but not the other (perhaps it chromatographs poorly or is not ionized). Or in a protein with several disulfide bonds, a few of them might be determined, but the others are refractory. Sometimes this can be overcome by using different proteases, whereby the resulting peptides might have more favorable chemical and physical properties for analysis. One aspect of disulfide chemistry that is often forgotten is the fact that once a protein starts losing secondary and tertiary structure (via proteolysis or addition of denaturant), both acid- and base-catalyzed scrambling can occur [84]. For proteolysis at  $\text{pH} > 7$  one needs to add a small amount of alkylating reagent to eliminate any catalytic amounts of thiol that might be present in the sample. The base-catalyzed scrambling drops off 10-fold per pH unit, which is why pepsin cleavage at pH 3 is often used for these purposes. The acid-catalyzed scrambling occurs in the presence of above 6M HCl and is not typically a condition used in protein chemistry. Finally, it should be noted that although CID does not typically break a disulfide bond, ETD favors cleavage of this bond [85, 86].

Proteolysis is normally thought of as something that is done to samples during a bottom-up proteomic analysis; however, it is also an important PTM that occurs in a variety of settings and has numerous biological purposes. One of the most common proteolytic events is the removal of the N-terminal initiator methionine. Secreted proteins and certain classes of membrane proteins possess secretory signal sequences at the N-terminus that are proteolyzed while entering the endoplasmic reticulum. There are cell surface proteases that clip other membrane proteins to release the extracellular domain; tumor necrosis factor being one of the more famous examples of a shed membrane protein [87]. There are many other examples of proteolysis occurring in a wide variety of biological processes ranging from blood clotting to processing polypeptide chains into smaller peptide hormones. It can be

relatively easy to establish that a proteolytic event occurred by, for example, identifying a transmembrane protein extracellular domain in some cultured cell supernatant [88]. Or sometimes one might identify a protein from a SDS-PAGE gel slice that is at a much lower molecular weight than would be predicted from the full sequence. However, identifying the specific site of proteolysis can be difficult if peptides from that region are hard to ionize or have unfortunate chromatographic properties, especially if the peptide containing the endogenous cleavage site is further modified, for example, by *O*-linked glycosylation. Obviously, the endogenous cleavage site has to be different from the protease specificity used to create peptides for LC-MS/MS (e.g., the C-terminus of a protein cannot be either arginine or lysine if trypsin was used). There are a few methods available that enrich for N- or C-terminal peptides that might be useful for these purposes [89–94].

### 1.2.3

#### **Ex Vivo Protein Modifications**

Aside from the numerous chemical modifications researchers do to proteins on purpose (e.g., reduction of disulfide bonds, alkylation of thiols, or various reactions that incorporate stable isotopes into modified peptides), there are several that occur by accident during sample handling. These modifications typically add mass and the corresponding shifts are measured by MS. What follows is a brief list of the more common ones.

Denaturation in urea is often accompanied by carbamylation of amino groups (either N-terminal or lysine), as well as other functional groups on other side-chains to a lesser extent [95]. Urea is in equilibrium with ammonium cyanate and it is the latter that is reactive with amines. Carbamylated peptides, like acetylated peptides, are stable to low-energy CID, which means that fragmentation of the peptide bonds will not result in loss of the carbamyl group. The key to limiting carbamylation is trying to limit the concentration of cyanate anion by making urea solutions fresh from solid urea, avoiding elevated temperatures, and using ion-exchange resins to deplete cyanate from the neutral urea solutions. In addition, use of amine-containing buffers (e.g., Tris) should also help scavenge cyanate. Acidification is often done to halt a tryptic digestion, but it will also limit further carbamylation by protonating amino groups.

Although there may be functional roles for *in vivo* oxidation of tryptophan and methionine, most often these modifications are observed as a result of sample handling. Exposure to oxidants can occur while running a SDS-PAGE gel [96, 97] or even from reaction of ozone from outside air with thin dry layers of samples during MALDI preparation [98]. The extent of modification can be limited by reducing exposure to oxygen (e.g., purging samples with argon before extensive digestion periods). Obviously, exposure to oxidizing chemicals such as sodium periodate (used in the so-called glyco-capture method [71]) or performic acid (used for cleaving disulfide bonds) will cause extensive oxidation [99]. Oxidation of methionine typically adds a single oxygen, and in low-energy CID neutral losses of 64 Da ( $\text{HSOCH}_3$ ) are often observed as satellite peaks below any fragment ion that contains the oxidized

methionine [96]. Even more extensive oxidation can lead to an additional oxygen (+32 Da) added onto the sulfur, but this is not seen as frequently. Oxidation of tryptophan is more complicated, and can result in mass increases of 3.9949, 15.9949, 19.9898, and 31.9898 Da [97].

Deamidation can occur at asparagines, glutamine, and carbamidomethylated cysteine residues. When glutamine is located at the N-terminus of a peptide (or protein) the alpha-amino group undergoes a nucleophilic attack of the side-chain amide resulting in the loss of ammonia and formation of a cyclic five-membered ring (pyroglutamic acid) [100, 101]. In buffers typically used for tryptic digestion, the half-life of this reaction is in the range of several hours to a day, so for a standard overnight tryptic digestion a substantial fraction of peptides with N-terminal glutamine will have converted. In a very similar fashion, N-terminal carbamidomethylated cysteine will also cyclize and lose ammonia to form (*R*)-5-oxoperhydro-1,4-thiazine-3-carbonyl residue [102]. The half-life for this reaction is also on the order of hours to days and is often seen in tryptic digests. N-Terminal asparagine does not undergo this reaction, since it would result in an unfavorable four-membered ring structure. However, when asparagine is not located at the terminus it can undergo a nucleophilic attack of the amide nitrogen on the C-terminal side of asparagine forming a succinimidyl intermediate that can then re-open as aspartic acid or isoaspartic acid [103]. The rate at which this reaction occurs is dependent on the steric hindrance introduced by the residue located C-terminal to the asparagine. Sequences containing Asn–Gly are particularly prone to this *ex vivo* modification. Internal glutamines can also deaminate, but the rate of reaction is orders of magnitude slower [63].

Even if purified “proteolytically correct” peptides enter a mass spectrometer, the mass spectrometer source may generate ions other than the desired intact protonated species. Either by design or accident, in-source CID [104] can occur when ions are accelerated with higher energy through regions of high pressure (e.g., use of high cone voltages for certain source designs). When these fragment ions are detected in a data-dependent scan mode, MS/MS spectra of these in-source fragments can be collected and a database search identifies them as “proteolytically incorrect” peptides. The most labile bonds are preferentially cleaved via in-source CID; for example, MS/MS are often collected on fragments containing an N-terminal proline (i.e., production of a  $\gamma$  ion via in-source cleavage at proline). Protons typically provide the positive charge for peptide ions; however, contaminated solvents or incomplete desalting can result in peptide charging via sodium, or other adventitious cations. Not only will this lead to incorrect mass determinations, but the MS/MS spectra will exhibit atypical fragment ions [105–107].

Proteolysis was mentioned earlier in the context of an *in vivo* post-translational event that is often of considerable biological interest; however, inadvertent proteolysis can also occur *ex vivo* through experimental mishandling. It is well known that cell lysis can release proteases from subcellular compartments and one typically disrupts cells only in the presence of a variety of protease inhibitors where the sample is worked-up at reduced temperature. In the case of trypsin it is thought that autolysis results in a protease that is still active, but with reduced specificity for arginine and lysine. Partial methylation of lysine side-chains within trypsin eliminates some of

these cleavage sites, thereby allowing for a prolonged use of trypsin. Even with precautions, a low level of nonspecific cleavages can occur [108]. The goal of achieving complete tryptic digestion has to be balanced against the increased level of nonspecific cleavage.

### 1.3

#### Sample Preparation and Data Acquisition

##### 1.3.1

##### Top-Down Versus Bottom-Up Proteomics

Bottom-up MS/MS methods are based on matching a single peptide to a single MS/MS spectrum. Of course, a given protein is likely to be digested into many different peptides and many of these will be identified, all of them pointing to the identification of the same gene product. The difficulty is that a single gene can give rise to many different proteins, either through gene splicing, proteolytic processing, or a variety of other PTMs. However, since the intact protein structure has been destroyed by proteolysis (trypsin), there is no way of reassembling the peptides into a 100% accurate determination of the protein present originally. This fact is one of the more compelling reasons for the promotion of the so-called top-down approach to proteomics [109]. Here, the intact protein is analyzed – measuring the masses and relative amounts of all of the protein variants and acquire structural data on each one individually. Clearly, this is the most logical route to take; however, the technical difficulties are significant and in many cases insurmountable. Typically, these experiments can only be done with the most expensive instrumentation (i.e., FT-MS), and one can only apply the technique to the most well-behaved proteins (soluble abundant ones that can be chromatographed in buffers suitable for ESI). Despite the difficulties, the top-down approach is becoming more popular and the identification of thyroglobulin extended the upper mass limit of the top-down to 669 kDa [110]. The bottom-up methods, where proteolytic peptides are analyzed, are likely to be applicable to the majority of biological problems; however, one needs to understand the limitations when attempting to jump from peptide identifications to protein identifications.

##### 1.3.2

##### Shotgun Versus Targeted Proteomics

Data-dependent shotgun analysis is the process whereby MS/MS spectra are acquired for the more abundant precursor ions over time as they elute from an HPLC column. These spectra are then analyzed as described below (Section 1.4.1), where the goal is to identify previously unknown proteins present in a sample. The problem with shotgun analysis for complex proteomic samples is that only the more abundant proteins are identified. To identify lower abundance proteins one needs to fractionate the proteins using, for example, SDS-PAGE [56], multi-dimensional



HPLC [27], gas-phase fractionation [111], isoelectric focusing [112], or extended gradients [113]. Sometimes combinations of these fractionation techniques are used such that a single sample will be subject to mass spectrometric analysis for several days to weeks. The goal is to increase the dynamic range over which protein identifications can be made; however, this process is subject to diminishing returns and the sample throughput is very slow.

In contrast, targeted proteomics is a much more sensitive method that has the goal of verifying the presence and quantity of known proteins within a sample. For each target, one needs to specifically monitor a few tryptic peptides that serve as surrogate measurements for the protein. Ideally these tryptic peptides would readily form from tryptic digestion, exhibit sharp chromatographic peaks, ionize easily, and not contain any confounding amino acid residues or sequences that could lead to variable quantitative results (no methionine or Asn–Gly, for example, that can variably oxidize or deamidate). The most sensitive way to perform targeted proteomics is to carryout SRM using a triple quadrupole (see Section 1.1.2). Setting up an SRM assay for targeted proteins involves determining which peptides are formed by tryptic cleavage and identifying those peptides that produce the most abundant precursor ions and their charge states, and then subjecting those precursors to CID and acquiring the MS/MS data. From these spectra one would choose the product ions to monitor – usually the more abundant  $\gamma$ -type ions, preferably at higher  $m/z$  than the precursor ion where there is less background. The assay is ready to use once several transitions (precursor–product ion pairs) have been established for each peptide to be monitored in a set of samples. Development of these SRM assays is time-consuming; however, there is an initiative called the SRM Atlas that has the goal of predetermining assays for every open reading frame from various species [114]. Such an atlas has already been completed for yeast [115], which allows for targeted SRM experiments to be performed without extensive assay development time.

### 1.3.3

#### **Enzymatic Digestion for Bottom-Up Proteomics**

The protease most frequently used is trypsin, which cleaves on the C-terminal side of arginine and lysine. This sounds like a simple rule, but there are a number of nuances. Usually the rule for trypsin also includes the prohibition of cleavages N-terminal to proline; however, there is growing evidence [116] that this cleavage reaction can occur with very slow kinetics. Sometimes trypsin will not cleave at certain arginine or lysine sites, which may be due to having stopped the proteolysis too soon – a process called “limited proteolysis” where the most susceptible bonds are cleaved first (at the “hinges” and “fringes” of a folded protein). Or sometimes trypsin cleavage is slowed or prevented by the presence of surrounding acidic residues. Also, it needs to be kept in mind that trypsin is not an exopeptidase. When there is a short series of contiguous arginine or lysine residues (e.g., the sequence ELVISKRRISQ-ING), trypsin will cleave at one of the sites, thereby producing two new peptides that contain additional cleavage sites at the N- and C-termini (e.g., ELVISKK and

RISQING). However, these potential cleavage sites at or near the termini of the resulting peptides are not amenable to further cleavage. Finally, it should be noted that trypsin is capable of nonspecific cleavages at a very low level [108], which is a problem that becomes worse with prolonged incubation times. Trypsin autolysis (self-digestion) results in a slightly damaged protease with reduced specificity. To prevent this there are a number of vendors that sell trypsin that has been partially methylated on lysine – the sites of self-immolation. Nonspecific cleavages can be a significant source of background when working with samples that possess a wide dynamic range of protein concentration (e.g., blood plasma). In addition to the very high abundance fully tryptic peptides (cleavage at arginine or lysine at each end of the peptide), the low level of semi-tryptic peptides (one end produced by nonspecific cleavage) will still be more abundant than the fully tryptic peptides derived from low-level proteins. Trypsin is not perfect.

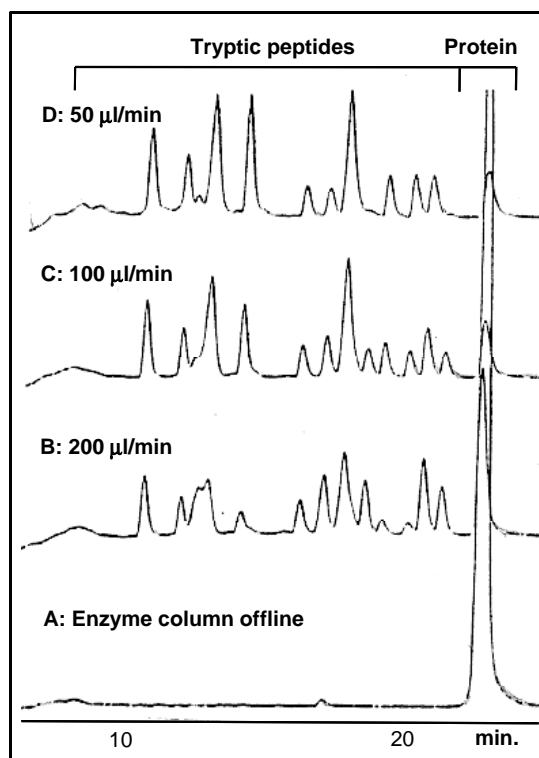
Other imperfect, but useful, enzymes include Lys-C, Lys-N, Asp-N, and Glu-C, which as their names imply, cleave on the C-terminal side of Lys, the N-terminal side of Lys, the N-terminal side of Asp, and the C-terminal side of Glu. Lys-C enzymes are commercially available from at least two biological sources (*Achromobacter lyticus* and *Lysobacter enzymogenes*) [117] and will generally produce larger peptides than trypsin. Similarly, Lys-N is an enzyme isolated from the mushroom *Grifola frondosa*, which cleaves on the N-terminal side of Lys [118, 119]. Asp-N protease cleaves at aspartic residues around 200 times faster than at glutamic acid, which means that some Glu-N activity will be seen, especially at higher enzyme/substrate ratios and with prolonged incubation time. Likewise, Glu-C will exhibit some Asp-C activity [120]. As with trypsin, one would expect to find low levels of nonspecific cleavages using these or any other enzyme. Most other enzymes, such as pepsin, chymotrypsin, subtilisin, or thermolysin, do not have reliable cleavage specificity. They can cleave at many different residues and will often produce peptides with ragged ends. There are chemical cleavage methods available, too, but the ones with the greatest specificity are those that cleave at the rarest amino acids (methionine, cysteine, and tryptophan), and therefore on average produce larger peptides. Large peptides can be good or bad. Production of a few larger peptides for each protein results in less complex mixtures for analysis, and theoretically would improve the ability to identify lower abundance proteins. On the other hand, large peptides can be more difficult to chromatograph, fragment, analyze, and identify.

#### 1.3.4

#### Liquid Chromatography and Capillary Electrophoresis for Mixtures in Bottom-Up

The analysis of peptide mixtures obtained from enzymatic digests of proteins is best performed by coupling liquid chromatography with MS (liquid chromatography mass spectrometry LC-MS). The most common approach utilizes reverse-phase (C18) columns with ESI for online analysis. Commercial columns are available ranging in size from narrow bore (1–2 mm inner diameter) to capillary (above 50  $\mu\text{m}$  inner diameter). Thus, users can match the column loading capacity with the sample size [121]. Greater overall sensitivity is achieved using the narrowest bore columns;

however, larger bore columns tend to be more robust and easier to use (more reproducible retention times, less plugging, and flow rates that are easier to manage). Fortunately, HPLC manufacturers have come out with suitable pumps and fittings that make it much easier to work with packed capillaries. Coupling capillary electrophoresis to MS (capillary electrophoresis mass spectrometry CE-MS) has been less popular due to limited sample loading compared to liquid chromatography. The advantages of capillary electrophoresis are lower sample consumption, shorter analysis time, and higher separation efficiencies. These benefits were shown in the analysis of a tryptic digest of human cerebrospinal fluid [122]. The high-throughput digestion of proteins is achieved by coupling of immobilized enzyme columns in tandem with the reverse-phase columns [123]. The interaction time of proteins with the immobilized enzyme phase is controlled by varying the flow rate through the enzyme column, which could be useful for digesting proteins resistant to proteolysis (Figure 1.8).



**Figure 1.8** HPLC tryptic map of horse cytochrome *c* using a  $2.1 \times 150$  mm Vydac-C18 column at a flow rate of 200 ml/min over 10 column volumes. A  $2.1 \times 30$  mm Trypsin-POROS column was equilibrated at  $50^\circ\text{C}$ . A

digestion buffer of 25 mM Tris-HCl, pH 8.5 containing 10 mM  $\text{CaCl}_2$  was used with flow rates of (B) 200, (C) 100, (D) 50 ml/min. The protein digestion increases by decreasing the flow rate of buffer through the enzyme column.

Although ESI is usually used, it is possible to couple liquid chromatography or capillary electrophoresis to MS using MALDI [124]. Peptides eluting from a reverse-phase capillary column are deposited off-line on a MALDI sample stage, and subsequently analyzed using appropriate software and robotics. This is potentially a high-throughput method where several HPLC and MALDI plate spotters could prepare plates to be analyzed by a single MALDI mass spectrometer. By decoupling the HPLC from the mass spectrometer in this manner, it is possible to interrogate an HPLC run several times, possibly performing MS/MS on various precursors, all at a leisurely pace. In practice, LC-MALDI-MS has not been very popular, due to the technical difficulty of making homogeneous sample-matrix spots. It can also be difficult to troubleshoot HPLC problems when the detector is off-line.

## 1.4

### Data Analysis of LC-MS/MS (or CE-MS/MS) of Mixtures

#### 1.4.1

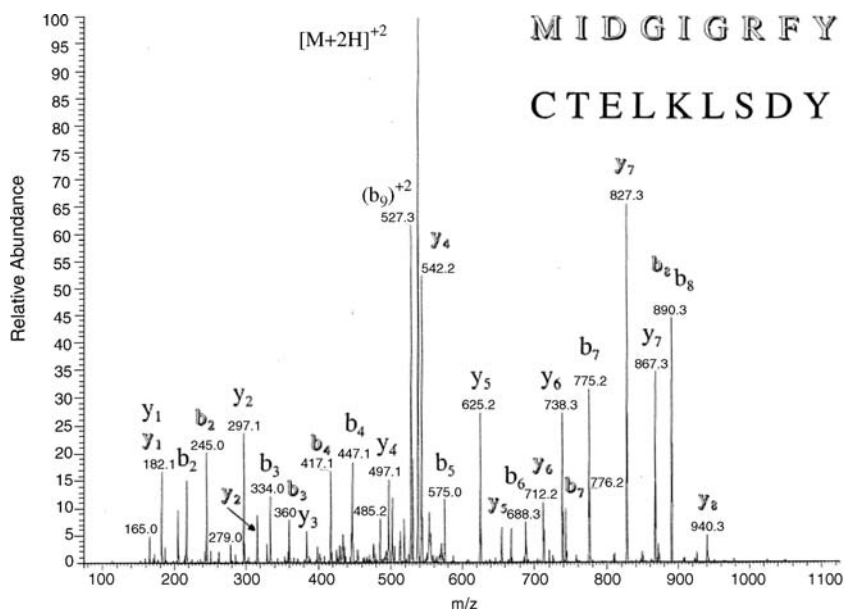
##### Identification of Proteins from MS/MS Spectra of Peptides

Mass mapping was the first high-throughput MS method developed for protein identification, where the general idea was to compare observed molecular weights of tryptic peptides with those calculated from a protein sequence database [125]. This procedure generally requires purified proteins (e.g., in-gel digests from two-dimensional gel spots) and is most rapidly performed using a simple MALDI-TOF instrument. However, for more complex samples containing more than two or three proteins, data-dependent shotgun analyses acquiring many MS/MS spectra have become typical. In order to analyze all of this data, each of the MS vendors has developed (or licensed) their own software for seamlessly moving from raw data files to protein identifications; however, most research groups do not find this solution satisfactory. Either the software is inadequate, not portable to different operating systems, or a single workflow is desired that can encompass data from different mass spectrometers regardless of the vendor. Hence, there has been a move towards open-source software solutions. The description that follows is based on the general flow used by one of the open-source packages, the Trans-Proteomic Pipeline (TPP), which involves (i) extracting MS/MS spectra from raw binary data files, (ii) performing database searches, (iii) validating the peptide to spectrum matches, and finally (iv) validating the protein identification. For more in-depth details on how to use the TPP, a tutorial has recently been published [126] and for the casual user an Internet version of the TPP is being developed at the Australian Proteomics Computational Facility ([www.apcf.edu.au](http://www.apcf.edu.au)).

Data files produced by mass spectrometers from different vendors have proprietary formats that need to be converted to a common open format. Some formats are flexible enough that they can capture most of the information contained in the original raw file (e.g., mzXML [127]), which is useful if subsequent processing of the

data involves MS spectra, in addition to MS/MS spectra. However, the file sizes for these open formats tend to be larger than the original raw binary file, so some laboratories favor smaller and simpler text file formats (e.g., *dta* or *mgf* files) that only contain fragment ion  $m/z$  and intensity values with headers containing limited information such as scan number, precursor  $m/z$ , and charge state. Several conversion programs have been written for each vendor's data files (e.g., ReAdW for conversion of Thermo raw files); however, there has been progress made within the ProteoWizard set of open-source tools and libraries [128] to support reading of multiple vendor files. In most cases, the converted files tend to be faithful reproductions of the original raw file; however, it seems likely that in the future conversion programs will optionally be able to perform some level of data enhancement. First, for low-resolution MS/MS spectra one could remove some of the noise via a moving window filter that, for example, only retains the four most intense ions within a 60  $m/z$  window. Using such a filter, an *mgf* file can be reduced in size by as much as 90%, and provide improved search results [129]. An exact conversion of the raw file to an open-source format associates the MS/MS spectra with the low intensity precursor mass measurement that triggered the MS/MS acquisition. A better approach is to take an intensity weighted average of the  $m/z$  measurements for all of the isotope peaks, for all precursor charges that are present, and for all of the mass spectra acquired across the chromatographic peak. This recalculation of the precursor mass is particularly useful when the single-stage mass spectrum level ( $MS^1$ ) is acquired at high resolution and mass accuracy [130]. Also, for high-resolution and high-mass-accuracy MS/MS spectra it seems that a significant improvement in database search results could be obtained by deisotoping the fragment ions (transforming isotope clusters into a single value corresponding to the  $^{12}C$  peak). Ideally, all of this data manipulation could occur at the point where the raw files are converted to an open format.

The next step is to perform a database search (also see Chapter 14). SEQUEST [25] was the first to perform a database search without any user-derived interpretation. The University of Washington, where SEQUEST was invented, gave Finnigan (now Thermo Corporation) an exclusive license to sell the software, along with the requirement that they vigorously defend the intellectual property. This briefly held up the development of alternative database search software, but others eventually came along. In addition to some for-profit software such as Mascot [24] and Phenyx [131], there are now a variety of freely available programs (e.g., Tandem [132], MyriMatch [133], OMSSA [134], and InsPecT [135]). With few exceptions, these programs use the precursor mass and tolerance as a filter to derive a list of candidate sequences from a protein sequence database. Mock spectra are made for each candidate sequence, these are compared to the real MS/MS spectrum, and scores are assigned to how well they match. The top-scoring match is the winner. Despite having identical purposes, and often similar algorithms, each program will produce slightly different results. There are at least two reasons for these differences. (i) Each program might process the real MS/MS spectrum differently (e.g., by de-noising or eliminating some of the fragment ions in various ways.) (ii) Each program will score the match between mock and real spectra with different equations, models, or



**Figure 1.9** Simultaneous CID of  $(M + H)^{2+}$  precursor ion containing two isobaric peptides with sequences of MIDGIGRFY and CTELKLSDY recorded on an ion-trap mass spectrometer. Both peptides were correctly identified by the PepSearch program and were correlated to the influenza A virus nucleoprotein.

methods. Each search engine has its own search result output format that needs to be converted to a single common format (e.g., pepXML) in order to be accessed in the next step – validation. To illustrate the accuracy of protein sequence identification through automated database searching, MS/MS spectrum obtained for a mixture containing two isobaric peptides was analyzed [136]. The CID mass spectrum shown in Figure 1.9 contains the *b*- and *y*-type ions for two different sequences, and despite this complexity, the program correctly identified both sequences associated with the isobaric peptides.

As just described, all MS/MS spectra will each have a top-scoring candidate sequence; however, the difficulty is determining whether it is correct [137]. Early on, validation was done manually by expert review or by using a simple score threshold that was assumed to reliably bisect correct assignments from incorrect ones. This approach was refined by the use of metrics that show how much better the top score for a given MS/MS spectrum was from all of the other candidate sequences. For SEQUEST, this was simply a score difference between the first and second ranked sequence candidates; later, expectation values were calculated. The latter were meant as estimates for how many times one would expect to achieve the first ranked score by chance. More rigorous validation methods were later developed for large MS/MS data sets (e.g., LC-MS/MS), which use either target-decoy or empirical Bayesian methods. It is now relatively common for proteomics researchers to estimate error rates by searching reversed or randomized databases along with the

targeted database [138]. The search results will then contain a number of matches to the randomized database, which are assumed to be false, and database search result scores can be matched to estimated error rates. Alternatively, the idea behind the TPP computer program Peptide Prophet [139] is that a histogram plot of the top scores for all spectra in a LC-MS/MS run (or any large collection of MS/MS spectra) is made from a composite of two distributions. The assumption is that there are two distributions – one for incorrectly identified spectra with a range of low scores and another for correctly identified spectra with a range of high scores. The mathematical best fit of two distributions is then used to determine error rates and probabilities. A good combination of the two approaches is to use the search results of a randomized database to model the distribution of scores for incorrect identifications, which is a concept that has been implemented within current versions of Peptide Prophet. In general, it is essential to have software that can objectively validate database search results, since expert reviews tend to vary with the physiological state of the expert (not to mention that some experts have delusions of adequacy).

The empirical Bayesian approach used by Peptide Prophet can incorporate additional information in order to modify the final probability determination. For example, those peptides that are formed with the anticipated tryptic cleavage specificity are more likely to be correct than those derived from completely nontryptic cleavages. High-mass-accuracy measurements of precursors permit a postsearch evaluation of how the calculated candidate sequence molecular weights cluster. Those sequences whose calculated molecular weights deviate by more than the average are less likely to be correct. Results from multiple search engines [140], presence of an anticipated motif (e.g., *N*-linked glycosylation), and HPLC retention times can all be included in a final probability determination to help with automated validation.

Shotgun bottom-up proteomics is intrinsically a peptide identification technique and protein identification can only be inferred from these identifications. At first, this step sounds like it should be simple, but database redundancies and protein homology often make it difficult to be certain. If a peptide sequence is shared between different proteins, how should one apportion peptide-spectrum match probabilities among the possible protein choices? In general, most protein validation software does this by using principles of parsimony to create the simplest and shortest protein list possible [141–143]. Although reality is rarely simple, this is really the only choice available.

#### 1.4.2

#### ***De Novo* Sequencing**

*De novo* sequencing refers to the process of deriving a peptide sequence directly from the MS/MS spectrum without recourse to any sequence database. Manual *de novo* sequencing can be mentally diverting (see <http://www.abrf.org/ResearchGroups/MassSpectrometry/EPosters/ms97quiz/abrfQuiz.html>); however, this is not practical when confronted with more than a handful of spectra. For larger numbers of

spectra one needs to use automated *de novo* sequencing programs. One of the first was Lutefisk [144], which has since been used to benchmark other *de novo* sequencing programs (PepNovo being a notable open-source example [145]). For a variety of reasons, deriving a single correct sequence exclusively from a MS/MS spectrum is often not possible, either manually or with a computer program:

- i) Some amino acids have identical or nearly identical masses – leucine and isoleucine, glutamine and lysine, and phenylalanine and oxidized methionine.
- ii) Cleavages may be absent between adjacent amino acids. Absence of cleavage between the first and second amino acids of a tryptic peptide is very common.
- iii) Some amino acids have the same mass as pairs of other amino acids (e.g., Gly–Gly is exactly the same as Asn).
- iv) If one can identify a series of ions whose mass differences delineate an amino acid sequence, it remains unclear whether the derived sequence is going from the N- to C-terminus or the other way around. In other words, it is not always clear whether a series of ions are all *b*- or *y*-type fragment ions.

For these reasons, *de novo* sequencing typically results in a short list of candidate sequences that each account for the data to varying degrees.

Why bother performing a *de novo* sequence determination when database search programs and validation tools are so fast and easy to use? Obviously, one reason would be if one was working with a species whose genome has not yet been sequenced [146]. Generally, this method involves generating a list of *de novo* sequences, and then submitting them to a homology search engine where the parameters have been optimized to account for the vagaries and problems associated with MS/MS-derived sequences (e.g., inability to distinguish leucine and isoleucine) [144]. A second reason for performing *de novo* sequencing is that it potentially provides further validation of a database search result. Database searching and *de novo* sequencing are quite orthogonal approaches, and agreement between the two should boost the likelihood of a correct identification [147]. A third reason is that one might be wondering about all of the unmatched spectra (often around 90%) in a typical LC-MS/MS experiment. For example, one could find that many peptides have been carbamylated due to bad urea or that the autosampler exhibits severe carryover problems from prior users studying a different species not present in the database that was searched. A fourth application of *de novo* sequencing is to help identify high-quality spectra, particularly ones that had not been matched to a database sequence. Finding the “sequenceable” spectra is the same as finding the “high-quality” spectra.

## 1.5

### MS of Protein Structure, Folding, and Interactions

The study of structural dynamics of proteins and noncovalent interactions ideally requires analytical methods that enable capturing events occurring over timescale from nanoseconds to seconds, while simultaneously monitoring specific sites

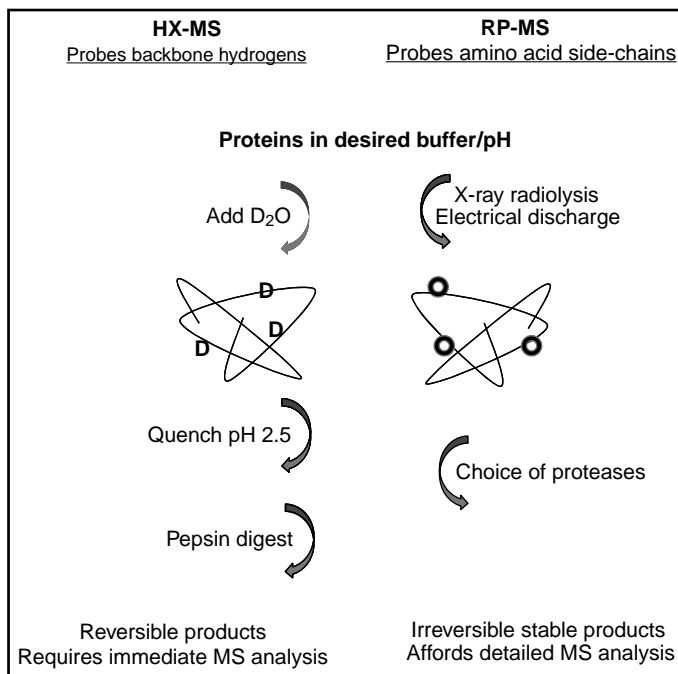


within the structure of each participant. Current analytical techniques do not provide these capabilities individually. The fast timescale for monitoring macromolecular motions are achieved by spectroscopy-based methods, which provide global structural information. Other techniques such as the X-ray diffractometry or nuclear magnetic resonance (NMR) spectroscopy with capabilities of defining the location of individual atoms do not provide the fast timescale required to capture many events. The process of forming quality crystals, and the need for high-purity samples with good solubility properties further limit the applicability of X-ray and NMR for many studies. The need for solution-based structural characterization of proteins and protein interactions prompted considerable recent attention to the development of chemical probes combined with the benefits of MS analysis (i.e., high-throughput identification of proteins at low levels). A recent book [148] provides a thorough and updated review of a wide range of techniques applied for structural elucidation of proteins and their interactions, and a brief summary of two methods providing high structural resolution is presented below.

### 1.5.1

#### **Methods to Mass-Tag Structural Features**

Hydrogen–deuterium exchange MS (hydrogen–deuterium exchange mass spectrometry HX-MS) is routinely used to probe protein structure, conformation, and dynamics [149–151]. This method measures hydrogen atoms located at peptide amide linkages (i.e., backbone amide hydrogens) with an exchange half-life of seconds to several weeks depending on their solvent accessibility. The HX-MS approach starts with proteins in a physiological buffer at room temperature to conserve their native structure (Scheme 1.1). Deuterium oxide ( $D_2O$ ) is added in excess of 10- to 20-fold and the onset of the exchange reaction is recorded. The exchange reaction is quenched at set time points by adjusting the pH to 2.5 (using an excess of protonated buffers) where the exchange rate drops to its minimum. Further reduction of the exchange rate is achieved by cooling the solution. The quenched timepoints can be snap-frozen in liquid nitrogen to be analyzed later, as no measureable back-exchange can be detected from frozen samples stored at  $-70^\circ C$ . At  $0^\circ C$  and pH 2.5, the deuterium label reverts to hydrogen in a back-exchange process with a half-life of approximately 1 h, providing sufficient time for MS analysis. After the hydrogen exchange is completed and the reaction is quenched, the protein solution is analyzed by LC-ESI to measure the overall uptake of deuterium. A portion of protein is digested with pepsin to monitor the localized deuterium uptake. The incorporation of deuterium as a function of time is plotted for the intact protein as well as the proteolytic peptides. The main disadvantages of the HX-MS are that the exchanged products are reversible with a limited lifetime and the possibility of the back exchange reactions introduces some error in measurements. Further, the low pH condition of the quench reaction limits the choice of proteases. The resolution of hydrogen exchange is limited to the size of the peptic peptides that are generated, as residue resolution based on CID is not



**Scheme 1.1** Experimental procedures for HX-MS and RP-MS.

possible due to proton scrambling [152]. However, there have been recent results indicating that scrambling does not occur when ETD is used for peptide fragmentation [153, 154], which could make it possible to measure hydrogen exchange rates for individual residues within a protein.

The protein structure could alternatively be evaluated through the solvent accessibility of the amino acid side-chains. The requirement for high structural resolution and the fast timescale of reactions prompted the use of hydroxyl radical as the ideal chemical probe that is similar in size to the water molecule with a diameter of  $2.5 \text{ \AA}^2$ . Time-resolved hydroxyl radical protein footprinting employing MS was developed over a decade ago by applying synchrotron X-ray radiolysis [155, 156] or an electrical discharge source [157] to effect the oxidation of proteins on millisecond timescales. These approaches, which are referred to as radical probe MS (radical probe mass spectrometry RP-MS), have since been successfully applied to the analysis of protein structure, protein folding, and protein–protein interactions [158]. Hydroxyl radicals induce oxidative modification of a number of amino acid side-chains in the range of  $10^9$  to  $10^{10} \text{ M}^{-1} \text{ s}^{-1}$ , which is sufficiently fast for studies of protein folding and interaction dynamics. Further, the reactive hydroxyl radical probe originates in water at physiological pH without the need for other chemicals. Other advantages of the oxidative labeling are that the products are

stable, and this affords the use of a wide range of proteases and the application of a number of MS experiments.

The RP-MS approach identifies the site of amino acid oxidation and this information combined with the quantitative measure of the level of oxidation is used to map the solvent accessibility of the side-chains across a protein's surface. For structural and conformational studies, oxidation is kept to approximately 30–50% for the whole protein, in order to avoid forming degraded and cross-linked products. The timescale and the extent of reactions could then be used to monitor the onset of oxidative damage of proteins in relation with various diseases and aging. The application of RP-MS to studies of the onset of damage was first reported in 2005 for the protein  $\alpha$ -crystallin, which demonstrated that different regions of a protein could exhibit different levels of susceptibility to oxidative damage [159]. These types of structural information are important in designing targeted therapeutics to prevent or control oxidative damage associated with a range of diseases.

A significant amount of information about a protein structure (i.e., solvent accessibility surface (SAS) of backbone hydrogen atoms and amino acid side-chains) is obtained through chemical labeling and MS protocols. These types of information prompted the development of a docking algorithm, PROXIMO, to propose structures for protein complexes based on those for their component molecules using RP-MS data [160]. The performance of the algorithm was successfully validated for a series of protein complexes, including the ribonuclease S-complex with several correctly identified conformers that deviated from the X-ray crystal structure with root mean square deviation values of 0.45 and 1.26 Å<sup>2</sup> (i.e., all within the 2.5 Å<sup>2</sup> SAS resolution of the RP-MS experimental structure). It could be envisioned that the application of chemical labeling in conjunction with computer algorithms will be valuable for solution-based structural analysis of proteins.

While the discovery of ESI has revolutionized the analysis of proteins and their noncovalently bound complexes, the question of whether or not the solution-based structures of proteins are preserved during the ESI process has been subject to numerous studies. A recent perspective [161] supports evidence for retention of native structure for some large proteins [162, 163]. Meanwhile, the authors propose that after the initial desolvation steps during the ESI process, structures of globular proteins of cytochrome *c* and ubiquitin undergo several transitions within picoseconds to seconds, which include collapse of the side-chains, unfolding and refolding steps that result in multiple conformers in the gas phase. Obviously, future studies for a range of proteins are required in order to validate this approach as a structural analysis tool.

As just discussed, various protein conformers could be generated during the ESI and these intermediate conformers are rarely isolated in the solution phase. Therefore, the gas-phase environment provides an ideal opportunity to investigate the subtle changes in protein structures during folding or binding transitions. Major advances in ion mobility MS (ion mobility mass spectrometry IMS) since the 1990s have made it possible to study small differences in structures of conformers in the gas phase based on their mobilities through a gas [164]. IMS has emerged as a powerful method for structural analysis of proteins and their complexes.

## 1.6

### Conclusions and Perspectives

MS of proteins has made major strides over the past few decades. Of key importance was the development of new methods for the ionization of peptides and proteins (FAB, MALDI, and ESI), as well as new high-accuracy and high-resolution mass analyzers. Improvements in computer speed and data storage capacity, plus the rapid accumulation of protein and DNA sequence databases over this time was critical for enabling what has become known as “shotgun proteomics.” The latter was also dependent on enhanced understanding of gas-phase peptide ion fragmentation and software tools for matching mass spectral data to database-derived sequences. All of this led to a considerable amount of irrational exuberance (e.g., claims of sequencing the human proteome [165]), which has now largely subsided to a point where mostly what remains are serious people studying real problems. For proteomics, the next big area seems to be targeted proteomics, which has an improved dynamic range over shotgun analysis. However, targeted proteomics is also in danger of being excessively promoted and has a number of problems that need to be solved (e.g., how to estimate false discovery rates, how to handle targeted peptides that elute in more than one peak, how to target peptides that may or may not be modified, how to resolve contradictory quantitative results from different peptides from the same protein, etc.). In short, the shotgun proteomics wave has crashed on the beach, the targeted proteomics wave is coming, but regardless of the level of enthusiasm, these tools will continue to be useful for those who know their limitations and how to use them properly.

The identification of gene products is only one facet of proteomics. Identification and quantitation of PTMs is important for full characterization of a protein or proteome. Noncovalent structural aspects of proteins (folding, solvent accessibility, binding sites, etc.) can also be determined using MS. For the most part, PTM and noncovalent analysis is simply a matter of basic protein chemistry that has been considerably enabled by MS. For example, in the olden days protein chemists measured hydrogen exchange rates by measuring tritium incorporation; with mass spectrometers one can now more easily and safely measure the incorporation of the stable heavy isotope of hydrogen instead. Instead of determining disulfide bonds by comparing electrophoresis migration before and after bond cleavage, one now measures the molecular weights. To summarize, protein chemistry requires MS.

### References

- 1 Langewiesche, W. (2005) The wrath of Khan. *The Atlantic Magazine*, (Nov), 62–85.
- 2 Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17, 994–999.
- 3 Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics*, 1, 376–386.

- 4 Yao, X., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. (2001) Proteolytic  $^{18}\text{O}$  labeling for comparative proteomics. *Analytical Chemistry*, **73**, 2836–2842.
- 5 Wool, A. and Smilansky, Z. (2002) Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting. *Proteomics*, **2**, 1365–1373.
- 6 Meng, C.K., Mann, M., and Fenn, J.B. (1988) Electrospray ionization of some polypeptides and small proteins. Proceedings 36th ASMS Conference, San Francisco, CA, pp. 771–772.
- 7 Mann, M., Meng, C.K., and Fenn, J.B. (1988) Parent mass information from sequences of peaks of multiply charged ions. Proceedings 36th ASMS Conference, San Francisco, CA, pp. 1207–1208.
- 8 Karas, M., Bachmann, D., and Hillenkamp, F. (1985) Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry*, **57**, 2935–2939.
- 9 Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes*, **78**, 53–68.
- 10 Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, **60**, 2299–2301.
- 11 Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988) Protein and polymer analyses up to  $m/z$  100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2**, 151–153.
- 12 McLafferty, F.W., Todd, P.J., McGilvery, D.C., and Baldwin, M.A. (1980) High-resolution tandem mass spectrometry (MS/MS) of increased sensitivity and mass range. *Journal of the American Chemical Society*, **102**, 3360–3363.
- 13 Yost, R.A. and Enke, C.G. (1978) Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society*, **100**, 2274–2275.
- 14 Douglas, D.J., Frank, A.J., and Mao, D. (2004) Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews*, **24**, 1–29.
- 15 March, R.E. (1997) An introduction to quadrupole ion trap mass spectrometry. *Journal of Mass Spectrometry*, **32**, 351–369.
- 16 Marshall, A.G., Hendrickson, C.L., and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews*, **17**, 1–35.
- 17 Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, **72**, 1156–1162.
- 18 Perry, R.H., Cooks, R.G., and Noll, R.J. (2008) Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrometry Reviews*, **27**, 661–699.
- 19 Cotter, R.J. (1994) Time-of-flight mass spectrometry, in *Basic Principles and Current State*, American Chemical Society, Columbus, OH, pp. 16–48.
- 20 Vestal, M.L. and Campbell, J.M. (2005) Tandem time-of-flight mass spectrometry. *Methods in Enzymology*, **402**, 79–108.
- 21 Morris, H.R., Paxton, T., Panico, M., McDowell, R., and Dell, A. (1997) A novel geometry mass spectrometer, the Q-TOF, for low-femtomole/attomole-range biopolymer sequencing. *Journal of Protein Chemistry*, **16**, 469–479.
- 22 Hager, J.W. and Yves Le Blanc, J.C. (2003) Product ion scanning using a Q-q-Q<sub>linear</sub> ion trap (QTRAP<sup>TM</sup>) mass spectrometer. *Rapid Communications in Mass Spectrometry*, **17**, 1056–1064.
- 23 Clauser, K.R., Baker, P., and Burlingame, A.L. (1999) Role of accurate mass measurement ( $\pm 10$ ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **71**, 2871–2882.
- 24 Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching

- sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
- 25 Eng, J.K., McCormack, A.L., and Yates, J.R. III, (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.
  - 26 Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Strittmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K.-K., Zhao, R., and Smith, R.D. (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11049–11054.
  - 27 Washburn, M.P., Wolters, D., and Yates, J.R. III, (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.
  - 28 Schwahn, A.B., Wong, J.W.H., and Downard, K.M. (2009) Subtyping of the influenza virus by high resolution mass spectrometry. *Analytical Chemistry*, **81**, 3500–3506.
  - 29 Schwahn, A.B., Wong, J.W.H., and Downard, K.M. (2010) Rapid differentiation of seasonal and pandemic H1N1 influenza through proteotyping of viral neuraminidase with mass spectrometry. *Analytical Chemistry*, **82**, 4584–4590.
  - 30 Covey, T.R., Bonner, R.F., Shushan, B.I., Henion, J., and Boyd, R.K. (1988) The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2**, 249–256.
  - 31 Mann, M., Meng, C.K., and Fenn, J.B. (1989) Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, **61**, 1702–1708.
  - 32 Senko, M.W., Beu, S.C., and McLafferty, F.W. (1995) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *Journal of the American Society for Mass Spectrometry*, **6**, 52–56.
  - 33 Zhang, Z. and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, **9**, 225–233.
  - 34 Horn, D.M., Zubarev, R.A., and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, **11**, 320–332.
  - 35 Senko, M.W., Beu, S.C., and McLafferty, F.W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **6**, 229–233.
  - 36 Chen, L., Sze, S.K., and Yang, H. (2006) Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, **78**, 5006–5018.
  - 37 Chen, L. and Yap, Y.L. (2008) Automated charge state determination of complex isotope-resolved mass spectra by peak-target Fourier transform. *Journal of the American Society for Mass Spectrometry*, **19**, 46–54.
  - 38 Maleknia, S.D. and Downard, K.M. (2005) Charge ratio analysis method: approach for the deconvolution of electrospray mass spectra. *Analytical Chemistry*, **77**, 111–119.
  - 39 Maleknia, S.D. and Downard, K.M. (2005) Charge ratio analysis method to interpret high resolution electrospray Fourier transform-ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry*, **246**, 1–9.
  - 40 Maleknia, S.D. and Green, D.C. (2010) eCRAM computer algorithm for implementation of the charge ratio analysis method to deconvolute electrospray ionization mass spectra. *International Journal of Mass Spectrometry*, **290**, 1–8.
  - 41 Bowie, J.H., Brinkworth, C.S., and Dua, S. (2002) Collision-induced

- fragmentations of the  $(M - H)^-$  parent anions of underivatized peptides: an aid to structure determination and some unusual negative ion cleavages. *Mass Spectrometry Reviews*, **21**, 87–107.
- 42 Papayannopoulos, I.A. (1995) The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrometry Reviews*, **14**, 49–73.
- 43 Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, **11**, 601.
- 44 Biemann, K. (1990) Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods in Enzymology*, **193**, 886–887.
- 45 Schlosser, A. and Lehmann, W.D. (2000) Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *Journal of Mass Spectrometry*, **35**, 1382–1390.
- 46 Wysocki, V.H., Tsaprailis, G., Smith, L.L., and Breci, L.A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, **35**, 1399–1406.
- 47 Yu, W., Vath, J.E., Huberty, M.C., and Martin, S.A. (1993) Identification of the facile gas-phase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix-assisted laser desorption time-of-flight mass spectrometry. *Analytical Chemistry*, **65**, 3015–3023.
- 48 Johnson, R.S., Martin, S.A., and Biemann, K. (1988) Collision-induced fragmentation of  $(M + H)^+$  ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes*, **86**, 137–154.
- 49 Johnson, R.S., Martin, S.A., Biemann, K., Stults, J.T., and Watson, J.T. (1987) Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical Chemistry*, **59**, 2621–2625.
- 50 Johnson, R.S. and Biemann, K. (1987) The primary structure of thioredoxin from *Chromatium vinosum* determined by high-performance tandem mass spectrometry. *Biochemistry*, **26**, 1209–1214.
- 51 Zubarev, R.A., Kelleher, N.L., and McLafferty, F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society*, **120**, 3265–3266.
- 52 Stensballe, A., Jensen, O.N., Olsen, J.V., Haselmann, K.F., and Zubarev, R.A. (2000) Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Communications in Mass Spectrometry*, **14**, 1793–1800.
- 53 Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9528–9533.
- 54 Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E.P., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., and Hunt, D.F. (2007) Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2193–2198.
- 55 Good, D.M., Wirtala, M., McAlister, G.C., and Coon, J.J. (2007) Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular and Cellular Proteomics*, **6**, 1942–1951.
- 56 Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, **379**, 466–469.
- 57 Furmanek, A. and Hofsteenge, J. (2000) Protein C-mannosylation: facts and questions. *Acta Biochimica Polonica*, **47**, 781–789.

- 58 Golks, A. and Guerini, D. (2008) The O-linked N-acetylglucosamine modification in cellular signalling and the immune system. "Protein modifications: beyond the usual suspects" review series. *EMBO Reports*, **9**, 748–753.
- 59 Kreppel, L.K., Blomberg, M.A., and Hart, G.W. (1997) Dynamic glycosylation of nuclear and cytosolic proteins: cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *Journal of Biological Chemistry*, **272**, 9308–9315.
- 60 Chalkley, R.J., Thalhammer, A., Schoepfer, R., and Burlingame, A.L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8894–8899.
- 61 Dell, A. and Morris, H.R. (2001) Glycoprotein structure determination by mass spectrometry. *Science*, **291**, 2351–2356.
- 62 Helenius, A. and Aebi, M. (2001) Intracellular functions of N-linked glycans. *Science*, **291**, 2364–2369.
- 63 Robinson, N.E., Robinson, Z.W., Robinson, B.R., Robinson, A.L., Robinson, J.A., Robinson, M.L., and Robinson, A.B. (2004) Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *Journal of Peptide Research*, **63**, 426–436.
- 64 Zaia, J. (2010) Mass spectrometry and glycomics. *Omic*s, **14**, 401–418.
- 65 Alvarez-Manilla, G., Warren, N.L., Atwood, J., Orlando, R., Dalton, S., and Pierce, M. (2010) Glycoproteomic analysis of embryonic stem cells: identification of potential glycomarkers using lectin affinity chromatography of glycopeptides. *Journal of Proteome Research*, **9**, 2062–2075.
- 66 Calvano, C.D., Zamboni, C.G., and Jensen, O.N. (2008) Assessment of lectin and HILIC based enrichment protocols for characterization of serum glycoproteins by mass spectrometry. *Journal of Proteomics*, **71**, 304–317.
- 67 Kaji, H., Saito, H., Yamauchi, Y., Shinkawa, T., Taoka, M., Hirabayashi, J., Kasai, K-ichi., Takahashi, N., and Isobe, T. (2003) Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nature Biotechnology*, **21**, 667–672.
- 68 McDonald, C.A., Yang, J.Y., Marathe, V., Yen, T.-Y., and Macher, B.A. (2009) Combining results from lectin affinity chromatography and glyco-capture approaches substantially improves the coverage of the glycoproteome. *Molecular and Cellular Proteomics*, **8**, 287–301.
- 69 Maleknia, S.D., Treuheit, M.J., Carlson, J.E., Halsall, H.B., and Costello, C.E. (1993) Analysis of heterogeneity of native a1-acid glycoprotein by MADLI-TOFMS. Proceedings of the 41st ASMS Conference, San Francisco, CA, pp. 81a–81b.
- 70 Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R., and Watts, J.D. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature Biotechnology*, **27**, 378–386.
- 71 Zhang, H., Li, X.-j., Martin, D.B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature Biotechnology*, **21**, 660–666.
- 72 Palumbo, A.M. and Reid, G.E. (2008) Evaluation of gas-phase rearrangement and competing fragmentation reactions on protein phosphorylation site assignment using collision induced dissociation-MS/MS and MS<sup>3</sup>. *Analytical Chemistry*, **80**, 9735–9747.
- 73 Alpert, A.J. (2008) Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Analytical Chemistry*, **80**, 62–76.
- 74 Ficarro, S.B., McClelland, M.L., Stukenberg, P.T., Burke, D.J.,



- Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnology*, **20**, 301–305.
- 75 McNulty, D.E. and Annan, R.S. (2008) Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Molecular and Cellular Proteomics*, **7**, 971–980.
- 76 Tsai, C.-F., Wang, Y.-T., Chen, Y.-R., Lai, C.-Y., Lin, P.-Y., Pan, K.-T., Chen, J.-Y., Khoo, K.-H., and Chen, Y.-J. (2008) Immobilized metal affinity chromatography revisited: pH/acid control toward high selectivity in phosphoproteomics. *Journal of Proteome Research*, **7**, 4058–4069.
- 77 Villén, J. and Gygi, S.P. (2008) The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature Protocols*, **3**, 1630–1638.
- 78 Welchman, R.L., Gordon, C., and Mayer, R.J. (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nature Reviews Molecular Cell Biology*, **6**, 599–609.
- 79 Peng, J. (2008) Evaluation of proteomic strategies for analyzing ubiquitinated proteins. *BMB Reports*, **41**, 177–183.
- 80 Peng, J., Schwartz, D., Elias, J.E., Thoreen, C.C., Cheng, D., Marsischky, G., Roelofs, J., Finley, D., and Gygi, S.P. (2003) A proteomics approach to understanding protein ubiquitination. *Nature Biotechnology*, **21**, 921–926.
- 81 Matic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S.C., Tatham, M.H., Hay, R.T., Lamond, A.I., Mann, M., and Vertegaal, A.C.O. (2008) *In vivo* identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an *in vitro* to *in vivo* strategy. *Molecular and Cellular Proteomics*, **7**, 132–144.
- 82 Kim, S.C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., Grishin, N.V., White, M., Yang, X.-J., and Zhao, Y. (2006) Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Molecular Cell*, **23**, 607–618.
- 83 Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.
- 84 Ryle, A.P. and Sanger, F. (1955) Disulphide interchange reactions. *Biochemical Journal*, **60**, 535–540.
- 85 Mikesch, L., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E.P., Shabanowitz, J., and Hunt, D.F. (2006) The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta*, **1764**, 1811–1822.
- 86 Chrisman, P.A., Pitteri, S.J., Hogan, J.M., and McLuckey, S.A. (2005)  $\text{SO}_2^{-}$  electron transfer ion/ion reactions with disulfide linked polypeptide ions. *Journal of the American Society for Mass Spectrometry*, **16**, 1020–1030.
- 87 Black, R.A., Rauch, C.T., Kozlosky, C.J., Peschon, J.J., Slack, J.L., Wolfson, M.F., Castner, B.J., Stocking, K.L., Reddy, P., Srinivasan, S., Nelson, N., Boiani, N., Schooley, K.A., Gerhart, M., Davis, R., Fitzner, J.N., Johnson, R.S., Paxton, R.J., March, C.J., and Cerretti, D.P. (1997) A metalloproteinase disintegrin that releases tumour-necrosis factor- $\alpha$  from cells. *Nature*, **385**, 729–733.
- 88 Guo, L., Eisenman, J.R., Mahimkar, R.M., Peschon, J.J., Paxton, R.J., and Black, R.A., and Johnson, R.S. (2002) A proteomic approach for the identification of cell-surface proteins shed by metalloproteases. *Molecular and Cellular Proteomics*, **1**, 30–36.
- 89 Abrahmsén, L., Tom, J., Burnier, J., Butcher, K.A., Kossiakoff, A., and Wells, J.A. (1991) Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. *Biochemistry*, **30**, 4151–4159.
- 90 Staes, A., Van Damme, P., Helsens, K., Demol, H., Vandekerckhove, J., and Gevaert, K. (2008) Improved recovery of proteome-informative, protein

- N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics*, **8**, 1362–1370.
- 91 Agard, N.J., Maltby, D., and Wells, J.A. (2010) Inflammatory stimuli regulate caspase substrate profiles. *Molecular and Cellular Proteomics*, **9**, 880–893.
- 92 Schilling, O., Barré, O., Huesgen, P.F., and Overall, C.M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature Methods*, **7**, 508–511.
- 93 Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature Methods*, **2**, 193–200.
- 94 Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A., Schilling, O., Kainthan, R.K., Starr, A.E., Foster, L.J., Kizhakkedathu, J.N., and Overall, C.M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature Biotechnology*, **28**, 281–288.
- 95 Stark, G.R., Stein, W.H., and Moore, S. (1960) Reactions of the cyanate present in aqueous urea with amino acids and proteins. *Journal of Biological Chemistry*, **235**, 3177–3181.
- 96 Swiderek, K.M., Davis, M.T., and Lee, T.D. (1998) The identification of peptide modifications derived from gel-separated proteins using electrospray triple quadrupole and ion trap analyses. *Electrophoresis*, **19**, 989–997.
- 97 Perdivara, I., Deterding, L.J., Przybylski, M., and Tomer, K.B. (2010) Mass spectrometric identification of oxidative modifications of tryptophan residues in proteins: chemical artifact or post-translational modification? *Journal of the American Society for Mass Spectrometry*, **21**, 1114–1117.
- 98 Cohen, S.L. (2006) Ozone in ambient air as a source of adventitious oxidation. A mass spectrometric study. *Analytical Chemistry*, **78**, 4352–4362.
- 99 Hirs, C.H.W. (1956) The oxidation of ribonuclease with performic acid. *Journal of Biological Chemistry*, **219**, 611–621.
- 100 Khandke, K.M., Fairwell, T., Chait, B.T., and Manjula, B.N. (1989) Influence of ions on cyclization of the amino terminal glutamine residues of tryptic peptides of streptococcal PepM49 protein: resolution of cyclized peptides by HPLC and characterization by mass spectrometry. *International Journal of Peptide and Protein Research*, **34**, 118–123.
- 101 Sanger, F., Thompson, E.O.P., and Kitai, R. (1955) The amide groups of insulin. *Biochemical Journal*, **59**, 509–518.
- 102 Geoghegan, K.F., Hoth, L.R., Tan, D.H., Borzilleri, K.A., Withka, J.M., and Boyd, J.G. (2002) Cyclization of N-terminal S-carbamoylmethylcysteine causing loss of 17 Da from peptides and extra peaks in peptide maps. *Journal of Proteome Research*, **1**, 181–187.
- 103 Geiger, T. and Clarke, S. (1987) Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimidyl-linked reactions that contribute to protein degradation. *Journal of Biological Chemistry*, **262**, 785–794.
- 104 Williams, J.D., Flanagan, M., Lopez, L., Fischer, S., and Miller, L.A.D. (2003) Using accurate mass electrospray ionization-time-of-flight mass spectrometry with in-source collision-induced dissociation to sequence peptide mixtures. *Journal of Chromatography A*, **1020**, 11–26.
- 105 Russell, D.H., McGlohon, E.S., and Mallis, L.M. (1988) Fast-atom bombardment-tandem mass spectrometry studies of organo-alkali-metal ions of small peptides. Competitive interaction of sodium with basic amino acid substituents. *Analytical Chemistry*, **60**, 1818–1824.
- 106 Grese, R.P., Cerny, R.L., and Gross, M.L. (1989) Metal ion-peptide interactions in the gas phase: a tandem mass spectrometry study of alkali metal cationized peptides. *Journal of the American Chemical Society*, **111**, 2835–2842.
- 107 Lee, S.-W., Kim, H.S., and Beauchamp, J.L. (1998) Salt bridge chemistry applied to gas-phase peptide sequencing: selective fragmentation of

- sodiated gas-phase peptide ions adjacent to aspartic acid residues. *Journal of the American Chemical Society*, **120**, 3188–3195.
- 108 Picotti, P., Aebersold, R., and Domon, B. (2007) The implications of proteolytic background for shotgun proteomics. *Molecular and Cellular Proteomics*, **6**, 1589–1598.
- 109 Sze, S.K., Ge, Y., Oh, H., and McLafferty, F.W. (2002) Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1774–1779.
- 110 Karabacak, N.M., Li, L., Tiwari, A., Hayward, L.J., Hong, P., Easterling, M.L., and Agar, J.N. (2009) Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Molecular and Cellular Proteomics*, **8**, 846–856.
- 111 Yi, E.C., Marelli, M., Lee, H., Purvine, S.O., Aebersold, R., Aitchison, J.D., and Goodlett, D.R. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis*, **23**, 3205–3216.
- 112 Cargile, B.J., Talley, D.L., and Stephenson, J.L. (2004) Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. *Electrophoresis*, **25**, 936–945.
- 113 Davis, M.T. and Lee, T.D. (1997) Variable flow liquid chromatography-tandem mass spectrometry and the comprehensive analysis of complex protein digest mixtures. *Journal of the American Society for Mass Spectrometry*, **8**, 1059–1069.
- 114 Moritz, R. (2010) The mechanics of detecting and quantifying the complete human proteome. Human Proteome World Congress 2010, Sydney, abstract OS073
- 115 Picotti, P., Lam, H., Campbell, D., Deutsch, E.W., Mirzaei, H., Ranish, J., Domon, B., and Aebersold, R. (2008) A database of mass spectrometric assays for the yeast proteome. *Nature Methods*, **5**, 913–914.
- 116 Rodriguez, J., Gupta, N., Smith, R.D., and Pevzner, P.A. (2008) Does trypsin cut before proline? *Journal of Proteome Research*, **7**, 300–305.
- 117 Jekel, P.A., Weijer, W.J., and Beintema, J.J. (1983) Use of endoproteinase Lys-C from *Lysobacter enzymogenes* in protein sequence analysis. *Analytical Biochemistry*, **134**, 347–354.
- 118 Taouatas, N., Heck, A.J.R., and Mohammed, S. (2010) Evaluation of metalloendopeptidase Lys-N protease performance under different sample handling conditions. *Journal of Proteome Research*, **9**, 4282–4288.
- 119 Hohmann, L., Sherwood, C., Eastham, A., Peterson, A., Eng, J.K., Eddes, J.S., Shteynberg, D., and Martin, D.B. (2009) Proteomic analyses using *Grifola frondosa* metalloendoprotease Lys-N. *Journal of Proteome Research*, **8**, 1415–1422.
- 120 Drapeau, G.R., Boily, Y., and Houmar, J. (1972) Purification and properties of an extracellular protease of *Staphylococcus aureus*. *Journal of Biological Chemistry*, **247**, 6720–6726.
- 121 Tomer, K.B., Moseley, M.A., and Deterding, L.J., and Parker, C.E. (1994) Capillary liquid chromatography/mass spectrometry. *Mass Spectrometry Reviews*, **13**, 431–457.
- 122 Wetterhall, M., Palmblad, M., Håkansson, P., Markides, K.E., and Bergquist, J. (2002) Rapid analysis of tryptically digested cerebrospinal fluid using capillary electrophoresis-electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry. *Journal of Proteome Research*, **1**, 361–366.
- 123 Maleknia, S.D., Mark, J.P., Dixon, J.D., Elicone, C.P., McGuinness, B.F., Fulton, S.P., and Afeyan, N.B. (1994) Real-time protein mapping utilizing immobilized enzyme columns. Proceedings of the 42nd ASMS Conference, Chicago, IL, pp. 304–305.
- 124 Peters, E.C., Brock, A., Horn, D.M., Phung, Q.T., Ericson, C., Salomon, A.R., Ficarro, S.B., and Brill, L.M. (2002) An

- automated LC-MALDI FT-ICR MS platform for high-throughput proteomics. *LC GC Europe*, **15**, 423–428.
- 125 Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 5011–5015.
- 126 Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J.K., Martin, D.B., Nesvizhskii, A.I., and Aebersold, R. (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, **10**, 1150–1159.
- 127 Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, **22**, 1459–1466.
- 128 Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, **24**, 2534–2536.
- 129 Renard, B.Y., Kirchner, M., Monigatti, F., Ivanov, A.R., Rappsilber, J., Winter, D., Steen, J.A.J., Hamprrecht, F.A., and Steen, H. (2009) When less can yield more – computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**, 4978–4984.
- 130 Cox, J. and Mann, M. (2009) Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *Journal of the American Society for Mass Spectrometry*, **20**, 1477–1485.
- 131 Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, **3**, 1454–1463.
- 132 Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
- 133 Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*, **6**, 654–661.
- 134 Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004) Open mass spectrometry search algorithm. *Journal of Proteome Research*, **3**, 958–964.
- 135 Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P.A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, **77**, 4626–4639.
- 136 Maleknia, S.D. (1996) Sequencing isobaric peptides by the application of MS<sup>nm</sup> analysis on an ion trap mass spectrometer. Proceedings of the 44th ASMS Conference, Portland, OR, pp. 703–704.
- 137 Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, **4**, 787–797.
- 138 Elias, J.E. and Gygi, S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, **4**, 207–214.
- 139 Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, **74**, 5383–5392.
- 140 Searle, B.C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, **10**, 1265–1269.
- 141 Zhang, B., Chambers, M.C., and Tabb, D.L. (2007) Proteomic parsimony

- through bipartite graph analysis improves accuracy and transparency. *Journal of Proteome Research*, **6**, 3549–3557.
- 142 Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, **75**, 4646–4658.
- 143 Feng, J., Naiman, D.Q., and Cooper, B. (2007) Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Analytical Chemistry*, **79**, 3901–3911.
- 144 Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, **11**, 1067–1075.
- 145 Frank, A. and Pevzner, P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, **77**, 964–973.
- 146 Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical Chemistry*, **73**, 1917–1926.
- 147 Taylor, J.A. and Johnson, R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, **73**, 2594–2604.
- 148 Downard, K.M. (ed.) (2007) *Mass Spectrometry of Protein Interactions*, Wiley Interscience Series on Mass Spectrometry, John Wiley & Sons, Inc., Hoboken, NJ.
- 149 Zhang, Z. and Smith, D.L. (1993) Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Protein Science*, **2**, 522–531.
- 150 Smith, D.L., Deng, Y., and Zhang, Z. (1997) Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry. *Journal of Mass Spectrometry*, **32**, 135–146.
- 151 Wales, T.E. and Engen, J.R. (2006) Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, **25**, 158–170.
- 152 Johnson, R.S., Krylov, D., and Walsh, K.A. (1995) Proton mobility within electrosprayed peptide ions. *Journal of Mass Spectrometry*, **30**, 386–387.
- 153 Rand, K.D., Zehl, M., Jensen, O.N., and Jørgensen, T.J.D. (2009) Protein hydrogen exchange measured at single-residue resolution by electron transfer dissociation mass spectrometry. *Analytical Chemistry*, **81**, 5577–5584.
- 154 Pan, J., Han, J., Borchers, C.H., and Konermann, L. (2008) Electron capture dissociation of electrosprayed protein ions for spatially resolved hydrogen exchange measurements. *Journal of the American Chemical Society*, **130**, 11574–11575.
- 155 Maleknia, S.D., Brenowitz, M., and Chance, M.R. (1999) Millisecond radiolytic modification of peptides by synchrotron X-rays identified by mass spectrometry. *Analytical Chemistry*, **71**, 3965–3973.
- 156 Maleknia, S.D., Ralston, C.Y., Brenowitz, M.D., Downard, K.M., and Chance, M.R. (2001) Determination of macromolecular folding and structure by synchrotron x-ray radiolysis techniques. *Analytical Biochemistry*, **289**, 103–115.
- 157 Maleknia, S.D., Chance, M.R., and Downard, K.M. (1999) Electrospray-assisted modification of proteins: a radical probe of protein structure. *Rapid Communications in Mass Spectrometry*, **13**, 2352–2358.
- 158 Maleknia, S.D. and Downard, K.M. (2001) Radical approaches to probe protein structure, folding, and interactions by mass spectrometry. *Mass Spectrometry Reviews*, **20**, 388–401.
- 159 Shum, W.-K., Maleknia, S.D., and Downard, K.M. (2005) Onset of oxidative damage in  $\alpha$ -crystallin by radical probe mass spectrometry. *Analytical Biochemistry*, **344**, 247–256.
- 160 Gerega, S.K. and Downard, K.M. (2006) PROXIMO – a docking new algorithm to model protein complexes using data from

- radical probe mass spectrometry. *Bioinformatics*, **22**, 1702–1709.
- 161** Breuker, K. and McLafferty, F. (2008) Stepwise evolution of protein native structure with electrospray into the gas phase,  $10^{12}$  to  $10^2$  s. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 18145–18152.
- 162** Heck, A.J.R. and van den Heuvel, R.H.H. (2004) Investigation of intact protein complexes by mass spectrometry. *Mass Spectrometry Reviews*, **23**, 368–389.
- 163** Benesch, J.L.P. and Robinson, C.V.R. (2006) Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Current Opinion Structure Biology*, **16**, 245–251.
- 164** Bohrer, B.C., Merenbloom, S.I., Koeniger, S.L., Hilderbrand, A.E., and Clemmer, D.E. (2008) Biomolecule analysis by ion mobility mass spectrometry. *Annual Reviews of Analytical Chemistry*, **1**, 293–327.
- 165** Service, R.F. (2000) Proteomics: can Celera do it again? *Science*, **287**, 2136–2138.

## 2

# X-Ray Structure Determination of Proteins and Peptides

Andrew J. Fisher

### 2.1

#### Introduction

##### 2.1.1

#### Light Microscopy

Before Robert Hooke's work on light microscopy and the publication of his microscopic images in 1665, no one knew that organisms were built from individual units, which Hooke called "cells" [1]. The microscope had a profound effect in biology that still resonates today by revealing the inner workings of the cell and helping biologists gain a better understanding of cellular architecture and function. Even some 350 years later, the microscope is still a valuable tool in today's arsenal of biology and chemistry research instruments.

As powerful as the light microscope is, it is still incapable of resolving individual atoms that make up the repertoire of complex molecules in biology. This is because the resolution limit of a microscope is limited by the wavelength of the incident radiation, which is scattered by the specimen under inspection. The scattered light is "collected" by the objective lens, which refocuses the light into an inverse image. With the proper placement of other lenses within the microscope, the image can be magnified resulting in the microscopist seeing greater detail or resolution. The simplified equation used to define the resolving power of a light microscope is given as:

$$d = \frac{0.612 \cdot \lambda}{n \cdot \sin \theta} \quad (2.1)$$

where  $\lambda$  is the wavelength of radiation used,  $n$  is the refractive index of the media between the specimen and the objective lens (1.0 for air), and  $\theta$  is the angle of scattering between the direct light beam and the limit of the objective lens. The denominator ( $n \cdot \sin \theta$ ) is often called the numerical aperture. The variable  $d$  is the resolution of an optical microscope, which is defined as the shortest distance

between two points on a specimen that can still be optically distinguished. The goal of microscopy is to make  $d$  (resolution) as small a number as possible. One way to decrease  $d$  is to increase the terms in the denominator, like using immersion oil between the specimen and the objective lens, which has a refractive index of around 1.5. Another method is by moving the lens closer to capture more scattered light. Alternatively, one can increase the resolution (decrease  $d$ ) by using a shorter wavelength of radiation; however, this is restricted by the refraction of visible light (i.e., light's ability to be bent by lenses). Using a microscope with blue light radiation of 400 nm (4000 Å) and a numerical aperture of 1.37 ( $n = 1.5$ ,  $\theta = 65^\circ$ ) the resolving power would be around 180 nm (0.18  $\mu\text{m}$ ) or 1800 Å. While this resolution is good for identifying gross cellular structures, one would never be able to see individual atoms and molecular structure. Looking at this formula, one can easily see that the wavelength ( $\lambda$ ) is directly proportional to the resolution, so decreasing the wavelength to atomic scale would make it possible to decrease the resolution to this scale.

### 2.1.2

#### X-Rays and Crystallography at the Start

X-rays are electromagnetic radiation with the wavelength in the range 0.1–100 Å. However, X-rays cannot be refracted by lenses like visible light. Therefore, the scattered light would need to be recorded and “recombined” using computational techniques. Originally, the scattered X-rays were recorded on X-ray film, but today's modern detectors use charge-coupled devices (CCD) similar to those used in digital photographic cameras.

The discovery of X-ray diffraction by Max von Laue and the first recorded diffraction patterns from a salt crystal by Friedrich and Knipping in 1912 started a new science of X-ray crystallography. The first crystal structures of salts by W. L. Bragg [2, 3] and Laue [4] soon followed. This new technique had an immeasurable impact on chemistry. For the first time, structures of molecules were being revealed at the atomic level of detail. This new technique pushed chemistry to unexplored territory. Atomic details were giving clues into atomic radii, the bond order between atoms, as well as revealing possible mechanisms of chemical reactions by comparing the crystal structures of the reactants and products.

X-ray crystallography did not start impacting the field of biology until the late 1950s/early 1960s when Max Perutz and John Kendrew worked out the structures of hemoglobin and myoglobin, respectively. These two structures achieved a goal that Linus Pauling thought would be impossible [5]. The two structures had many noteworthy ramifications in biochemistry:

- i) The structures confirmed the  $\alpha$ -helix secondary structure does exist in proteins.
- ii) The structures revealed the overall fold of the proteins.
- iii) The structures revealed the atomic detail of how oxygen is carried in the blood and muscles by binding to the heme iron.



- iv) The structures revealed that over half the volume of the crystals is occupied by solvent, suggesting the structures represent those in the cell.
- v) The structures provided the first evidence that two distinct peptide sequences can fold into a similar three-dimensional structure given that hemoglobin and myoglobin are only 24% identical in primary amino acid sequence.
- vi) From these similarities, the authors deduced the two proteins developed from a common genetic precursor.

### 2.1.3

#### **X-Ray Crystallography Today**

Today, X-ray crystal structure determination of proteins is a common tool in biochemistry that can often determine atomic-resolution structures in a matter of weeks instead of years as before. The development of synchrotron radiation sources, sensitive X-ray detectors, software algorithms, and biotechnology advances has reduced the time and manpower required to determine the atomic-resolution structure.

Understanding the power of X-ray crystallography in biology was realized early on and was summed up nicely by Dorothy Hodgkin in her 1964 Nobel prize acceptance lecture for her use of X-ray techniques on the structures of biochemical substances, by saying “a great advantage of X-ray analysis as a method of chemical structure analysis is its power to show some totally unexpected and surprising structure with, at the same time, complete certainty.” Knowing the atomic-resolution structure of a biological macromolecule is one of the most powerful and insightful tools that a biochemist can use to not only gain a tremendous understanding of function, but also help chart a course for future experiments to test these insights. For this reason, the protein structure should not be regarded as the endpoint of analysis, but the starting point to help direct new experiments to elucidate the biological function.

Knowledge of the protein or enzyme’s three-dimensional structure not only plays a key role in understanding its biological function, but it can also be exploited to help design antagonists or inhibitors for drug development to ameliorate diseases faced by mankind. In the 1980s, great excitement in the pharmaceutical industry centered on using the active-site structure of a target enzyme to design inhibitors from the ground up. A shining example of structure-based drug design was finding inhibitors of the HIV protease, which are used to fight the AIDS epidemic. However, this *de novo* design of drugs proved very difficult for a variety of reasons, including insufficient computational docking and energy calculation routines starting with the static structures from crystallography. Today, X-ray crystallography still plays an important role in drug development, but often in the form of optimizing lead compounds, which were identified from high-throughput inhibitor screening and computational chemistry. Nevertheless, once computational algorithms improve in docking and energy calculations, and faster parallel computers are developed in the future, the industry may likely see a renaissance in the field of *de novo* structure-based drug design.

## 2.1.4

**Limitations of X-Ray Crystallography**

Despite the powerful method of X-ray crystallography, and its contributions to the advancement of both chemistry and biology, there are some limitations and weaknesses. The first limitation is that the sample must be crystallized. Today, this step is the least understood and therefore often the rate-limiting step in protein structure determination. Additionally, one must also ask if the process of protein crystallization alters the structure from that which occurs in solution or the cell. This answer came with the first protein structures, which revealed protein crystals average about 50% solvent volume and the protein concentration within the crystal is typically on the order of 10–40 mM. Additionally, comparison of many crystal structures to nuclear magnetic resonance (NMR) solution structures of the same protein reveal that crystallization of proteins has little consequence on the folded state.

Another limitation of X-ray crystal structure determination is that because data are recorded from X-rays scattered from the crystal, the image obtained is that of the “average” protein structure that occurs in the crystal (around  $10^{14}$  protein molecules contribute to scattering in a 200- $\mu\text{m}$  X-ray beam). Therefore, any dynamics or multiple conformations of loops or amino acid side-chains are not observed in the average structure. At times, this results in regions of protein structures that have no observable electron density and therefore the crystallographer cannot build an atomic model for this region of the structure. However, this lack of density does indeed give evidence that the disordered region is mobile or can adopt multiple conformations, which gives clues to potential biological function.

Another shortfall of X-ray crystallography is that because the electrons of the atoms scatter the X-rays, it is difficult to observe atoms with few electrons, like the hydrogen atom or protons. This limitation is especially apparent in trying to decipher catalytic mechanisms, because many enzymatic reactions in biology involve the transfer of hydrogens or protons in acid–base catalysis. The biochemist must infer the location of the protons by other theoretical and biochemical techniques. Hydrogen locations can be assumed simply based on geometry and bond distances between hydrogen-bonding partners, or more exhaustive techniques like site-directed mutagenesis, kinetic assays, pH profiles, or catalysis, or even the method of NMR. Another crystallography method that has been gaining some momentum in recent years is that of neutron diffraction. Unlike X-rays, scattering of neutrons is not proportional to atom number, but depends on nuclear characteristics like NMR. The hydrogen atoms scatters neutrons with a  $180^\circ$  phase change, which results in negative peaks in a neutron density map, making identification of hydrogens relatively easy. However, neutron sources are expensive and limited, and much larger crystals are usually required for accurate neutron diffraction experiments.

## 2.2 Growing Crystals

### 2.2.1 Why Crystals?

The first step in X-ray crystal structural analysis is to grow crystals of suitable order that results in diffraction of X-rays. It is this internal order that determines the quality and level of diffraction. This is why some protein crystals may look beautiful macroscopically with magnificently shaped facets and sharp edges, but the internal microscopic order may be very poor, resulting in either little or no diffraction of X-rays. This can arise from the globular nature of some proteins. If the protein has a nice robust globular or spherical shape, the proteins may fall out of solution too quickly, resulting in the packing of “spheres,” which will still result in flat facets with sharp edges, but the protein globules may pack in random orientations, resulting in no internal order resulting in little to no X-ray diffraction.

The crystal is required because it serves two purposes. The first function is to amplify the scattered X-rays. The structure from X-ray scattering can only be achieved if one can accurately record the scattered light. The scattering intensity is proportional to the number of electrons and the intensity of the incident radiation. Biological macromolecules are made of relatively light atoms like carbon, nitrogen, and oxygen, which do not scatter light strongly. However, if one can obtain approximately a quadrillion copies of a protein, all oriented in exactly the same way, the samples would scatter light coherently, resulting in a much stronger signal, which can be recorded accurately. In the early days of crystallography, protein crystals had to be larger than today because the X-ray intensities were weaker. Diffraction data were typically collected from crystals around 0.2–1 mm in size. Today, with the ability to achieve many orders of magnitude more intense X-ray beams with synchrotrons, accurate data can be recorded on crystals as small as 10  $\mu\text{m}$  and lower.

The second function of the crystal is to accurately orient the sample with high precision within the X-ray beam. To get a true three-dimensional structure of the protein, scattering data has to be recorded from all three dimensions of the sample. Crystals are large and stable enough to be physically manipulated by mounting and orienting precisely in the X-ray beam. As will be discussed in more detail below, X-ray diffraction images are recorded by exposing the crystal numerous times, each at a slightly different crystal (and protein) orientation, and the data combined and processed to generate a three-dimensional electron density map.

### 2.2.2 Basic Methods of Growing Protein Crystals

As mentioned above, the method of growing crystals today is the least understood process in X-ray crystal structure determination and it is often the rate-limiting step

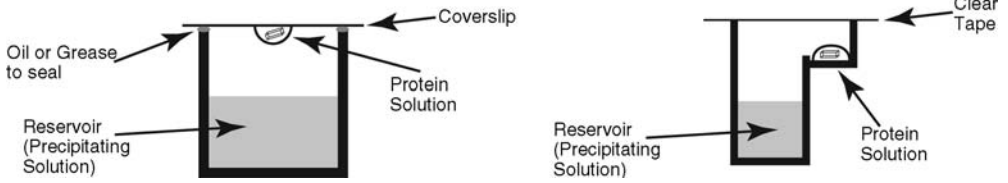
of structure determination. This is because the solubility of macromolecules depends on their interactions with solvent. Water surrounds the protein and interacts with surface charges, dipoles, hydrogen-bond donors, hydrogen-bond acceptors, and alkyl hydrophobic groups. Every unique protein has a novel shape and surface, and therefore interacts differently with water, resulting in different solubility properties. Thus, no single condition can be employed to crystallize all or most proteins.

Fundamentally, the process of crystallization is to achieve a state of supersaturation of the solute (protein) to bring the solute out of solution under the proper conditions that results in a highly ordered crystal lattice. Presently, finding the conditions to grow highly ordered crystals is simply trial and error. The chance to reach success is increased by exploring more conditions. However, this requires more protein and time to explore more conditions, both of which have seen advances over recent years. Proteins can be readily biosynthesized in tens of milligram quantities through the advent of molecular biology and inserting the protein gene into an ectopic expression host like *Escherichia coli*. Additionally, the development of robotics has made it possible to automate crystallization experiments, allowing the ability to explore thousands of conditions and saving many hours of manual crystallization setups.

As proteins have surface charges, this can be thought of as a polyvalent charge and, traditionally, the solubility can be thought of in terms of the Debye–Hückel theory of solubility [6]. This theory describes the solubility of ions in ionic solution, and it has been found that many solutes experience a salting-in and salting-out effect. The solubility of many solutes decreases with increasing ionic strength of a more soluble salt. This practice works nicely when proteins are the solute. The concept of salting-out has been historically used as a common protein purification step using ammonium sulfate to fractionate cell lysate. In fact, ammonium sulfate is the most common salt used to grow protein crystals. Additionally, the solubility of some proteins generally decreases in low-salt conditions and this has also been found to be successful in growing protein crystals, by simply dialyzing the protein against water.

Another common technique to reach supersaturation of the protein is to add an organic hydrophilic polymer at high concentrations. Poly(ethylene glycol) (PEG) is a common organic polymer that comes in a variety of different molecular weight ranges, all of which have been successfully used for growing protein crystals. The basic principles of achieving supersaturation using either salts or organic molecules are similar; in that they have a strong affinity for water molecules, which causes the protein solute molecules to interact together rather than with water causing the protein solute to fall out of solution.

Diffraction-quality crystals are typically grown at a slow rate to help maximize the solute–solute interactions, which is necessary to pay the entropic cost of a highly ordered crystal lattice. One of the common techniques used to slowly reach the state of supersaturation is to utilize vapor diffusion. In this method, the protein solution (typically around 10 mg/ml) is diluted 1 : 1 with a precipitating solution in a drop, which is then equilibrated against the precipitating solution in a sealed chamber.

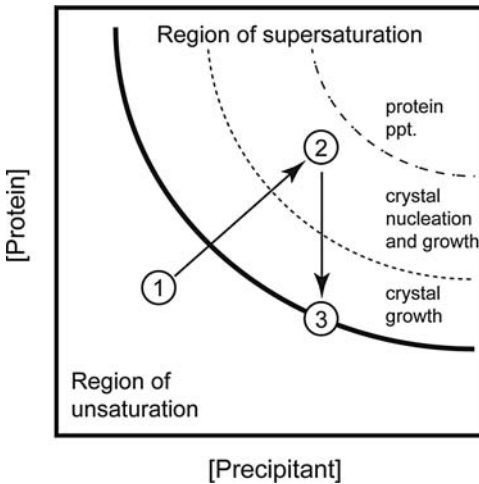


**Figure 2.1** Schematic illustration of vapor diffusion crystallization setups. Left shows the typical “hanging drop” experiment where around 1–5  $\mu\text{l}$  of protein solution is suspended on a glass coverslip over around 1 ml reservoir

precipitating solution. Right shows the sitting drop where around 1–2  $\mu\text{l}$  of protein solution is sealed in a small chamber against around 50–100  $\mu\text{l}$  or reservoir.

The drops can be suspended on a glass coverslip over the reservoir precipitating solution in what is called “hanging drop” crystallization or the protein solution drop can be placed on a surface that is sealed next to the reservoir solution in the “sitting drop” method (Figure 2.1).

In both methods, the protein solution drop is sealed in a small chamber to slowly establish equilibrium between the two solutions through vapor diffusion. The protein starts out in a low concentration of precipitant below the supersaturation stage (point 1 in Figure 2.2). Over the course of time, water will slowly diffuse out of the protein solution drop into the reservoir to establish an equilibrium. This increases the concentration of both the protein and precipitating agent in the drop to the point of supersaturation. Usually, the region of supersaturation can be divided



**Figure 2.2** Phase diagram of crystal growth. The protein is mixed with precipitant at initial concentrations (point 1) in a vapor diffusion experiment. After time, water diffuses out of the protein drop into the higher concentrated reservoir solution increasing the concentration

of both protein and precipitant (point 2). After crystal growth is nucleated, the crystal grows decreasing the concentration of the protein until it reaches the unsaturated state where the crystal stops growing (point 3).

into subregions – that which supports crystal nucleation and growth, and that which would support crystal growth, but not nucleation. If the protein and precipitant concentration reaches the crystal nucleation region slowly (point 2 in Figure 2.2), then crystal growth will start. After time, the protein concentration in solution will decrease, because the protein is withdrawing into the growing crystal, until the protein concentration drops below the state of supersaturation and thus the crystal stops growing (point 3 in Figure 2.2). If the region of crystal nucleation is reached too quickly, showers of microcrystals may form or simply the protein may fall out of solution as an amorphous precipitate. If the proper conditions are found, crystals can appear in as little time as overnight or it may take many weeks or months to grow crystals.

To maximize chances of successful crystal growth, many different crystallization conditions must be explored. Today, robotics can help assist in rapidly and accurately pipetting both precipitating and protein solutions into plastic trays and coverslips, which are commercially available. Additionally, many different premade precipitating solutions are also commercially available, often in a 96-well format, to easily pipette into 96-well crystallization trays. Once sealed, the trays are left undisturbed in a low-vibration, stable-temperature incubator or room. Temperature can often affect the rate and thermodynamics of crystal growth, and can be a parameter that can be explored to achieve successful crystal growth: either growing the crystals at a different constant temperature, or changing the temperature slowly during the course of crystallization, which may possibly help nucleate crystal growth. The crystallization experiments are then checked periodically under a microscope to examine the results.

Another method of crystal growth that has gained popularity recently is the batch or microbatch method. This method uses fewer materials in both precipitating solution and protein because it can be achieved in lower volumes. This method does not equilibrate against a reservoir buffer; it simply mixes the protein and precipitant solution together under oil. The philosophy of this method is to reach point 2 in Figure 2.2 at the initial state, then crystals will grow without equilibrating against a larger reservoir buffer. The advantage this technique is the simplicity and speed with both robotics and manual screens. Additionally, in microbatch experiments, very small volumes of sample can be used without drying out completely, which can happen with vapor diffusion. Volumes in the nanoliter range have been successful.

The different precipitation conditions that are explored consist typically of three components, a precipitating agent, a salt, and a buffer. The precipitating agent can be high concentrations of a salt like ammonium sulfate, lithium sulfate, sodium chloride, phosphate, malonate, citrate or high concentrations of organic polymers like PEG. Additionally, smaller organic molecules can also bring proteins out of solution, such as 2-methyl-2,4-pentanediol (MPD), pentaerythritol propoxylate or ethoxylate, as well as many alcohols like ethanol or propanol. The additional salt, or counterions, found in the crystallization conditions, can help mediate protein–protein contacts in the crystal, as well as bind in an active-site pocket to maintain structural stability. The buffer is also important since different pHs will change the protein surface charges because of titratable functional groups on the

protein surface. Altering the protein surface will have a major effect on the solubility and crystallization properties.

### 2.2.3

#### **Protein Sample**

Protein crystallization was a technique used by biochemists long before X-ray analysis of protein crystals. Early biochemists used crystallization as a way to isolate or purify a particular protein of interest. In growing protein crystals today for X-ray analysis, it is best to eliminate the unknown factor that can influence the rate-limiting step of structure determination. Therefore, it is best to crystallize as pure a sample as possible. A good rule of thumb is to have the protein sample at least 95% pure and at a high concentration (around 10 mg/ml). Additionally, the sample should be in low salt and buffer conditions, or preferably water if the sample is stable, so the precipitating concentrations will dominate the crystallization condition.

Any substrates, products, inhibitors, or cofactors that can bind to the protein should be included, which may help stabilize or lock the protein into a single conformation. Flexible proteins or samples that can adopt different conformational states can be difficult to crystallize. Furthermore, subsequent small-molecule binding partners can also be soaked into protein crystals prior to data collection to gain structural insight in binding and function. However, one must keep in mind that the crystallized enzyme still possesses catalytic activity and substrates will likely be turned into products if all the necessary components are present. Consequently, nonfunctional structural analogs are ideal tools to trap the active site in a precatalytic Michaelis-like “ES” state. These structures provide tremendous insight into the catalytic mechanism and the functional role of active-site residues.

Multidomain proteins with flexible linkers can be difficult to crystallize. This is because, for a well-ordered crystal that can diffract X-rays, the crystal must be made up of repeating units all in the same conformation. Therefore, large multidomain proteins may prove difficult to crystallize and many laboratories utilize the “divide and conquer” strategy. In this approach, individual protein domains, which have often been identified by sequence analysis, are cloned separately, expressed, crystallized, and the structure determined. Utilization of other biophysical techniques can be employed to piece together how the individual domains may be spatially related in the complete full-length protein. Cryoelectron microscopy is an ideal method to view the larger complex at lower resolution, where placement of individual atomic structures into the electron microscopy model provides significant structural detail that cannot be obtained by either single method [7].

### 2.2.4

#### **Preliminary Crystal Analysis**

Once crystals appear, some information can be gained from characterization and analysis of crystals that can help in structure determination. Examining the crystals under a polarizing microscope can help determine crystal quality, symmetry, and

possible twinning. In the polarizing microscope, the crystals are placed between the pair of polarizers – the polarizer and analyzer, which are rotated  $90^\circ$  from each other to block transmittance of light. Most protein crystals will rotate plane-polarized light, so when inspected in the polarizing microscope, the crystals will appear bright in the dark field. All noncubic crystals are anisotropic or birefringent and will rotate plane-polarized light. Given that proteins rarely crystallize in cubic space groups, if the crystals do not rotate plane-polarized light, there is a good chance they lack order or may be salt crystals. However, be sure to check different orientations of the crystals because the crystal can be uniaxial, in that the crystal does not rotate plane-polarized light along axes of three-, four-, or six-rotational symmetry. These crystals belong to the trigonal, tetragonal, and hexagonal space groups, respectively, and the axis viewed that does not rotate polarized light corresponds to the unique  $c$ -axis of the unit cell. Biaxial crystals can rotate light in all orientations, and belong to the triclinic, monoclinic, and orthorhombic space groups.

Twinned crystals are crystals that grow with different parts oriented differently within a single crystal. This may be detected by inspection of the crystals in a polarizing microscope. The different regions of the crystals may rotate plane-polarized light at different angles. This can be seen when the analyzer of the polarizing microscope is rotated while looking for different regions of the crystals to appear light, while other regions within the same crystal appear dark. Often it can be seen as one-half of the crystal light while the other half is dark. However, not all twinned crystals exhibit this phenomenon. Therefore, simple inspection of the crystal in a polarizing microscope before X-ray analysis can help not only to determine the crystal space group, but also quality.

Proteins typically have similar densities because they are made up of only 20 different kinds of amino acids with similar densities and similar internal packing densities within the protein core. This was first identified by Brian Matthews in the late 1960s [8]. From surveying 116 different protein crystals, he found that most protein crystals have similar densities. This can be expressed in a volume per molar mass parameter ( $V_M$ ). The formula to determine the  $V_M$  is given as:

$$V_M = \frac{V}{nM} \quad (2.2)$$

where  $V$  is the volume of the unit cell in  $\text{\AA}^3$  (determined from unit cell parameters),  $n$  is the number of protein molecules in the unit cell, and  $M$  is the molecular mass of the protein in Daltons. Matthews found that  $V_M$  (which has units of  $\text{\AA}^3/\text{Da}$ ) ranged from around 2.0 to 3.5 for most protein crystals. This corresponds to a range of solvent content of around 38–65% as calculated by:

$$V_{\text{sol}} = 1 - \frac{1.23}{V_M} \quad (2.3)$$

where  $V_{\text{sol}}$  is fractional volume of solvent in the crystal.

This finding testifies that protein crystals are not void of aqueous solvent and in fact average about 50% solvent by volume. Therefore the crystal structures determined do represent the structure of the protein in aqueous solution, within



the crowded cell. Additionally, this  $V_M$  parameter, which is now known as the Matthews coefficient, can also help in identifying oligomeric states of the protein within the crystals. Given that  $V_M$  has a typical range of values and the molar mass of the crystallized protein is known, the equation for the Matthews coefficient can be used to determine “ $n$ ” or the number of polypeptides within the unit cell. Therefore, if the protein forms a large oligomeric state like dimers, it is not unusual for the dimer to form one asymmetric unit (defined below) within the unit cell. Knowing the number of polypeptides per asymmetric unit is essential to structure determination too.

### 2.2.5

#### Mounting Crystals for X-Ray Analysis

Since protein crystals are grown in aqueous solutions and contain around 50% solvent, they need to stay hydrated during data collection. J. D. Bernal and Dorothy Crowfoot (later Hodgkin) first discovered this in 1934 when they were working on crystals of pepsin, supplied by Professor T. Svedberg’s laboratory [9]. The method they used to maintain crystal hydration was to seal the crystal in a thin-walled glass capillary with excess mother liquor (reservoir solution), which was still very common until the early-mid 1990s when the method of freezing protein crystals gained popularity.

Cryocrystallography, or freezing crystals to near liquid nitrogen temperatures during data collection, was developed to improve two aspects of crystallography. (i) In theory the colder temperature reduces atomic vibration, which results in stronger X-ray scattering, especially at higher scattering angles (higher resolution), thus improving data quality and resolution. (ii) X-rays are ionizing radiation, which knocks electrons out of orbitals, creating reactive species. The chemistry resulting from these reactive radicals can disrupt protein–protein lattice contacts, which leads to disordered crystals. This is often reflected in diffraction quality deteriorating over time while the crystal is exposed in the X-ray beam. Therefore, multiple crystals are required to collect a complete diffraction data set, which can reduce data quality because of slight variations between crystals. By cooling the crystals to ultra-low cryogenic temperatures, the chemistry is slowed down tremendously and multiple data sets can be collected from a single crystal.

To determine the protein structure, one needs to record or measure the X-rays diffracted from the protein crystal, and not ice crystals that often result in freezing crystals directly in the mother liquor. Cryoprotectants are needed to prevent ice formation upon rapidly freezing the protein crystals. High concentrations of low-molecular weight organic molecules can serve as ideal cryoprotectants. Some of the common cryoprotectants include ethylene glycol, glycerol, low-molecular-weight PEGs (100–1000), MPD, sucrose, or alcohols. Typical concentrations of these cryoprotectants range from 20 to 40% (by volume) in addition to the mother liquor conditions. If the crystals are grown from high-salt concentrations, increasing the salt concentration to around 4 to 5 M may also serve to prevent ice crystals upon freezing. Transferring crystals slowly or rapidly to the final cryoprotectant concentrations have

both worked successfully. One needs to experiment to find the optimal way to transfer crystals that does not affect crystal diffraction quality. Another common freezing technique that is gaining popularity is to drag the protein crystal through very thick oil, like Paratone-N. This serves to pull the mother liquor off the crystal surface, but the oil does not penetrate the crystal. The water molecules within the crystal solvent channels do not form ice crystals upon freezing possibly because there is not a large enough critical mass to nucleate an ice crystal or the abundant hydrophilic protein surface within the crystal serves as a type of antifreeze, functioning similar to the antifreeze proteins seen in nature.

After the crystals are cryoprotected, they are mounted by suspending the crystals in a small loop of thread typically made of nylon (10–20  $\mu\text{m}$  thick). The loops themselves can range from 50  $\mu\text{m}$  to 1 mm in diameter to suit the size of the crystal being mounted. Surface tension of the liquid holds the crystal suspended in the loop, which is situated typically at the end of a copper pin. The loop is positioned in the path of the X-ray beam with a cold stream of nitrogen gas (90–100 K) gently blowing over the crystal loop. The pin that holds the loop disrupts the laminar flow of nitrogen gas, which can cause ambient moisture to collect on the pin in the form of ice, which adds to the turbulence in the stream. For this reason the pins are made of a good thermal conductor such as copper, so the base of the pin at ambient temperature is warm enough to melt the ice and prevent it from snowballing into a larger aggregate that can cover the crystal.

## 2.3

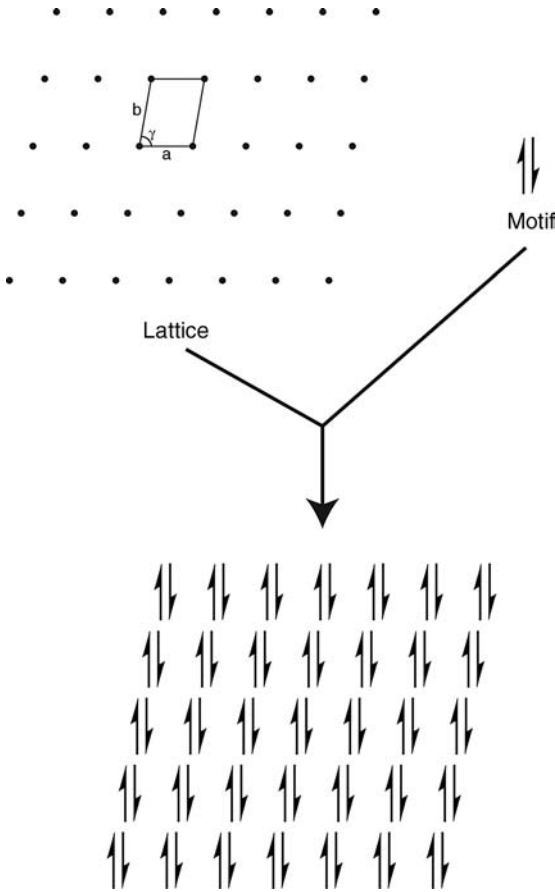
### Symmetry and Space Groups

#### 2.3.1

##### Crystals and the Unit Cell

Crystals are made up of a motif that is repeated in a lattice that extends in all three dimensions. The motif is the object that was crystallized and the lattice is a set of points that display only specific types of symmetry that can fill three-dimensional space. Both components contribute to scattering X-rays. The lattice causes diffraction of X-rays, so the scattered X-ray pattern is not a continuous function (as observed from a single object), but only occurs at discrete maxima, or reflections, whose spacing is dictated by the lattice spacing. The motif or protein causes a change in intensity distribution of the scattered X-rays based on its three-dimensional structure.

The smallest repeating unit in a crystal is called a *unit cell* and can be thought of as a parallelepiped-shaped block, which when repeated by translation in all three dimensions produces the volume of the crystal. The *lattice* is an arrangement of points such that each point is in the same environment and repeats infinitely in all three dimensions. Figure 2.3 illustrates the concept of how the lattice and motif are combined to generate a crystal. The position of the unit cell with respect to the lattice is arbitrary and can be the choice of the crystallographer. For example, the



**Figure 2.3** Illustration of crystal (bottom) being composed of a motif (top right) repeated in a lattice (top left). One two-dimensional unit cell is outlined in the lattice.

origin of the unit cell in Figure 2.3 may be the tip of the left arrow pointing up and therefore each lattice point will fall on the tip of the left arrow. Another crystallographer may choose the tail of the left arrow for the origin of the unit cell. The overall repeating structure does not change, as long as each vertex of the unit cell is in an identical environment. Additionally, the structure motif does not change, only the coordinates of atoms within the motif. Therefore, the crystallographer sets the origin of the unit cell when he/she sets the coordinates of the first atoms.

The symmetry of the lattice only has a finite number of possibilities to pack a motif and completely fill three-dimensional space. Mathematical derivations for the possible types of symmetry were actually determined in the nineteenth century, many years before X-ray crystallography. The only types of rotational symmetry that are possible in a three-dimensional lattice are 1-, 2-, 3-, 4- and 6-fold symmetry. This is also seen in two-dimensional lattices, and explains why you cannot tile a floor with

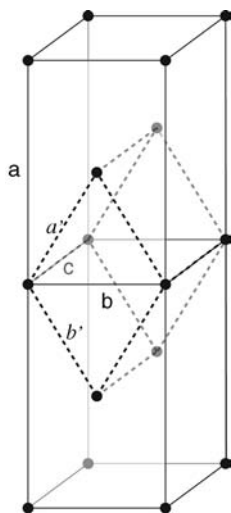
**Table 2.1** The seven crystal systems.

Name	Rotational symmetry	Bravais lattice	Lattice restrictions
Triclinic	1-fold	P	$a \neq b \neq c; \alpha \neq \beta \neq \gamma \neq 90^\circ$
Monoclinic	2-fold	P, C	$a \neq b \neq c; \alpha = \gamma = 90^\circ, \beta > 90^\circ$
Orthorhombic	2-, 2-, 2-fold	P, C, I, F	$a \neq b \neq c; \alpha = \beta = \gamma = 90^\circ$
Trigonal (Rhombohedral setting)	3-fold	P (R)	$a = b \neq c; \alpha = \beta = 90^\circ, \gamma = 120^\circ$ $a = b = c; \alpha = \beta = \gamma \neq 90^\circ$
Tetragonal	4-fold	P, I	$a = b \neq c; \alpha = \beta = \gamma = 90^\circ$
Hexagonal	6-fold	P	$a = b \neq c; \alpha = \beta = 90^\circ, \gamma = 120^\circ$
Cubic	2- and 3-fold	P, I, F	$a = b = c; \alpha = \beta = \gamma = 90^\circ$

5-fold symmetric pentagons and completely cover the floor; there will always be gaps. This does not mean you cannot crystallize a motif that displays 5-fold symmetry; in fact this is observed, but the 5-fold symmetric motif resides at each lattice point. For example, in Figure 2.3, the motif displays 2-fold symmetry (double arrows), but one can easily imagine a pentad of arrows arranged with 5-fold symmetry being placed at each of the lattice points.

The unit cell, which defines the smallest repeating parallelepiped unit of the crystal, is outlined in Figure 2.3 in a two-dimensional lattice. The three-dimensional unit cell is defined by dimensions of the parallelepiped  $a$ ,  $b$ , and  $c$  (units of Å), which define the edges, and the angles  $\alpha$ ,  $\beta$ , and  $\gamma$  (units of degrees) between the edges. The restricted rotational symmetry results in only seven possible crystal systems, which also restricts the unit cell dimensions (Table 2.1). The orthorhombic crystal system displays three mutually perpendicular 2-fold axes, and the cubic crystal system displays both 2-fold along the axis, and a 3-fold symmetry along the body diagonal of the cube. Some of the trigonal crystal systems can display and be indexed in a rhombohedral setting, with the 3-fold axis being coincident with the body diagonal of the rhombohedron. The lattice crystal shown in Figure 2.3 would fall in the two-dimensional version of the monoclinic crystal system, which is called oblique, because it displays 2-fold symmetry.

In addition to the seven crystal systems, the unit cells can also display certain *centering*, like face-centered (F) orthorhombic, where the motif can also fall on each face of the parallelepiped, or body-centered (I), where it falls in the center of the unit cell. There are 14 possible centering or Bravais lattices, with primitive (P) displaying no centering. Each of the nonprimitive centered cells (C (end-centered on C-face), I, and F) can also be indexed into a lower-symmetry primitive cell, but if the structure is determined in the lower-symmetry primitive cell, it ignores the higher symmetry, which is technically incorrect. Figure 2.4 shows how a C-centered orthorhombic cell can be indexed in the lower-symmetry primitive triclinic cell. Two orthorhombic cells are outlined with solid lines with the  $a$ ,  $b$ , and  $c$  edges labeled; notice the lattice point on the A–B face, which designates C-centered. Shown in the dashed line is the primitive cell that can be indexed from the C-centered orthorhombic, with the  $a$  and  $b$  edges labeled in italics with primes; the  $c$  axis is the same for both cells.



**Figure 2.4** Two cells of a C-centered orthorhombic cell outlined with solid lines. Also shown is the lower symmetry triclinic primitive cell outlined with dashed lines. The primitive  $a'$  and  $b'$  axes are labeled, with the  $c$  axis coincident with the orthorhombic cell.

### 2.3.2

#### Point Groups

In addition to the 14 Bravais lattices and seven crystal systems that a lattice can adopt to fill three-dimensional space, within the unit cell other symmetry elements can operate about a point of symmetry, called *point groups*. These not only include the five types of rotational symmetry (1, 2, 3, 4 and 6), but also include other symmetries like mirror symmetry, where two molecules of opposite hands can pack across a plane. Additionally, there is inversion symmetry (center of symmetry), also where racemic mixtures of molecules can assemble about a central inversion point. However, crystallographers working with biological samples never see their compounds crystallized with this symmetry because biological samples contain only a single pure enantiomer. Combining the rotational symmetry, inversion symmetry, and mirror symmetry results in 32 possible point groups.

Translational symmetry can also be found within a unit cell in addition to point symmetry. An example commonly seen for biological samples is termed a screw axis. The screw axis rotates and translates a motif from part of the unit cell to another part of the unit cell. A screw axis can be designated generically as  $N_m$  where there is an  $N$ -fold rotation followed by an  $m/N$  translation in a unit cell. For example, a  $2_1$  screw axis is where the motif is rotated by  $180^\circ$  about the axis followed by translation of  $1/2$  the unit cell. Repeating this twice places the motif in the same starting position in the adjacent unit cell.

Another type of translational symmetry is called a glide plane. This involves taking a motif, reflecting it across a mirror plane followed by translation of half a unit cell.

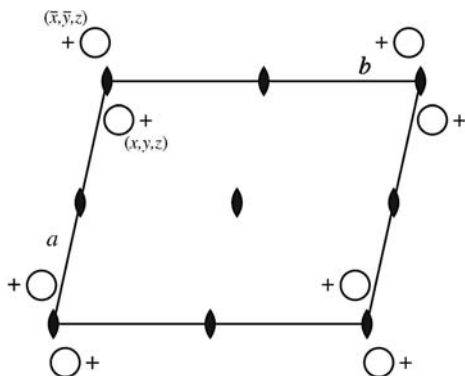
Again, this symmetry is only observed when a racemic mixture of enantiomers is crystallized together, which is not seen with biological specimens.

### 2.3.3

#### Space Groups

Combining all the point symmetries with translational symmetry and the 14 possible Bravais lattices results in a possible 230 unique space groups. The majority of which involves symmetry that will result in inversion and can only occur when crystallizing racemic mixtures. That is, there are 230 ways a physical specimen can pack together in three dimensions filling all space producing a crystal. Of the 230 space groups, only 65 enantiomeric space groups are possible when crystallizing a single pure enantiomer as in biological samples. Surprisingly some space groups are much more common, possibly due to packing contacts. It turns out that more than 63% of biological samples crystallize in only six space groups. The most common space group, which accounts for over 20% of the crystals, is  $P2_12_12_1$ , where the first letter designates the Bravais lattice, followed by symmetry observed along the three axes, in this case, a  $2_1$  screw axis runs parallel to all three unit cell edges.

The 230 space groups along with symmetry elements and diagrammatic illustrations of the space group are published in a book called the *International Tables for Crystallography* [10]. An example of what a space group diagram may look like in the *International Tables for Crystallography* is shown in Figure 2.5, which would represent the three-dimensional equivalent for the two-dimensional crystal in Figure 2.3.



**Figure 2.5** Schematic representation of space group  $P2$  as seen in the *International Tables for Crystallography*. The circle represents the asymmetric unit and the large rectangle the entire unit cell. The black oval represents the crystallographic 2-fold axis perpendicular to the page. Note there is a 2-fold axis along each unit cell vertex as well as halfway along each cell edge and the center of the unit cell. The unit cell contains two asymmetric units within the unit cell. Equivalent coordinate positions are drawn

next to the asymmetric units closest to the origin. For every atom observed at  $x, y, z$ , an equivalent atom in the symmetry related asymmetric unit is at atomic position of  $-x, -y, z$ . (convention is to put negative sign above variable). The plus sign next to the circle indicates it lies above the plane of the page. For the two-dimensional plane group diagram that corresponds to Figure 2.3, the figure will be identical, except the plus sign and the  $z$ -coordinate will be removed.

## 2.3.4

**Asymmetric Unit**

The smallest repeating unit within the unit cell is called the *asymmetric unit*. This differs from the unit cell in generating the crystal, because the asymmetric unit can not be simply translated in all three directions to produce the crystal, which defines the parallelepiped unit cell. The asymmetric unit in Figure 2.3 is only one arrow (e.g., pointing up). This single arrow cannot generate the entire lattice by simple translation. However, what relates one arrow to another arrow is a 2-fold rotation symmetry perpendicular to the page. Also, the 2-fold axis operates not only on one arrow to another, the 2-fold can actually operate on the entire lattice, thus this lattice displays 2-fold rotational symmetry, which is allowed for oblique plane groups. Therefore, the two-dimensional lattice in Figure 2.3 would have the designated plane group (two-dimensional space group) of p2 (plane groups are designated by lower case letters). Thus, technically speaking, the repeating motif (asymmetric unit) in Figure 2.3 is really just a single arrow, where the other arrow is generated by 2-fold rotational symmetry.

The asymmetric unit is the unique part of the crystal and can contain multiple copies of protein or sample within. For proteins, the minimum sample contained in the asymmetric unit is a single monomer, whereas DNA or RNA may contain half a duplex in the asymmetric unit, where a 2-fold rotation generates the other strand of the duplex. Therefore, the structure of only the asymmetric unit is required to determine the complete crystal structure. Nevertheless, for complete biological analysis, the crystallographer must keep in mind the symmetry of the unit cell, which may generate a dimer, trimer, and so on, resulting in biologically important functional interpretations.

## 2.4

**X-Ray Scattering and Diffraction**

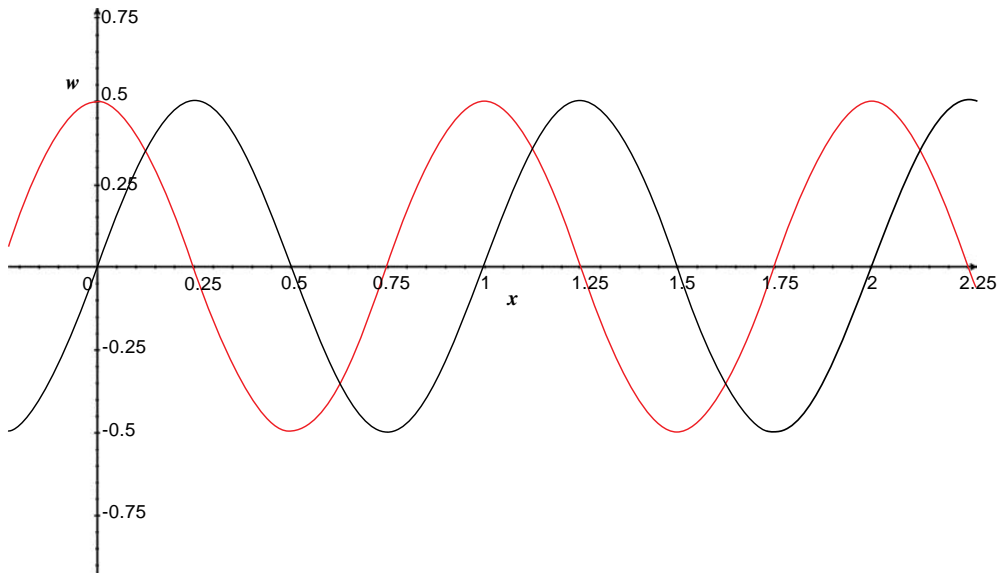
## 2.4.1

**X-Rays and Mathematical Representation of Waves**

X-ray diffraction utilizes the wave nature of light. X-rays are electromagnetic radiation, which contain both an electric and magnetic field components that oscillate perpendicular to each other. The distance between wave peaks or crests defines the wavelength. The height or magnitude of the wave peak defines its amplitude. The wave of the electric and magnetic components as they propagate through time and space can be mathematically represented as a cosine function:

$$w_0 = f \cos\left(\frac{2\pi x}{\lambda}\right) \quad (2.4)$$

where  $f$  represents the amplitude,  $x$  represents the distance propagated (same units of  $\lambda$ ),  $\lambda$  is the wavelength, and  $2\pi/\lambda$  converts it to radians. The term  $w_0$  represents the wave at time zero for a reference wave. This wave is plotted in red in Figure 2.6. Now if



**Figure 2.6** Plot of two waves, the reference wave in red with an amplitude ( $f$ ) equal to 0.5 and a wavelength ( $\lambda$ ) of 1.0. The black curve represents a wave with the same amplitude and wavelength as the reference wave, but shifted along  $x$  by 0.25 of a unit ( $x_1$  in Eq. (2.5)).

we consider another wave having a positive displacement by  $x_1$ , then this wave can be represented by the equation:

$$w_1 = f \cos\left(\frac{2\pi x_0}{\lambda} - \frac{2\pi x_1}{\lambda}\right) \quad (2.5)$$

An example of this displaced wave is plotted in black in Figure 2.6, where  $x_1$  is 0.25 (of a unit). Equation (2.5) can also be represented by the simplified equation of:

$$w_1 = f \cos(\omega t - \varphi_1) \quad (2.6)$$

where  $\omega t$  is the angular frequency of ( $2\pi\nu$ ) with  $\nu$  equal to the frequency and  $t$  is time. The phase shift of the displaced wave is represented by  $\varphi_1$ , which is equal to  $(2\pi x_1/\lambda)$ . The displaced wave ( $w_1$ ) would overlap with the reference wave ( $w_0$ ) exactly if  $\varphi_1$  is equal to even multiples of  $\pi$  (e.g.,  $2\pi$ ,  $4\pi$ , ...), and would exactly be out of phase or register for odd multiples of  $\pi$ .

Two different waves can be summed together into one combined wave, which itself can also be described as a cosine function also in the form of:

$$W = F \cos(\omega t - \varphi) \quad (2.7)$$

where  $F$  is the amplitude of the combined wave, and  $\varphi$  is the phase shift of the combined wave, with respect to the reference incident wave. Using the geometric law of cosines, Eq. (2.7) can also be represented by the equation:

$$W = F \cos(\cos \omega t \cos \varphi + \sin \omega t \sin \varphi) \quad (2.8)$$



Using this Eq. (2.8), the cosine and sine terms of the combined wave with phase  $\varphi$  can be separated out from the reference wave ( $\omega t$ ). Therefore, a single wave combined from multiple waves can be separated into a sum of cosine and sine terms:

$$F \cos \varphi = \sum_n f_n \cos \varphi_n \quad F \sin \varphi = \sum_n f_n \sin \varphi_n \quad (2.9)$$

where  $F$  is the magnitude of the single wave summed from  $n$  individual waves,  $f_n$  is the amplitude of each individual wave, and  $\varphi_n$  is the phase of each individual wave. The sum of the cosine and sine terms of the individual waves can be represented by  $A$  and  $B$ , respectively. Additionally, the magnitude of the combined wave can be calculated by the square root of the summed cosine and sine terms squared:

$$\begin{aligned} |F| &= \sqrt{A^2 + B^2} \\ A &= \sum_n f_n \cos \varphi_n \quad \text{and} \quad B = \sum_n f_n \sin \varphi_n \end{aligned} \quad (2.10)$$

The tangent of the phase of the combined wave can be calculated by the dividing the sine terms into the cosine terms:

$$\tan \varphi = \frac{\sum_n f_n \sin \varphi_n}{\sum_n f_n \cos \varphi_n} = \frac{B}{A} \quad (2.11)$$

This applies to X-ray scattering because, as we will see in Section 2.4.2, each individual atom will scatter X-rays with different amplitude and phase, and the overall combined scattered wave that is recorded can be thought of summing the scattering from all the individual atoms.

A single wave can also be represented in complex Vector form in two-dimensional real and imaginary space ( $i = \sqrt{-1}$ ), where the vector magnitude represents the wave magnitude and the vector direction is represented by the phase. This representation is called an Argand diagram, which is illustrated in Figure 2.7.

The vector representation can be mathematically expressed as:

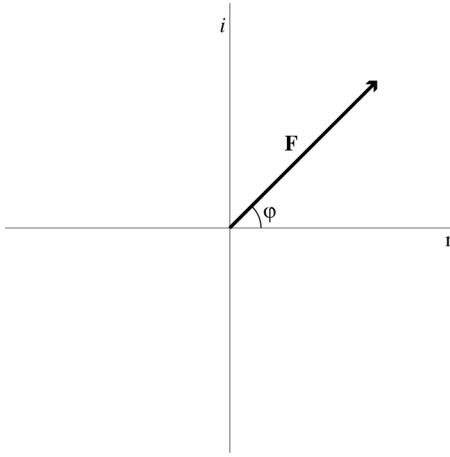
$$\mathbf{F} = |F|\cos \varphi + i|F|\sin \varphi \quad (2.12)$$

where  $\mathbf{F}$  is the vector representing the wave, and  $|F|$  is the magnitude and  $\varphi$  the phase of the vector. This vector representation of waves makes it easier to visualize the summation of multiple waves, simply by carrying out vector addition, which is illustrated in Figure 2.8.

Using a mathematical relationship, Eq. (2.12) can also be represented in exponential form:

$$\mathbf{F} = |F|\cos \varphi + i|F|\sin \varphi = |F|e^{i\varphi} \quad (2.13)$$

This exponential form describes the integrated form of the Fourier transform function. We will utilize this shorthand notation later in the chapter.



**Figure 2.7** Argand diagram showing the vector representation of a wave. The real and imaginary axes are horizontal and vertical, respectively. The magnitude of the wave is represented by the magnitude of the vector  $|F|$  and the phase by  $\varphi$ . Here, the phase of the wave is  $45^\circ$  or 0.125 cycles.

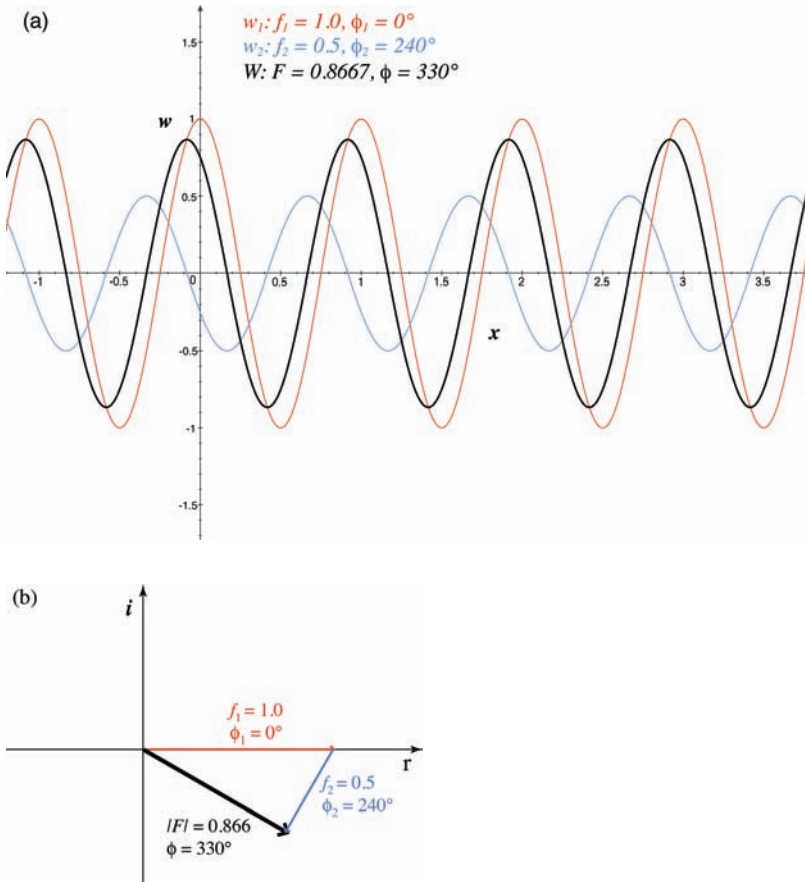
#### 2.4.2

##### Interaction of X-Rays with Matter

The X-ray has an electronic and magnetic component. When a charged particle is placed in an electric field it experiences a force. If the electric field oscillates at a frequency, the charged particle will experience an oscillating force, which will cause it to oscillate at the same frequency. A charged particle that oscillates produces a source of electromagnetic radiation with the same frequency as the oscillation, which is the same frequency as the incident radiation. This type of scattering is called Thomson scattering, which is the major contributor of scattered X-rays with matter. The intensity of the scattered X-ray is proportional to the charge-to-mass ratio squared, so the electrons will contribute six orders of magnitude more intensity than protons, which allows us to ignore proton contribution. Since different atoms have different numbers of electrons, the scattered intensity is proportional to the number of electrons squared,  $I \propto z^2$ . This assumes the electrons behave as free electrons, which is a good approximation for light atoms, but the inner core electrons of heavy atoms cause some alterations that will be discussed below. Therefore, in simple terms, X-ray scattering can be thought of X-ray absorption and re-emission at the same wavelength. However, the intensity of the scattered X-ray decreases with increasing scattering angle.

The scattering of X-rays from the volume of a single atom can be represented by the Fourier transform:

$$\mathbf{F}(\mathbf{s})_{\text{atom}} = \int_0^r \rho(\mathbf{r}) e^{2\pi i(\mathbf{r} \cdot \mathbf{s})} d\mathbf{r} \quad (2.14)$$



**Figure 2.8** Summation of waves. (a) Summation of waves 1 (red) and 2 (blue) to the combined wave shown in black. The amplitude and phase parameters of the waves are listed. The wavelength for all waves are the same (1.0).

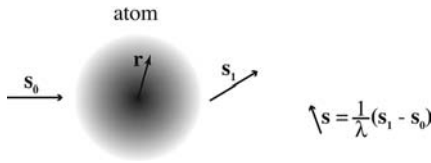
(b) The waves can be represented in vector form in an Argand diagram where the individual wave vectors are colored coded as in (a) with the same parameters.

where  $\rho(\mathbf{r})$  is the electron density and  $\mathbf{s}$  represents the vector of the scattered X-ray with respect to the incident X-ray scaled by  $1/\lambda$ . The vector  $\mathbf{r}$  represents the distance from the atomic center as shown in Figure 2.9. Integrating this over the volume of the atom will result in a magnitude corresponding to the number of electrons within the atom.

Equation (2.14) can be approximated with the equation:

$$\mathbf{F}(\mathbf{s})_{\text{atom}} = f_{\text{atom}} e^{2\pi i(\mathbf{r} \cdot \mathbf{s})} \quad (2.15)$$

where  $f_{\text{atom}}$  corresponds to the scattering factor of the atom. This value starts at the number of electrons for the atom, but drops exponentially as the scattering angle increases. The amount of thermal vibration decreases these values as well, especially at higher scattering angles. This value is known as the temperature factor or  $B$ -value,

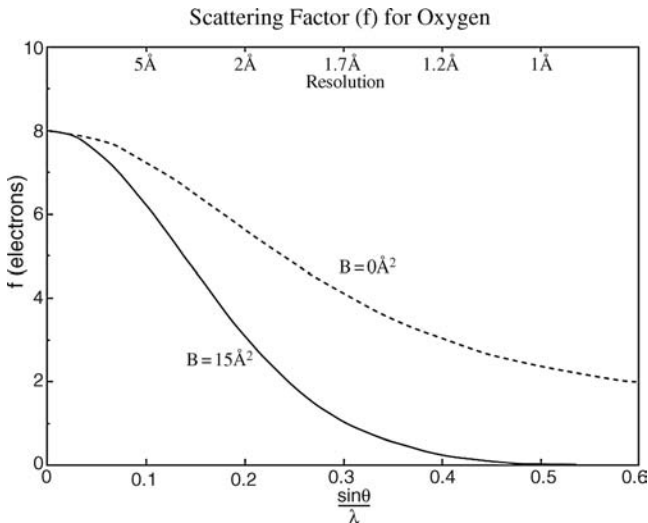


**Figure 2.9** Scattering of X-rays from an atom.  $s_0$  and  $s_1$  represent unit vectors of the incident and scattered X-ray, respectively.  $r$  is a distance vector with origin at the nucleus and extends to the atomic radius. The vector  $s$  represents the difference between the scattered and incident X-ray. These variables correspond to Eq. (2.14) above.

which has units of  $\text{\AA}^2$ . Figure 2.10 shows a plot of how the scattering factor decreases with scattering angle for an oxygen atom.

If we look at scattering from a molecule within the unit cell, Eq. (2.14) can still be used to describe the scattering, but now  $r$  covers the entire unit cell instead of a single atom (Eq. (2.17)). The choice of the origin of  $r$  represents the origin of the reference wave and all other scattered waves, are relative to the origin of  $r$ . Therefore, the origin is usually chosen as the origin of the unit cell and the magnitude of  $r$  simply represents the atomic coordinates  $(x, y, z)$  from the origin. However, now we integrate over all atoms within the unit cell. This equation can be simplified by the summation:

$$F(\mathbf{s}) = \sum_{j \text{ atoms}} f_j e^{2\pi i(\mathbf{r} \cdot \mathbf{s})} \quad (2.16)$$



**Figure 2.10** Plot of scattering factor for oxygen as a function of scattering angle at two different  $B$  values. Note at a scattering angle of zero, the scattering factor corresponds

to the number of electrons for oxygen. This value decreases for both  $B = 0$  and  $15 \text{ \AA}^2$ , but falls off faster with higher  $B$ -values.

where we sum over the scattering of all  $j$  atoms in the unit cell and  $f_j$  represents the scattering factor for each atom (which changes with the different types of atoms). The vector  $\mathbf{r}$  corresponds to the vector between the unit cell origin and the atomic coordinate for atom  $j$ . The vector  $\mathbf{s}$  is the same as described above.

The summation corresponding to  $\mathbf{F}(\mathbf{s})$  represents summing of the waves scattered from the  $j$  atoms in the cell and is called the structure factor.  $\mathbf{F}(\mathbf{s})$  represents a single wave with both a magnitude and a phase angle. For a single molecule or unit cell, there are no restrictions for the vector  $\mathbf{s}$  and  $\mathbf{F}$  is a continuous function as scattering angle changes. We will see below that the vector  $\mathbf{s}$  takes on certain point values for a repeating unit cell in a crystal due to diffraction.

Equation (2.14) is the form of a Fourier transform, which is a mathematical function that can be used to describe any function by summing cosine terms. The remarkable relationship of the Fourier Transform is the ease in calculating the inverse transform. Therefore, if the Fourier transform is described by the equation:

$$\mathbf{F}(\mathbf{s}) = \int_{\text{Unit Cell}} \rho(\mathbf{r})e^{2\pi i(\mathbf{r} \cdot \mathbf{s})} d\mathbf{r} \quad (2.17)$$

The inverse Fourier transform is:

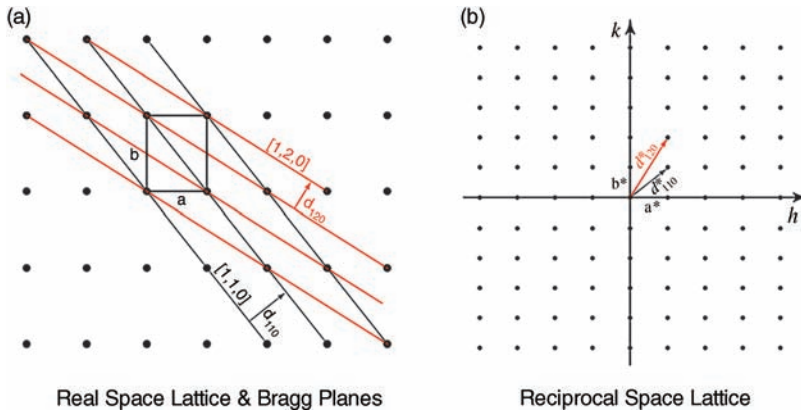
$$\rho(\mathbf{r}) = \int_{\text{Scattering Volume}} \mathbf{F}(\mathbf{s})e^{-2\pi i(\mathbf{r} \cdot \mathbf{s})} d\mathbf{s} \quad (2.18)$$

where we now integrate over the volume of scattering space  $\mathbf{s}$  (which will correspond to reciprocal space). Also note that there is now a minus sign before the  $2\pi$ . Therefore, if we know both the magnitude and phase of the scattered wave  $\mathbf{F}(\mathbf{s})$ , we can calculate the electron density  $\rho(\mathbf{r})$ , or in essence the atomic structure! However, the “phase problem” of X-ray crystallography is that there are currently no means possible to directly measure the phase of the scattered X-rays. We can, however, directly measure the magnitude of the scattered wave by the recorded intensity of the scattered X-ray. The intensity is proportional to  $|\mathbf{F}|^2$ . Once the phase can be calculated by other physical means, the crystallographer can unlock the puzzle of the atomic-resolution structure.

### 2.4.3

#### Crystal Lattice, Miller Indices, and the Reciprocal Space

The crystal is made up of a motif and lattice. Sets of parallel planes can cut through an ordered lattice that cut the unit cell an integral number of times. An analogy of seeing the sets of planes would be to look at an orchard of trees that are planted in a lattice. If one looks down the column of trees in one direction, one can see they line up in a set of lines with a specific spacing between parallel columns of trees. By looking at a different direction, one can see the trees line up to define another column of trees with a different spacing between the parallel sets of trees. If the lattice is very well ordered, planes (or lines in two dimensions) can be drawn with ever decreasing plane



**Figure 2.11** Real space lattice and corresponding reciprocal space lattice. (a) Real space lattice outlined by unit cell  $a$  and  $b$  (with  $c$  coming out of the page), along with two sets of planes corresponding to Miller index  $(1,1,0)$  and  $(1,2,0)$  in black and red, respectively. Vectors

indicating the interplane spacing ( $d$ ). (b) Reciprocal lattice corresponding to lattice in (a). Shown is only the  $a^*$  and  $b^*$  axis ( $h$  and  $k$ ), because  $c^*$  projects out of the page. Also shown are the  $d^*$  spacing for the sets of planes  $(1,1,0)$  and  $(1,2,0)$  with a magnitude of  $1/d$ .

spacing. Figure 2.11(a) shows a lattice defined by unit cell parameters  $a$  and  $b$ , with  $c$  coming out of the page and is not shown. Drawn on the lattice are two sets of planes. One set of planes cuts the  $a$  and  $b$  axis once, while the other set of parallel planes cuts the  $a$  axis once and the  $b$  axis twice within one unit cell repeat. Each set of planes can be defined by the Miller index  $(h,k,l)$ , which defines how many times the planes cut the unit cell axis. The planes corresponding to  $[1,1,0]$  and  $[1,2,0]$  are illustrated in Figure 2.11(a). The zero value for  $l$  indicates the planes are parallel to the  $c$  axis. Also drawn in Figure 2.11(a) is the vector perpendicular to the sets of planes, whose length corresponds to the interplane spacing ( $d_{hkl}$ ). Note that  $(1,1,0)$  interplane spacing is greater than the  $(1,2,0)$  set of planes. The  $(1,0,0)$  set of planes (not shown) would correspond to planes parallel to both the  $b$  and  $c$  axes, with the interplane spacing corresponding to unit cell distance of  $a$ .

Each set of planes can be plotted in a reciprocal lattice where the interplane spacing vector ( $d_{hkl}$ ) is plotted with the same direction, but the magnitude is now  $1/d_{hkl}$  or  $d_{hkl}^*$ . The reciprocal unit cell would correspond to  $d_{100}^*$ ,  $d_{010}^*$ ,  $d_{001}^*$ , which defines the  $a^*$ ,  $b^*$ , and  $c^*$  axis, respectively. The angles that define the reciprocal cell are  $\alpha^*$ ,  $\beta^*$ , and  $\gamma^*$ . However, one cannot simply state that the magnitude of  $a^*$  equals  $1/a$ . It does when the angle is  $90^\circ$  between axes, but the general conversion is given by the equations:

$$a^* = \frac{bc \sin \alpha}{V}; \quad b^* = \frac{ac \sin \beta}{V}; \quad c^* = \frac{ab \sin \gamma}{V}$$

where :

$$V = \frac{1}{V^*} = abc \sqrt{1 - \cos^2 \alpha - \cos^2 \beta - \cos^2 \gamma + 2 \cos \alpha \cos \beta \cos \gamma}$$

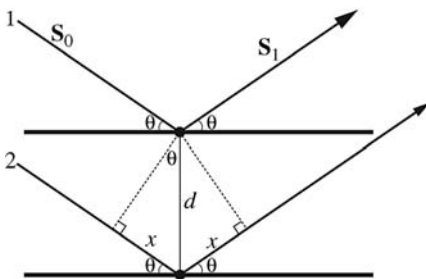
Each point in the reciprocal lattice has a Miller index  $(h,k,l)$  and represents an entire set of parallel planes that divides up the real space lattice. The distance of the point from the origin of the reciprocal lattice  $(0,0,0)$  is  $1/d_{hkl}$  for the set of planes in the real space. The direction of the vector between the origin and the point is perpendicular to the set of planes in real space. The reciprocal lattice is used to geometrically explain the X-ray diffraction from a crystal and will be explained in more detail later.

#### 2.4.4

#### X-Ray Diffraction from a Crystal: Bragg's Law

For an ordered crystal, if an atom falls on a plane that cuts through the unit cell, the atom is repeated for all unit cells and will fall on the same set of parallel planes. Thus, the planes will have a repeating pattern of point scatterers on each plane. If each atom scatters X-rays as illustrated in Figure 2.12, then only specific angles of scatter will constructively interfere to produce an additive signal or diffraction. When light waves are exactly in phase from different point scatterers, they constructively interfere to produce a strong combined light wave. If lightwaves 1 and 2 in Figure 2.12 are in phase in the incident waves, then for them to constructively interfere after scattering from atoms on the planes, wave 2 from the bottom plane has to be in phase with wave 1 from an atom on the top plane. Inspection of the two waves scattered from two different planes reveals that wave 2 has to travel an additional distance  $(x + x)$ . In other words, if the distance  $2x$  is equal to an integral number of wavelengths, the scattered wave from the bottom plane will be exactly in phase with the scattered wave from the top plane. Constructive interference will occur resulting in a diffraction maximum.

This difference in distance  $(2x)$  must equal an integral number of wavelengths, which can be represented by the equation of  $2x = n\lambda$ , where  $n$  is an integer. Using trigonometry,  $x = d \cdot \sin\theta$ , where  $d$  is the interplane spacing, and  $\theta$  is the angle between the incident X-ray and the plane. As a result, diffraction will occur when the



**Figure 2.12** Scattering of X-rays off two parallel planes. Two planes with an interplane spacing of  $d$  contain point scatterers above each other. Wave 2 scattered from the bottom plane must travel an extra distance  $(2x)$  than wave 1 from the top. If the scattered waves are in

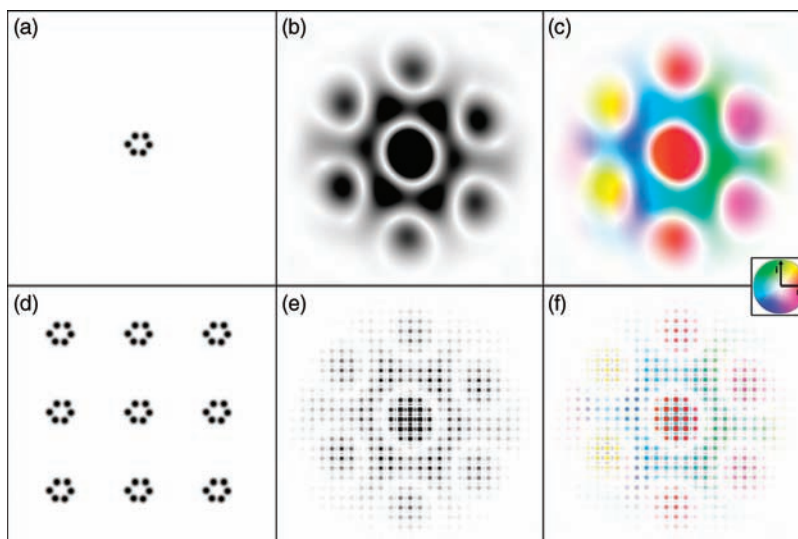
phase ( $2x$  equals integral number of wavelengths), then constructive interference occurs and a diffraction maximum can be measured.  $s_0$  and  $s_1$  represent unit vector direction of the incident and diffracted beam respectively.

all the variables ( $d$ ,  $\lambda$ , and  $\theta$ ) are at a value that satisfies the equation known as Bragg's law.

$$n\lambda = 2d \cdot \sin \theta \quad (2.19)$$

Even though Bragg's law equation was derived based on point scatterers directly above each other in Figure 2.12, the scatterers can lie anywhere on the plane separated by distance  $d$  and Bragg's law will still hold. Also, note that the diffracted X-ray beam is at the angle of  $2\theta$  with respect to the incident X-ray. When  $n$  is greater than 1, it corresponds to the higher-order diffraction maxima, but this is also equivalent to  $n = 1$  with the corresponding  $d$  spacing equal to  $d/n$  in the Bragg's law equation. Therefore, each lattice point in reciprocal space can be indexed with a unique Miller index  $(h,k,l)$  corresponding to  $n = 1$ . For example in Figure 2.11, the lattice point  $(2,0,0)$  for  $n = 1$  corresponds to a  $d$  spacing of half of  $d_{100}$  spacing, which is equivalent to drawing planes parallel to the  $b$ - $c$  plane that divide the  $a$  axis twice.

The consequence of the crystal lattice scattering of X-rays is a diffraction lattice, or discontinuous X-ray scattering. Thus, the scattered X-rays are only observed at distinct diffraction points. While we lose some information, recall one main reason for the crystal is to amplify the signal of the scattered X-ray. The overall diffraction pattern is made up of a convolution of scattering from the motif (protein) and scattering from the lattice, resulting in separate diffraction points. Figure 2.13 illustrates the difference between the scattering from a crystal lattice of motifs



**Figure 2.13** Illustration of X-ray scattering from a single motif and a crystal lattice of the motif. Top panels show the motif in (a), recorded X-ray scattering pattern only showing intensities in (b) and the scattering pattern with phase information in (c). The bottom panels

show X-ray scattering of the same motif but arranged in a crystal lattice. Note the intensity and phase information is the same as the single motif, but is a discontinuous function as a consequence of diffraction from the lattice.



compared to that of a single motif. The left-hand panels show the two-dimensional structure of a motif. The center panels show the recorded X-ray scattering; intensity only with no phase information. The panels on the far right show the scattering with phase information as a color spectrum, the insert shows the color conversion on an Argand diagram.

Ideally if one can record the X-ray scattering from a single motif more information would be recorded, but the signal from X-ray scattering off a single molecule is too weak to detect with present detectors. One way to increase the signal is to use a much more intense X-ray beam. Construction and experiments are underway now to create very intense X-ray free-electron lasers that might be able to record scattering from a single molecule [11, 12]. However, many challenges must still be overcome, like determining orientation of the molecule to accurately represent three-dimensional scattering space, how to obtain data from only one single molecule at a time (like using an aerosol), the effects of hydration or dehydration on scattering and protein stability in the aerosol, and engineering very sensitive X-ray detectors. One of the main hurdles is the intensity of the X-ray beam required for these experiments is so great that inelastic scattering would contribute significantly to ionization of the sample, leading to plasma formation and a Coulombic explosion. Therefore, the exposure and data collection would have to happen in the femtosecond timescale in order to capture the X-ray scattering before the protein starts to obliterate.

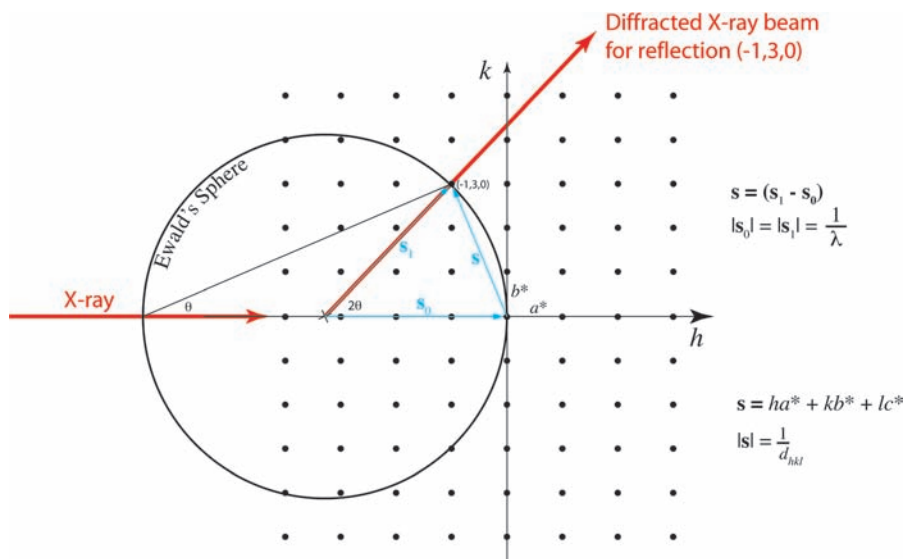
#### 2.4.5

##### **Bragg's Law in Reciprocal Space**

If one draws a sphere in the reciprocal lattice with radius of  $1/\lambda$  ( $\lambda$  is wavelength of the X-ray), such that the right edge of the circle lies on the reciprocal space origin and the left edge lies on a vector from the direction of X-rays, a simplified model of X-ray diffraction geometry can be constructed (Figure 2.14). A right-triangle can be drawn for any reciprocal lattice point that falls on the surface of the sphere (as reflection  $(-1, 3, 0)$  in Figure 2.14). The hypotenuse of the triangle would be coincident with the X-ray beam along the diameter of the sphere with length of  $2/\lambda$ . Another side of the triangle would correspond to a vector between the reciprocal lattice origin and the lattice point. The final side of the triangle would run from the lattice point and the left edge of the sphere. If we call the angle between this final side and the direct X-ray beam  $\theta$ , then:

$$\sin \theta = \frac{1/d_{hkl}}{2/\lambda} = \frac{\lambda}{2d_{hkl}} \quad (2.20)$$

This equation is indeed Bragg's law for lattice point  $hkl$  and  $n = 1$ . Additionally, the angle between the center of the sphere and the lattice point that falls on the sphere would be  $2\theta$ , which corresponds to the angle of the diffracted beam with respect to the incident beam. Therefore, the reciprocal lattice drawn with this sphere of reflection, known as Ewald's sphere, beautifully illustrates the geometry of diffraction for the different planes that "slice" through the unit cell. Keep in mind that each



**Figure 2.14** Bragg's law in reciprocal space. Shown is reciprocal space lattice from Figure 2.11. Drawn on the lattice is Ewald's sphere of reflection with a radius of  $1/\lambda$  ( $x$  marks the center of sphere). The vector  $\mathbf{s}$  can be defined in general by  $h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ , where  $\mathbf{a}^*$ ,

$\mathbf{b}^*$ , and  $\mathbf{c}^*$  are the basis vectors for the reciprocal space lattice. For the vector  $\mathbf{s}$  drawn in the figure,  $h = -1$ ,  $k = 3$ ,  $l = 0$ .  $\mathbf{s}_0$  and  $\mathbf{s}_1$  represent the vectors of incident and diffracted beam, respectively with a magnitude of  $1/\lambda$ .

reciprocal lattice point represents a set of real-space Bragg planes with the vector from the reciprocal lattice origin to the lattice point representing the normal to the real-space planes with a magnitude equal to the reciprocal of the interplane spacing ( $1/d_{hkl}$ ). Thus, one can think of the reciprocal space lattice occurring at the point of the crystal in the X-ray beam in order to determine or predict the diffraction pattern of a given crystal with known unit cell parameters and orientation. Conversely, there are algorithms that can determine three-dimensional reciprocal lattice dimensions and orientation given the diffraction pattern, and a known distance between the crystal and the detector.

In Figure 2.14 above, reflection  $(-1,3,0)$  falls on the sphere and therefore satisfies Bragg's law and would give rise to a diffracted beam illustrated with the bold red arrow. As shown in Figure 2.14, reflection  $(0,3,0)$  does not fall on Ewald's sphere and therefore does not satisfy Bragg's law at the current crystal orientation. Nevertheless, keep in mind that the reciprocal space orientation is derived from the real-space crystal lattice, so as the crystal is rotated, it has the effect of rotating the reciprocal lattice by the same amount. Thus, in order to bring the  $[0,3,0]$  real-space planes into an angle that satisfies Bragg's law (to record the intensity scattered off these planes), the crystal needs to be rotated around  $20^\circ$  counter-clockwise around an axis perpendicular to the page. Ewald's sphere remains fixed as defined by the incident X-ray beam, so only the lattice rotates.

## 2.4.6

**Fourier Transform Equation from a Lattice**

The vector  $\mathbf{s}$  in Figure 2.14 is the same vector in Eqs. (2.16)–(2.18) and represents the component of scattering space. Owing to the crystal lattice, X-ray scattering can only be measured at certain values of  $\mathbf{s}$  (at specific reciprocal lattice points, or diffraction maxima that satisfy Bragg's law). This makes the scattering space discontinuous (Figure 2.13e). The vector  $\mathbf{s}$  can only be measured at general points:

$$\mathbf{s} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \quad (2.21)$$

where  $\mathbf{a}^*$ ,  $\mathbf{b}^*$ , and  $\mathbf{c}^*$  represent a set of basis vectors for reciprocal space. Likewise, the vector  $\mathbf{r}$  also takes on certain values that correspond to specific atomic positions:

$$\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c} \quad (2.22)$$

where  $x$ ,  $y$ , and  $z$  are fractional unit-less coordinates that range from 0 to 1, and  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{c}$  are a vector representation of the unit cell, whose magnitudes are unit cell parameters. Therefore the dot product of  $\mathbf{r} \cdot \mathbf{s}$  in Eqs. (2.16)–(2.18) equals:

$$\mathbf{r} \cdot \mathbf{s} = hx + ky + lz \quad (2.23)$$

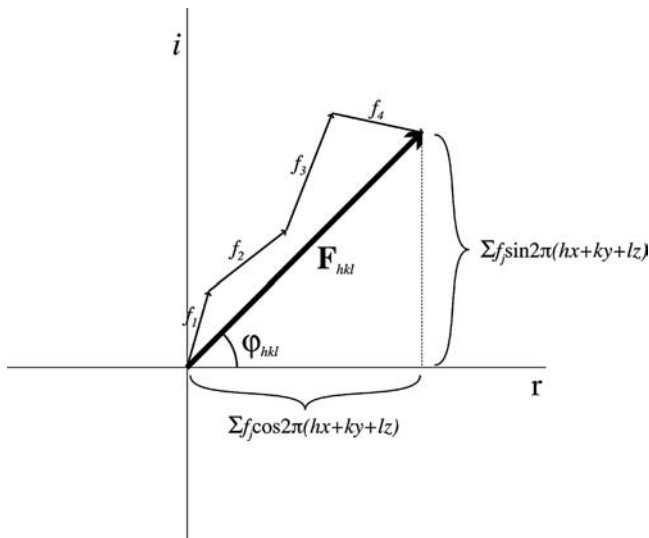
because the dot product of  $\mathbf{a} \cdot \mathbf{a}^* = \mathbf{b} \cdot \mathbf{b}^* = \mathbf{c} \cdot \mathbf{c}^* = 1$  and all dot products with mixed vectors are zero (e.g.,  $\mathbf{a} \cdot \mathbf{b}^* = \mathbf{a} \cdot \mathbf{c}^* = \dots = 0$ ). Therefore the structure factor Eq. (2.16) now becomes:

$$\mathbf{F}(hkl) = \sum_{j \text{ atoms}} f_j e^{2\pi i(hx + ky + lz)} \quad (2.24)$$

or in the form of cosine and sine addition:

$$\mathbf{F}(hkl) = \sum_{j \text{ atoms}} f_j [\cos 2\pi(hx + ky + lz) + i \sin 2\pi(hx + ky + lz)] \quad (2.25)$$

The structure factor  $\mathbf{F}(hkl)$  is equivalent to adding together the scattering waves of all the individual atoms in the unit cell to one overall combined wave, whose intensity is recorded on the X-ray detector. This can be illustrated in an Argand diagram in Figure 2.15 below, which shows a sample structure consisting of only four atoms. Shown is the structure factor for one single  $hkl$  reflection, which diffracted off the  $hkl$  Bragg planes. This diagram can be repeated for the tens of thousands of diffraction reflections that are recorded for a typical protein crystal. The diagram is drawn assuming we know the position of the four individual atoms so we can properly draw the individual atomic scattering factors with both a magnitude and direction, leading to an overall summed wave  $\mathbf{F}(hkl)$  with a known phase  $\varphi(hkl)$ , the tangent of which ( $\tan\varphi$ ) is equal to the sum of the sine terms divided into the sum of the cosine terms (Eq. (2.11)). In experimental structure determination, the recorded intensity only gives the amplitude of the structure factor ( $|\mathbf{F}|$ ) without the phase.



**Figure 2.15** Argand diagram illustrating a structure factor for the single reflection,  $hkl$ , from a unit cell consisting of four known atoms. This assumes the atomic coordinates are known for the four atoms to draw the individual scattering factors with known magnitude and direction.

#### 2.4.7

#### Friedel's Law and the Electron Density Equation

Inspection of the structure factor Eq. (2.25) reveals that the structure factor for reflection  $-h, -k, -l$  will have the same amplitude as reflection  $h, k, l$  but negative phase ( $\varphi_{hkl} = -\varphi_{-h-k-l}$ ) because the sine term will be negative, while the cosine term is the same ( $A_{hkl} = A_{-h-k-l}$  and  $B_{hkl} = -B_{-h-k-l}$ , where  $A$  and  $B$  represent the summation of the cosine and sine terms, respectively). Another way to think of this is to consider a set of planes like  $[1,0,0]$  and  $[-1,0,0]$  that divide a unit cell in the crystal. These planes are parallel to the  $b-c$  plane and cut the unit cell along  $a$  once. They both describe the same set of planes so one would expect the intensity to be identical. However, what distinguishes them is the reciprocal space. The two reflections are centrosymmetric in that they are related by a  $180^\circ$  rotation of the crystal to intersect Ewald's sphere to satisfy Bragg's law. This relationship is known as Friedel's law, which gives the relationship that the intensity of reflection of  $hkl$  is the same for  $-h-k-l$  ( $I_{hkl} = I_{-h-k-l}$  which results in  $|\mathbf{F}_{hkl}| = |\mathbf{F}_{-h-k-l}|$ ).

Given the structure factor equation from a lattice above, which represents a Fourier Transform, the inverse Fourier transform for the electron density equation now becomes:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l \mathbf{F}(hkl) e^{-2\pi i(hx + ky + lz)} \quad (2.26)$$

where the function becomes a triple summation of all reflections of reciprocal space, and the value of  $1/V$  ( $V$  = unit cell volume) is added to give units of electron density ( $e^-/\text{\AA}^3$ ). Keep in mind that the value of  $\mathbf{F}(hkl)$  represents a vector with a magnitude and phase, which can be split up to the magnitude component and phase component:

$$\mathbf{F}(hkl) = |\mathbf{F}_{hkl}|e^{i\varphi} \quad (2.27)$$

where  $\varphi$  is the phase angle in radians. This form of the structure factor can be incorporated in the electron density Eq. (2.25) to give:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |\mathbf{F}_{hkl}| e^{-2\pi i(hx + ky + lz - \varphi_{hkl})} \quad (2.28)$$

where  $\varphi_{hkl}$  is the phase in cycles (0–1) for the reflection  $hkl$ . Therefore the electron density at a specific point in the unit cell ( $xyz$ ) is calculated by summing overall all the scattering reciprocal space reflections ( $hkl$ ). This requires the repetitive summing of tens of thousands of terms for proteins, which would take many man-hours of time without the aid of a computer!

The electron density equation still has the imaginary component of “ $i$ ” in the formula, which must be eliminated to calculate a true real function. The electron density Eq. (2.29) can be broken up into the summation of cosine and sine terms to give:

$$\begin{aligned} \rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |\mathbf{F}_{hkl}| \cdot [\cos 2\pi(hx + ky + lz - \varphi_{hkl}) \\ + i \sin 2\pi(hx + ky + lz - \varphi_{hkl})] \end{aligned} \quad (2.29)$$

The electron density equation requires the summation of all scattering space or reciprocal space, which corresponds to all reflections. However, given Friedel’s law of  $|\mathbf{F}_{hkl}| = |\mathbf{F}_{-h-k-l}|$  and  $\varphi_{hkl} = -\varphi_{-h-k-l}$  and the trigonometric relationship of  $\sin(-x) = -\sin(x)$  when one sums over all the terms, the Friedel related reflections results in canceling out the sine term in Eq. (2.29) to give the following real function for electron density:

$$\rho(xyz) = \frac{2}{V} \sum_{\frac{h}{2}} \sum_k \sum_l |\mathbf{F}_{hkl}| \cdot \cos 2\pi(hx + ky + lz - \varphi_{hkl}) \quad (2.30)$$

where the summation is now over half the reflections (positive  $h$  values) because the Friedel related reflections cancel the sine term, and the factor of 2 is incorporated because  $\cos(x) = \cos(-x)$  so the cosine term needs to be doubled when summing half the reflections.

To calculate the real-function electron density map, both the magnitude and the phase of each diffraction reflection must be known to calculate the correct electron density or structure. The magnitude is easy to experimentally obtain because it equals the square root of the recorded intensity. The phase problem is more challenging to determine and different methods will be outlined below.

## 2.5

### Collecting and Processing Diffraction Data

#### 2.5.1

##### Data Collection Strategy

Before we can calculate the electron density maps, the diffraction intensities must first be measured. As discussed in Figure 2.14, one way to visualize which Bragg planes satisfy Bragg's law is to imagine a reciprocal lattice construction occurring at the crystal. First, one needs to know the crystal unit cell parameters and the orientation before one can plan a strategy of rotating the crystal such that all unique reciprocal lattice points (reflections) intersect Ewald's sphere. Typically, the crystallographer puts a crystal in the X-ray beam and takes a couple of setting diffraction patterns with the crystal rotated  $90^\circ$  between shots. Computer algorithms today can accurately calculate a reciprocal space lattice with dimensions and orientation from these setting diffraction images.

The next step is to determine the space group crystal class to plan a data collection strategy. The crystal class can usually be obtained by the unit cell parameters from the setting shots. However, occasionally the computer algorithm may assume the wrong crystal class because of small errors. For example, if one has a monoclinic crystal with the  $\beta$  angle very close to  $90^\circ$ , the software may assume the crystal is orthorhombic ( $\alpha = \beta = \gamma = 90^\circ$ ). The correct crystal class will become apparent upon merging the collected data together. Additionally, if the setting shots give parameters of  $a = b \neq c$ ,  $\alpha = \beta = 90^\circ$ ,  $\gamma = 120^\circ$ , the crystal may be trigonal or hexagonal since both have the same lattice restrictions. Therefore, it is best to assume the lowest symmetry when planning a data collection strategy or to process the data before removing the crystal from the X-ray instrument.

Diffraction data are typically collected using what is called the "oscillation method," where the crystal is oscillated back and forth by a small amount ( $1^\circ$ ) while being exposed to X-rays, to record the maximum number of reflections intersecting Ewald's sphere without causing overlaps. The data are typically recorded on CCD area detectors, which are two-dimensional flat surfaces (like film). Too large an oscillation angle will cause many more reflections to intersect Ewald's sphere, which will cause overlapped reflections on the area detector making it difficult or impossible to record the intensity from each single reflection.

Typically, the data processing software fits a histogram profile to the individual pixels (from the CCD) of each reflection "peak." The program then integrates the volume under the profile to accurately quantitate the overall intensity from the reflection. The software "knows" exactly where to "look" for each reflection on a specific diffraction image because it can be accurately predicted based on the orientation and parameters of the reciprocal space lattice (using the reciprocal space lattice). If the predicted reflection positions do not match up accurately with the observed, this high root-mean-squared (rms) deviation parameter can be a flag indicating a wrong crystal class was chosen or the crystal may be twinned giving rise to two lattices in very similar orientations.

## 2.5.2

**Symmetry and Scaling Data**

The symmetry within the crystal is reflected in the symmetry of the diffraction pattern. Figure 2.13 shows scattering from a molecule with approximate 6-fold symmetry, which is evident in the scattering reciprocal space (Figure 2.13b and e). For exact symmetry, the scattering reciprocal space will display exact intensity distributions, therefore to save time or crystal degradation, one does not need to collect data from all reciprocal space. Friedel's law also indicates we do not need to collect all reciprocal space and only need to collect a hemisphere of reciprocal space for triclinic crystals that contain no symmetry. (Later we will see that slight variations of Friedel's law break down for heavy atoms at certain X-ray energies, whose scattering information can be used to calculate phases, so Friedel related reflections need to be collected.) Monoclinic crystals with 2-fold symmetry only need to collect a fourth of reciprocal space and orthorhombic crystals with 222 symmetry only requires one-eighth of reciprocal space because the other reflections are theoretically identical.

With a known crystal lattice (and reciprocal lattice) orientation, algorithms are developed to automatically plan data collection strategies to rotate the crystal during data collection to bring all reciprocal space points (which represent planes) into Ewald's sphere and thus measure the X-ray diffraction intensity. Ideally, one needs to record the reflection intensity more than once from symmetry-related reflections. The data are then scaled and merged together into a unique region of reciprocal space that is used for subsequent calculations. The parameter that measures the intensity similarity between symmetry-related reflections is the  $R_{\text{merge}}$  value, which is calculated by:

$$R_{\text{merge}} = \frac{\sum_{hkl} \sum_i |\bar{I}_{hkl} - I_{hkl}(i)|}{\sum_{hkl} \sum_i I_{hkl}(i)} \quad (2.31)$$

where  $\bar{I}_{hkl}$  is the average intensity of the symmetry-related  $hkl$  reflection group and  $I_{hkl}(i)$  is the intensity of each individual reflection in the specific  $hkl$  reflection group. The  $R_{\text{merge}}$  value should be less than around 8% for precisely determined data, but higher values are acceptable for weakly diffracting crystals, or low-resolution crystals. If the wrong crystal class is chosen, (e.g., the computer selected orthorhombic instead of monoclinic with  $\beta \approx 90^\circ$ ) the  $R_{\text{merge}}$  will be large (above 30%) and the data should be reprocessed with a different crystal class.

## 2.6

**Solving the Structure (Determining Phases)**

## 2.6.1

**Molecular Replacement**

Often, the proteins under study will display amino acid sequence similarity to other proteins whose structure may have been previously determined. If a structure is

known, the structure factor ( $\mathbf{F}_{hkl}$ ) can be calculated from the atomic coordinates of the homologous protein using Eq. (2.25) and illustrated in Figure 2.15. If the protein under study displays greater than 30% sequence identity to another known protein, it will likely have the same overall protein fold and phases calculated from the known structure can be utilized in the electron density equation:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l \left| \mathbf{F}_{\text{obs}}(hkl) \right| e^{-2\pi i(hx + ky + lz - \varphi_{\text{calc}}(hkl))} \quad (2.32)$$

In this equation, the hybrid summation is calculated using the structure factors observed from the unknown structure ( $\mathbf{F}_{\text{obs}}(hkl)$ ) with the phases determined from the similar known structure ( $\varphi_{\text{calc}}(hkl)$ ). This process is called molecular replacement [13]. The resultant electron density map will contain features of both the known and unknown structure. The crystallographer then interprets features of the electron density map and builds a model of the unknown structure by altering the known model. This is done with the help of a computer graphics program like COOT [14] or O [15], which can display atomic models with electron density maps. Ideally it is best to truncate the nonconserved residues to alanine (or the entire protein) before phases are calculated. This way all observed electron density corresponding to amino acid side-chains will come from the experimentally observed scattering data and not be biased by the known protein model that is used to calculate phases.

To calculate the phases from the known structure that is applied to the unknown protein, the known structure must be situated in the same orientation and position within the unit cell of the unknown protein. There are a number of programs that can do this (e.g., PHASER [16], EPRM [17], AMoRe [18], CNS [19], GLRF [20]). These programs perform two searches – first a rotation search or function followed by a translation search. The rotation function finds the correlation of the reciprocal space calculated from the known structure superimposed and rotated on the reciprocal space of the unknown structure. Thus, the rotation function measures the correlation as a function of rotation angles, which can be described in simplified form of:

$$R(\phi, \psi, \kappa) = \sum_{hkl} \sum_{h'k'l'} \left| \mathbf{F}_{hkl} \right|^2 \left| \mathbf{F}_{h'k'l'}(\phi, \psi, \kappa) \right|^2 G_{hkl,h'k'l'} \quad (2.33)$$

where  $\mathbf{F}_{hkl}$  is the unknown reciprocal lattice that is fixed, and  $\mathbf{F}_{h'k'l'}(\phi, \psi, \kappa)$  is the reciprocal lattice magnitudes from the known structure rotated by the polar angles  $\phi$ ,  $\psi$ , and  $\kappa$ . The  $G$  function is an interpolation function to scale reciprocal lattice magnitudes because rotated lattice points will not fall exactly on the fixed reciprocal lattice points from the unknown structure. When the rotation function ( $R(\phi, \psi, \kappa)$ ) has a high value, the rotated lattice magnitudes of the known structure match the magnitudes of the unknown structure. The resultant angles of rotation ( $\phi, \psi, \kappa$ ) are applied to rotate the known structure to position it in the same orientation as the unknown structure.

Once the orientation of the known structure is matched to the unknown structure, the next step is to find where within the unit cell the protein is situated. This is done by what is called the translation search. With today's speedy



computers, the quickest way to determine the translation function is simply trial and error. This is accomplished by taking the known protein (with correct orientation determined in the rotation function) and simply translating, by user-defined steps, along all three directions within the unit cell and calculating a correlation coefficient at each step. The correct position in the unit cell is determined when there is a high correlation of the calculated magnitudes, based on the known structure in the unknown unit cell, compared to observed magnitudes that were recorded from the unknown structure. The correlation coefficient CC is defined as:

$$CC = \frac{\sum_{hkl} \left( |F(\text{obs})|^2 - \overline{|F(\text{obs})|^2} \right) \left( |F(\text{calc})|^2 - \overline{|F(\text{calc})|^2} \right)}{\sqrt{\sum_{hkl} \left( |F(\text{obs})|^2 - \overline{|F(\text{obs})|^2} \right)^2 \sum_{hkl} \left( |F(\text{calc})|^2 - \overline{|F(\text{calc})|^2} \right)^2}} \quad (2.34)$$

The molecular replacement method will likely be the most common technique to determine protein structures (if not the only method) in the future because of the ever-increasing number of protein structures being determined. Eventually, all of protein folding space will be covered, so there will likely be a similar structure known when initiating structural studies of a new protein either from a new organism, to investigating catalytic details, examining details of protein–protein interactions, or understanding how chemical regulators may interact and regulation function.

## 2.6.2

### Isomorphous Replacement

The method of isomorphous replacement was the most common method of phase determination for novel protein structures up to about the late 1990s. The first protein structure determinations of hemoglobin and myoglobin were determined with this method. While the practice is not as common in the present day, it is introduced here to establish a foundation of principles for the more common method of multiwavelength anomalous dispersion (MAD) discussed in Section 2.6.3.

Isomorphous replacement, as the name suggests, is substituting an element onto a protein structure without altering its overall structure. Then phase information can be obtained by comparing the substituted isomorphous structure to the native structure. Proteins contain thousands of relatively light atoms like hydrogen, carbon, nitrogen, and oxygen, so to record a measurable difference in X-ray scattering while still maintaining an isomorphous structure a small strongly scattering center has to be introduced. This is achieved by introducing a much heavier atom on the protein surface by soaking native protein crystals with heavy atom reagents. The most common heavy atoms used for this method include complexes with platinum, gold, mercury, uranium, iodine, and lead. While some of these metals can form covalent complexes with the proteins (especially mercury), many will bind to the protein surface through amino acid residues that serve as

coordination ligands to the metal. The metal ions will bind to residues according to the hard–soft theory of acids and bases, where “soft” refers to polarizable and “hard” are not polarizable. Most of the common metal ions are “soft” metals so they prefer the softer ligands like sulfides of cysteines and imidazole of histidine. The hard metals will bind to the ligands like carboxylate groups of aspartic and glutamic acids.

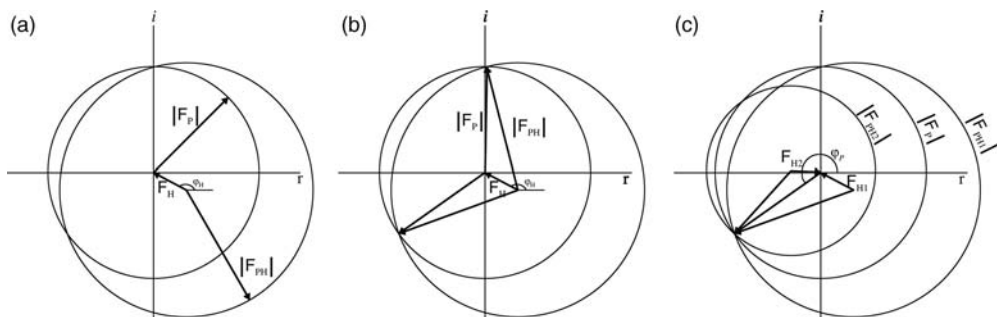
Since the protein structure is unknown, the surface residues are not known, and therefore the process of making isomorphous derivatives is a trial and error procedure. Many crystals are required to screen the numerous heavy metal compounds. Typically protein crystals are soaked in low concentration of heavy metal complex (e.g., 1 mM) for a few hours to overnight. Then data are collected on the soaked crystal and compared to the native nonsoaked crystal. First, the diffraction quality is determined to see the limit of diffraction resolution. Crystals that diffract poorly will be of little help. The metal ion can disrupt crystal contacts, resulting in poor diffraction. Additionally, the metal soaks can cause subtle changes in the unit cell due to lattice contact changes. If the lattice parameter changes more than a small percentage, the crystals are no longer isomorphous and will not yield accurate phase information. If the crystals diffract well with isomorphous unit cell parameters, the data are compared to native data by matching the individual reflection intensities to determine a derivative or isomorphous  $R$ -factor:

$$R_{\text{der}} = R_{\text{iso}} = \frac{\sum_{hkl} |I_{\text{der}}(hkl) - I_{\text{nat}}(hkl)|}{\sum_{hkl} I_{\text{nat}}(hkl)} \quad (2.35)$$

where we compare the average intensity for each derivative-soaked crystal to the native nonsoaked crystal. If the  $R_{\text{iso}}$  is low (below 15%) it is likely that no derivatization took place because the two data sets scale together fairly reasonably. If  $R_{\text{iso}}$  is high (above 15%) there is some indication that the protein was altered by metal ions binding to the surface. (Nonisomorphism also results in high  $R_{\text{iso}}$ , but it is usually more apparent at high resolution so the crystallographer needs to inspect how  $R_{\text{iso}}$  changes with resolution.)

Recall that the structure factor ( $\mathbf{F}(hkl)$ ) is determined by summing up the scattering of all the atoms. This also will include the summation of a heavy atom, which can be separated into two vectors – one for the protein and one for the heavy atom:  $\mathbf{F}_{\text{PH}}(hkl) = \mathbf{F}_{\text{P}}(hkl) + \mathbf{F}_{\text{H}}(hkl)$ , where  $\mathbf{F}_{\text{PH}}$  represents protein + heavy atom. To obtain phase information that can be applied to the protein structure factors, the position of the heavy atoms is essential because one can calculate both the vector magnitude and phase for  $\mathbf{F}_{\text{H}}$ .

By subtracting the structure factor magnitude of the protein ( $|\mathbf{F}_{\text{P}}|$ ) from the structure factor magnitude of the protein + heavy atom ( $|\mathbf{F}_{\text{PH}}|$ ) one can approximate the structure factor magnitude for the heavy atom alone ( $|\mathbf{F}_{\text{H}}|$ ). This now becomes a simplified small-molecule-type problem because we are considering scattering only from a few heavy atoms. One technique of solving few-atom structures is the Patterson method. Patterson maps are calculated by using only squared structure



**Figure 2.16** Harker diagram showing the phase determination from isomorphous replacement. (a) Construct of a protein and protein soaked with heavy atom. Only the magnitudes of the structure factors are known ( $|F_P|$  and  $|F_{PH}|$ ), so they are drawn as circles with a radius corresponding to the magnitude. The vector with both magnitude and phase is drawn

for  $F_H$ . (b) The circles intersect at two points ( $90^\circ$  and  $220^\circ$ ), both of which would satisfy the vector addition. The ambiguity can be solved by introducing another circle construct from another derivative ( $F_{PH2}$ ), which in this example reveals the phase angle for the protein is around  $220^\circ$  (or 0.611 cycles).

factor magnitude values with no phase information. The Patterson function is calculated by:

$$P(uvw) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{\text{obs}}(hkl)|^2 e^{-2\pi i(hu + kv + lw)} \quad (2.36)$$

Here, the Patterson map uses coordinates designated  $(u, v, w)$  because it is not a real electron density map. The Patterson maps give peaks that correspond to the interatomic distances with a peak height corresponding to the product of the number of electrons from each atom for the specific distance. For example, if a one-dimensional map is considered with symmetry such that every atom at  $x$  also occurs at  $-x$ : if a carbon is situated at  $x$  and  $-x$  then the distance between these atoms would be  $2x$  with a peak height of 12 electrons. So if inspection of the Patterson map reveals that a peak with height about 12 occurs at  $u = 0.3$ , then we can determine that  $0.3 = 2x$ , so the carbon  $x$  coordinate is 0.15 (and  $-0.15$  or 0.85).

The heavy atom positions soaked into the protein are determined by Patterson methods, but using the structure factor magnitude of  $(|F_{PH}| - |F_P|)^2$ , which approximates to  $(|F_H|)^2$ . Once the heavy atom positions are known, then both the magnitude and phase of the  $F_H(hkl)$  vector can be calculated for each reflection (based on heavy atoms positions only). This does not give any information on the phase for the protein ( $F_P(hkl)$ ) directly but by knowing the vector for  $F_H(hkl)$  we can determine the phase using a Harker diagram (Figure 2.16). In Figure 2.16(a), the magnitude of both  $F_P$  and  $F_{PH}$  are known, but not the phase, so they are drawn as circles with radius equal to magnitude. The  $F_H$  vector is drawn with known magnitude and phase ( $\varphi_H$ ) because its atomic position was determined from the Patterson map. The  $F_H$  vector is drawn with its head at the origin because the  $F_P$  vector is drawn with its tail starting at the origin and the  $F_{PH}$  vector starts at the tail of the  $F_H$  vector. Now the two circles

intersect at two points (around  $90^\circ$  and around  $220^\circ$  in Figure 2.16a), both of which would satisfy the vector addition of  $\mathbf{F}_{\text{PH}} = \mathbf{F}_{\text{P}} + \mathbf{F}_{\text{H}}$  (Figure 2.16b). This ambiguity can be resolved if another heavy atom derivative was generated and another circle drawn with  $\mathbf{F}_{\text{PH}_2}$  and  $\varphi_{\text{H}_2}$  for the second derivative (Figure 2.16c). Now the phase for the protein alone ( $\mathbf{F}_{\text{P}}$ ) can be determined to be around  $220^\circ$  in this example because it satisfies the vector summation for both derivatives. Now the phase was determined for one single reflection. The computer program repeats these Harker diagrams for each of the thousands of reflections recorded for a typical protein crystal.

### 2.6.3

#### MAD

The advent of synchrotron radiation, sensitive area detectors, and molecular biology protocols have combined to foster another method of phase determination that has been commonly used since the 1990s. This method, called MAD, exploits the physics that the inner core electrons of heavy atoms do not scatter X-rays as the outer electrons. The inner core electrons of the heavy atoms have a wave function frequency that falls in the range of X-ray frequencies. If an electron has an orbital frequency comparable to the frequency of the X-ray radiation, the electron cannot be considered a free electron and some absorption will occur resulting in anomalous scattering. This anomalous scattering results in a phase shift of  $90^\circ$  from the electrons at the inner core. The overall scattering factor vector can be broken down to a normal scattering part and an anomalous scattering with a real and imaginary phase component:

The diagram illustrates the decomposition of the anomalous scattering factor vector  $f_{\text{anom}}$ . A vector  $f_{\text{anom}}$  is shown originating from the origin and pointing into the first quadrant. A horizontal vector  $f_o$  is drawn below it, representing the normal scattering factor. A vertical vector  $\Delta f''$  is drawn from the tip of  $f_o$  to the tip of  $f_{\text{anom}}$ . A horizontal vector  $\Delta f'$  is drawn from the tip of  $f_o$  to the tip of  $f_{\text{anom}}$ . The diagram shows that  $f_{\text{anom}}$  is the sum of  $f_o$ ,  $\Delta f'$ , and  $\Delta f''$ .

$$f_{\text{anom}} = f_o + \Delta f' + i\Delta f''$$

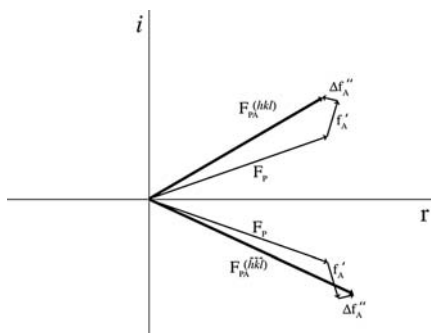
or

$$f_{\text{anom}} = f' + i\Delta f''$$

Here,  $f_o$  is the normal scattering factor without anomalous scattering. The anomalous scattering component can be broken down into a real component ( $\Delta f'$ ), which is usually negative, and an imaginary component  $\Delta f''$  ( $90^\circ$  phase shift). The  $f_{\text{anom}}$  would correspond to the scattering factor of the heavy atom experiencing anomalous scattering. The vectors of the anomalous scattering component ( $\Delta f'$  and  $\Delta f''$ ) can change magnitude independent of each other at different X-ray energies. The largest values occur at and very near the absorption edge for the heavy atom.

When a heavy atom with anomalous scattering is mixed with lighter atoms that do not scatter anomalously, Friedel's law breaks down because of the anomalous phase shift (Figure 2.17). Therefore, the intensities of reflection  $hkl$  are slightly different than those for reflection  $-\bar{h}\bar{k}\bar{l}$ . Since the  $f'$  and  $f''$  components are typically a few electrons, this difference in intensity is very small, but the more sensitive CCD X-ray detectors that are routinely used for data collection can measure a difference in intensities. However, highly redundant data sets are usually required to accurately measure these small differences.

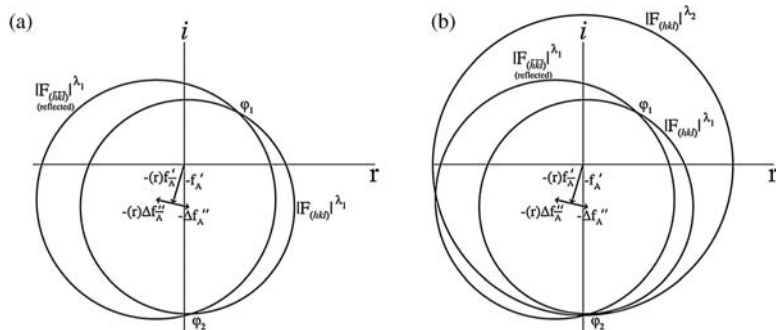
Like the isomorphous replacement method, the position of the anomalous scatterer can be determined using a difference Patterson map. Here the Patterson



**Figure 2.17** Argand diagram showing the breakdown of Friedel's law with an anomalous scattering heavy atom. The vectors below the real axis show the components of Friedel's law for  $(-h, -k, -l)$ . The structure factor for

protein alone ( $F_P$ ) has the same magnitude but negative phase for the Friedel paired reflection, but the heavy atom anomalous scattering results in two different magnitudes for the  $F_{PA}$  vectors ( $|F_{PA}(hkl)| \neq |F_{PA}(-h-k-l)|$ ).

map is calculated with the term  $|\Delta F_{\text{ano}}|^2$  which is equal to  $|F_{PA}(hkl) - F_{PA}(-h-k-l)|^2$  (where A represents the heavy atom anomalous scatterer). Solving the Patterson map will determine the position of the heavy atom that possesses anomalous scattering. Once the position of the heavy atom is determined the vectors from the anomalous scatterer can be constructed and a series of circles can be made for a Harker diagram to determine the phase of the protein (Figure 2.18). However, instead of solving the phase ambiguity by determining a new derivative, the vectors



**Figure 2.18** Harker diagram using MAD. (a) Known scattering vectors calculated from the heavy atom position determined from an anomalous difference Patterson map. The vectors for the positive  $hkl$  are drawn as the negative vector. The vectors for the Friedel (Bijvoet) mate  $-h-k-l$  (illustrated with the negative vector after it was reflected ( $r$ ) across the horizontal real axis because  $\varphi_{hkl} = -\varphi_{-h-k-l}$ ). The magnitude circles are drawn for the  $F_{hkl}$  and  $F_{-h-k-l}$  with the center at the end of their

respective anomalous  $-\Delta f''$  vector. The vectors result in a possibility of two phases ( $\varphi_1$  and  $\varphi_2$ ) that satisfy the vector addition of  $F_P = F_{PA} - F_A$  (where A represent anomalous scattering atom). (b) Incorporated onto the existing vectors is now the structure factor from a remote wavelength ( $\lambda_2$ ) where the anomalous scattering is minimized ( $F_P$ ) drawn with the center at origin. The three circles intersect at one point that reveals  $\varphi_2$  is the true phase for the protein.

from the anomalous scattering can be altered by collecting data at a different X-ray energy (wavelength). This has the effect of changing the vector magnitude because the  $f'$  and  $f''$  components will change at different energy. Also, since data is collected from the same crystal, the data sets are from perfectly isomorphous crystals, eliminating nonisomorphous errors.

The technique of MAD has become the method of choice for novel protein structure determination because of the availability of synchrotron radiation laboratories across the globe. Not only do they produce many orders of magnitude greater flux, but they also produce a “white” radiation, allowing the experimenter to easily change the wavelength of the X-rays during data collection through the use of a crystal monochromator. Typically, three data sets are collected from a single crystal containing an anomalous scatterer, first one data set is collected to maximize the  $f''$  component, then a remote wavelength with both  $f'$  and  $f''$  are minimized, and finally at a wavelength that maximizes the  $f'$  component. The phase problem can theoretically be solved by the first two data sets (and often are), but the third data set is collected and is equivalent to another derivative that can improve phases. However, crystals being exposed to X-rays for a long time required to collect three complete highly redundant data sets can cause radiation damage to the crystal and sometimes the last data set may cause problems in structure solution.

This MAD technique, however, still requires the incorporation of a heavy atom that displays anomalous scattering. The lightest element that does display some anomalous scattering is sulfur, but it is weak and its absorption edge is at a very low energy. However, progress is being made to use sulfur in cysteines to solve protein structures [21]. The element below sulfur in the periodic table, selenium, has a k-shell electron absorption edge at 0.9795 Å, which is a wavelength readily available in the range of synchrotron energies. This element displays chemistry very similar to sulfur and is found naturally in one form of the amino acid methionine (selenomethionine (Se-Met)). Therefore, with modern molecular biology techniques, the gene corresponding to the protein of interest can be cloned into a heterologous expression host like *E. coli*. The bacteria can be auxotrophic for Met<sup>-</sup> synthesis, but is not required. This is because normal heterotrophic bacteria can shut down the amino acid synthesis pathway by addition of amino acids to the media. Thus, about 15 min before induction of gene expression, many amino acids are added to the media with Se-Met and the bacterial incorporates Se-Met very efficiently into the expressed protein [22]. The protein, of course, needs methionine in the protein sequence for this technique to work. A good rule of thumb is to have at least one methionine for every 100 residues.

## 2.7

### Analyzing and Refining the Structure

#### 2.7.1

##### Electron Density Interpretation and Model Building

After the phases are determined using one of the methods described above, the phases are applied to Eq. (2.30) to calculate an electron density map. This map needs

to be analyzed and interpreted to build an atomic model structure that fits into the electron density. While there are some excellent algorithms recently written that automate this labor-intensive task [23, 24], human interpretation is still necessary to confirm results and identify significant structural findings. These programs are also only successful with moderately high-resolution data (better than 2.3 Å). In lower resolution electron density maps, the model is built by “hand” on a computer with the aid of computer software that can display electron density and an atomic model, which can be manipulated. The two most common programs used today for protein model building are O [15] and COOT [14].

Even with the first electron density maps of the myoglobin and hemoglobin, the protein amino acid sequence is essential to interpret the electron density. The first step in electron density interpretation is to match an electron density segment that correlates to the protein sequence. Large bulges of electron density protruding off a main-chain help identify large aromatic residues and are often used to initiate the model building. Once the sequence of the protein has been anchored in a region of the electron density map, the crystallographer then walks along the continuous path of electron density corresponding to the main-chain backbone and fits both the main-chain and side-chain atoms into the electron density “envelope.”

Many times, regions of electron density may not be visible due to disorder either from mobility or multiple conformations that averages the electron density out to very low levels. This is common for side-chains of lysine and arginine residue on the protein surface, but can correspond to entire stretches of residues corresponding to a loop. While the loop segment may not be visible in the electron density, this lack of structure indicates flexibility, which may be important for interpreting function.

The human interpretation of electron density maps and modeling in an atomic protein structure does leave room for errors. As the electron density maps are not truly atomic resolution, small positional errors in atomic placement are common, but can be corrected or minimized by molecular refinement. However, there are some rare cases where low-resolution maps, combined with the competitive rush to publish, have resulted in grossly negligent incorrect structures. One should always keep this in mind when viewing a structure. Many journals are requiring not only the submission of the atomic coordinates into the Protein Data Bank (PDB; [www.pdb.org](http://www.pdb.org)), but also the raw diffraction data in the form of structure factor intensities so other researchers can calculate the electron density map themselves to measure the level of agreement between the structure and the electron density.

### 2.7.2

#### **Protein Structure Refinement**

As described above, the placement of individual atoms within the electron density map is not perfect because the resolutions of typical protein crystals are not truly atomic (electron density peaks that correspond to individual atoms). However, the atomic structure of individual amino acid residues is known with great certainty, so placement of individual residues helps guide and place the individual atoms. Even so, there is some error in atomic positions that needs to be refined to better fit the electron density and diffraction data.

Given the atomic positions of the place atoms in the model, a “diffraction pattern” can be calculated using Eq. (2.25). The process of refinement is typically to perform a least squares minimization between the calculated diffraction pattern based on your atomic model (in the form of structure factor amplitudes  $|F_{\text{calc}}(hkl)|$ ) and the observed diffraction pattern ( $|F_{\text{obs}}(hkl)|$ ). The function that is minimized (and often reported) is the *R*-factor:

$$R = \frac{\sum_{hkl} |F_{\text{obs}}(hkl)| - k |F_{\text{calc}}(hkl)|}{\sum_{hkl} |F_{\text{obs}}(hkl)|} \times 100\% \quad (2.37)$$

where *k* is simply a scale factor to put the observed and calculated diffraction data on the same scale.

The algorithm will slightly shift the atomic positions to minimize this function. Determining the gradient or slope of the difference electron density map at each atom deciphers which direction the atoms shift. The difference electron density map is calculated by using coefficients of  $||F_{\text{obs}}| - |F_{\text{calc}}||$ . Inspection of this map, after a round of refinement, also helps identify large differences in atomic positions that need to be manually adjusted. This map also helps identify atoms lacking in the original model. It is used to position well-ordered solvent molecules as well, which show up as small spheres of electron density near the surface of the protein.

Typical protein structures contain thousands of atoms, and the three atomic coordinates are being adjusted in refinement (*x,y,z*) as well as the individual atomic temperature value (*B*-value) resulting in the concurrent refinement of four independent variables per atom. This results in a lower observation ( $F_{\text{obs}}$ ) to refinement variable (*x, y, z, and B*) ratio and can lead to large errors due to the simultaneous adjustment of thousands of atoms. Consequently, statistical minimization of the *R*-factor alone may result in sizable atomic shifts resulting in unreasonable molecular geometry. To help maintain more realistic geometry, the atom positions are restrained to prevent large atomic shifts. Therefore, refinement is often a minimization of multiple functions that counter-balance ideal geometry with *R*-factor minimization. Typically, other functions that are utilized to maintain ideal geometry include restraint terms like:

$$\text{Bond distances : } \sum_j w_d (d_j^{\text{ideal}} - d_j^{\text{model}})^2$$

$$\text{Bond angles : } \sum_j w_\chi (\chi_j^{\text{ideal}} - \chi_j^{\text{model}})^2$$

$$\text{Planarity : } \sum_k \sum_i w_p(i, k) (\mathbf{m}_k \cdot \mathbf{r}_{i,k} - d_k)^2$$

$$\text{Nonbonded contacts } \sum_j w_N (d_j^{\text{ideal}} - d_j^{\text{model}})^2$$



The term “ $w$ ” in these summations is a weighting term for the individual geometric parameters. One can increase these weights to give “perfect” ideal geometry, but the  $R$ -factor will not be minimized, or conversely if the weights are too low, the  $R$ -factor minimization will dominate the refinement resulting in a low  $R$ -factor, but an atomic model with geometry that might be unreasonable. Therefore, the crystallographer adjusts the weights to give both reasonable  $R$ -factor and geometry. Typically two geometric parameters that are often reported are rms deviation (rmsd) of bond distances and bond angles from ideality. Typically, these values should be below 0.02 Å for bond distances and 2° for bond angles.

The major limitation of the least-squares method of molecular refinement is the minimization follows a downhill (gradient) trajectory towards the minimum and cannot hurdle a small barrier near a local minimum to reach a true global minimum. To avoid being trapped in local minima, molecular dynamics or simulated annealing was introduced to allow small uphill shifts to overcome barriers in order to reach the true minimum [25]. In principle, the simulated annealing algorithm theoretically heats up the protein model to a large temperature (2500 K or above), and slowly cools the structure while minimizing the  $R$ -factor and other geometric and energy terms to achieve a global minimum.

The method of molecular dynamics in protein structure refinement made a significant contribution to crystallography and appreciably helped refine many difficult structures. However, one of the major limitations of this method is that if the initial starting structure has some large errors or is incorrect, the simulated annealing method can “press” a wrong structure to appear correct by giving reasonable geometry and  $R$ -factor statistics. To circumvent this problem, the concept of cross-validation was introduced to monitor the progress and accuracy of refinement [26]. In the cross-validation method a small random subset of observed structure factors are flagged (typically 5%) and excluded during the refinement minimization. Then the  $R$ -factor is calculated for both the 95% of the data (working set) for which the target functions were minimized as well as the 5% data excluded from the minimization. The  $R$ -factor calculated from the small cross-validated data is called the  $R$ -free value. Both values are typically reported today. If the initial starting structure had significant errors, the  $R$ -free value will not decrease much and typically increases, even though the  $R$ -factor decreases. Structures with a  $R$ -free value greater than 35% will likely have gross structural errors. If the structure is correct and refines well, both the  $R$ -factor and  $R$ -free will decrease, with the  $R$ -free being typically around 5% greater than the  $R$ -factor. For a well-refined complete structure the  $R$ -factor ranges from the high teens to low twenties percent.

### 2.7.3

#### **Protein Structure Validation**

During and after refinement, the crystallographer should also validate the structure. Geometric and stereochemistry validation has been described above. Many software algorithms are available to analyze and validate your protein structure to help identify possible errors. Reporting the rmsd of ideal geometry is often reported, but

inspection of large deviations from normal or ideal geometry in relation to the protein sequence or protein structure can give clues to possible regions in the structure that might need more attention in refinement. Some programs (e.g., WHATCHECK) can calculate the “Z-score” for the protein residues, which measures how parameters deviate from the mean. Large deviations or outliers from the standard deviations can be easily seen and should be inspected in the model.

Another common validation of protein structure is to plot the main-chain backbone  $\varphi$  and  $\psi$  torsion angles in a Ramachandran plot [27]. This two-dimensional plot reveals the favorably allowed regions of the protein backbone torsion angles. Ideally, your protein model should have more than 90% of the residues in the most favored region. If a string of residues falls outside the allowed regions, inspection of this region of the structure is warranted for additional refinement. The program PROCHECK [28] can run a series of diagnostic tools on the protein structures, including Ramachandran plots, to identify possible regions of the protein structure that needs further inspection and possible “tweaking” and refinement.

After the crystallographer is satisfied with the protein structure, refinement and validation, the three-dimensional Cartesian coordinates, B-values, and occupancy, for all the atoms are deposited into the PDB. Additionally, more journals are also requiring, as they should, the structure factor amplitudes data to be deposited as well. Upon submission, the structures and data are also scrutinized a final time by the PDB for validity before being released to share with the scientific community.

## References

- Hooke, R. (1665) *Micrographia: Or Some Physiological Descriptions of Minute Bodies, Made by Magnifying Glasses, with Observations and Inquires Thereupon*, Royal Society, London.
- Bragg, W.L. (1913) Diffraction of short electromagnetic waves by a crystal. *Proceedings of the Cambridge Philosophical*, 43–57.
- Bragg, W.L. (1913) The structure of some crystals as indicated by their diffraction of X-rays. *Proceedings of the Royal Society of London Series A*, **89**, 248–277.
- Friedrich, W., Knipping, P., and Laue, M. (1913) Interference appearances in X-rays. *Annalen Der Physik*, **41**, 971–988.
- Pauling, L. and Niemann, C. (1939) The structure of proteins. *Journal of the American Chemical Society*, **61**, 1860–1867.
- Debye, P. and Hückel, E. (1923) The theory of electrolytes I. The lowering of the freezing point and related occurrences. *Physikalische Zeitschrift*, **24**, 185–206.
- Rossmann, M.G., Morais, M.C., Leiman, P.G., and Zhang, W. (2005) Combining X-ray crystallography and electron microscopy. *Structure*, **13**, 355–362.
- Matthews, B.W. (1968) Solvent content of protein crystals. *Journal of Molecular Biology*, **33**, 491–497.
- Bernal, J.D. and Crowfoot, D. (1934) X-ray photographs of crystalline pepsin. *Nature*, **133**, 794–795.
- Hahn, T., Shmueli, U., Wilson, A.J.C., and International Union of Crystallography (1984) *International Tables for Crystallography*, Kluwer, Dordrecht.
- Hajdu, J. (2000) Single-molecule X-ray diffraction. *Current Opinion in Structural Biology*, **10**, 569–573.
- Gaffney, K.J. and Chapman, H.N. (2007) Imaging atomic structure and dynamics with ultrafast X-ray scattering. *Science*, **316**, 1444–1448.

- 13 Rossmann, M.G. (ed.) (1972) The molecular replacement method, in *A Collection of Papers on the Use of Non-Crystallographic Symmetry*, Gordon & Breach, New York.
- 14 Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallographica D*, **60**, 2126–2132.
- 15 Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991) Improved methods for the building of protein models in electron density maps and the location of errors in these models. *Acta Crystallographica A*, **47**, 110–119.
- 16 McCoy, A. (2007) Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallographica D*, **63**, 32–41.
- 17 Kissinger, C.R., Gehlhaar, D.K., and Fogel, D.B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallographica D*, **55**, 484–491.
- 18 Navaza, J. (2001) Implementation of molecular replacement in amore. *Acta Crystallographica D*, **57**, 1367–1372.
- 19 Brünger, A., Adams, P., Clore, G., DeLano, W., Gros, P., Grosse-Kunstleve, R., Jiang, J., Kuszewski, J., Nilges, M., Pannu, N., Read, R., Rice, L., Simonson, T., and Warren, G. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallographica D*, **54**, 905–921.
- 20 Tong, L. and Rossmann, M.G. (1990) The locked rotation function. *Acta Crystallographica A*, **46**, 783–792.
- 21 Sarma, G. and Karplus, P. (2006) In-house sulfur SAD phasing: a case study of the effects of data quality and resolution cutoffs. *Acta Crystallographica D*, **62**, 707–716.
- 22 Doublé, S. (1997) Preparation of selenomethionyl proteins for phase determination. *Methods in Enzymology*, **276**, 523–530.
- 23 Perrakis, A., Morris, R., and Lamzin, V.S. (1999) Automated protein model building combined with iterative structure refinement. *Nature Structural Biology*, **6**, 458–463.
- 24 Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Adams, P.D., Read, R.J., Zwart, P.H., and Hung, L.W. (2008) Iterative-build omit maps: map improvement by iterative model building and refinement without model bias. *Acta Crystallographica D*, **64**, 515–524.
- 25 Brünger, A.T., Kuriyan, J., and Karplus, M. (1987) Crystallographic R factor refinement by molecular dynamics. *Science*, **235**, 458–460.
- 26 Brünger, A.T. (1992) The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–474.
- 27 Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, **7**, 95–99.
- 28 Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993) Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291.



### 3

## Nuclear Magnetic Resonance of Amino Acids, Peptides, and Proteins

*Andrea Bernini and Pierandrea Temussi*

### 3.1

#### Introduction

Nuclear magnetic resonance (NMR) is only a few decades old, yet NMR spectroscopy is beyond doubt the most widely used and most useful analytical tool for chemistry and biology, surpassing the applications of IR spectroscopy and rivaling X-ray crystallography in biochemical applications.

In the 1930s, Isidor Isaac Rabi, a physicist of the Columbia University, showed that some atomic nuclei behave like little magnets. Rabi experimented on beams of nuclei passing through magnetic fields; when these beams were irradiated with electromagnetic waves of the right frequency, nuclei possessing a magnetic moment could absorb energy, flipping from a lower energy state to another higher energy state. In 1946, two teams of researchers in Harvard and Stanford demonstrated independently that the same phenomenon could be detected also in condensed matter (liquids and solids). At the Massachusetts Institute of Technology (Harvard), E. Purcell, H. Torrey, and R. Pound used solid paraffin to study a collection of hydrogen nuclei. When the paraffin sample was irradiated at a fixed radiofrequency and, at the same time, immersed in a magnetic field of several thousand gauss, the paraffin's protons were forced into two allowed magnetic alignments. The energy difference between these two states depends on the strength of the magnetic field. When the energy difference between the two states reached the energy of the radiofrequency photons, Purcell, Torrey and Pound observed a sharp increase in the absorption of radiation corresponding to nuclei flipping from a lower energy state to a higher energy state. F. Bloch, W. Hansen, and M. Packard of Stanford University in California used water as a collection of protons, wrapping a coil that delivered radiofrequency radiation around a vessel containing water immersed in an adjustable magnetic field. Also, the Stanford group increased the magnetic field strength to reach resonance. However, these physicists did not measure absorption, but detected re-emission of resonant radiation (a "resonance") using a second coil placed perpendicular to the first.

At the time of the discovery of NMR, its main application appeared to be the accurate determination of the nuclear magnetic moments of all the elements in the periodic table. However, physicists soon realized that the numbers they obtained

for a given nucleus were strongly influenced by the chemical environment and regarded this “chemical” influence mainly as a nuisance. Magnetic moments were different not only for the same nuclei in different compounds, such as paraffin and water, but even for different chemical groups within the same molecule. The first example was ethyl alcohol that showed three separate resonances for the protons of  $\text{CH}_3$ ,  $\text{CH}_2$ , and  $\text{OH}$  groups. Soon after this, many chemists realized the enormous potential of NMR and the new spectroscopy bloomed.

### 3.1.1

#### Active Nuclei in NMR

Nuclei have a property called angular momentum. The total angular momentum of a nucleus is generally represented by the symbol  $\mathbf{I}$  and is called “nuclear spin.” Associated with each nuclear spin is a nuclear magnetic moment that produces magnetic interactions with its environment. The actual value of the nuclear magnetic moment, described by the quantum number called “nuclear spin” ( $I$ ), depends ultimately on the relative numbers of nuclear particles (neutrons and protons). The rules generating the value of  $I$  for different nuclei are complex, but their knowledge is not necessary for the spectroscopist using NMR in chemistry or biology. However, it is simple enough, if one looks at the periodic table, to extract the following general rules. The spin of nuclei with an odd mass number is half-integral: for instance,  $^1\text{H}$ ,  $^3\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{19}\text{F}$ , and  $^{31}\text{P}$  all have spin  $1/2$ . It is a fortunate circumstance that most of these nuclei are of great biological relevance, because NMR spectroscopy, as we know it, is far simpler for this type of nucleus. The spin of nuclei with an even mass number and odd charge number is integral: for instance,  $^2\text{H}$  and  $^{14}\text{N}$ . Nuclei with an even number of both charge and mass have a spin quantum number of zero.

### 3.1.2

#### Energy Levels and Spin States

All spectroscopic techniques are based on the assumption that molecules can populate a set of *energy levels*. Spectral lines originate from *transitions* between these energy levels, whereas intensities of the lines are proportional to the population difference of the levels (or states). The relationship between the frequency ( $\nu$ ) of the photon causing the transition and the energy gap between the two levels ( $\Delta E$ ) is the well known  $\Delta E = h\nu$ , where  $h$  is Planck’s constant.

The energy gap determines also the distribution of the molecules of the sample among the accessible levels. The key factor is the relation between the energy of each state and the thermal energy,  $kT$ , where  $T$  is the temperature in Kelvin and  $k$  is Boltzmann’s constant. If two states  $E_1$  and  $E_2$  are separated by an energy gap  $\Delta E$ , the ratio of their populations is given by:

$$p_1/p_2 = \exp(-\Delta E/kT)$$

In NMR,  $\Delta E$  is small and thus the two states have almost identical populations. This explains why NMR is not a very sensitive technique; however, the energy gap  $\Delta E$

depends on the strength of the magnetic field and one of the big advantages of operating at very high magnetic fields is the increase in sensitivity.

In optical spectroscopies, the energy levels are a unique feature of each molecule, but in NMR they are also dependent on the strength of the magnetic field. This is not the only peculiarity that puts NMR spectroscopy in a different category. Relaxation parameters are also fundamental in NMR spectroscopy, because relaxation influences both line shapes and intensities of NMR signals, providing information about structure and dynamics of molecules.

As previously stated, NMR spectroscopy is made possible by the existence of a nuclear property known as *spin*. According to quantum mechanics, nuclear spin is characterized by a nuclear spin quantum number,  $I$ . Most nuclei of interest for peptides and proteins have  $I = 1/2$ . A spin-half nucleus, like  $^1\text{H}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$ , when interacting with a magnetic field, gives rise to two energy levels characterized by quantum number  $m$ , which is restricted to the values  $-I$  to  $+I$  in integer steps. So, in the case of  $I = 1/2$ , there are only two values of  $m = +1/2$  and  $m = -1/2$ , traditionally called  $\alpha$  (or spin up) and  $\beta$  (or spin down).

The energies of the levels depend on  $m$ :  $E_m = m\nu_0$ , where  $\nu_0$  is called the *Larmor frequency*.

It is important to note that, in this expression, as it is often the case with spectroscopists, the energies are expressed in frequency units. This is allowed because frequency differs from energy only by a proportionality constant. The Larmor frequency depends on three parameters: a quantity known as the *gyromagnetic ratio*,  $\gamma$  (characteristic of any given nucleus); the chemical shift,  $\delta$  (that reflects the actual electronic environment of the nucleus in the molecule under study); and the strength of the magnetic field,  $B_0$ :

$$\nu_0 = -(1/2\pi)\gamma(1 + \delta)B_0$$

The energies of the two levels corresponding to the  $m$  values  $\alpha$  and  $\beta$  are:  $E_\alpha = +(1/2)\nu_0$  and  $E_\beta = -(1/2)\nu_0$ . The selection rule for an NMR transition states that the quantum number  $m$  *can only change by one unit*:

$$\Delta m = m(\text{initial state}) - m(\text{final state}) = \pm 1$$

In the case of one spin  $1/2$ , this selection rule amounts to a change in  $m$  between the states  $\alpha$  and  $\beta$  equal to  $(+1/2 - (-1/2)) = 1$ . The corresponding frequency is:

$$\nu_{\alpha\beta} = E_\beta - E_\alpha = (1/2)\nu_0 - (+1/2)\nu_0 = -\nu_0$$

### 3.1.3

#### Main NMR Parameters (Glossary)

##### 3.1.3.1 Chemical Shift

The spins of a given isotope resonate at the same frequency if they experience the same environment. In reality, different molecules and even different parts of the same molecule have different electronic distributions. The magnetic field experienced by the molecule induces a precession of electrons around the magnetic field

direction and thus, in turn, generates a tiny local magnetic field, which counters a portion of the external field.

The actual frequency of a given nucleus, arising from variations in the electron distribution, is called the “chemical shift.” It gives us information on the local electronic environment. The chemical shift ( $\delta$ ) is expressed as the ratio between the difference in electron shielding to a reference nucleus and the reference frequency, expressed in units of ppm (parts per million):  $\delta = \Delta\nu/\nu_{\text{ref}}$ . The chemical shift  $\delta$ , expressed in this way, is independent of magnetic field strength. That is, the resonances in ppm remain the same when measured at different field strengths. The reference compound should be stable and characterized by an unchanged chemical shift value over the normal ranges of temperature and pH values. A compound commonly used for  $^1\text{H}$ -NMR reference in aqueous solutions is 3-(trimethylsilyl) propane sulfonic acid sodium salt. However, in studies of amino acids, peptides and proteins, it is often preferable to reference all signals directly to the water resonance. During the last few years, chemical shifts have been used extensively as a valuable structural tool in protein structure determination.

### 3.1.3.2 Scalar Coupling Constants

NMR signals are generally split into more or less complex multiplets owing to a parameter called the coupling constant,  $J$ . The terms “scalar coupling” and “ $J$ -coupling” are frequently used interchangeably as synonyms. Strictly speaking, scalar coupling is the isotropic part (independent of the molecular orientation) of the  $J$ -coupling. The  $J$ -coupling arises from the indirect interaction, mediated by bonding electrons, between any two nuclear spins. The magnitude of  $J$ -coupling is typically larger when nuclei are separated by a small number of bonds, typically two or three, and declines sharply with more bonds.  $J$ -couplings can provide very useful conformational information. The most widespread use is that of vicinal scalar coupling constants (i.e., between atoms separated from each other by three covalent bonds) because their values can be analyzed in terms of the so-called Karplus equation to yield dihedral angles. In turn, dihedral angles can be used as structural restraints in calculations of peptide or protein structure.

### 3.1.3.3 NOE

The nuclear Overhauser effect (NOE) is the transfer of spin polarization from one spin population to another via cross-relaxation. The original Overhauser effect was about polarization transfer between electron and nuclear spins, but NOEs are now mostly used for transfer between nuclear spins. NOE is observed through space; thus, all atoms that are in proximity to each other, within a certain distance, give a NOE. In all protocols for protein structure determination, NOEs are the most important parameters because they provide accurate distance information between pairs of hydrogen atoms separated by less than 5 Å. Distance information is usually grouped into three different groups: 1.8–2.5 (strong), 1.8–3.5 (medium), and 1.8–5.0 Å (weak). Short-range NOEs are mostly valuable for defining secondary structure elements whereas long-range NOEs allow reconstruction of tertiary structure.



### 3.1.3.4 RDC

Use of residual dipolar couplings (RDCs) is a fairly recent addition to the tools for structure determination of proteins. In solids, NMR spectra are dominated by internuclear dipole–dipole couplings. In liquids, these couplings are typically averaged out, due to random molecular tumbling. However, it may be possible to observe small residual couplings when there is a small degree of alignment with the external magnetic field. The RDCs are observed in heteronuclear correlation spectra as small changes of the splitting caused by one-bond  $J$  couplings between directly bound nuclei.

The RDCs give information on long-range order that is complementary to that given by NOEs. The addition of RDCs has often improved the precision as well as the accuracy of NMR structures.

## 3.2

### Amino Acids

#### 3.2.1

##### Historical Significance

In early applications of NMR to structural biology, several researchers used NMR spectra of amino acids as an aid to recognize typical features of protein spectra when trying to assign resonances of protein residues. Since the use of an isolated amino acid as a reference for polypeptide assignment purposes is limited by the large influence, on the resonance values, of the charged amino and carboxyl groups [1, 2], capped amino acids and subpeptides were initially proposed as references [1, 2], until the resonances measured from flexible short peptides became a *de facto* standard [3, 4]. These are usually referred to as “random coil” chemical shifts, corresponding to reference shifts for each residue in the absence of secondary or tertiary structure interactions. As the number of NMR protein structures increased, statistical data for resonances of chained amino acids became available and the influence of secondary structure on chemical shift changes was inferred [5–7]. The “secondary structure shift” was then defined as the difference between the observed chemical shift and the appropriate random coil value, and was proposed as a method for fast protein structure characterization [8, 9]. Up-to-date statistics are maintained at the Biological Magnetic Resonance Data Bank ([www.bmrb.wisc.edu](http://www.bmrb.wisc.edu)). Random coil, as well as statistics-derived chemical shift values, have been extended from proton to other nuclei constituting amino acids, particularly the NMR-active isotopes  $^{15}\text{N}$  and  $^{13}\text{C}$ , and taken together they constitute the base of resonance identification and assignment methods employed for protein structure determination.

#### 3.2.2

##### Amino Acids Structure

Amino acids are 2-amino carboxylic acids that differ only by the moiety at C2, usually referred to as the side-chain, as opposed to the main-chain, formed by the -N-C-CO-

skeleton. The presence of the amino and carboxyl functions allows, indeed, the chaining into a polymer, generally referred to as a polypeptide, by formation of amide bonds.

Accordingly, amino acids are often called the “building blocks” of proteins. Amino acids are characterized by a chiral center on the  $\alpha$ -carbon (C2). The 20 amino acids commonly found in proteins are also called naturally occurring (or proteinogenic) and are invariably in the *S*-configuration (except glycine, 2-aminoacetic acid, which is non-chiral, and cysteine, which is *R*-configured due to a change in priority due to the presence of sulfur), but many others are known, occurring also in several biological processes. Amino acids are classically grouped as polar, charged, aliphatic, and aromatic, according to chemico-physical properties of the side-chain, although its nature may be more complex (e.g., the lysine side-chain shows both charged and hydrophobic character because, in addition to the amino group, it contains a long hydrophobic chain, and the imidazolyl moiety of histidine is both aromatic, polar, and/or charged at specific pHs). Recent texts adopt a more general classification dividing amino acids in hydrophobic, hydrophilic, and amphipathic, leaving deduction of finer details to the observation of chemical structure (Table 3.1).

### 3.2.3

#### Random Coil Chemical Shift

The NMR parameters that most characterize amino acids are the so-called *random coil* chemical shifts of their protons. These reference values were first tabulated as the  $^1\text{H}$  frequencies for an amino acid X measured in an aqueous solution of the tetrapeptide GGXA [3], where G and A are the one-letter codes for glycine and alanine, respectively (Table 3.2). Corresponding values are tabulated also for other solvents [10, 11] and for additional nuclei [12, 13], and are referred to as random coil because it is assumed that a short polypeptide does not adopt any stable structure. Other values have been proposed since these early studies [14–16].

The comparison of measured chemical shifts with their random coil counterparts is commonly used to identify secondary structure elements in folded proteins and to reveal the presence of regions with residual structure in unfolded states [5, 17].

Necessarily, chemical shifts derived from the simple GGXA peptides do not reflect any sequence dependence. To account for sequence-dependent effects, induced on Ala by the preceding residue X, backbone correction factors have been published for the  $^{15}\text{N}$  chemical shift of the Ala4 residue in a set of H-GGXA-OH peptides [18]. In a later study, Sykes *et al.* [19] measured random coil chemical shifts for a series of peptides of formulae Ac-GGXAGG-NH<sub>2</sub> and Ac-GGXPPG-NH<sub>2</sub>, and investigated the effect on the chemical shifts of residue X ( $^{13}\text{C}$  and  $^1\text{H}$ ) on the following residue (A or P, respectively). Although accurate, such a method is really time-consuming since obtaining the sequence dependence of the chemical shifts of all dipeptide sequences would require data to be measured on 400 peptides, and to investigate the sequence-dependent effect on the preceding and the following residues in a tripeptide sequence, it would be necessary to measure data on 8000 peptides of this type. Thus, a simplified approach, which combines the methods of Wishart *et al.* [19]

**Table 3.1** The 20 common amino acids with major properties and structures.

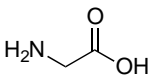
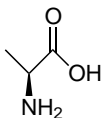
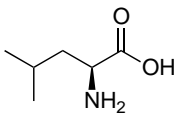
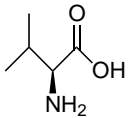
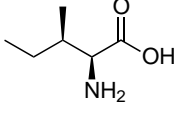
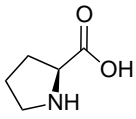
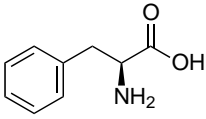
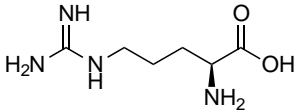
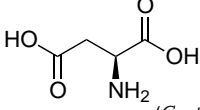
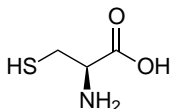
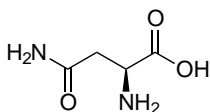
Legend	Structure		
<b>Amino acid name</b>		<b>Glycine</b>	
Abbreviation; short		GLY; G	
Molecular weight (MW) (Da)		MW: 75.07	
Isoelectric point (pI)		pI: 5.97	
pK <sub>a</sub> of side-chain labile proton (pK <sub>asc</sub> )		pK <sub>asc</sub> : -	
Formula		C <sub>2</sub> H <sub>5</sub> N <sub>1</sub> O <sub>2</sub>	
Hydrophobic amino acids			
<b>Alanine</b>		<b>Leucine</b>	
ALA; A		LEU; L	
MW: 89.09		MW: 131.17	
pI: 6.00		pI: 5.98	
pK <sub>asc</sub> : -		pK <sub>asc</sub> : -	
C <sub>3</sub> H <sub>7</sub> N <sub>1</sub> O <sub>2</sub>		C <sub>6</sub> H <sub>13</sub> N <sub>1</sub> O <sub>2</sub>	
<b>Valine</b>		<b>Isoleucine</b>	
VAL; V		ILE; I	
MW: 117.15		MW: 131.17	
pI: 5.96		pI: 5.94	
pK <sub>asc</sub> : -		pK <sub>asc</sub> : -	
C <sub>5</sub> H <sub>11</sub> N <sub>1</sub> O <sub>2</sub>		C <sub>6</sub> H <sub>13</sub> N <sub>1</sub> O <sub>2</sub>	
<b>Proline</b>		<b>Phenylalanine</b>	
PRO; P		PHE; F	
MW: 115.13		MW: 165.19	
pI: 6.30		pI: 5.48	
pK <sub>asc</sub> : -		pK <sub>asc</sub> : -	
C <sub>5</sub> H <sub>9</sub> N <sub>1</sub> O <sub>2</sub>		C <sub>9</sub> H <sub>11</sub> N <sub>1</sub> O <sub>2</sub>	
Hydrophilic amino acids			
<b>Arginine</b>		<b>Aspartic acid</b>	
			(Continued)

Table 3.1 (Continued)

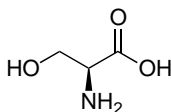
ARG; R  
MW: 174.20  
pI: 11.15  
pK<sub>asc</sub>: 12.48  
C<sub>6</sub>H<sub>14</sub>N<sub>4</sub>O<sub>2</sub>  
**Cysteine**



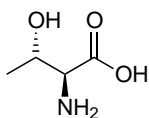
CYS; C  
MW: 121.16  
pI: 5.02  
pK<sub>asc</sub>: 8.18  
C<sub>3</sub>H<sub>7</sub>N<sub>1</sub>O<sub>2</sub>S<sub>1</sub>  
**Asparagine**



ASN; N  
MW: 132.12  
pI: 5.41  
pK<sub>asc</sub>: -  
C<sub>4</sub>H<sub>8</sub>N<sub>2</sub>O<sub>3</sub>  
**Serine**

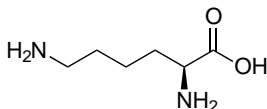


SER; S  
MW: 105.09  
pI: 5.68  
pK<sub>asc</sub>: -  
C<sub>3</sub>H<sub>7</sub>N<sub>1</sub>O<sub>3</sub>  
**Threonine**



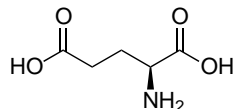
THR; T  
MW: 119.12  
pI: 5.64  
pK<sub>asc</sub>: -  
C<sub>4</sub>H<sub>9</sub>N<sub>1</sub>O<sub>3</sub>  
**Amphipathic amino acids**

**Lysine**

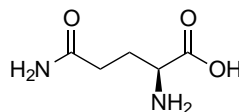


LYS; K  
MW: 146.19  
pI: 9.59  
pK<sub>asc</sub>: 10.54

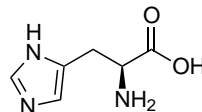
ASP; D  
MW: 133.10  
pI: 2.77  
pK<sub>asc</sub>: 3.90  
C<sub>4</sub>H<sub>7</sub>N<sub>1</sub>O<sub>4</sub>  
**Glutamic acid**



GLU; E  
MW: 147.13  
pI: 3.22  
pK<sub>asc</sub>: 4.07  
C<sub>5</sub>H<sub>9</sub>N<sub>1</sub>O<sub>4</sub>  
**Glutamine**

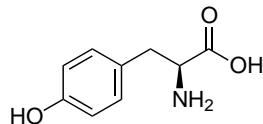


GLN; Q  
MW: 146.15  
pI: 5.65  
pK<sub>asc</sub>: -  
C<sub>5</sub>H<sub>10</sub>N<sub>2</sub>O<sub>3</sub>  
**Histidine**



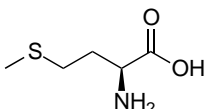
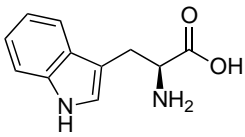
HIS; H  
MW: 155.16  
pI: 7.47  
pK<sub>asc</sub>: 6.04  
C<sub>6</sub>H<sub>9</sub>N<sub>3</sub>O<sub>2</sub>

**Tyrosine**



TYR; Y  
MW: 181.19  
pI: 5.66  
pK<sub>asc</sub>: 10.46

Table 3.1 (Continued)

$C_6H_{14}N_2O_2$ <b>Methionine</b>		$C_9H_{11}N_1O_3$ <b>Tryptophan</b>	
MET; M MW: 149.21 pI: 5.74 p <i>K</i> <sub>asc</sub> : - $C_5H_{11}N_1O_2S_1$		TRP; W MW: 204.23 pI: 5.89 p <i>K</i> <sub>asc</sub> : - $C_{11}H_{12}N_2O_2$	

and Braun *et al.* [18], has been developed by Schwarzingner *et al.* [20]. These authors analyzed data for a series of model peptides of sequence Ac-GGXGG-NH<sub>2</sub> to determine the variations in the chemical shifts of the four glycine residues caused by substitutions at position X. Complete assignments have been made for all resonances in these peptides in 8 M urea at pH 2.3 and the changes caused by changing the central residue X have been analyzed. A complete tabulation of correction factors for the random coil chemical shifts of the most sequence-dependent nuclei, <sup>15</sup>N, <sup>1</sup>H<sup>N</sup>, and <sup>13</sup>CO, was thus produced. The sample conditions were chosen to match those used in NMR studies of early folding of polypeptides. The usefulness of the correction factors for chemical shift calculations was demonstrated in the case of urea-denatured apomyoglobin. Such correction factors are also implemented in a module [20] for the program NMRView [21].

The importance of measuring backbone chemical shifts in unfolded states has recently been further emphasized by the recognition that proteins containing natively unfolded regions may represent up to one-third of eukaryotic proteomes and play a variety of essential biological roles [22]. Furthermore, it has also been pointed out that several amyloidogenic proteins associated with neurodegenerative diseases are natively unfolded [23].

A recent advance in associating random coil chemical shift values to amino acid sequences of proteins is the CamCoil approach [24]. Here, the relationship between amino acid sequences and chemical shifts is mapped using the flexible loop regions in native states as a model of the random coil state, enabling discrimination of the dependence of the chemical shifts on the primary structure of proteins from the effects associated with the secondary and tertiary structures.

### 3.2.4

#### Spin Systems

Chemical shifts originating from backbone atoms, common to all amino acids, fall in a narrow window when the peptides are in the random coil state, but spread over wider frequency ranges under the influence of secondary structural features.

**Table 3.2** Random coil chemical shift values (ppm), in denaturing conditions, for  $^1\text{H}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$  nuclei of a residue X embedded in linear pentapeptides GGXGG [20].

<b>Ala</b>					
H	8.35	C	178.5	N	125.0
H $\alpha$	4.35	C $\alpha$	52.8		
H $\beta$	1.42	C $\beta$	19.3		
<b>Arg</b>					
H	8.39	C	177.1	N	121.2
H $\alpha$	4.38	C $\alpha$	56.5		
H $\beta$ 2	1.91	C $\beta$	30.9		
H $\beta$ 3	1.79	C $\gamma$	27.3		
H $\gamma$ 2	1.68	C $\delta$	43.6		
H $\gamma$ 3	1.64	C $\zeta$	159.7		
H $\delta$ 2	3.2				
H $\delta$ 3	3.2				
H $\epsilon$	7.2				
<b>Asn</b>					
H	8.51	C	176.1	N	119.0
H $\alpha$	4.79	C $\alpha$	53.3		
H $\beta$ 2	2.88	C $\beta$	39.1		
H $\beta$ 3	2.81	C $\gamma$	177.3		
H $\delta$ 21	7.59				
H $\delta$ 22	7.01				
<b>Asp</b>					
H	8.56	C	175.9	N	119.1
H $\alpha$	4.82	C $\alpha$	53.0		
H $\beta$ 2	2.98	C $\beta$	38.3		
H $\beta$ 3	2.91	C $\gamma$	177.4		
<b>Cys (oxidized)</b>					
H	8.54	C	175.5	N	118.7
H $\alpha$	4.76	C $\alpha$	55.6		
H $\beta$ 2	3.29	C $\beta$	41.2		
H $\beta$ 3	3.02				
<b>Cys (reduced)</b>					
H	8.44	C	175.3	N	118.8
H $\alpha$	4.59	C $\alpha$	58.6		
H $\beta$ 2	2.98	C $\beta$	28.3		
H $\beta$ 3	2.98				
<b>Gln</b>					
H	8.44	C	176.8	N	120.5
H $\alpha$	4.38	C $\alpha$	56.2		
H $\beta$ 2	2.17	C $\beta$	29.5		
H $\beta$ 3	2.01	C $\gamma$	34.0		
H $\gamma$ 2	2.39	C $\delta$	180.5		
H $\gamma$ 3	2.39				
H $\epsilon$ 21	7.5				
H $\epsilon$ 22	6.91				
<b>Glu</b>					
H	8.40	C	176.8	N	120.2
H $\alpha$	4.42	C $\alpha$	56.1		

Table 3.2 (Continued)

H $\beta$ 2	2.18	C $\beta$	28.9		
H $\beta$ 3	2.01	C $\gamma$	32.9		
H $\gamma$ 2	2.50	C $\delta$	180.0		
H $\gamma$ 3	2.50				
<b>Gly</b>					
H	8.41	C	174.9	N	107.5
H $\alpha$ 2	4.02	C $\alpha$	45.4		
<b>His</b>					
H	8.56	C	175.1	N	118.1
H $\alpha$	4.79	C $\alpha$	55.4		
H $\beta$ 2	3.35	C $\beta$	29.1		
H $\beta$ 3	3.19	C $\epsilon$ 1	136.4		
H $\epsilon$ 1	8.61	C $\delta$ 2	120.2		
H $\delta$ 2	7.31	C $\gamma$	131.4		
<b>Ile</b>					
H	8.17	C	177.1	N	120.4
H $\alpha$	4.21	C $\alpha$	61.6		
H $\beta$	1.89	C $\beta$	38.9		
H $\gamma$ 12	1.48	C $\gamma$ 1	27.5		
H $\gamma$ 13	1.19	C $\gamma$ 2	17.5		
H $\gamma$ 2	0.93	C $\delta$ 1	13.2		
H $\delta$ 1	0.88				
<b>Leu</b>					
H	8.28	C	178.2	N	122.4
H $\alpha$	4.38	C $\alpha$	55.5		
H $\beta$ 2	1.67	C $\beta$	42.5		
H $\beta$ 3	1.62	C $\gamma$	27.1		
H $\gamma$	1.62	C $\delta$ 1	25.0		
H $\delta$ 1	0.93	C $\delta$ 2	23.3		
H $\delta$ 2	0.88				
<b>Lys</b>					
H	8.36	C	177.4	N	121.6
H $\alpha$	4.36	C $\alpha$	56.7		
H $\beta$ 2	1.89	C $\beta$	33.2		
H $\beta$ 3	1.77	C $\gamma$	25.0		
H $\gamma$ 2	1.47	C $\delta$	29.3		
H $\gamma$ 3	1.42	C $\epsilon$	42.4		
H $\delta$ 2	1.68				
H $\delta$ 3	1.68				
<b>Met</b>					
H	8.42	C	177.1	N	120.3
H $\alpha$	4.52	C $\alpha$	55.8		
H $\beta$ 2	2.15	C $\beta$	32.9		
H $\beta$ 3	2.03	C $\gamma$	32.3		
H $\gamma$ 2	2.63	C $\epsilon$	17.0		
H $\gamma$ 3	2.64				
H $\epsilon$	2.11				

(Continued)

Table 3.2 (Continued)

<b>Phe</b>					
H	8.31	C	176.6	N	120.7
H $\alpha$	4.65	C $\alpha$	58.1		
H $\beta$ 2	3.19	C $\beta$	39.8		
H $\beta$ 3	3.04	C $\gamma$	139.2		
H $\delta$ 1	7.28	C $\delta$ 1	132.0		
HE1	7.38	C $\epsilon$ 1	131.5		
H $\zeta$	7.33	C $\zeta$	130.0		
H $\epsilon$ 2	7.38	C $\epsilon$ 2	131.5		
H $\delta$ 2	7.28	C $\delta$ 2	132.0		
<b>Pro (trans conformation)</b>					
H $\alpha$	4.45	C	177.8		
H $\beta$ 2	2.29	C $\alpha$	63.7		
H $\beta$ 3	1.99	C $\beta$	32.2		
H $\gamma$ 2	2.04	C $\gamma$	27.3		
H $\gamma$ 3	2.04	C $\delta$	49.8		
H $\delta$ 2	3.67				
H $\delta$ 3	3.61				
<b>Pro (cis conformation)</b>					
H $\alpha$	4.6	C $\alpha$	63.0		
H $\beta$ 2	2.39	C $\beta$	34.8		
H $\beta$ 3	2.18	C $\gamma$	24.9		
H $\gamma$ 2	1.95	C $\delta$	50.4		
H $\gamma$ 3	1.88				
H $\delta$ 2	3.60				
H $\delta$ 3	3.54				
<b>Ser</b>					
H	8.43	C	175.4	N	115.5
H $\alpha$	4.51	C $\alpha$	58.7		
H $\beta$ 2	3.95	C $\beta$	64.1		
H $\beta$ 3	3.90				
<b>Thr</b>					
H	8.25	C	175.6	N	112.0
H $\alpha$	4.43	C $\alpha$	62.0		
H $\beta$	4.33	C $\beta$	70.0		
H $\gamma$ 2	1.22	C $\gamma$ 2	21.6		
<b>Trp</b>					
H	8.22	C	177.1	N	122.1
H $\alpha$	4.7	C $\alpha$	57.6		
H $\beta$ 2	3.34	C $\beta$	29.8		
H $\beta$ 3	3.25	C $\delta$ 1	127.4		
H $\epsilon$ 1	10.63	C $\gamma$	111.7		
H $\delta$ 1	7.28	C $\epsilon$ 3	122.2		
H $\epsilon$ 3	7.65	C $\zeta$ 3	124.8		
H $\zeta$ 3	7.18	CH2	121.1		
HH2	7.26	C $\zeta$ 2	114.8		
H $\zeta$ 2	7.51	C $\epsilon$ 2	139.0		
		C $\delta$ 2	129.6		
<b>Tyr</b>					
H	8.26	C	176.7	N	120.9



Table 3.2 (Continued)

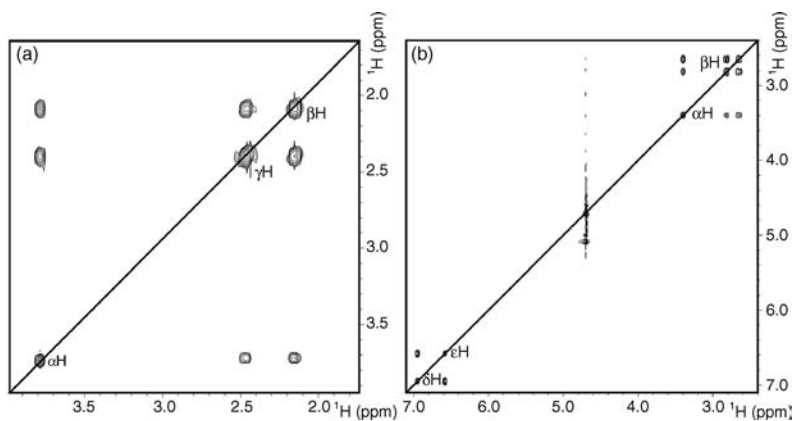
H $\alpha$	4.58	C $\alpha$	58.3		
H $\beta$ 2	3.09	C $\beta$	38.9		
H $\beta$ 3	2.97	C $\gamma$	130.8		
H $\delta$ 1	7.15	C $\delta$ 1	133.3		
H $\epsilon$ 1	6.86	C $\epsilon$ 1	118.3		
H $\epsilon$ 2	6.86	C $\zeta$	157.5		
H $\delta$ 2	7.15	C $\epsilon$ 2	118.3		
		C $\delta$ 2	133.3		
<b>Val</b>					
H	8.16	C	177.0	N	119.3
H $\alpha$	4.16	C $\alpha$	62.6		
H $\beta$	2.11	C $\beta$	32.8		
H $\gamma$ 1	0.96	C $\gamma$ 1	21.1		
H $\gamma$ 2	0.96	C $\gamma$ 2	20.3		

Amino acids are reported as three-letter code and side-chain atoms are labeled with classical numbering based on Greek letters.

In particular, it has been found that the  $^1\text{H}$ -NMR chemical shift of the  $\alpha$ -proton of all amino acids experiences an upfield shift (with respect to the random coil value) when in a helical conformation [9, 25] and a comparable downfield shift when in an extended conformation [9]. Conversely, side-chain chemical diversity is reflected in NMR spectra, where  $^1\text{H}$ -J-coupling gives rise to different spectral patterns, typical of different spin systems. A spin system is a group of spins that are connected through scalar (through-bond) spin-spin coupling. In biological molecules this coupling is usually observed for nonlabile protons separated by three or less covalent bonds. For such reason, most amino acids, with only five exceptions, show a unique spin system in which all nuclei are correlated.

The five exceptions are the aromatic side-chains and that of methionine. In the case of amino acids with an aromatic side-chain (Y, W, H, and F), the spin systems are neatly separated in two halves because their aromatic rings are not magnetically connected to the  $\alpha\text{CH}-\beta\text{CH}_2$  segment. Similarly, in the case of methionine (M), the  $\epsilon\text{CH}_3$  is separated from the  $\alpha\text{CH}-\beta\text{CH}_2-\gamma\text{CH}_2$  segment by a sulfur atom that prevents efficient transmission via bonding electrons.

The spin system pattern is a property retained by amino acids in spectra of peptides and proteins (the peptide linkage keeps protons of chained residues more than three bonds apart), whereas chemical shifts are influenced by secondary and tertiary structure (spatial folding) and deviate from random coil values. Spin systems are easily revealed by correlation spectroscopy (COSY) or total correlation spectroscopy (TOCSY) types of two-dimensional spectra as cross-peak connectivities between resonances (Figure 3.1). Spin systems are, indeed, at the basis of amino acid identification and resonance assignment for polypeptides, as described below, and are usually described in terms of the classical notation of John A. Pople or graphically, as illustrated in Figure 3.2.



**Figure 3.1** Spin system patterns as observed in TOCSY spectra (400 MHz). (a) Spin system of glutamine: here all proton resonances (on-diagonal) are connected to each other by cross-peaks (off-diagonal), due to the fact that each pair is no more than three bonds

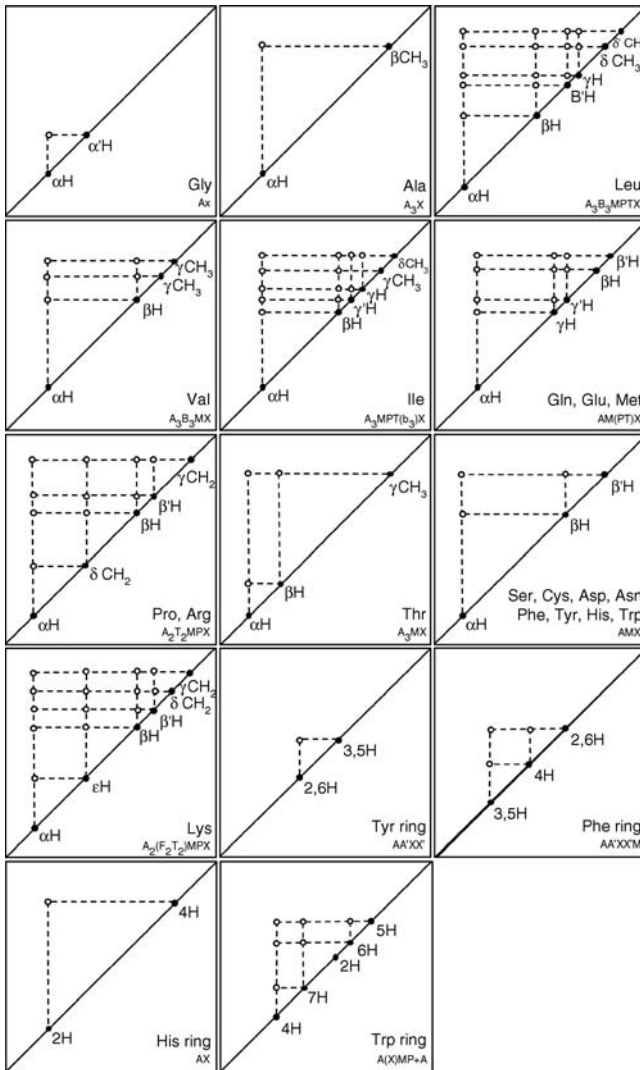
apart. (b) TOCSY spectrum of a tyrosine: aliphatic ( $\alpha$  and  $\beta$ ) and aromatic ( $\delta$  and  $\epsilon$ ) on-diagonal resonances show cross-peaks forming two distinct spin systems because of their protons being more than three bonds apart.

### 3.2.5

#### Labile Protons

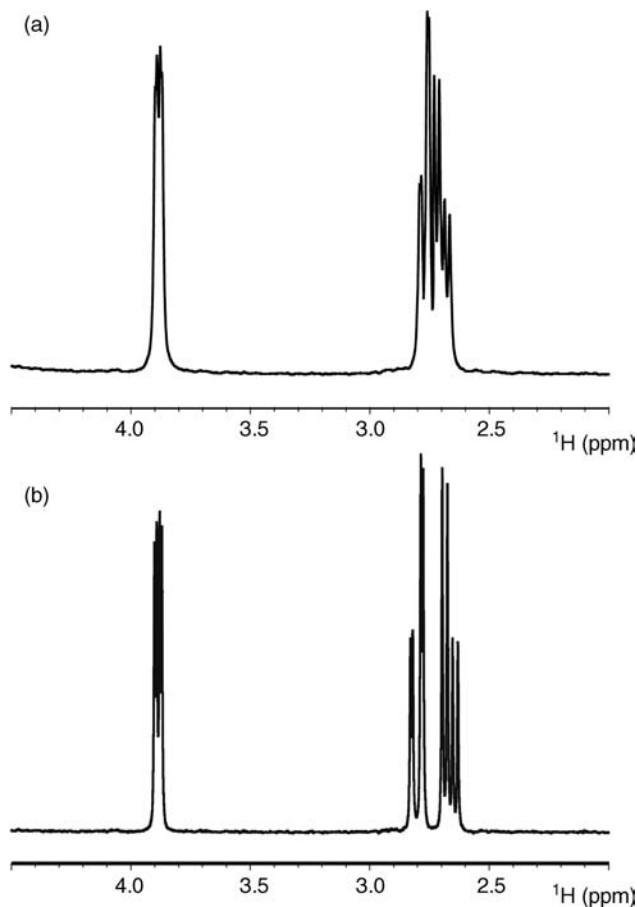
Amino acids carry protons exchangeable with the solvent, referred to as labile protons. Their intrinsic exchange rate is characteristically modulated by pH and temperature. Such behavior is well illustrated by the graphical representation of the simulated pH dependence of the intrinsic exchange rate ( $K_{\text{intr}}$ ), as reported by Wagner and Wüthrich [26].

As increasing the exchange rate increases the line broadening, signal observation of labile protons is feasible only if the exchange rate is slow on the NMR timescale. Practically, on high field spectrometers, an exchange rate smaller than  $10^3 \text{ min}^{-1}$  is necessary for proton observation. Thus, only the intrinsic exchange rates of the labile protons of side-chains of R, K, N, Q, W, and backbone amides allow for signal observation and only in the slightly acidic pH range. The pH range 3–5 is ideal to achieve best experimental results, but it is not always consistent with the mimic of reasonable biological environments. It is also clear that, for an isolated amino acid, as well as for the N- or C-terminal residues of a polypeptide chain, the amino and the acidic protons are not observable due to the fast exchange rate over all the pH range. However, the exchange rate of labile protons is largely influenced by spatial folding (due to hydrogen bonding, solvent exclusion), thus it plays important roles in the study of peptides and proteins. See Figure 3.3.



**Figure 3.2** Schematic patterns of the spin systems for nonlabile protons of the 20 amino acids. Filled circles represent proton resonances at their relative chemical shift; open circles represent  $J$ -coupled resonances cross-peaks pattern for three-bond connections as expected to be observed in TOCSY type spectra. The insets report the spin system type

in the Pople notation, with the letter representing the highest field resonance preceding the others in the alphabet; neighboring letters in the alphabet represent nuclei with strong coupling, while non-neighboring letters represent weak coupling; a group of  $n$  equivalent spins is represented as  $A_n$ .



**Figure 3.3** NMR  $^1\text{H}$  one-dimensional spectra of an aspartyl residue acquired at 90 (a) and 400 MHz (b). Signal resolution improvement in (b) is apparent.

### 3.2.6

#### Contemporary Relevance: Metabolomics

The growing field called “metabolomics” detects and quantifies the low-molecular-weight molecules, known as metabolites (constituents of the metabolome), produced by active, living cells under different conditions and times in their life cycles. The words “metabolomics” and “metabonomics” are often used interchangeably, although a consensus is beginning to develop as to the specific meaning of each. The goals of metabolomics are to catalog and quantify the myriad small molecules found in biological fluids under different conditions. Metabonomics is the study of how the metabolic profile of a complex biological system changes in response to stresses like disease, toxic exposure, or dietary change. NMR is playing an important role in metabolomics because of its ability to observe mixtures of small molecules in

living cells or in cell extracts with little or no pretreatment of the samples. NMR and mass spectrometry are the two most often used analytical methods for metabolite profiling because of their high resolution and rich data content. Although mass spectrometry is the more sensitive technique, NMR provides broad coverage of the metabolome by detecting all of the metabolites present in the biofluid simultaneously, with excellent reproducibility and only limited sample pretreatment [27–31].

Since amino acids are important metabolites present in several biological fluids, they are one of the subjects of NMR metabolomics. For example, several of the inborn errors of metabolism are associated with the accumulation of amino acids as metabolites in serum and urine.  $^1\text{H}$ -NMR studies on urinary excretion of diagnostic amino acids such as phenylalanine in phenylketonuria, branched chain amino acids (leucine, valine, isoleucine) in maple syrup urine disease, *N*-acetyl aspartic acid in Canavan disease, and tyrosine and *N*-acetyl tyrosine in tyrosinemia type I have been reported [32, 33]. Figure 3.4 shows a typical NMR spectrum of a full-term normal newborn urine where, among many other metabolites, amino acids can be observed. It must be reported that metabolomics analysis of NMR spectra does usually imply use of a complex statistical approach in place of direct signal observation.

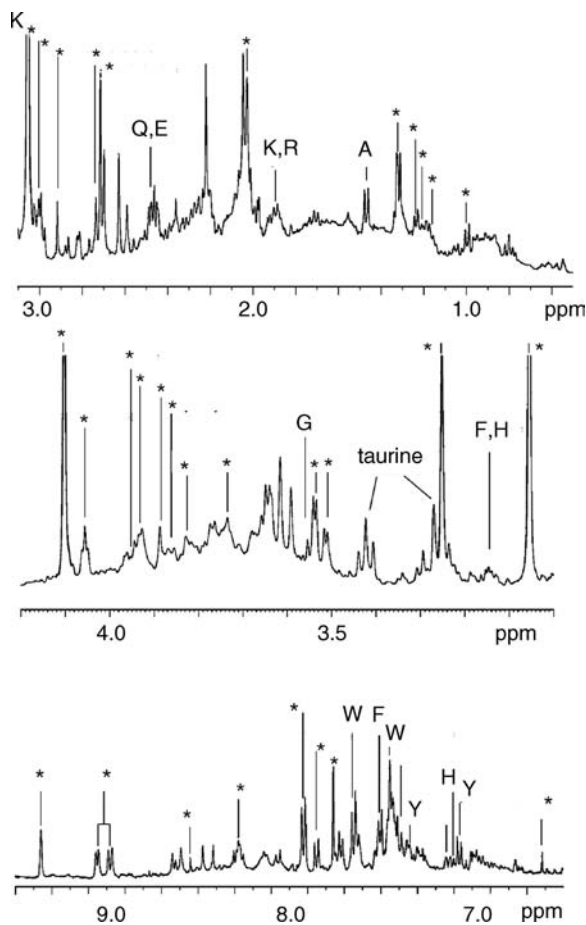
### 3.3

#### Peptides

##### 3.3.1

##### Historical Significance

NMR of peptides played an important role during the first years of structural biology. When assignment of protein resonances could rely only on basic homonuclear bidimensional experiments it was important to have, as guidelines, typical correlation spectra of each amino acid residue and tables of characteristic chemical shifts (see previous section). This approach is well illustrated in the book of K. Wüthrich [4] that shows the essential points of the protocol ever since followed by all NMR structural biologists in the determination of protein structure. As discussed above, while the patterns of side-chains (dictated by coupling constants) could be transferred directly from spectra of simple amino acids to those of proteins, in the case of chemical shifts it was necessary to resort to small peptides to mimic the changes brought about by the presence of next neighbors in the polypeptide chain. The use of the chemical shift, as an aid in protein structure determination, was more or less put aside in subsequent years when powerful multidimensional heteronuclear experiments were introduced at a very high pace, but has been resumed in the last 15 years in various forms, although it is more convenient nowadays to use chemical shift data extracted from protein spectra rather than resort to model peptides [34].



**Figure 3.4** NMR <sup>1</sup>H spectrum at 400 MHz of the urine of a full-term normal newborn. Resonances from amino acids are labeled by the single-letter code of the amino acid residue; other assigned resonances are

simply labeled by an asterisk. Taurine is also indicated, because although strictly speaking not an amino acid, it is often called in scientific literature. (Adapted from [32].)

### 3.3.2

#### Oligopeptides as Models for Conformational Transitions in Proteins

Polypeptides and oligopeptides, usually of uniform composition (i.e., made of a single type of residue), have been extensively studied in the past to understand basic structural aspects of biopolymers. Model peptides were used mainly to study conformations that characterize the secondary structure of proteins, particularly the tendency, in small sequences, to assume the  $\alpha$ -helix and  $\beta$ -sheet conformations, regarded as “building blocks” of proteins. Model polyamino acids were mainly used to study cooperative phenomena, in particular conformational transitions such as the

helix  $\rightarrow$  coil transition. It is probably worth noting that polypeptides composed of a single amino acid residue are better discussed here rather than in Section 3.4 on "Proteins" since, like small peptides, they differ from proteins mainly by the inability to sustain self standing structures. Most of the studies on synthetic polyamino acids were performed by means of optical spectroscopies, but there is an interesting group of NMR studies.

Most of the polyamino acids used to study helix  $\rightarrow$  coil transitions were hydrophobic, because water-soluble polyamino acids have very little tendency to assume nonrandom conformations. Thus, polyamino acids studied to observe the transition were either those built using residues with natural hydrophobic side-chains or those with hydrophilic side-chains modified to become hydrophobic (e.g., by esterification of the carboxyl groups of glutamic acids). Helicity was favored by dissolving the hydrophobic polyamino acids in suitable organic solvents [35] and the transition to random coil was induced by additions of small percentages of strong acids. In all cases the  $\alpha$ -helical resonance was found centered around 4.1 ppm and moved downfield by about 0.4 ppm upon addition of the helix breaker cosolvent.

The phenomenon of interconverting resonances attributable to helical and random coil conformations was observed for the first time by J. Ferretti [36], who considered the two limiting positions as typical of  $\alpha$ -helix and random coil conformations. Although this interpretation was prevailing it was not unchallenged [37–40], particularly because, in some samples, the  $\alpha$ CH and NH resonances give rise to double rather than single peaks [41, 42]. Bradbury *et al.* [25] showed convincingly that the assignment of resonances to helix and random coil conformations was genuine, whereas the occasional observation of noninterconverting double peaks can be explained with the polydispersity of the synthetic polymers [43]. The possibility of a major role of protonation in the helix  $\rightarrow$  coil transition [41, 42] was also ruled out by the observation that dimethylsulfoxide (DMSO), a helix breaker with no labile protons, behaves like strong acids [35]. The detection of characteristic peaks for the two conformations of the  $\alpha$ CH resonance also in  $^{13}\text{C}$  spectra confirmed the generality of the phenomenon [44].

One of the first questions addressed in studies of model peptides was whether it is possible to determine a minimum length (necessary and sufficient) for an isolated  $\alpha$ -helix. Most of the studies on synthetic peptide oligomers originated in the laboratory of M. Goodman, who pioneered the synthetic methods to produce many different peptide oligomers. A paper by Temussi and Goodman on a small family of synthetic oligopeptides is paradigmatic of this type of approach and represents a sort of "swan song" of continuous-wave NMR applications to peptides, just before the advent of modern pulse spectrometers [45]. This paper dealt with NMR studies of oligomers of benzyloxycarbonyl- $\gamma$ -ethyl glutamate in mixtures of chloroform and trifluoroacetic acid. Titration of the chloroform solutions of oligopeptides with trifluoroacetic acid showed that some of the NH resonances are involved in strong intramolecular hydrogen bonds, a clear indication of a seeding for helical structure, a sort of "nascent helix." The nature of these observations was strongly backed by the fact that structuring becomes increasingly clear in going from the tetramer to the heptamer.

Although it is difficult to observe very stable short helices in water, it is often possible to detect the tendency to become helical: the concept of a nascent helix was introduced explicitly by P. Wright and J. Dyson in an extensive NMR study of immunogenic peptides [46, 47]. The stability of short isolated helices forms the basis for hierarchical models of protein folding [48].

Contrary to studies on the helix  $\rightarrow$  coil transition, that were very popular in the 1970s but have been superseded by the vast literature on actual proteins, studies of model peptides, regarded as indicative of small isolated building blocks of proteins (the  $\alpha$ -helix and the  $\beta$ -sheet), have attracted much attention also in recent years, mainly because of their relevance for protein folding. Prominent among studies on isolated  $\beta$ -sheets are the NMR investigations performed by L. Serrano and his group. These researchers showed by NMR that a 16-residue peptide, corresponding to the second  $\beta$ -hairpin of the B1 domain of protein G, assumed a  $\beta$ -hairpin conformation when dissolved in water [49]. Such a finding hinted at the possible relevance of  $\beta$  structural elements in the early steps of protein folding. Even more spectacular was the design of a 20-residue peptide (named Betanova) that can exist as a triple-stranded  $\beta$ -sheet [50]. It is very unfortunate that the authors had to partially retract this work, by admitting that the population of Betanova was overestimated [51] after Hilario and Keiderling [52] showed that the data had been overinterpreted. In fact, according to Kuznetsov *et al.* [53] the amount of  $\beta$ -sheet in Betanova is vanishingly small. However, these authors were able to describe authentic cases of small peptides capable of assuming isolated  $\beta$ -sheet conformation. All of these cases require the insertion in the sequence of nonproteic residues to strengthen the conformational tendency [54].

### 3.3.3

#### **Bioactive Peptides**

It is now well known that natural peptides have a huge variety of biological functions, mainly as hormones or signaling molecules. To quote just a few of the most important bioactive peptides, it is sufficient to recall opioids, antibiotics, immunostimulants, calcitonins, tachykinins, and vasoactive intestinal peptides. The interest in bioactive peptides was ignited in 1975 by the discovery of enkephalin [55]: the importance of this discovery prompted a very large number of studies whose goal was to find new bioactive peptides, soon identified as ideal lead compounds for drug design. In turn, there was also a parallel bloom of conformational studies on opioid peptides and on many other bioactive peptides. The goal of these studies was to find the so-called bioactive conformation of the peptides (i.e., the conformation adopted by the peptides inside their receptor). Given that most bioactive peptides interact with important receptors, this search was motivated by the hope of modifying the structure of the peptides to obtain useful drugs. The main drawback of peptides as drugs, apart from the possible difficulty in their production, is their extreme sensitivity to proteolytic enzymes. The modifications ranged from simple substitutions of natural residues with other natural residues, aimed at increasing their activity, to the incorporation of non-natural residues, aimed at increasing their



resistance to proteolysis, up to synthesizing peptidomimetic compounds. In spite of the enormous number of conformational studies of bioactive peptides, the attempts to find “bioactive” conformations essentially failed, because small peptides rarely adopt a well-defined conformation in solution and, moreover, their bioactive conformation is adopted only in the unique environment of the active site of the receptor. The search for the bioactive conformation has been aptly defined “an elusive goal” [56]. Notwithstanding this, these studies have yielded many interesting results, mainly on the conformational tendencies of peptides. Several ingenious ways have been devised to circumvent the intrinsic difficulties originating from conformational flexibility. Probably, the most important one is the choice of the right media, but combinations of experimental approaches and theoretical calculations also play a relevant role.

#### 3.3.4

##### **Choice of the Solvent**

Among the first NMR studies dealing with the conformation of bioactive peptides are probably two papers describing the solution structure of enkephalins, soon after their discovery [57, 58]. These studies, as most early structural studies on bioactive peptides, were performed in aqueous solutions and were largely based on the assumption that observable NMR parameters reflected the structure of a single conformer. This assumption may be true for peptides dissolved in some peculiar environment, but it is certainly not true in water. In fact, all linear peptides of short sequence are too flexible to assume a single structure, or even a limited number of conformations, in aqueous solution [59]. The conformational flexibility of linear peptides is reflected by NMR parameters: typically, NH chemical shifts are all close to those of random coil conformations and  $J_{\text{NH}C\alpha_{\text{H}}}$  scalar couplings are distributed around the average value of 6.5 Hz, also typical of random coil conformations. Even more important, when studying small peptides in solution, it is impossible to observe diagnostic NOEs. Apart from sequential and a few intraresidue ones, no medium- or long-range NOEs are observed in spectra of small linear peptides. This difficulty originates in part from the intrinsic flexibility of small peptides and in part from the sheer size of these molecules, corresponding to very short rotational correlation times. The latter problem found a solution with the introduction of the rotational nuclear Overhauser effect spectroscopy (ROESY) experiment [60], that circumvents problems connected to unfavorable correlation times, allowing the measurement of a larger number of nuclear Overhauser effects. However, it was soon found that in small peptides observable ROEs are still of low diagnostic value.

One of the first observations of diagnostic NOEs in the spectra of small peptides was reported in a study of enkephalins at low temperatures, in solvents of elevated viscosity [61, 62]. The experimental observation that it is possible to induce NOEs emphasizes the importance of the environment in changing the conformational state of the peptide, but it also raises the question whether it is meaningful, from a biological point of view, to search for *any* environment that favors the existence of

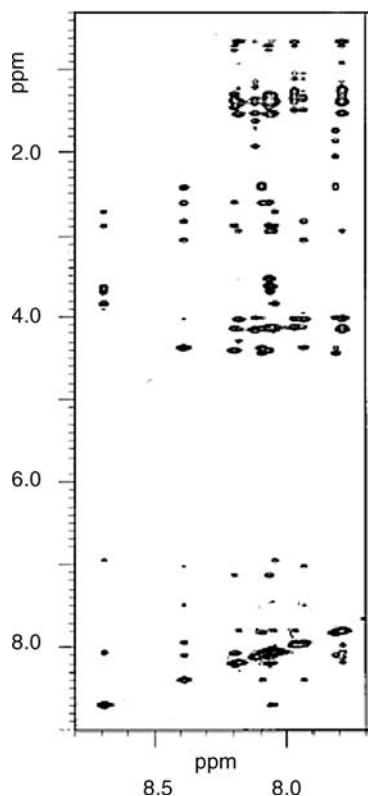
stable conformers. In principle, finding a suitable structuring environment is not too different from finding the right crystallization conditions for X-ray diffraction studies. However, the main goal of conformational investigations on bioactive peptides is not to find the best structuring conditions per se but to study peptides in biologically relevant environments. Bioactive peptides can be found in several biological environments that can be grouped in three categories: transport fluids, membranes and receptor's active sites. The big question in solution studies of small peptides is whether we can simulate biological environments and, possibly, get diagnostic NMR parameters at the same time.

#### 3.3.4.1 Transport Fluids

All transport fluids of course are essentially aqueous solutions containing a wide range of solutes, which differ in different biological environments. Small peptides are essentially disordered in water and this fact was well documented by all attempts to measure NOEs in peptide solutions in the decade 1976–1986. It came as a surprise when Gupta *et al.* [63] reported the observation of large NOEs among aromatic protons of Tyr1 and Phe4 of enkephalin in water, although these effects were not accompanied by any other parameter indicating a stable structure. This claim was only a false-positive, due to a trivial technical error, but it stimulated the search for the right conditions leading to measurable NOEs.

Motta *et al.* were able to show that in a mixture of DMSO and water it was indeed possible to observe both intra- and inter-residue NOEs [61, 62]. Aqueous solutions containing variable amounts of DMSO have been dubbed cryoprotective mixtures (cryomixtures), along with other mixtures of water and hydrophilic organic solvents (e.g., alcohols and dimethylformamide) employed in numerous biochemical and crystallographic studies on proteins [64]. They were shown to be fully biocompatible, and their ability to protect or even induce structure on peptides and proteins is akin to that of osmolytes – protective solutions found in many organisms living in extreme conditions [65, 66]. Cryomixtures can be used in a wide range of temperatures and, at a given low temperature, they may have a dielectric constant identical to that of water at room temperature. This unique feature makes them possible candidates for a mimic of transport fluids. From the viewpoint of structure determination of small peptides, the main effect of cryomixtures is to change the correlation time of the solutes, owing to their slower motion in a viscous medium, allowing the measurement of NOEs. This technical aspect is reminiscent of the use of the ROESY experiment, instead of the NOESY. Figure 3.5 shows the rich NOESY of dynorphin A at low temperature in a DMSO/water cryomixture [67].

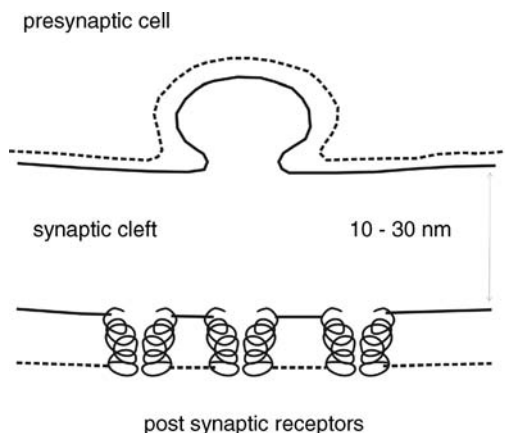
However, the action of cryomixtures is not limited to the influence of viscosity on NMR parameters, but can be likened to a “conformational sieve,” because high viscosities can favor folded (more compact) conformers over disordered ones [68]. It had been pointed out that the interaction of nanomolar solutions of *disordered* bioactive peptides with their receptors amounts to overcoming an entropic barrier, because the dilution of the bioactive conformation into a very large number of disordered conformations lowers the effective concentrations even further [69]. Temussi *et al.* postulated that the viscosity of these fluids contributes, in addition



**Figure 3.5** Partial NOESY spectrum of dynorphin A at 278 K in DMSO/water (80:20, v/v).

to the membrane catalysis proposed by Schwyzer [69], to overcoming the entropic barrier to the transition state of peptide–receptor interaction, by selecting ordered conformations prior to receptor interaction. A further point of interest, in the use of these media, comes from the fact that, contrary to the dilute aqueous solutions generally employed in NMR studies, they have a viscosity close to that of cytoplasm. Viscosities of cytoplasm range from 5 to 30 cP [70] and may play an important role in cell communication processes [71].

Many bioactive peptides exert their action at synapses. After being excreted at the presynaptic cell, in order to reach receptors located on the postsynaptic membrane they have to cross a narrow cleft (100–300 Å) occupied by the intersynaptic fluid – an aqueous solution whose viscosity can be even higher than that of cytoplasm, because of the ordering effect of membrane heads and of unstirred layer phenomena [72]. Figure 3.6 shows a cartoon of a hypothetical synaptic cleft: it can be surmised that the narrow cleft may have very high local viscosity. The use of cryomixtures, with viscosities mimicking that of the synaptic cleft, yielded the measurement of the first NOEs of enkephalins, that, in turn, allowed sophisticated calculations on the composition of conformers in equilibrium [133].



**Figure 3.6** Cartoon of a synaptic cleft.

### 3.3.4.2 Membranes

Nearly all receptors of bioactive peptides are membrane receptors (e.g., the seven transmembrane helices of G-protein-coupled receptors of opioids). Although receptors terminals face the outer environment, the contiguity of membranes explains, in part, the interest in the study of the conformational preferences of bioactive peptides in media similar to the membrane environment. Among the media generally employed to mimic inner membrane environment, the most popular ones are probably aqueous micellar solutions. Many of the older studies are in micelles of sodium dodecylsulfate (SDS) – a detergent that is easy to find in perdeuterated form [73–76]. There are also several studies in micelles of phospholipids [77–87]. It is difficult to assess the biological relevance of these studies, because most bioactive peptides do not cross the lipidic phase of the membranes when interacting with the receptors. The only aspect that the inside of micelles has in common with the receptors is the fact that the cavities of receptors active sites are generally apolar.

An obvious application of media mimicking membranes interior is the study of peptides naturally found in membranes. Although short peptides are not normally found in natural membranes, many synthetic peptides derived from transmembrane helices have been studied.

Numerous studies in membrane-mimetic solvents demonstrated that peptides corresponding to single transmembrane helices of bacteriorhodopsin [88–93], rhodopsin [94–98], Ste2p, the  $\alpha$ -factor receptor [96, 99–104], and the adenosine A2 receptor [105, 106] do assume a helical conformation *in vitro*. A very accurate high-resolution NMR study on an 80-residue fragment of the yeast  $\alpha$ -factor receptor Ste2p (G31–T110) showed that it is possible to observe even a specific tertiary structure if the peptide contains more than one transmembrane domain [107]. This fragment, containing a short stretch of the N-terminus, TM domain 1, the first intracellular loop, TM domain 2, and a short stretch of the first extracellular loop, was biosynthesized with [ $^{15}\text{N}$ ], [ $^{15}\text{N}$ ,  $^{13}\text{C}$ ], and [ $^{15}\text{N}$ ,  $^{13}\text{C}$ ,  $^2\text{H}$ ] uniform isotope labeling. These

and additional specific labeling led to a nearly complete assignment of backbone and side-chain resonances in 1-palmitoyl-2-hydroxy-*sn*-glycero-3-[phospho-*rac*-(1-glycerol)] micelles. The conformation, determined without introducing any artificial restraints, shows a well-defined secondary and a few interhelical contacts consistent with the fact that the protein is folded in micelles into a helical hairpin that splays apart at the termini [107].

Another natural application of membrane-mimetic media is that of antimicrobial peptides (AMP), particularly cationic AMPs. Naturally occurring AMPs differ in size (ranging from 10 to 50 residues), sequence, and three-dimensional structure, but share net positive charge and amphipathicity [108]. Many of them adopt an  $\alpha$ -helical amphipathic structure in membrane-mimetic media, which is considered to be a prerequisite for their lytic activity [109–111]. Members of the dermaseptin family, isolated from the skin secretion of the *Phyllomedusinae* tree frogs, are among the most studied antimicrobial peptides and represent a convenient subset of AMPs, from the point of NMR solution studies.

Dermaseptins exert a lytic action on bacteria, protozoa, yeast, and filamentous fungi at micromolar concentrations, but unlike polylysines, show little hemolytic activity. The mechanism of action against bacteria is linked to the propensity of these peptides to form amphipathic helices. However, it is not easy to understand the capacity of dermaseptins to discriminate between mammalian and microbial cells. NMR studies of dermaseptins have been very helpful in all attempts to build a consensus model for their action, although the relationship between the three-dimensional structure of the membrane bound dermaseptins and the proposed mechanisms of membrane-perturbing action remains controversial. NMR studies in trifluoroethanol (TFE)/water mixture or SDS micelles showed that dermaseptin B2 adopts different  $\alpha$ -helical structures depending on the environment [112]. On the other hand, dermaseptin S3 does not adopt a stable  $\alpha$ -helical structure in TFE/water, but has several turn-like regions, typical of a nascent helix [113]. Dermaseptin S4, characterized by weak antibacterial activity, but strong hemolytic and antiprotozoan effects, is so highly aggregated in aqueous solution that it prevents detailed NMR analysis [114].

The isolation of dermaseptin DS 01, a new dermaseptin from the skin secretion of *Phyllomedusa oreades*, has posed an interesting problem related to the structure–activity relationship of this class of antimicrobial peptides. DS 01 seems to share properties of the two classes (S and B): it is active against bacteria without substantial hemolytic activity (like dermaseptins S), but it is also active against higher microorganisms (like dermaseptins B) [111]. Castiglione-Morelli *et al.* [111] undertook a detailed structural characterization of DS 01 in media that mimic the membrane environment, both in the lipid phase (mixtures of TFE and water) and on the membrane surface (SDS micelles). They found that DS 01 has a high tendency to assume a helical conformation in both environments and concluded that the ability of DS 01 to interact also with the membrane of protozoa is probably linked to the high helical propensity of its sequence, whereas the lack of hemolytic activity is probably due to the presence of an uncharged residue in the fifth position.

### 3.3.4.3 Receptor Cavities

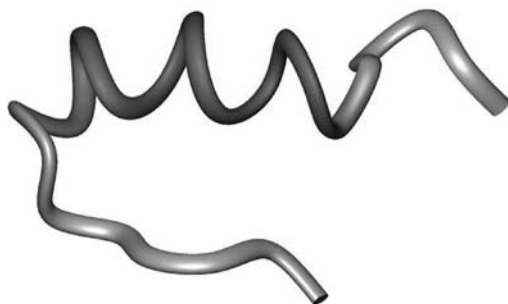
In general, receptor active sites are apolar cavities, with one or few charged groups to anchor the agonists. For instance, in the case of opioids, the active site is a largely hydrophobic cavity, lined by the several aromatic side-chains hosting a negatively charged moiety interacting with the basic site common to all opioids, both peptidic or alkaloidic ([115] and references therein). A large number of NMR studies in alcohols or in aqueous alcoholic mixtures of different bioactive peptides were performed with the idea of reproducing, at least in part, the apolar environment of receptors active sites [86, 87, 116–119]. Alcohols *per se* are not apolar molecules, but their radicals are typical hydrocarbon moieties and, as such, can mimic the apolar walls of the active site by presenting an apolar face to the peptides. For instance, according to Rajan *et al.* [119], mixtures of water and fluorinated alcohols (e.g., hexafluoroacetone hydrate (HFA)), can surround the helix with a sort of “Teflon coating.”

Most bioactive peptides are so intrinsically flexible that even in alcohol/water mixtures they are completely random, but these environments proved very useful in the case of some longer bioactive peptides, such as human  $\beta$ -endorphin – a peptide composed of 31 residues, with an N terminal part coincident with the sequence of Leu-enkephalin. It was shown that this peptide has very little tendency to assume an ordered structure in water but a strong tendency to assume a helical structure in its address domain (from P<sup>13</sup> to Y<sup>27</sup>) in mixtures of water and alcohols (see Section 3.3.6.2).

The random coil conformation the N-terminal part of endorphin (YGGFL) is consistent with the many unsuccessful attempts to observe structured enkephalins, but it is at variance with the behavior of the similar sequence (YGGFM) present in the C-terminal part of enkelytin – an antibiotic peptide [120]. Although coming from proenkephalin A, a precursor of enkephalin, enkelytin has not opioid but antibiotic activity. An NMR study showed that synthetic PEAP-209–237 is unstructured in water, but folds into a largely helical conformation in TFE/water [121]. As observed by these authors, theirs is the first observation of a helical structure for the sequence of enkephalin. It is interesting that the ability to assume an ordered conformation is paralleled by the absence of any binding (of enkelytin) to opioid receptors. It is possible to hypothesize that the N-terminal charge is essential both for binding and to prevent helicity.

Another interesting application of this type of solvent was the attempt to validate a possible mechanism of secretion of Chaperonin 10 (Cpn10) from *Mycobacterium tuberculosis*. Cpn10 is secreted outside the live bacillus, and accumulates both in the bacterial wall and in the matrix of the phagosomes [122]. Although the N-terminal portion of the Cpn10 was, indeed, identified as possessing an amphiphilic helical character, available crystallographic structures of Cpn10 oligomers showed no  $\alpha$ -helices.

To understand, on a structural basis, whether any conformational changes might take place when the protein interacts with the cytosolic membrane phospholipids or goes through the highly hydrophobic interior of the mycobacterium cytosolic membrane, the solution structure of a synthetic peptide reproducing the 1–25



**Figure 3.7** Solution structure of Cpn10<sup>1-25</sup> in a 50% HFA aqueous mixture (the  $\alpha$ -helix is shown in dark grey [119]). (Image produced with MOLMOL [123].)

fragment of Cpn10 was studied by NMR under conditions of reduced solvent polarity. Indeed, while the peptide Cpn10<sup>1-25</sup> was shown not to be structured in water, in 95% methanol and 50% HFA, the segment encompassing residues 5–16 showed a well structured  $\alpha$ -helix (Figure 3.7). The NMR data collected in 50% HFA showed that, in the absence of intersubunit interactions, a structural transition from  $\beta$ -strand to  $\alpha$ -helix does occur in the N-terminal region of the monomeric protein and that the stability of these helices increases as the hydrophobic or acidic environments increase.

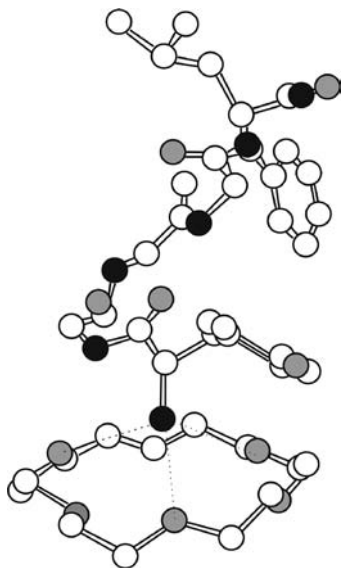
As in the case of endorphin, it might be argued that the structural transition observed for the Cpn10 N-terminal region of *M. tuberculosis* can be attributed to the  $\alpha$ -helical stabilizing effects of fluorinated solvents such as TFE, but experimental evidence from several peptides show that regions without helical propensity do not form helices even in 100% TFE.

It is difficult to judge whether the environment seen by peptides when dissolved in hydroalcoholic mixtures is truly apolar, because alcohols are not apolar molecules. On the other hand, it is not possible to study peptides in truly apolar solvents since they are not soluble enough for most spectroscopic techniques. Solubility is precluded by the presence, on peptides, of charged groups that is essential, in general, for an interaction with the receptor [124].

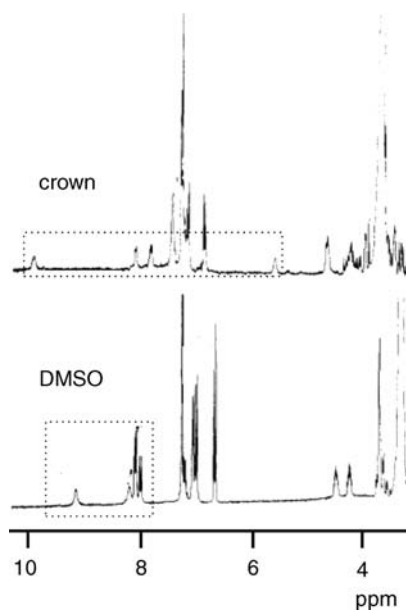
A possible way to circumvent this difficulty was proposed by Temussi *et al.* [125].

It was shown that complexation of the  $-\text{NH}_3^+$  group of enkephalin amides with a crown ether allows dissolution of the complex in truly apolar solvents at concentrations consistent with spectroscopic measurements. The binding of the charged group to the ether can be likened to the binding of the same group to the anionic subsite of the receptor, whereas the apolar solvent can play the role of the hydrophobic cavity. Figure 3.8 shows a schematic model of the complex between Leu-enkephalin amide and a crown ether. The improvement of the NMR spectrum in  $\text{CDCl}_3$ , as a crown ether complex, with respect to that of the free amide in DMSO (Figure 3.9), is indeed spectacular.

These experimental conditions favor folded conformations of the family of  $\beta$ -turns, but in the case of enkephalin there is still too much residual flexibility to determine a precise structure [125–127].



**Figure 3.8** Schematic model of the complex between Leu-enkephalin amide and a crown ether. Carbons are represented as open balls, oxygens as gray balls, and nitrogens as black balls. The dotted lines connect the ammonium group of Leu-enkephalin amide with three oxygens of the ether.



**Figure 3.9** Comparison of the one-dimensional spectra of Leu-enkephalin amide in DMSO and in  $\text{CDCl}_3$  (as a complex with a crown ether). The relative ranges of NH resonances are boxed.



### 3.3.5

#### Ensemble Calculations

In the case of flexible molecules, the analysis of NMR data can be greatly facilitated by combination with ensemble calculations. The main difficulty in the determination of a three-dimensional structure of small peptides stems from the fact that experimentally measured NOE intensities do not originate from a single structure, but represent averages over individual contributions. Therefore, procedures routinely employed in the determination of protein structure by NMR, require specific adaptations for peptides. Mierke *et al.* [128] modified standard restrained molecular dynamics (RMD) calculations, using pseudo potentials based on NMR data, and time averaged restraints. Brüschweiler *et al.* [129] set up a procedure, dubbed MEDUSA, to compensate for the fact that, in an ensemble of conformations, individual conformations might violate some of the NOE distance restraints. In their calculations, pairs of exchanging conformations were considered and then the best combinations in terms of structural similarity were delineated [129]. Several alternative methods of different levels of sophistication for analyzing NMR parameters of small peptides have been proposed by several other groups, notably Cicero *et al.* [130] and Bonvin and Brünger [129]. These methods basically generate a set of more or less reliable conformations using currently available force fields and then select conformations only if their energies are below an arbitrary threshold, and at the same time have significantly different topologies and are consistent with a subset of the NMR restraints. The main disadvantage of these methods is that the selection of the conformations is essentially arbitrary, implying that the populations become fitting parameters, rather than thermodynamic variables. To overcome this difficulty, Meirovitch *et al.* [132–135] developed a statistical mechanics methodology for treating flexibility and used it to analyze NMR data. The first step of this methodology is also based on an extensive conformational search using the local torsional deformation method for identifying a large number of the low energy structures not more than 2–3 kcal above the global energy minimum. Comparison of these structures allows clustering into a subset of significantly different structures. Each of these structures then becomes a seed for a canonical Monte Carlo calculation that spans its vicinity in multidimensional space. The free energy of the corresponding sample is calculated with the local states method from which the relative populations of these regions (microstates) are obtained. This methodology was applied initially to the linear peptide Leu-enkephalin (H-Tyr-Gly-Gly-Phe-Leu-OH) described by the potential energy function ECEPP [136].

### 3.3.6

#### Selected Examples from the Major Fields of Bioactive Peptides

##### 3.3.6.1 Aspartame

L-Aspartyl-phenylalanine methyl ester (dubbed aspartame) was the first sweet dipeptide, discovered by serendipity in the 1960s [137]. Its solution structure was particularly difficult to determine, not only because of the predictable flexibility, but

also because it is not likely to observe diagnostic NMR parameters from a molecule composed by only two residues. The approach of Lelj *et al.* [138] was a combination of NMR measurements and exhaustive molecular mechanics calculations explicitly used to interpret the NMR data. This approach was an absolute novelty in the early 1960s, particularly because it included an explicit attempt to account for conformational equilibria via an estimation of entropic contributions.

In turn, the combination of the solution structure of aspartame [138] with several observations on more rigid compounds, led to a detailed quasiplanar model of the active site of the sweet receptor. The main features of this model can be summarized as follows: (i) the active site of the receptor is a flat cavity with one side partially accessible even during the interaction with the agonist, (ii) the lower part of the cavity hosts the AH-B entity complementary to that of the sweet molecule, and (iii) the upper part is hydrophobic and plays an important role in the case of very active sweeteners. This is often referred to as the “Temussi model” [139, 140].

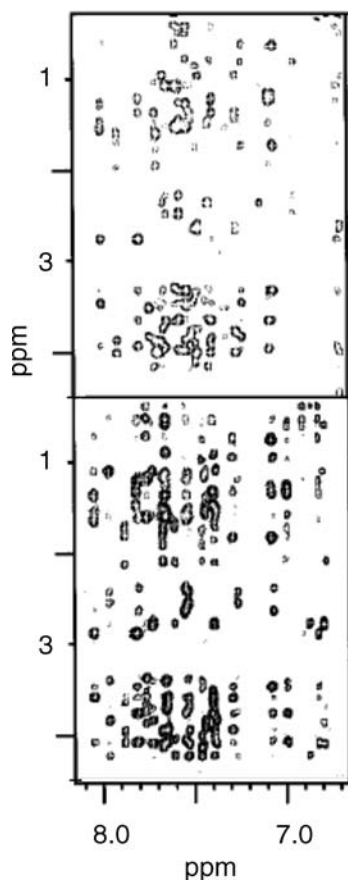
### 3.3.6.2 Opioids

The conformation of endogenous opioids has been studied extensively since the discovery of enkephalins, in the hope of finding a perfect analgesic, but these studies have proven very difficult. Structural studies of small peptides in general are hampered by their intrinsic flexibility and opioid peptides are no exception.

Ideally, a direct structural study of the complex between a peptide and its receptor should yield the bioactive conformation, but such a study is not possible since opioid receptors are large membrane proteins, difficult to study by standard structural techniques. Thus, in spite of their intrinsic limitations, conformational studies of opioid peptides are still important for drug design and also for indirect receptor mapping.

Solution studies of opioid peptides are so many that it is not possible to describe in detail all or even a substantial part of them. It is instructive to consider the study of one of them,  $\beta$ -endorphin. The N-terminal part of human  $\beta$ -endorphin, a 31mer peptide, coincides with the sequence of enkephalin.  $\beta$ -Endorphin as a whole has very little tendency to assume an ordered structure in water but its C-terminal domain (from P<sup>13</sup> to Y<sup>27</sup>) has a strong tendency to assume a helical structure in hydroalcoholic mixtures, particularly when the alcohol is a fluoroalcohol (Figure 3.10).

Therefore, the structure of  $\beta$ -endorphin represents also an interesting example of the role played by sequence, even in very strong helix-inducing media [139]. In turn, the nature and location of the secondary structure elements may reveal something of the function of each segment. In the case of  $\beta$ -endorphin, the fact that the initial 12 amino acids do not show any tendency to go helical even in a strong helix-inducing solvent like HFA/water (50 : 50 v/v) hints that the N-terminal message domain must remain very flexible to favor an induced fit interaction with the receptor active site, whereas the regular helical structure of the C-terminal domain may be well suited to interact with stable elements of secondary structure of the apolar cavity of the seven transmembrane helices receptor. Saviano *et al.* [141] hypothesized a “two-point” attachment involving an interaction of the helical part of  $\beta$ -endorphin (the address domain) with either an extracellular loop or with one or more of the transmembrane



**Figure 3.10** Comparison of the NOESY spectra of  $\beta$ -endorphin in two helix-promoting mixtures: 30:70 (v/v) TFE/water (top panel) and in 50% (v/v) HFA/water (bottom panel).

helices and the (triggering) interaction of the message domain (YGGF) with the receptor subsite common to all opioid receptors [141].

### 3.3.6.3 Transmembrane Helices

Among the many synthetic peptides derived from transmembrane helices a special case is that of  $A\beta(1-42)$ , the peptide involved in the formation of amyloid plaques of people affected by Alzheimer's disease.

Plaques found in the brains of Alzheimer's disease patients are built up by fibrils derived from the aggregation of peptides known as  $\beta$ -amyloid ( $A\beta$ ), which have sequences ranging from 39 to 43 residues, the major form found in plaques being  $A\beta(1-42)$ .  $A\beta$  peptides originate from cleavage of a common precursor called amyloid precursor protein [142], a glycoprotein made of three parts: the extracellular N-terminal region, a single hydrophobic transmembrane region, and the cytoplasmic C-terminal domain. The N-terminal region of  $A\beta(1-42)$  derives from the extracellular

domain of the precursor, whereas its C-terminal region derives from the membrane-spanning domain [141]. Several NMR studies both on A $\beta$ (1–40) and A $\beta$ (1–42) in different solvents mimicking the interface between aqueous and apolar phases have been reported. Notably, in SDS micelles [144, 145] and in helix-promoting solvents such as TFE/water mixtures [146].

These studies proved the presence of two helical regions, connected by a more flexible and disordered link; however, there was no consensus on the length and position of the helical stretches, nor on the structural features of the link region.

Crescenzi *et al.* [147] studied the structure of A $\beta$ -P(1–42) in several media that can create apolar microenvironments mimicking the lipid phase of membranes. The most detailed structure was obtained using aqueous mixtures of a fluorinated alcohol, hexafluoroisopropanol. This structure is boomerang-shaped, with its second helix (residues 28–38) corresponding to the transmembrane region of amyloid precursor protein, in very good agreement with a theoretical model proposed for membrane-bound A $\beta$ (1–40) [148]. A further point of interest of this structure is its similarity with the structure of the fusion domain of influenza hemagglutinin (HA<sub>fd</sub>), determined in detergent micelles [149].

#### 3.3.6.4 Cyclopeptides

Cyclization is one of the most used approaches to stabilize the structure of flexible bioactive peptides and turn them into useful drugs. This approach is well represented in nature where it is customary to find small cyclic peptides with powerful biological activity. Prominent among naturally occurring cyclic peptides are small, disulfide-rich peptide toxins found in venomous animals such as spiders, scorpions, and mollusks [150–152]. The structure of these peptides is so stabilized by the presence of two or more disulfide bridges that their NMR parameters can be treated with standard NMR methods used to determine protein structure (see Section 6.4).

A very interesting naturally occurring cyclic peptide that, in spite of cyclization retains great conformational flexibility is cyclolinopeptide. Cyclolinopeptide A, a cyclic nonapeptide of sequence cyclo(Pro-Pro-Phe-Phe-Leu-Ile-Ile-Leu-Val) present in linseed, was one of the first peptides isolated from natural sources [153]. Soon after its synthesis it was the object of several structural studies, motivated in part by the hope of a rapid solution determination, favored by the cyclic nature and by the presence of two Pro residues that ought to reduce the accessible conformational space even further [154–156].

Unfortunately, the peptide proved as flexible as linear peptides. The solution structure came many years later thanks to a unique feature of this peptide with respect to most natural peptides: it is soluble in apolar solvents. A study in chloroform over a very wide temperature range showed that raising the temperature, with respect to room temperature, leads only to a sharpening of the peaks resulting from an average of the conformational ensemble, whereas cooling the sample to very low temperature leads to the appearance of several new peaks. At 214 K it is possible to analyze the NMR data in terms of a single conformer whose parameters coincide with those of the solid-state structure [157].

### 3.4 Proteins

#### 3.4.1

##### **An Alternative to or a Validation of Diffractometric Methods?**

NMR is considered a valid alternative to diffraction methods for protein structure determination, even if it is limited to systems of small size. In the early days of NMR protein structure determination, when many crystallographers were skeptical about the validity of the new method, the coincidence of NMR and crystallographic results was considered a *validation* of the NMR method. Gerhard Wagner [158] has aptly commented that the substantial agreement between NMR and X-ray structures of a given protein should not be viewed as a validation of NMR, but rather as a validation of X-ray methods because in solution there are fewer interactions among molecules of the same species than in the solid state.

In the 1960s, the application of NMR to structural problems of simple organic molecules was already flourishing, but applications to systems of biological interest were very rare, mainly because the simple mono-dimensional techniques available at the time were inadequate to study molecules containing hundreds of protons with closely spaced resonances and also because of the intrinsic poor sensitivity of the technique. Two major innovations of NMR spectroscopy, introduced mainly by Richard Ernst, changed the field completely, allowing direct structural studies of proteins. The first innovation was the use of Fourier transformation NMR coupled to pulse spectrometers. This new way to record NMR spectra allowed, among other advantages, accumulations of several experiments and thus a great gain in sensitivity. The second decisive innovation was the introduction of two-dimensional NMR spectroscopy that, among other advantages, led to a dramatic increase in resolution.

The first protein structures were published in the mid-1980s, mainly from the laboratory of Kurt Wüthrich. They were all very small proteins, typically of molecular weight less than 10 kDa. The size limit was pushed up to 40 kDa when multidimensional heteronuclear NMR methods were developed and nowadays further technical improvements, notably transverse relaxation-optimized spectroscopy (TROSY) and RDC, coupled with sophisticated isotope enrichment schemes have pushed it even further.

However, it is clear that size is an intrinsic limitation for NMR structure determinations, but also that NMR can allow many investigations that are difficult or altogether not allowed by diffraction methods.

#### 3.4.2

##### **Protein Spectra**

Compared to those originating from peptides, spectra from proteins obviously look much more crowded because, owing to the larger molecular size, resonances

originating from many hundreds of nuclear spins overlap to such an extent that one-dimensional spectra are utterly uninterpretable. Moreover, the slow diffusional motion affects the spin relaxation significantly, leading to broader lines with respect to those observed in the spectra of smaller molecules, thus increasing even further the problem of signal overlapping. On the other hand, protein spectra benefit from a larger dispersion of resonances with respect to the crowding around random coil values, typical of small peptides, stemming naturally from the presence of secondary and tertiary structures. Secondary structures, notably  $\beta$  structures, cause a large spread of some resonances, particularly the diagnostically crucial NH resonances. Folding into tertiary structure causes some peptide segments to be excluded from contact with the solvent and favors the relocation in space of other peptide segments in such a way that their nuclei experience many different chemical microenvironments. The most common example of the latter is the “ring current shift” experienced by nuclei located, due to folding, close to aromatic rings. Protons lying above the ring plane experience a tiny paramagnetic field that shifts the signal upfield [159]. The very presence of ring current shifted peaks in a one-dimensional NMR spectrum of a protein can be taken as a clear indication of correct folding. Having a large dispersion of chemical shifts induced by local conformations is of great help for sequence specific identification of the most abundant residues, because their resonances, though very similar, are differentially shifted apart.

In addition, the spatial folding of proteins leads to other peculiarities of NMR signals: due to a “protection” effect from the structure, the exchange rate of labile protons can be several orders of magnitude slower than the intrinsic exchange rates, making the corresponding resonance observable in NMR spectra under conditions where they could not be seen in small compounds.

### 3.4.3

#### **Wüthrich's Protocol**

The determination of protein structure by NMR is essentially based on a protocol introduced by Kurt Wüthrich [160]. It can be summarized as follows.

- i) Sample preparation, including isotope labeling.
- ii) Recording one-, two-, three-, and (occasionally) four-dimensional NMR spectra.
- iii) Sequence-specific resonance assignments.
- iv) Collection of the main conformational constraints:
  - a) chemical shifts (secondary structure information)
  - b) NOEs (providing proton–proton distances within 5 Å)
  - c) RDCs (providing long-range distance information).
- iv) Model building.
- v) Structure refinement.

#### 3.4.3.1 Sample Preparation

Typical samples are generally made of about 0.3–0.5 ml of protein solution in the concentration range from 0.1 to 1 mM. Although, occasionally, some proteins are studied as extracted by traditional biochemical methods (i.e., with all their nuclei at natural abundance), the great majority of proteins studied for structure determination are overexpressed in bacteria or other organisms using recombinant DNA techniques and are routinely enriched with  $^{15}\text{N}$ ,  $^{13}\text{C}$ , and/or  $^2\text{H}$ . Both  $^{15}\text{N}$  and  $^{13}\text{C}$  have spin 1/2, thus making it possible to do more advanced experiments in which these nuclei are manipulated. After purification, the protein is dissolved in a medium as close as possible to the native environment and inserted in the NMR spectrometer inside a thin walled glass tube with a diameter of 5–3 mm.

#### 3.4.3.2 Recording NMR Spectra

As already mentioned, one of the main difficulties in NMR studies of proteins comes from the superposition of signals. In principle each nucleus “sees” a distinct chemical environment and thus has a distinct chemical shift, but the active nuclei of proteins can be several thousand and their one-dimensional spectra look like inextricable overlaps of thousands of resonances. The difficulty was overcome, in most practical cases, by the introduction of multidimensional experiments, which correlate the frequencies of distinct nuclei. Multinuclear NMR experiments used in the determination of protein structures belong broadly to two categories – one in which magnetization is transferred through chemical bonds, giving information on the identity of nuclei (chemical shifts) and one where the transfer is through space, giving information on conformational constraints.

Apart from the one-dimensional experiment, in the case of isotope-labeled proteins, a quick check of the “health” of the protein sample can be made with a two-dimensional heteronuclear single quantum correlation spectrum (HSQC). Commonly, the heteronucleus is  $^{15}\text{N}$  and thus in a HSQC spectrum it is possible to detect a cross-peak for each proton bound to a nitrogen. The  $^{15}\text{N}$ -HSQC is regarded as the fingerprint of a protein: it allows evaluation of whether the number of peaks corresponds to the number of residues and thus identify possible problems due to multiple conformations or heterogeneity. Recent technical advances [161] allow the recording of an HSQC of a protein in a few seconds, greatly helping to determine the feasibility of subsequent longer and more elaborate experiments.

#### 3.4.3.3 Sequential Assignment

A necessary, albeit not sufficient, step in protein structure determination by NMR is the sequential assignment of most if not all resonances (i.e., associating a chemical shift to each atom, for the backbone, and, possibly, for all side-chains). When only unlabeled proteins were available, the procedure was painfully long and tedious. It was necessary to use sequentially and recursively the information of correlation spectroscopy (basically different types of COSY and TOCSY) and NOE spectroscopy (nuclear Overhauser effect spectroscopy NOESY).

COSY and TOCSY experiments transfer magnetization through chemical bonds (scalar coupling) between protons separated by three or less covalent bonds.

Thus, in a homonuclear correlation spectroscopy, an  $\alpha$ -proton transfers magnetization to the  $\beta$ -protons and so forth for all protons connected by a continuous chain of protons, as it is the case for amino acid side-chains. As already mentioned in Section 3.2, the peptide linkage keeps protons of chained residues more than three bonds apart, thus in  $^1\text{H}$  scalar spectroscopy protons of a given residue show only their own spin system. Eventually, these experiments allow the complete determination of the spin systems corresponding to residue types. NOESY experiments, on the other hand, transfer magnetization through space, showing cross-peaks for all protons that are sufficiently close in space. Since neighboring residues are necessarily close in space, it is possible to assign NOESY cross-peaks to pairs of protons belonging to different spin systems and thus connect spin systems in a sequential order, identifying peptide segments. When unique sequence segments are encountered, an unambiguous sequence specific assignment is obtained. Such stretches of assigned spin systems are then extended at both ends by iteration of many steps of TOCSY and NOESY connectivity analysis, aiming to cover the full protein sequence with specific assignment.

Severe overlap between peaks limited the application of this homonuclear procedure to very small proteins. Use of heteronuclear scalar coupling has improved the assignment process because transferring magnetization directly across peptide bonds allows an easier observation of the connection of sequential spin systems with respect to sequential NOE. Labeling a protein with  $^{15}\text{N}$ , but particularly with both  $^{13}\text{C}$  and  $^{15}\text{N}$  facilitates sequential assignment to the extent that it can be considered semiautomatic. The most obvious experiments, consisting basically of  $^{15}\text{N}$ -HSQC planes expanded along a carbon dimension, are called HNCOC, HNCACO, HNCA, HNCOCA, HNCACB, and CBCACONH.

HNCOC and HNCA experiments have very similar pulse sequences, if one only exchanges  $\alpha$ -carbon pulses for carbonyl pulses. In the HNCOC spectrum it is possible to assign the carbonyl carbon shifts that correspond to each HSQC peak and to the one previous to that one. The HNCA and HNCOCA work similarly; it is only necessary to exchange the  $\alpha$ -carbons for the carbonyls. The HNCACB and the CBCACONH contain both the  $\alpha$ -carbon and the  $\beta$ -carbon. This procedure is far less ambiguous than the old method based on COSY–NOESY connections. When the sequential assignment is completed it is also possible to assign the side-chains resonances using HCCH-TOCSY – an experiment similar to TOCSY resolved in an additional carbon dimension.

#### 3.4.3.4 Conformational Constraints

**3.4.3.4.1 Chemical Shifts** Since each chemical shift is uniquely determined by its environment, both steric and electronic, it has always been tempting to use the chemical shift data as the main or one of the main sources for structure determination. With the accumulation of structural data some methods have met considerable success, particularly for the prediction of secondary structure. Several authors have shown convincingly that the deviations of  $^{13}\text{C}$  chemical shifts of  $\alpha$ -carbons and, partially, even  $\beta$ -carbons, correlate well with  $\alpha$ -helix or  $\beta$ -sheet conformations, the



basic building blocks of all protein folds [17, 19]. Recently the group of Ad Bax has proposed a more ambitious use of backbone chemical shifts. The program TALOS, using a database containing  $^{13}\text{C } \alpha$ ,  $^{13}\text{C } \beta$ ,  $^{13}\text{C}'$ ,  $^1\text{H } \alpha$ , and  $^{15}\text{N}$  chemical shifts for a large number of high-resolution X-ray crystal structures of proteins, tries to predict the most likely dihedral angles from backbone chemical shifts [162].

The database is searched for triplets of adjacent residues with secondary chemical shifts and sequence similarity providing the best match to the query triplet of interest. TALOS yields the 10 triplets that have the closest chemical shift and sequence similarity to those of the query sequence. If the central residues in these 10 triplets show similar backbone angles, it is possible to use their averages as angular restraints for the protein whose structure is being studied. TALOS, in general, can predict the backbone angles for about 70% of the residues

**3.4.3.4.2 NOEs** The central constraints in NMR protein structure determination are those derived from NOEs. NOEs provide distance information between pairs of hydrogen atoms separated by less than 6 Å. The intensity of cross-peaks in NOESY experiments can be converted to a maximum distance between the nuclei, because the intensity of the peak is proportional to the distance to the minus sixth power. This relationship is not exact, so usually the distance information is grouped into three ranges: 1.8–2.5 (strong), 1.8–3.5 (medium), and 1.8–5.0 Å (weak). The lower bound does not come from the relationship, but simply from the knowledge of the van der Waals repulsion range. Although the distances are not very precise, they constitute powerful constraints when used in the appropriate model building procedures. Short-range NOEs are most valuable to define secondary structure elements, whereas the long-range NOEs give crucial information on the tertiary structure [160].

**3.4.3.4.3 RDCs** If the protein molecules in solution are partially aligned, spatially anisotropic dipolar couplings are no longer completely averaged and thus it is possible to observe a RDC between pairs of spins. The original method of partial alignment proposed by Bax and Tjandra [163] for structure refinement employed a dilute liquid crystalline phase made by a mixture of dihexanoyl phosphatidylcholine and dimyristoyl phosphatidylcholine that in water forms disk-shaped particles, often referred to as bicelles. Other popular media are made by addition of bacteriophages and stretched polyacrylamide gels. Addition of magnetically aligned Pf1 filamentous bacteriophage as a cosolute was introduced by Hansen *et al.* [164] who described the technique as allowing alignment of macromolecules over a wide range of temperature and solution conditions.

Strain-induced alignment in a gel had been extensively used to study the properties of polymer gels but was proposed also for RDCs by Sass *et al.* [165] and by Tycko *et al.* [166]. The advantages of gels are that they allow the unrestricted scaling of alignment over a wide range and can be used for aqueous as well as organic solvents, a feature that turned out to be useful for peptides [167]. RDCs complement NOEs because they give long range information. RDCs can be considered orientational restraints; accordingly, they give information about the relative orientation of parts of

the molecule, that can even be far apart in the structure. There are unique advantages in the use of RDCs: for instance, they are ideal for the rapid determination of the relative orientations of units of known structures in proteins [168] and they can be detected even in large molecules for which it is often difficult to record NOEs due to spin diffusion.

When a complete set of RDCs, involving  $^1\text{H}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$  nuclei, is available, it is even possible to determine molecular structures without recourse to NOE restraints. However, in general, it is not practical to measure nearly all possible RDC; it is preferable to use them to refine structures determined on the basis of conventional constraints. As a rule of thumb, for an accurate refinement of a protein structure, it is necessary to collect a number of RDCs double the number of residues in the protein [168].

#### 3.4.3.5 Model Building

Historically, distance geometry and RMD can be considered the two most common approaches that NMR spectroscopists have used to generate structures of proteins. Distance geometry methods can be roughly grouped in two categories: those using the matrix algorithm [169, 170] and those based on the variable target function approach [171]. Starting from incomplete sets of distance constraints, distance geometry methods attempt to determine consistent ensembles of three-dimensional structures. It is not surprising that in the early days of NMR structural determination the quality of the ensembles was low. The main reason being that the constraints were incomplete since it was not possible to determine all interproton distances from two-dimensional NOEs and also because the distance constraints are not very precise.

The most popular alternative to distance geometry is RMD, based on force fields taking into account pseudoenergy terms from NMR-derived restraints [172]. RMD programs attempt to drive the structure toward a conformation that will minimize the violation of the restraints during an annealing cycle. Generally, RMD methods, during the dynamical simulated annealing, use a simplified force field in which bond length, bond angle and repulsive van der Waals terms are retained [173].

Occasionally, it is possible to adopt a hybrid method [174]. Initial structures are generated by distance geometry; then the resulting set of conformers are refined using RMD.

#### 3.4.4

##### Recent Developments

As the molecular size of a protein increases, the overlap of signals in the NMR spectra becomes more and more severe. A drastic, albeit costly, way to overcome this problem is to reduce the number of observable resonance lines by selective labeling of only parts of the sequence. This can be achieved by a proper choice of isotope segmental labeling schemes [175], including chemical ligation [176–178] and intein-based methods [179–181]. However, the increase in transverse relaxation due to the increase of the molecular volume cannot be circumvented by these approaches and the huge line broadening observed in the spectra of higher-molecular-weight proteins poses a

severe limitation. Since major sources of relaxation are the omnipresent hydrogen atoms, their replacement by deuterons [182, 183] substantially reduces transverse relaxation, resulting in increased resolution and significant sensitivity gains. Partial deuteration, such as 70% of C–H moieties, allows for a significant reduction of transverse relaxation and, at the same time, favors sequential resonance assignments, and collection of structural and functional information from  $^1\text{H}$  dependant spectra as already described [184–186]. However, deuteration alone is not sufficient to extend the application of solution NMR above the size limit of 50 kDa. Only the introduction of an innovative NMR experiment, TROSY [187], allowed the reduction of the effects of relaxation to such an extent that satisfactory line widths and sensitivity can be achieved in NMR experiments with very large molecules. TROSY makes use of the fact that cancellation of transverse relaxation effects can be achieved for one of the four multiplet components observed for  $^{15}\text{N}$ – $^1\text{H}$  or  $^{13}\text{C}$ – $^1\text{H}$  moieties and leads to the exclusive observation of the narrow component of the multiplet. To apply the TROSY technique, at least two different interfering relaxation mechanisms must contribute to relaxation. The interference between two relaxation mechanisms can be additive or subtractive; in the latter case, the effective relaxation is reduced. In the case of a  $^{15}\text{N}$  labeling,  $^1\text{H}$  nuclei couple to  $^{15}\text{N}$  nuclei (scalar coupling), and the  $^1\text{H}$ -NMR spectrum of such an amide moiety consists of two lines representing protons attached to  $^{15}\text{N}$  nuclei with spin up and protons attached to  $^{15}\text{N}$  nuclei with spin down, relative to the externally applied magnetic field. In the spectrum of a large protein, the two lines have different line widths, which directly demonstrates the relaxation interference. In conventional NMR experiments, the two lines are collapsed by a technique called “decoupling,” but at the cost of averaging the relaxation rates, thus attenuating the signal because of the contribution of the more rapidly relaxing resonance line. The TROSY technique exclusively selects the slowly relaxing resonance line, eliminating the faster relaxing resonance. Thus, TROSY disregards half of the potential signal; in large molecules, however, this is more than compensated for by the slower relaxation during the pulse sequence and the acquisition. The two interfering relaxation mechanisms in the case of the amide proton are dipole–dipole relaxation between the proton and nitrogen spins, and the chemical shift anisotropy of the protons. The dipole–dipole interaction is independent of the static magnetic field, whereas the chemical shift anisotropy increases with larger magnetic fields. In experiments with  $^1\text{H}$  and  $^{15}\text{N}$  nuclei, the line with the slower relaxation rate for both nuclei is selected in a relaxation-optimized experiment. TROSY is not limited to amide moieties in biological macromolecules; some important applications use C–H groups in aromatic rings [188]. TROSY works best with deuterated proteins of molecular size greater than 20 kDa and is especially suited for application to protonated amide groups [189–191]. Theory predicts that the extent of the cancellation effect at the basis of the TROSY experiment is dependent on the polarizing magnetic field: at  $^1\text{H}$ -NMR frequencies in the range 900–1000 MHz it may be nearly complete [187], but TROSY yields significantly narrower spectral linewidths and improved sensitivity for observation of  $^{15}\text{N}$ – $^1\text{H}$  groups already at 750 MHz. For technical details on the TROSY sequence implementation see Pervushin [192], Venters *et al.* [193], and Wider [194], while interesting applications can be found

in McKenna *et al.* [195], Garcia-Herrero *et al.* [196], Renault *et al.* [197], and Frueh *et al.* [198].

As the size of the protein increases, obtaining well-resolved spectra is not the only problem arising, as the number of scalar peaks, NOE connectivities, and RDCs also increase, making the data volume to be analyzed very difficult and time consuming.

Several steps in the protein structure determination by NMR are exceedingly time consuming and repetitive. Accordingly, it is not surprising that, since the early days of structure determination, there have been attempts at automation of peak peaking, sequence-specific assignment, and NOE assignment, with the aid of dedicated computer software. Most algorithms mirror the previously described strategy developed by Wüthrich [4, 199]. Commonly used algorithms for automated analysis of resonance assignments generally follow the scheme: (i) register peak lists in comparable dimensions (registering/aligning), (ii) group resonances into spin systems (grouping), (iii) identify amino acid type of spin systems (typing), (iv) find and link sequential spin systems into segments (linking), and (v) map spin system segments onto the primary sequence (mapping). Programs implementing such steps are typically categorized by the algorithm implemented for the mapping step. These include simulated annealing/Monte Carlo algorithms such as MONTE [200] and PASTA [201]; genetic algorithms such as GARANT [202, 203]; and exhaustive search algorithms such as TATAPRO [204], MAPPER [205], and PACES [206]. CAMRA [207] performs a heuristic comparison based on predicted chemical shifts derived from homologous proteins, whereas IBIS [208] and AutoAssign [209] are heuristic best-first algorithms.

Following the sequence-specific assignment of spectra, the peak picking, assignment, and integration of NOESY resonances are the next step subject to automation. Such procedures are usually carried out alternately to structure determination steps to get feedback for errors in assignments deriving from overlapping, ambiguities, and so on. The most used methods use different strategies. NOAH [210] (implemented in DYANA [211]) temporarily ignores cross-peaks with too many assignment possibilities and generates independent distance constraints for each of the assignment possibilities of the remaining low-ambiguity peaks, aiming to determine an initial, although distorted, correct folding, to be refined with subsequent introduction of less trusted NOEs. ARIA [212, 213] uses ambiguous distance constraints. Auto-Structure [214] identifies iteratively self-consistent NOE contact patterns, without using a three-dimensional structure model, but delineating secondary structures; the results are then fed to the DYANA program. KNOWNOE [215] is a knowledge driven Bayesian algorithm for resolving ambiguities in NOE assignment.

#### 3.4.5

##### **Selected Structures**

It is not possible, within the boundaries of a short chapter, to quote a significant number of structures solved by NMR methods, because the protein structures

reported in the Protein Data Bank (PDB; www.pdb.org) already number more than 9000. We arbitrarily chose only two recent structures: one is that of a small protein whose structure was solved employing typical methods, but with the additional interesting feature of direct carbon detection, whereas the other one emphasizes the possibility to reach the size limits of NMR.

#### 3.4.5.1 Superoxide Dismutases

Superoxide dismutases (SODs) are a class of enzymes that catalyze the dismutation of superoxide into oxygen and hydrogen peroxide. The monomers have a molecular weight of about 16 kDa. Accordingly, the size of this protein is well within the limits of NMR methods. This protein has always attracted the attention of researchers in many fields, as demonstrated by the more than 200 different structures deposited in the PDB at the time of writing. Most of these were determined by X-ray diffraction methods, but also a few NMR structures have been released in the last years, providing insights into the solution behavior of this metalloprotein. For instance, the study by Mori *et al.* [216] on *SodCII*-encoded monomeric Cu,Zn-SOD from *Salmonella enterica* resulted in the first solution structure of a natural and fully active monomeric SOD. This study provides novel insights into the functional differences between monomeric and dimeric bacterial Cu,Zn-SODs, in turn helping to explain the convergent evolution toward a dimeric structure in prokaryotic and eukaryotic enzymes of this class.

Sequence-specific assignment of backbone and side-chain resonances was obtained using HSQC, HNCA, HNCOC, HNCACB, CBCA(CO)NH, HBHA(CO)NH, (H)CCH-TOCSY, and  $^1\text{H}$ ,  $^{15}\text{N}$ -NOESY-HSQC. In addition to the usual set of two- and three-dimensional  $^1\text{H}$ -detected experiments, direct detected CACO, CBCACO, CON, and CC-COSY experiments were used in this case to confirm proline sequential assignments, and provided complete resonance assignment of aspartic acid/asparagine and glutamic acid/glutamine side-chains and even carbons of aromatic rings. NOEs were measured from two-dimensional NOESY, three-dimensional  $^{13}\text{C}$ -resolved NOESY, and three-dimensional  $^{15}\text{N}$ -resolved NOESY to derive distance restraints. A total of 3291 such restraints resulted after optimization.

The ratio between  $\text{H}_\text{N}(i)\text{-H}\alpha(i)$  and  $\text{H}_\text{N}(i)\text{-H}\alpha(i - 1)$  NOEs taken from HSQC-NOESY experiments was used to calculate backbone dihedral angles, to an overall 138  $\varphi$  and 131  $\psi$  values to be used as restraints.

The resulting structure (Figure 3.11) showed the characteristic  $\beta$ -barrel common to the whole enzyme family. The general shape of the protein is quite similar to that of *Escherichia coli* Cu,Zn-SOD, although some differences are observed mainly in the active site. *SodCII* presents a more rigid conformation with respect to the engineered monomeric mutants of the human Cu,Zn-SOD, even though significant disorder is still present in the loops shaping the active site.

#### 3.4.5.2 Malate Synthase G

Recent studies have demonstrated that NMR can solve structures far larger than *Sod*, as in the case of the solution global fold of the monomeric 723-residue (82 kDa) enzyme malate synthase G (MSG) from *E. coli* [217]. To achieve such a result nearly all



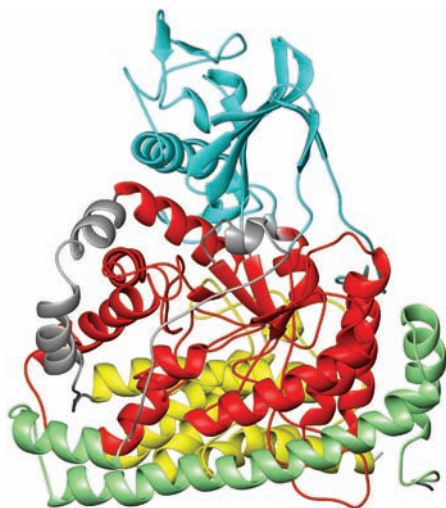
**Figure 3.11** Ribbon representation of the NMR-solved structure of the *SodCII*-encoded monomeric Cu,Zn-SOD from *S. enterica* (PDB ID: 24KW); zinc and copper ions are reported in magenta and green, respectively. (Picture made with MOLMOL [121].)

of the aforementioned techniques (i.e., selective isotope labeling, TROSY, RDCs, TALOS, and four-dimensional spectra) have been combined. It was already known from crystallographic studies that MSG is composed of four domains, including (i) a centrally located  $\beta 8/\alpha 8$  core, (ii) an N-terminal  $\alpha$ -helical domain ( $\alpha$ -helical clasp) linked to the first strand of the barrel by a long extended loop, (iii) an  $\alpha/\beta$  domain appended to the molecular core, and (iv) the C-terminal end of the enzyme consisting of a five-helix “plug” connected to the barrel by an extended loop (Figure 3.12). The core folds to form a triose phosphate isomerase (TIM) barrel arranged such that the eight strands form a parallel  $\beta$ -sheet that wraps in a cylinder surrounded by the eight  $\alpha$ -helices. The NMR structure clearly reproduces the topological features of the enzyme, including the direction of the polypeptide chain, the domain organization, the position and the orientation of the helical elements, and the locations of most of the  $\beta$ -strands.

### 3.4.5.3 Interactions

A unique, powerful feature of NMR spectroscopy is its ability to characterize protein interaction with other molecules under physiological conditions at atomic detail, even if the interactions are weak and transient. Moreover, the other molecules involved can be of any type, like other proteins, flexible peptides, small organic compounds, and nucleic acids. Thus, NMR has assumed a unique role in the investigation of protein interactions in many fields.

The most widely used approach for probing protein–ligand interfaces by NMR spectroscopy is the chemical shift perturbation (CSP) experiment, generally based on correlation  $^{15}\text{N}$ -HSQC spectra. Its utility and popularity are due to the straightforward nature of the technique and the high sensitivity of the experiment, which can be recorded in 20–30 min on a typical protein sample (0.2 mM). The addition of a (unlabeled) binding partner causes changes in the environment of the backbone



**Figure 3.12** Ribbon representation of MSG from *E. coli* as determined by solution NMR (PDB ID: 1Y8B) with color highlights of the structural domains:  $\beta 8/\alpha 8$  barrel

(core, red),  $\alpha$ -helical clasp (N-terminal, green),  $\alpha/\beta$  domain (cyan), and C-terminal domain (yellow). (Image produced with MOLMOL [121].)

amides and therefore of the chemical shifts of nuclei at the binding interface. These changes can be mapped onto the surface of the labeled protein, provided that its structure is known either from NMR or crystallography. In the latter case the interaction study can be particularly rapid because only a sequential assignment is necessary, without the need for a (time consuming) *de novo* structure determination.

Intermolecular NOEs are also a potential tool for interaction studies, although practical disadvantages (spectral overlapping, weak intensity, assignment ambiguity, dependence on tight binding) make this tool harder to use, especially for screening purposes.

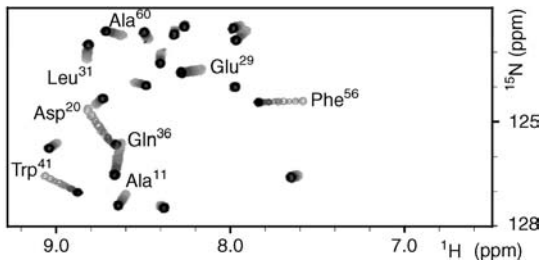
Interactions can be monitored by the use of the paramagnetic relaxation effect (PRE). The PRE effect arises from the large magnetic dipolar interaction that exists between the unpaired electron in a paramagnetic center and a nearby NMR-active nucleus, which results in an increase in the relaxation rate of the latter [218, 219]. The magnitude of the effect is proportional to  $r^{-6}$  for an electron–nucleus distance  $r$ , and, due to the large magnetic moment of the electron, the PRE effect can be observed at distances extending up to 25–35 Å, far beyond the limits of NOEs. Thus paramagnetic tagging of proteins has been proposed as a probe for interaction [220, 221]. Common paramagnetic tags are usually nitroxide moieties [222] covalently bound to the protein (i.e., cysteine-thiol linked) or coordinated lanthanide ions [223]. PRE data are typically measured as an increase in the transverse relaxation rate of signals in a  $^{15}\text{N}$ -HSQC spectrum following introduction of the paramagnetic tag on one of the binding partners. In turn, this rate increase can then be used to calculate the distance between the paramagnetic moiety and the affected nucleus of the other partner [224, 225]. A powerful tool to analyze the results of interactions is the use of appropriate docking

programs. At present the most popular software to this end is HADDOCK [226, 227], mainly because this molecular docking tool has been designed to accept NMR-style constraints deriving from CSP, NOE, RDC, or PRE. For details on structure determination of complexes by NMR see the following subsection.

**3.4.5.3.1 Complexes** A special case of interaction is that of stable complexes, once again of proteins with other biomacromolecules, but also with small compounds. Obviously, the spectrum of a protein–ligand complex is not the sum of the spectra of the free compounds, because their interaction causes involved nuclei to experience a new microenvironment, leading to a change in their chemical shifts. The appearance of the spectrum of a complex is dominated by the exchange regime: when the exchange between the free and bound state of a nucleus is fast on the NMR timescale, a single, weight-averaged resonance (both in frequency and line shape) of the two forms will be observed, while in the case of slow exchange both resonances of the free and bound forms, with weighted intensities, will be reflected by their chemical shift. Classification of fast and slow exchange regimes is based on whether the condition  $k \gg \delta_A - \delta_B$  or  $k \ll \delta_A - \delta_B$  is matched, respectively, where  $k$  is the kinetic constant and  $\delta_A$  and  $\delta_B$  are the chemical shift of the free and bound form of an interacting nucleus. In terms of binding constant, as a rule of thumb, it can be said that slow exchange is characterized by  $K_d < \mu\text{M}$  (tight binding) while fast exchange is characterized by  $K_d > \mu\text{M}$  (weak binding), but many conditions of intermediate exchange may also apply, usually involving severe broadening of peaks, thus making spectral interpretation very tough. In all cases titration of the protein with the ligand is needed for evaluation of the exchange regime and to assign resonances for the bound form, assuming that assignments for the free forms were already available. We chose, from recent literature, an example of a fast exchange regime and another of a complex in the slow exchange regime.

As an example of a protein–peptide complex and of a fast exchange regime, it is interesting to quote the binding of peptides derived from the tyrosine kinase interacting protein (Tip) of Herpesvirus saimiri to the SH3 domain of the T-cell-specific tyrosine kinase Lck (LckSH3). The activation of Lck by Tip is considered as a key event in the transformation of human T-lymphocytes during herpes viral infection. Schweimer *et al.* [228] investigated the interaction of some proline-rich Tip peptides with the Lck SH3 domain starting with the structural characterization of the unbound interaction partners by NMR and continuing with the investigation of the interaction surfaces of the binary complex. Titrations of a 0.8 mM solution of  $^{15}\text{N}$ -labeled LckSH3 with unlabeled Tip(173–185) or Tip(168–187) were carried out to a 4-fold excess of peptides while acquiring a  $^{15}\text{N}$ -HSQC spectrum at each step. Superposition of the spectra corresponding to all steps shows that some resonances shift continuously upon titration (Figure 3.13) revealing that the fast exchange regime is dominating the process. Use of closely spaced titration steps allowed for resonance shifts to be followed accurately without the need for reiteration of the assignment for the bound form. The largest changes in chemical shifts were observed for three stretches of the SH3 chain (S18–G21, Q36–W41, and F53–N57) corresponding to the RT loop, the n-src loop, and a helical turn connecting strands  $\beta_4$  and  $\beta_5$ , respectively.





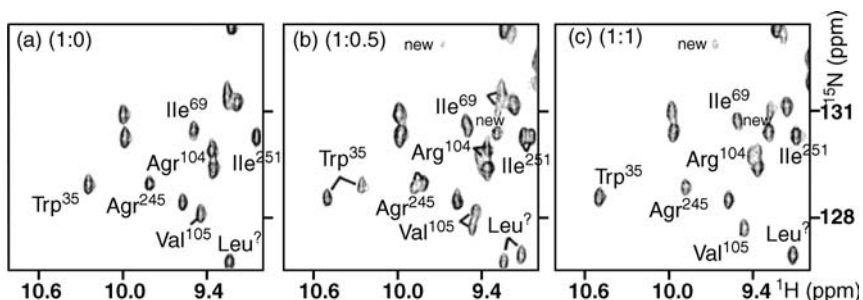
**Figure 3.13** NMR titration. Superposition of  $^{15}\text{N}$ -HSQC partial spectra of LckSH3 (light gray spots) upon gradual addition of Tip (173–185). Resonances belonging to the spectrum at the final step of the titration

(4-fold molar excess of Tip) are shown in black. Cross-peaks experiencing large change in chemical shift are labeled with the corresponding residue (three-letter code).

These findings are consistent with typical binding surfaces found in many other interactions of proline-rich peptides with SH3 domains [229].

The behavior of resonances in the slow exchange regime is well illustrated by an NMR study of the interaction between a base excision repair protein and a double-stranded DNA oligomer. It was known that the binding between *E. coli* formamido pyrimidine-DNA glycosylase (Fpg) and a double-stranded DNA oligomer containing 1,3-propanediol (13-PD) is very tight, as indicated by a  $K_D = 2.9$  nM [230]. Monitoring the titration of the protein with the oligonucleotide by  $^{15}\text{N}$ -HSQC spectra uncovers interacting residues and shows the exchange regime to be slow [231].

Looking at Figure 3.14 it is apparent that resonances of the side-chain amide of Trp<sup>35</sup> ( $\epsilon$ 1) and of the backbone amides of Ile<sup>69</sup>, Arg<sup>104</sup>, Val<sup>105</sup>, Ile<sup>251</sup>, and Arg<sup>245</sup> split into pairs when the ratio between the protein and the nucleotide is 1:0.5, because both resonances of the free and bound forms are observable. At the 1:1 ratio the signals of the free Fpg virtually disappear whereas the cross-peaks originating from the Fpg/13-PD complex dominate the spectrum. In addition, it is possible to



**Figure 3.14** Partial  $^{15}\text{N}/^1\text{H}$  HSQC spectra of perdeuterated Fpg in the presence of various amounts of 13-PD. (a) Free Fpg, (b) Fpg in the presence of approximately 0.5 equivalents of 13-PD, and (c) Fpg at a 1:1 molar ratio with 13-PD.

observe the sudden appearance of new peaks for the bound form in the slow exchange regime, making full assignment much harder.

A very interesting aspect of this study is the experimental evidence that slow and fast (or intermediate) exchange can coexist. Part of the structure of free Fpg undergoes intermediate timescale motion that is not quenched by tight DNA binding. These observations are consistent with the fact that the motion at the DNA binding surface of Fpg may be functional to the search for DNA damage and its catalytic functions.

The above examples show the extent to which the use of labeled proteins facilitate surface mapping in interactions with ligands, because the absence of the ligand resonances makes spectral interpretation really simple. On the contrary, when we wish to study the structure of a bound ligand in detail, we must be able to assign the signals of the ligand within the complex and determine structural constraints belonging to those signals (e.g., NOEs). This is better achieved by retaining resonances from the ligand only, and the following techniques are usually employed:

- i) Deuteration.
- ii) Isotope editing/filtering.
- iii) Transferred NOE (trNOE).

**Deuteration** Replacement of nonlabile protons by deuterons in a protein (perdeuteration) eliminates the signals of the receptor while leaving resonances from the ligand unaffected; standard NMR experiments can then be carried out on the bound form [232, 233]. This approach has been superseded by isotope filtering because deuteration of a recombinant protein is rather expensive with respect to  $^{15}\text{N}/^{13}\text{C}$  labeling and, even more important, it does not remove exchangeable protons on the receptor protein (i.e., amide protons) from spectra collected in water. Applications to large proteins have also been demonstrated; Hsu and Armitage [234] have resolved the structure of the complex between a potent immunosuppressor, cyclosporin A (CsA), and the ubiquitous and highly conserved 17.7 kDa immunophilin, cyclophilin (CyP). Fully deuterated CyP was produced by overexpressing the human CyP gene in *E. coli* grown on deuterated algal hydrolyzate in 98%  $\text{D}_2\text{O}$ . As only the CsA molecule is protonated in the CsA–CyP complex, it was possible to perform a complete sequential assignment of the bound drug using standard two-dimensional proton NMR experiments.

**Isotope Editing/Filtering** Labeling of the receptor (or ligand) with  $^{13}\text{C}/^{15}\text{N}$  allows the use of an “isotope filtering/editing” NMR experiment to select for signals in the spectrum from protons that are either bonded to  $^{13}\text{C}/^{15}\text{N}$  (editing) or  $^{12}\text{C}/^{14}\text{N}$  (filtering). The advantage, with respect to perdeuteration, is that, from the same sample, NOESY spectroscopy can yield three sets of signals in separate spectra, making the determination of the full structure possible (provided that the complex is in the slow exchange regime): NOEs between protons on the ligand, NOEs between protons on the protein, and NOEs between ligand protons and

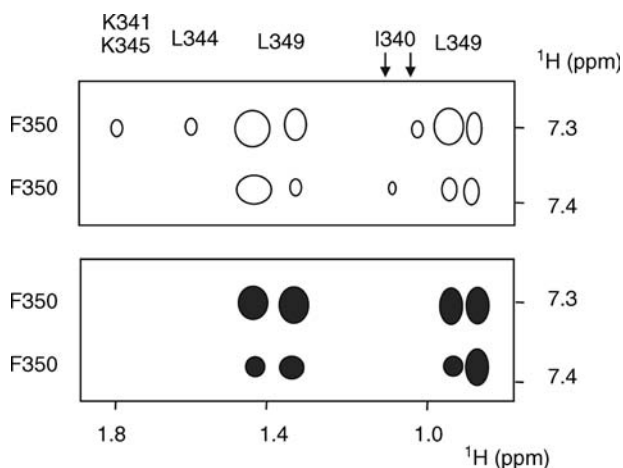
receptor protons. An example of such a strategy is found in the work of Rustandi *et al.* [235] reporting the structure of the calcium binding protein S100B $\beta\beta$  ( $^{13}\text{C}$ ,  $^{15}\text{N}$  labeled) in complex with the negative regulatory domain of p53 (unlabeled). Resonance assignment and intramolecular NOEs measurements were carried out for the S100B and of p53 peptide by use of, respectively, two-dimensional  $^{13}\text{C}$ -edited and  $^{12}\text{C}$ -filtered spectra

Then a three-dimensional  $^{13}\text{C}$ -edited/ $^{12}\text{C}$ -filtered NOESY was employed for measurement of intermolecular NOEs. The full structure of the complex was obtained revealing p53 peptide  $\alpha$ -helix occupying a pocket regulated by loop 2 (the hinge) of S100B $\beta\beta$ .

**trNOE** The trNOE effect, originally described by Balaram *et al.* [236], combines the NOE between adjacent spins in the ligand with chemical exchange between bound and free forms. While large molecules induce large negative NOEs, small molecules induce smaller positive NOEs. However, when a small ligand binds to a protein, it starts tumbling as a large complex and sign inversion of NOE signal occurs. If the ligand is in fast exchange between its bound and free form, a transfer of negative NOE may occur from the bound to the population of the free ligand molecule. Thus, in the NOESY spectrum negative signals appear for ligands that bind to the protein, while signals from the unbound form are positive or disappear. There are several advantages in the use of trNOE: signals are obtained from regular NOESY spectra; no isotope labeling is necessary; large proteins, usually giving resolution problems in NMR, are welcome because they induce strong negative NOEs; signals arising from the protein are usually not observed if it is large enough, alternatively background signals can be suppressed by a  $T_2$  or  $T_{1\rho}$ . It is apparent that this technique is well suited for cases where the binding protein is really large and spectra cannot be easily obtained, thus neither can the spectra of the protein/ligand complex. The regular NOESY spectra acquired allow for the bound form of the ligand structure to be derived and knowledge of the protein structure (e.g., from crystallography) or the knowledge of binding residues (e.g., from mutagenesis experiments) makes reconstruction of the complex possible.

The applicability to weak binding molecules, the condition for the necessary fast exchange regime to apply, make trNOE an alternative to diffraction studies of such systems, whereas the same condition makes cocrystallization not easily possible. Moreover, the reported characteristics make trNOE experiment suitable for screening of compound mixtures for binders [237–239].

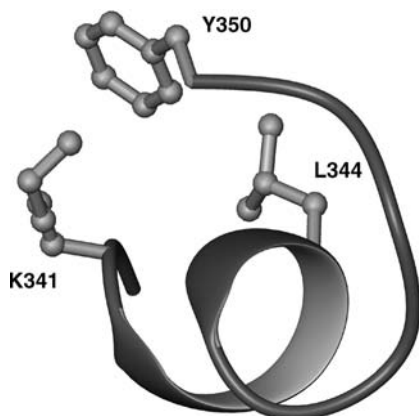
trNOE can of course be used not only for screening of compounds in the search for potential ligands of proteins, but also to examine the protein counterpart of a given binder. A particularly fascinating example is furnished by the work of Anderson *et al.* [240] who have shown how the structure of the Gta(340–350) peptide bound to photoactivated rhodopsin ( $R^*$ ) can be dependent on irradiation of the sample. The study was conducted by use of Gta(340–350) peptide derivatives and one of these (carrying a terminal carboxamide) clearly showed long-range NOEs in the presence of photoactivated rhodopsin, while only sequential NOEs are present in the case of the



**Figure 3.15** Schematic representation of selected regions from the two-dimensional NOESY spectra of Gt $\alpha$  (340–350)-carboxamide derivatized peptide in the presence of rhodopsin, showing aromatic ( $x$ -axis)–aliphatic ( $y$ -axis) NOE cross-peaks. The open cross-peaks correspond to the light-activated state (top panel); the black cross-peaks correspond to the dark-adapted state

(bottom panel). Both spectra show strong NOEs resulting from the close proximity between the aromatic ring protons of F350 and the neighboring side-chain protons of L349. However, only the light-activated state shows NOEs between F350 and residues at the other end of the peptide sequence: I340, K341/345, and L344, inferring head-to-tail conformation for the peptide.

dark-adapted receptor (Figure 3.15), inferring a head-to-tail conformation for the active, bound form of the peptide (Figure 3.16). trNOE NMR was crucial in suggesting a cation– $\pi$  interaction stabilizing the structure between the  $\epsilon$ -amine of Lys341 and the aromatic ring of the C-terminal residue, Phe350.



**Figure 3.16** The R<sup>1</sup>-bound NMR structure of Gt $\alpha$ (340–350)-carboxamide derivatized peptide represented as ribbon. Side-chains of Y350, K341, and L344 are highlighted to show their proximity giving rise to the NOEs discussed in Figure 3.15. (Image produced with MOLMOL [123].)

### 3.5 Conclusions

This chapter gives a succinct account of the state of the art in NMR studies of amino acids, peptides, and proteins. We have shown that NMR spectroscopy is an ideal analytical and structural technique for the study of these compounds, albeit with large differences among the three classes.

In the case of amino acids, there is no significant structural application, because they are well-characterized, small-molecular-weight organic compounds, with no tendency to adopt stable conformations in solution. The main application in the field of amino acids is analytical, with a special emphasis on the (noninvasive) detection of amino acids in whole samples, from body fluids to whole organisms, both unicellular and complex. Most of these studies are usually grouped under the heading of metabolomics (i.e., the systematic study of the chemical fingerprints of cellular processes). Owing to its moderate sensitivity, particularly with respect to gas chromatography and mass spectrometry, NMR spectroscopy is often considered as an ancillary technique in metabolomics. However, it must be stressed that the specific ability to recognize the constitution of metabolic components, including many that were not sought, is greatly superior in NMR with respect to other analytical methods. Thus, not only proteic amino acids, but also many other natural amino acids have been identified in the metabolome.

At the other extreme, in terms of complexity, stand the proteins, for which the main application of NMR is structural. Although, in the early days, NMR structural determinations were somewhat looked upon as makeshift by crystallographers, it can be said that now NMR structural methods can compete successfully with diffraction methods, at least for macromolecules smaller than about 50 kDa. Although the resolution of NMR structures is generally lower than that of diffraction methods, NMR has the advantage that no tiresome crystallization procedure is required and the environment can be made closer to the natural one. In addition, there are structural areas where NMR offers decisive advantages with respect to other techniques: the biggest advantage of NMR as a structural tool is the easiness with which it is possible to address interaction problems.

Applications on peptides are often indistinguishable from those on proteins, because NMR structural techniques employed to study peptides are the same as those used to solve protein structures. The main difficulty resides in the fact that order, in small flexible peptides, exists only at very short range: it is often possible to define structural relationships between adjacent residues but the persistence of these structures is very short in time. This situation is curiously reminiscent of the different information and different approaches one has learned in the study of the main aggregation states of matter: amino acids are studied as a collection of independent entities, much as gas molecules; the same amino acid residues, in peptides, behave as partially related entities (like molecules in a liquid), whereas in proteins residues are tightly correlated like molecules in the solid phase of matter.

This metaphor pertains to the relevance recently acquired by studies on the so-called “intrinsically unstructured proteins” and also because these proteins play a

relevant role in several neurodegenerative diseases. From the point of view of structural NMR studies, these proteins behave exactly like small bioactive peptides, as it is emphasized by the difficulty of finding their *structure*. Therefore, it can be hoped that the numerous ways, summarized in this chapter, to circumvent the difficulties in the NMR study of peptides, can be valuable also for the more trendy studies of Intrinsically Unstructured Proteins.

## References

- 1 Von Dreele, P.H., Brewster, A.I., Scheraga, H.A., Ferger, M.F., and Du Vigneaud, V. (1971) *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 1028–1031.
- 2 Von Dreele, P.H., Scheraga, H.A., Dyckes, D.F., Ferger, M.F., and Du Vigneaud, V. (1972) *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 3322–3326.
- 3 Bundi, A. and Wüthrich, K. (1979) *Biopolymers*, **18**, 285–297.
- 4 Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York.
- 5 Wishart, D.S., Sykes, B.D., and Richards, F.M. (1991) *Journal of Molecular Biology*, **222**, 311–333.
- 6 Szilagyi, L. and Jardetzky, O. (1989) *Journal of Magnetic Resonance*, **83**, 441–449.
- 7 Pastore, A. and Saudek, V. (1990) *Journal of Magnetic Resonance*, **90**, 165–176.
- 8 Wishart, D.S., Sykes, B.D., and Richards, F.M. (1991) *FEBS Letters*, **293**, 72–80.
- 9 Wishart, D.S., Sykes, B.D., and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.
- 10 Bundi, A., Grathwohl, C., Hochmann, J., Keller, R.M., Wagner, G., and Wüthrich, K. (1975) *Journal of Magnetic Resonance*, **18**, 191–198.
- 11 Merutka, G., Dyson, H.J., and Wright, P.E. (1995) *Journal of Biomolecular NMR*, **5**, 14–24.
- 12 Grathwohl, C. and Wüthrich, K. (1974) *Journal of Magnetic Resonance*, **13**, 217–225.
- 13 Richarz, R. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2133–2141.
- 14 Neri, D., Billeter, M., Wider, G., and Wüthrich, K. (1992) *Science*, **257**, 1559–1563.
- 15 Thanabal, V., Omecinsky, D.O., Reily, M.D., and Cody, W.L. (1994) *Journal of Biomolecular NMR*, **4**, 47–59.
- 16 Plaxco, K.W., Morton, C.J., Grimshaw, S.B., Jones, J.A., Pitkeathly, M., Campbell, I.D., and Dobson, C.M. (1997) *Journal of Biomolecular NMR*, **10**, 221–230.
- 17 Spera, S. and Bax, A. (1991) *Journal of the American Chemical Society*, **113**, 5490–5492.
- 18 Braun, D., Wider, G., and Wüthrich, K. (1994) *Journal of the American Chemical Society*, **116**, 8466–8469.
- 19 Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S., and Sykes, B.D. (1995) *Journal of Biomolecular NMR*, **5**, 67–81.
- 20 Schwarzingler, S., Kroon, G.J.A., Foss, T.R., Wright, P.E., and Dyson, H.J. (2000) *Journal of Biomolecular NMR*, **18**, 43–48.
- 21 Johnson, B.A. and Blevins, R.A. (1994) *Journal of Biomolecular NMR*, **4**, 603–614.
- 22 Dyson, H.J. and Wright, P.E. (2005) *Nature Reviews Molecular Cell Biology*, **6**, 197–208.
- 23 Chiti, F. and Dobson, C.M. (2006) *Annual Review of Biochemistry*, **75**, 333–366.
- 24 De Simone, A., Cavalli, A., Hsu, S.-T.D., Vranken, W., and Vendruscolo, M. (2009) *Journal of the American Chemical Society*, **131**, 16332–16333.
- 25 Bradbury, E.M., Cary, P., Crane-Robinson, C., Paolillo, L., Tancredi, T., and Temussi, P.A. (1971) *Journal of the American Chemical Society*, **93**, 5916–5918.
- 26 Wagner, G. and Wüthrich, K. (1979) *Journal of Molecular Biology*, **134**, 75–94.

- 27 Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., and Willmitzer, L. (2000) *Nature Biotechnology*, **18**, 1157–1161.
- 28 Lindon, J.C., Holmes, E., and Nicholson, J.K. (2004) *Progress in Nuclear Magnetic Resonance Spectroscopy*, **45**, 109–143.
- 29 van der Greef, J. and Smilde, A.K. (2005) *Journal of Chemometrics*, **19**, 376–386.
- 30 Wagner, S., Scholz, K., Donegan, M., Burton, L., Wingate, J., and Völkel, W. (2006) *Analytical Chemistry*, **78**, 1296–1305.
- 31 Pan, Z. and Raftery, D. (2007) *Analytical and Bioanalytical Chemistry*, **387**, 525–527.
- 32 Constantinou, M.A., Papakonstantinou, E., Spraul, M., Sevastiadou, S., Costalos, C., Koupparis, M.A., Shulpis, K., Tsantili-Kakoulidou, A., and Mikros, E. (2005) *Analytica Chimica Acta*, **542**, 169–177.
- 33 Engelke, U.F.H., Liebrand-van Sambeek, M.L.F., de Jong, J.G.N., Leroy, J.G., Morava, E., Smeitink, J.A.M., and Wevers, R.A. (2004) *Clinical Chemistry*, **50**, 58–66.
- 34 Wang, Y. and Jardetsky, O. (2002) *Journal of the American Chemical Society*, **124**, 14075–14084.
- 35 Bradbury, E.M., Crane-Robinson, C., Paolillo, L., and Temussi, P. (1973) *Journal of the American Chemical Society*, **95**, 1683–1684.
- 36 Ferretti, J.A. (1967) *Chemical Communications*, 1030–1032.
- 37 Markley, J.L., Meadows, D.H., and Jardetzky, O. (1967) *Journal of Molecular Biology*, **27**, 25–40.
- 38 Bradbury, E.M., Crane-Robinson, C., and Rattle, H.W.E. (1967) *Nature*, **216**, 862–864.
- 39 Ferretti, J.A. and Paolillo, L. (1969) *Biopolymers*, **7**, 155–171.
- 40 Bradbury, J.H. and Fenn, M.D. (1969) *Australian Journal of Chemistry*, **22**, 357–371.
- 41 Joubert, F.J., Lotan, N., and Scheraga, H.A. (1970) *Biochemistry*, **9**, 2197–2211.
- 42 Klotz, I.M. and Tam, J.W.O. (1971) *Journal of the American Chemical Society*, **93**, 1313–1315.
- 43 Ullman, R. (1970) *Biopolymers*, **9**, 471–487.
- 44 Paolillo, L., Tancredi, T., Temussi, P. A., Trivellone, E., Bradbury, E. M. and Crane-Robinson, C. (1972) *Journal of the Chemical Society, Chemical Communications*, 335–336.
- 45 Temussi, P.A. and Goodman, M. (1971) *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 1767–1772.
- 46 Dyson, H.J., Rance, M., Houghten, R.A., Wright, P.E., and Lerner, R.A. (1988) *Journal of Molecular Biology*, **201**, 201–217.
- 47 Wright, P.E., Dyson, H.J., and Lerner, R.A. (1988) *Biochemistry*, **27**, 7167–7175.
- 48 Baldwin, R.L. and Rose, G.D. (1999) *Trends in Biochemical Sciences*, **24**, 26–33.
- 49 Blanco, F.J., Rivas, G., and Serrano, L. (1994) *Nature Structural Biology*, **1**, 584–590.
- 50 Kortemme, T., Ramírez-Alvarado, M., and Serrano, L. (1998) *Science*, **281**, 253–256.
- 51 López, M., Lacroix, E., Ramírez-Alvarado, M., and Serrano, L. (2001) *Journal of Molecular Biology*, **312**, 229–246.
- 52 Hilario, J. and Keiderling, T.A. (2001) *Biophysical Journal*, **80**, 557a.
- 53 Kuznetsov, S.V., Hilario, J., Keiderling, T.A., and Ansari, A. (2003) *Biochemistry*, **42**, 4321–4332.
- 54 Hilario, J., Kubelka, J., and Keiderling, T.A. (2003) *Journal of the American Chemical Society*, **125**, 7562–7574.
- 55 Hughes, J., Smith, T.W., Kosterlitz, H.W., Fothergill, L.A., Morgan, B.A., and Morris, H.R. (1975) *Nature*, **258**, 577–579.
- 56 Rose, G.D., Gierasch, L.M., and Smith, J.A. (1985) *Advances in Protein Chemistry*, **37**, 1–109.
- 57 Jones, C.R., Gibbons, W.A., and Garsky, V. (1976) *Nature*, **262**, 779–782.
- 58 Roques, B.P., Garbay-Jaureguiberry, C., Oberlin, R., Anteuinis, M., and Lala, A.K. (1976) *Nature*, **262**, 778–779.

- 59 Temussi, P.A., Picone, D., Castiglione-Morelli, M.A., Motta, A., and Tancredi, T. (1989) *Biopolymers*, **28**, 91–107.
- 60 Bothner-By, A.A., Stephens, R.L., Lee, J., Warren, C.D., and Jeanloz, R.W. (1984) *Journal of the American Chemical Society*, **106**, 811–813.
- 61 Motta, A., Tancredi, T., and Temussi, P.A. (1987) *FEBS Letters*, **215**, 215–218.
- 62 Motta, A., Picone, D., Tancredi, T., and Temussi, P.A. (1987) *Journal of Magnetic Resonance*, **75**, 364–370.
- 63 Gupta, G., Sarma, M.H., Sarma, R.H., and Dhingra, M.M. (1986) *FEBS Letters*, **198**, 245–250.
- 64 Douzou, P. and Petsko, G.A. (1984) *Advances in Protein Chemistry*, **36**, 245–361.
- 65 Santoro, M.M., Liu, Y., Khan, S.M.A., Hou, L.X., and Bolen, D.W. (1992) *Biochemistry*, **31**, 5278–5283.
- 66 Matthews, S.J. and Leatherbarrow, R.J. (1993) *Journal of Biomolecular NMR*, **3**, 597–600.
- 67 Spadaccini, R., Crescenzi, O., Picone, D., Tancredi, T., and Temussi, P.A. (1998) *Journal of Peptide Science*, **5**, 306–312.
- 68 Amodeo, P., Motta, A., Picone, D., Saviano, G., Tancredi, T., and Temussi, P.A. (1991) *Journal of Magnetic Resonance*, **95**, 201–207.
- 69 Schwyzer, R. (1987) *Peptides* 86 (ed. D. Theodoropoulos), de Gruyter, Berlin, pp. 7–23.
- 70 Pollard, E.C. (1976) *The Aqueous Cytoplasm* (ed. A.D. Keith), Dekker, New York, pp. 1–22.
- 71 Chapmann, D. and Peel, W.E. (1976) *The Aqueous Cytoplasm* (ed. A.D. Keith), Dekker, New York, pp. 137–178.
- 72 Barry, P.H. and Diamond, J.M. (1984) *Physiological Reviews*, **64**, 763–872.
- 73 Fiori, S., Renner, C., Cramer, J., Pegoraro, S., and Moroder, L. (1999) *Journal of Molecular Biology*, **291**, 163–175.
- 74 Yan, C., Digate, R.J., and Guiles, R.D. (1999) *Biopolymers*, **49**, 55–70.
- 75 Picone, D., D'Ursi, A., Motta, A., Tancredi, T., and Temussi, P.A. (1990) *European Journal of Biochemistry*, **192**, 433–439.
- 76 Zetta, L., Consonni, R., De Marco, A., Longhi, R., Manera, E., and Vecchio, G. (1990) *Biopolymers*, **30**, 899–909.
- 77 Deber, C.M. and Behnam, B.A. (1984) *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 61–65.
- 78 Rinaldi, F., Lin, M., Shapiro, M.J., and Petersheim, M. (1997) *Biophysical Journal*, **73**, 3337–3348.
- 79 Tessmer, M.R., Meyer, J.-P., Hruby, V.J., and Kallick, D.A. (1997) *Journal of Medicinal Chemistry*, **40**, 2148–2155.
- 80 Tessmer, M.R. and Kallick, D.A. (1997) *Biochemistry*, **36**, 1971–1981.
- 81 Carpenter, K.A., Wilkes, B.C., Weltrowska, G., and Schiller, P.W. (1996) *European Journal of Biochemistry*, **241**, 756–764.
- 82 Kallick, D.A., Tessmer, M.R., Watts, C.R., and Li, C.-Y. (1995) *Journal of Magnetic Resonance*, **109**, 60–65.
- 83 Segawa, M., Ohno, Y., Doi, M., Ishida, T., and Iwashita, T. (1995) *International Journal of Peptide and Protein Research*, **46**, 37–46.
- 84 Schwyzer, R. (1986) *Biochemistry*, **25**, 6335–6342.
- 85 Kallick, D.A. (1993) *Journal of the American Chemical Society*, **115**, 9317–9318.
- 86 Bazzo, R., Tappin, M.J., Pastore, A., Harvey, T.S., Carver, J.A., and Campbell, I.D. (1988) *European Journal of Biochemistry*, **173**, 139–146.
- 87 Marion, D., Zaslhoff, M., and Bax, A. (1988) *FEBS Letters*, **227**, 21–26.
- 88 Barsukov, I.L., Abdulaeva, G.V., Arseniev, A.S., and Bystrov, V.F. (1990) *European Journal of Biochemistry*, **192**, 321–327.
- 89 Grabchuk, I.A., Orekhov, V.Y., and Arseniev, A.S. (1996) *Pharmaceutica Acta Helveticae*, **71**, 97–102.
- 90 Lomize, A.L., Pervushin, K.V., and Arseniev, A.S. (1992) *Journal of Biomolecular NMR*, **2**, 361–372.
- 91 Pervushin, K.V., and Arseniev, A.S. (1992) *FEBS Letters*, **308**, 190–196.
- 92 Pervushin, K.V., Arseniev, A.S., Kozhich, A.T., and Ivanov, V.T. (1991) *Journal of Biomolecular NMR*, **1**, 313–322.



- 93 Pervushin, K.V., Orekhov, V.Y., Popov, A.I., Musina, L.Y., and Arseniev, A.S. (1994) *European Journal of Biochemistry*, **219**, 571–583.
- 94 Chopra, A., Yeagle, P.L., Alderfer, J.A., and Albert, A.D. (2000) *Biochimica et Biophysica Acta*, **1463**, 1–5.
- 95 Katragadda, M., Chopra, A., Bennett, M., Alderfer, J.L., Yeagle, P.L., and Albert, A.D. (2001) *Journal of Peptide Research*, **58**, 79–89.
- 96 Arshava, B., Taran, I., Xie, H., Becker, J.M., and Naider, F. (2002) *Biopolymers*, **64**, 161–176.
- 97 Xie, H., Ding, F.-X., Schreiber, D., Eng, G., Liu, S.-f., Arshava, B., Arevalo, E., Becker, J.M., and Naider, F. (2000) *Biochemistry*, **39**, 15462–15474.
- 98 Reddy, A.P., Tallon, M.A., Becker, J.M., and Naider, F. (1994) *Biopolymers*, **34**, 679–689.
- 99 Arshava, B., Liu, S.-F., Jiang, H., Breslav, M., Becker, J.M., and Naider, F. (1998) *Biopolymers*, **46**, 343–357.
- 100 Estephan, R., Englander, J., Arshava, B., Samples, K.L., Becker, J.M., and Naider, F. (2005) *Biochemistry*, **44**, 11795–11810.
- 101 Naider, F., Arshava, B., Ding, F.-X., Arevalo, E., and Becker, J.M. (2001) *Biopolymers*, **60**, 334–350.
- 102 Naider, F., Khare, S., Arshava, B., Severino, B., Russo, J., and Becker, J.M. (2005) *Biopolymers*, **80**, 199–213.
- 103 Neumoin, A., Arshava, B., Becker, J., Zerbe, O., and Naider, F. (2007) *Biophysical Journal*, **93**, 467–482.
- 104 Valentine, K.G., Liu, S.-F., Marassi, F.M., Veglia, G., Opella, S.J., Ding, F.-X., Wang, S.-H., Arshava, B., Becker, J.M., and Naider, F. (2001) *Biopolymers*, **59**, 243–256.
- 105 Thevenin, D., Roberts, M.F., Lazarova, T., and Robinson, C.R. (2005) *Biochemistry*, **44**, 16239–16245.
- 106 Thevenin, D. and Lazarova, T. (2008) *Protein Science*, **17**, 1188–1199.
- 107 Neumoin, A., Cohen, L.S., Arshava, B., Tantry, S., Becker, J.M., Zerbe, O., and Naider, F. (2009) *Biophysical Journal*, **96**, 3187–3196.
- 108 Galanth, C., Abbassi, F., Lequin, O., Ayala-Sanmartin, J., Ladram, A., Nicolas, P., and Amiche, M. (2009) *Biochemistry*, **48**, 313–327.
- 109 Shai, Y. (1999) *Biochimica et Biophysica Acta*, **1462**, 55–70.
- 110 Tossi, A., Sandri, L., and Giangaspero, A. (2000) *Biopolymers*, **55**, 4–30.
- 111 Castiglione-Morelli, M.A., Cristinziano, P., Pepe, A., and Temussi, P.A. (2005) *Biopolymers*, **80**, 688–696.
- 112 Lequin, O., Bruston, F., Convert, O., Chassaing, G., and Nicolas, P. (2003) *Biochemistry*, **42**, 10311–10323.
- 113 Shalev, D.E., Mor, A., and Kustanovich, I. (2002) *Biochemistry*, **41**, 7312–7317.
- 114 Feder, R., Dagan, A., and Mor, A. (2000) *Journal of Biological Chemistry*, **275**, 4230–4238.
- 115 Pogozheva, I.D., Lomize, A.L., and Mosberg, H.I. (1998) *Biophysical Journal*, **75**, 612–634.
- 116 Sonnichsen, F.D., Van Eyk, J.E., Hodges, R.S., and Sykes, B.D. (1992) *Biochemistry*, **31**, 8790–8798.
- 117 Verheyden, P., De Wolf, E., Jaspers, H., and Van Binst, G. (1994) *International Journal of Peptide and Protein Research*, **44**, 401–409.
- 118 Reymond, M.T., Huo, S., Duggan, B., and Wright, P.E., and Dyson, H.J. (1997) *Biochemistry*, **36**, 5234–5244.
- 119 Rajan, R., Awasthi, S.K., Bhattacharjya, S., and Balaram, P. (1997) *Biopolymers*, **42**, 125–128.
- 120 Goumon, Y., Strub, J.-M., Moniatte, M., Nullans, G., Poteur, L., Hubert, P., Van Dorsselaer, A., Aunis, D., and Metz-Boutigue, M.-H. (1996) *European Journal of Biochemistry*, **235**, 516–525.
- 121 Kieffer, B., Dillmann, B., Lefevre, J.-F., Goumon, Y., Aunis, D., and Metz-Boutigue, M.-H. (1998) *Journal of Biological Chemistry*, **273**, 33517–33523.
- 122 Fossati, G., Izzo, G., Rizzi, E., Gancia, E., Modena, D., Moras, M.L., Niccolai, N., Giannozzi, E., Spiga, O., Bono, L., Marone, P., Leone, E., Mangili, F., Harding, S., Errington N., Walters, C., Henderson, B., Roberts, M.M., Coates, A. R., Casetta, B. and Mascagni, P. (2003) *Journal of Bacteriology*, **185**, 4256–67.
- 123 Koradi, R., Billeter, M., and Wüthrich, K. (1996) *Journal of Molecular Graphics*, **14**, 51–55.

- 124 Crescenzi, O., Fraternali, F., Picone, D., Tancredi, T., Balboni, G., Guerrini, R., Lazarus, L.H., Salvadori, S., and Temussi, P.A. (1997) *European Journal of Biochemistry*, **247**, 66–73.
- 125 Temussi, P.A., Tancredi, T., Pastore, A., and Castiglione-Morelli, M.A. (1987) *Biochemistry*, **26**, 7856–7863.
- 126 Castiglione-Morelli, M.A., Lejl, F., Pastore, A., Salvadori, S., Tancredi, T., Tomatis, R., Trivellone, E., and Temussi, P.A. (1987) *Journal of Medicinal Chemistry*, **30**, 2067–2073.
- 127 Beretta, C.A., Parrilli, M., Pastore, A., Tancredi, T., and Temussi, P.A. (1984) *Biochemical and Biophysical Research Communications*, **121**, 456–462.
- 128 Mierke, D.F., Scheek, R.M., and Kessler, H. (1994) *Biopolymers*, **34**, 559–563.
- 129 Brüscheiler, R., Blackledge, M., and Ernst, R.A. (1991) *Journal of Biomolecular NMR*, **1**, 3–11.
- 130 Cicero, D.O., Barbato, G., and Bazzo, R. (1995) *Journal of the American Chemical Society*, **117**, 1027–1033.
- 131 Bonvin, A.M.J.J. and Brünger, A.T. (1996) *Journal of Biomolecular NMR*, **7**, 72–76.
- 132 Meirovitch, H., Meirovitch, E., and Lee, J. (1995) *Journal of Physical Chemistry*, **99**, 4847–4854.
- 133 Meirovitch, E. and Meirovitch, H. (1996) *Biopolymers*, **38**, 69–88.
- 134 Meirovitch, H. and Meirovitch, E. (1996) *Journal of Physical Chemistry*, **100**, 5123–5133.
- 135 Baysal, C. and Meirovitch, H. (1999) *Biopolymers*, **50**, 329–344.
- 136 Sippl, M.J., Nemethy, G., and Scheraga, H.A. (1984) *Journal of Physical Chemistry*, **88**, 6231–6233.
- 137 Mazur, R.H., Schlatter, J.M., and Goldkamp, A.H. (1969) *Journal of the American Chemical Society*, **91**, 2684–2691.
- 138 Lejl, F., Tancredi, T., Temussi, P.A., and Toniolo, C. (1976) *Journal of the American Chemical Society*, **98**, 6669–6675.
- 139 Walters, D.E., Pearlstein, R.A., and Krimmel, C.P. (1986) *Journal of Chemical Education*, **63**, 869–872.
- 140 Walters, D.E. (1995) *Journal of Chemical Education*, **72**, 680–683.
- 141 Saviano, G., Crescenzi, O., Picone, D., Temussi, P., and Tancredi, T. (1999) *Journal of Peptide Science*, **5**, 410–422.
- 142 Selkoe, D.J. (1994) *Annual Review of Neuroscience*, **17**, 489–517.
- 143 Lichtenthaler, S.F., Wang, R., Grimm, H., Uljon, S.N., Masters, C.L., and Beyreuther, K. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 3053–3058.
- 144 Coles, M., Bicknell, W., Watson, A.A., Fairlie, D.P., and Craik, D.J. (1998) *Biochemistry*, **37**, 11064–11077.
- 145 Shao, H., Jao, S.-c., Ma, K., and Zagorski, M.G. (1999) *Journal of Molecular Biology*, **285**, 755–773.
- 146 Sticht, H., Bayer, P., Willbold, D., Dames, S., Hilbich, C., Beyreuther, K., Frank, R.W., and Rösch, P. (1995) *European Journal of Biochemistry*, **233**, 293–298.
- 147 Crescenzi, O., Tomaselli, S., Guerrini, R., Salvadori, S., D’Ursi, A., Temussi, P. A. and Picone, D. (2002) *European Journal of Biochemistry*, **269**, 5642–5648.
- 148 Durell, S.R., Guy, H.R., Arispe, N., Rojas, E., and Pollard, B.H. (1994) *Biophysical Journal*, **67**, 2137–2145.
- 149 Han, X., Bushweller, J.H., Cafiso, D.S., and Tamm, L.K. (2001) *Nature Structural Biology*, **8**, 715–720.
- 150 Craik, D.J., Clark, R.J., and Daly, N.L. (2007) *Expert Opinion on Investigational Drugs*, **16**, 595–604.
- 151 Livett, B.G., Gayler, K.R., and Khalil, Z. (2004) *Current Medicinal Chemistry*, **11**, 1715–1723.
- 152 Olivera, B.M. (2006) *Journal of Biological Chemistry*, **281**, 31173–31177.
- 153 Kaufmann, H.P. and Tobschirbel, A. (1959) *Chemische Berichte*, **92**, 2805–2809.
- 154 Naider, F., Benedetti, E., and Goodman, M. (1971) *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 1195–1198.
- 155 Brewster, A.I. and Bovey, F.A. (1971) *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 1199–1202.
- 156 Tonelli, A.E. (1971) *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 1203–1207.

- 157 Di Blasio, B., Rossi, F., Benedetti, E., Pavone, V., Pedone, C., Temussi, P.A., Zanotti, G., and Tancredi, T. (1989) *Journal of the American Chemical Society*, **111**, 9089–9098.
- 158 Wagner, G. (1997) *Nature Structural Biology*, **4** (Suppl.), 841–844.
- 159 Williamson, M.P. and Asakura, T. (1997) *Methods in Molecular Biology*, **60**, 53–69.
- 160 Wüthrich, K. (1990) *Journal of Biological Chemistry*, **265**, 22059–22062.
- 161 Gal, M., Kern, T., Schanda, P., Frydman, L., and Brutscher, B. (2009) *Journal of Biomolecular NMR*, **43**, 1–10.
- 162 Cornilescu, G., Delaglio, F., and Bax, A. (1999) *Journal of Biomolecular NMR*, **13**, 289–302.
- 163 Bax, A. and Tjandra, N. (1997) *Journal of Biomolecular NMR*, **10**, 289–292.
- 164 Hansen, M.R., Mueller, L., and Pardi, A. (1998) *Nature Structural Biology*, **5**, 1065–1074.
- 165 Sass, H.-J., Musco, G., Stahl, S.J., Wingfield, P.T., and Grzesiek, S. (2000) *Journal of Biomolecular NMR*, **18**, 303–309.
- 166 Tycko, R., Blanco, F.J., and Yoshitaka, I. (2000) *Journal of the American Chemical Society*, **122**, 9340–9341.
- 167 Pavone, L., Crescenzi, O., Tancredi, T., and Temussi, P.A. (2002) *FEBS Letters*, **513**, 273–276.
- 168 Bax, A. and Grishaev, A. (2005) *Current Opinion in Structural Biology*, **15**, 563–570.
- 169 Crippen, G.M. and Havel, T.F. (1988) *Distance Geometry and Molecular Conformation*, Research Studies Press, Taunton.
- 170 Havel, T.F. (1991) *Progress in Biophysics and Molecular Biology*, **56**, 43–78.
- 171 Güntert, P., Braun, W., and Wüthrich, K. (1991) *Journal of Molecular Biology*, **217**, 517–530.
- 172 Brünger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., and Warren, G.L. (1998) *Acta Crystallographica D*, **54**, 905–921.
- 173 Nilges, M., Clore, G.M., and Gronenborn, A.M. (1988) *FEBS Letters*, **239**, 129–136.
- 174 Nilges, M., Clore, G.M., and Gronenborn, A.M. (1988) *FEBS Letters*, **229**, 317–324.
- 175 Skrisovska, L., Schubert, M., and Allain, F.H. (2010) *Journal of Biomolecular NMR*, **46**, 51–65.
- 176 Xu, R., Ayers, B., Cowburn, D., and Muir, T.W. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 388–393.
- 177 Skrisovska, L. and Allain, F.H.-T. (2008) *Journal of Molecular Biology*, **375**, 151–164.
- 178 Kent, S.B.H. (2009) *Chemical Society Reviews*, **38**, 338–351.
- 179 Otomo, T., Teruya, K., Uegaki, K., Yamazaki, T., and Kyogoku, Y. (1999) *Journal of Biomolecular NMR*, **14**, 105–114.
- 180 Yagi, H., Tsujimoto, T., Yamazaki, T., Yoshida, M., and Akutsu, H. (2004) *Journal of the American Chemical Society*, **126**, 16632–16638.
- 181 Oeemig, J.S., Aranko, A.S., Djupsjöbacka, J., Heinämäki, K., and Iwai, H. (2009) *FEBS Letters*, **583**, 1451–1456.
- 182 Grzesiek, S., Anglister, J., Ren, H., and Bax, A. (1993) *Journal of the American Chemical Society*, **115**, 4369–4370.
- 183 Yamazaki, T., Lee, W., Arrowsmith, C.H., Muhandiram, D.R., and Kay, L.E. (1994) *Journal of the American Chemical Society*, **116**, 11655–11666.
- 184 LeMaster, D.M. and Richards, F.M. (1988) *Biochemistry*, **27**, 142–150.
- 185 Torchia, D.A., Sparks, S.W., and Bax, A. (1988) *Biochemistry*, **27**, 5135–5141.
- 186 Gardner, K.H. and Kay, L.E. (1998) *Annual Review of Biophysics and Biomolecular Structure*, **27**, 357–406.
- 187 Pervushin, K., Riek, R., Wider, G., and Wüthrich, K. (1997) *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 12366–12371.
- 188 Pervushin, K., Riek, R., Wider, G., and Wüthrich, K. (1998) *Journal of the*

- American Chemical Society, **120**, 6394–6400.
- 189 Salzmann, M., Pervushin, K., Wider, G., Senn, H., and Wüthrich, K. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 13585–13590.
- 190 Shan, X., Gardner, K.H., Muhandiram, D.R., Rao, N.S., Arrowsmith, C.H., and Kay, L.E. (1996) *Journal of the American Chemical Society*, **118**, 6570–6579.
- 191 Gardner, K.H., Zhang, X., Gehring, K., and Kay, L.E. (1998) *Journal of the American Chemical Society*, **120**, 11738–11748.
- 192 Pervushin, K. (2000) *Quarterly Reviews of Biophysics*, **33**, 161–197.
- 193 Venters, R.A., Thompson, R., and Cavanagh, J. (2002) *Journal of Molecular Structure*, **602**, 275–292.
- 194 Wider, G. (2002) *IEEE Transactions on Applied Superconductivity*, **12**, 740–745.
- 195 McKenna, S.A., Lindhout, D.A., Kim, I., Liu, C.W., Gelev, V.M., Wagner, G., and Puglisi, J.D. (2007) *Journal of Biological Chemistry*, **282**, 11474–11486.
- 196 Garcia-Herrero, A., Peacock, R.S., Howard, S.P., and Vogel, H.J. (2007) *Molecular Microbiology*, **66**, 872–889.
- 197 Renault, M., Saurel, O., Czaplicki, J., Demange, P., Gervais, V., Löhr, F., Réat, V., Piotto, M., and Milon, A. (2009) *Journal of Molecular Biology*, **385**, 117–130.
- 198 Frueh, D.P., Arthanari, H., Koglin, A., Walsh, C.T., and Wagner, G. (2009) *Journal of the American Chemical Society*, **131**, 12880–12881.
- 199 Wagner, G. and Wüthrich, K. (1982) *Journal of Molecular Biology*, **155**, 347–366.
- 200 Lukin, J.A., Gove, A.P., Talukdar, S.N., and Ho, C. (1997) *Journal of Biomolecular NMR*, **9**, 151–166.
- 201 Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G., and Kessler, H. (1998) *Journal of Biomolecular NMR*, **11**, 31–43.
- 202 Bartels, C., Billeter, M., Güntert, P., and Wüthrich, K. (1996) *Journal of Biomolecular NMR*, **7**, 207–213.
- 203 Bartels, C., Güntert, P., Billeter, M., and Wüthrich, K. (1997) *Journal of Computational Chemistry*, **18**, 139–149.
- 204 Atreya, H.S., Sahu, S.C., Chary, K.V.R., and Govil, G. (2000) *Journal of Biomolecular NMR*, **17**, 125–136.
- 205 Güntert, P., Salzmann, M., Braun, D., and Wüthrich, K. (2000) *Journal of Biomolecular NMR*, **18**, 129–137.
- 206 Coggins, B.E. and Zhou, P. (2003) *Journal of Biomolecular NMR*, **26**, 93–111.
- 207 Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sönnichsen, F.D., and Sykes, B.D. (1998) *Journal of Biomolecular NMR*, **12**, 395–405.
- 208 Hyberts, S.G. and Wagner, G. (2003) *Journal of Biomolecular NMR*, **26**, 335–344.
- 209 Zimmerman, D., Kulikowski, C., Wang, L., Lyons, B., and Montelione, G.T. (1994) *Journal of Biomolecular NMR*, **4**, 241–256.
- 210 Mumenthaler, C., Güntert, P., Braun, W., and Wüthrich, K. (1997) *Journal of Biomolecular NMR*, **10**, 351–362.
- 211 Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997) *Journal of Molecular Biology*, **273**, 283–298.
- 212 Linge, J.P., Habeck, M., Rieping, W., and Nilges, M. (2003) *Bioinformatics*, **19**, 315–316.
- 213 Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., and Nilges, M. (2007) *Bioinformatics*, **23**, 381–382.
- 214 Huang, Y.J., Tejero, R., Powers, R., and Montelione, G.T. (2006) *Proteins*, **62**, 587–603.
- 215 Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.-P., and Kalbitzer, H.R. (2002) *Journal of Biomolecular NMR*, **23**, 271–287.
- 216 Mori, M., Jiménez, B., Piccioli, M., Battistoni, A., and Sette, M. (2008) *Biochemistry*, **47**, 12954–12963.
- 217 Tugarinov, V., Choy, W.-Y., and Orekhov, V.Y., and Kay, L.E. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 622–627.

- 218 Bloembergen, N. and Morgan, L.O. (1961) *Journal of Chemical Physics*, **34**, 842–850.
- 219 Bernini, A., Venditti, V., Spiga, O., Ciutti, A., Prischi, F., Consonni, R., Zetta, L., Arosio, I., Fusi, P., Guagliardi, A., and Niccolai, N. (2008) *Biophysical Chemistry*, **137**, 71–75.
- 220 Battiste, J.L. and Wagner, G. (2000) *Biochemistry*, **39**, 5355–5365.
- 221 Iwahara, J., Anderson, D.E., Murphy, E.C., and Clore, G.M. (2003) *Journal of the American Chemical Society*, **125**, 6634–6635.
- 222 Kosen, P.A. (1989) *Methods in Enzymology*, **177**, 86–121.
- 223 Otting, G. (2008) *Journal of Biomolecular NMR*, **42**, 1–9.
- 224 Iwahara, J., Schwieters, C.D., and Clore, G.M. (2004) *Journal of the American Chemical Society*, **126**, 5879–5896.
- 225 Iwahara, J., Tang, C., and Clore, G.M. (2007) *Journal of Magnetic Resonance*, **184**, 185–195.
- 226 Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003) *Journal of the American Chemical Society*, **125**, 1731–1737.
- 227 de Vries, S.J., van Dijk, A.D.J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A.M.J.J. (2007) *Proteins*, **69**, 726–733.
- 228 Schweimer, K., Hoffmann, S., Bauer, F., Friedrich, U., Kardinal, C., Feller, S.M., Biesinger, B., and Sticht, H. (2002) *Biochemistry*, **41**, 5120–5130.
- 229 Larson, S.M. and Davidson, A.R. (2000) *Protein Science*, **9**, 2170–2180.
- 230 Castaing, B., Fourrey, J.L., Hervouet, N., Thomas, M., Boiteux, S., and Zelwer, C. (1999) *Nucleic Acids Research*, **27**, 608–615.
- 231 Buchko, G.W., McAteer, K., Wallace, S.S., and Kennedy, M.A. (2005) *DNA Repair*, **4**, 327–339.
- 232 LeMaster, D.M. (1990) *Quarterly Reviews of Biophysics*, **23**, 133–174.
- 233 Matthews, S. (2004) *Methods in Molecular Biology*, **278**, 35–45.
- 234 Hsu, V.L. and Armitage, I.M. (1992) *Biochemistry*, **31**, 12778–12784.
- 235 Rustandi, R.R., Baldisseri, D.M., and Weber, D.J. (2000) *Nature Structural Biology*, **7**, 570–574.
- 236 Balaram, P., Bothner-By, A.A., and Breslow, E. (1973) *Biochemistry*, **12**, 4695–4704.
- 237 Casset, F., Peters, T., Etzler, M., Korchagina, E., Nifant'ev, N., Pérez, S., and Imberty, A. (1996) *European Journal of Biochemistry*, **239**, 710–719.
- 238 Pellecchia, M., Sem, D.S., and Wuthrich, K. (2002) *Nature Reviews. Drug Discovery*, **1**, 211–219.
- 239 Meyer, B. and Peters, T. (2003) *Angewandte Chemie (International Edition in English)*, **42**, 864–890.
- 240 Anderson, M.A., Ogbay, B., Arimoto, R., Sha, W., Kisselev, O.G., Cistola, D.P., and Marshall, G.R. (2006) *Journal of the American Chemical Society*, **128**, 7531–7541.



## 4

### Structure and Activity of *N*-Methylated Peptides

Raymond S. Norton

#### 4.1

##### Introduction

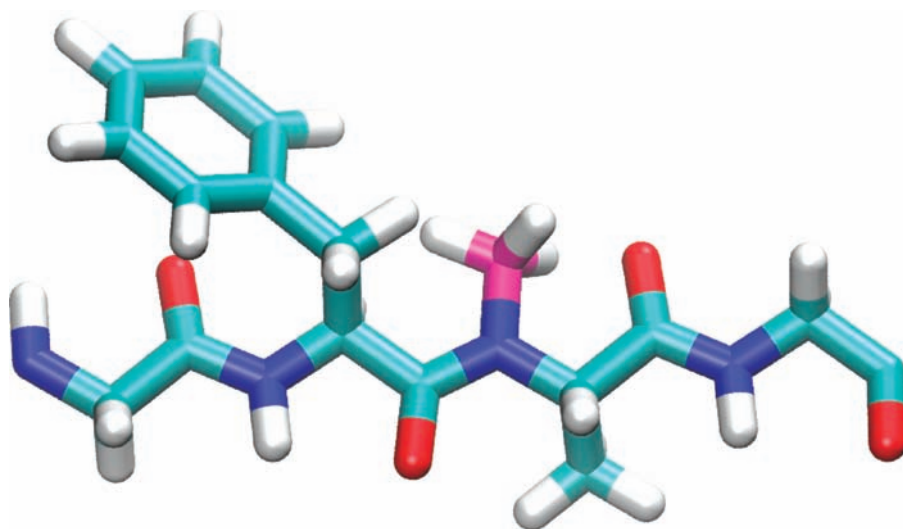
Peptides typically display high potency and target selectivity, making them valuable leads in the development of new therapeutics. Indeed, many peptides have made the transition to clinical use, including cyclosporine (cyclosporin A) [1, 2], gonadotropin-releasing hormone (also known as luteinizing-hormone releasing hormone) agonists and antagonists [3, 4], somatostatin analogs [3, 4], exenatide [5], ziconotide [6], and glatiramer acetate [7], to name just a few. Nonetheless, converting lead peptides to drugs represents a considerable challenge. Many peptides lack oral bioavailability as a consequence of their susceptibility to proteolysis in the gut, inefficient transport across the intestinal wall, proteolytic degradation in the bloodstream, and rapid clearance by the kidney.

Experience with the 35-residue polypeptide ShK toxin highlights the problem of renal clearance. This peptide, and analogs thereof [8], are potent immunosuppressants [9] that are of interest as therapeutic leads for the treatment of multiple sclerosis and other autoimmune diseases. One analog of ShK composed entirely of *D*-amino acids possessed a structure essentially identical to that of ShK [10], but was resistant to proteolysis [11]. This analog blocked the target potassium channel with nanomolar affinity and inhibited human T cell proliferation. Its immunogenicity was not tested, but it is reasonable to assume that if it could not be processed it is unlikely to be displayed by antigen-presenting cells. Despite these favorable attributes, the circulating half-life of *D*-allo-ShK was only slightly longer than that of ShK, implying that renal clearance was the major determinant of its plasma level. One potential strategy to circumvent this problem would be to encapsulate the peptide in a slow-release formulation that provides both predictable rates of release into the bloodstream [12, 13], and protection from peptidases and proteases. Another approach to prolong plasma half-life would be to couple the peptide to poly(ethylene glycol) (PEG) [14–16] or other partners [17, 18].

The first peptide therapeutic designed to target voltage-gated calcium channels (or indeed any ion channel), Prialt® (ziconotide) [19], illustrates several of the

challenges facing peptide therapeutics. This 25-residue peptide, approved for severe chronic pain [20, 21], is a synthetic version of  $\omega$ -conotoxin MVIIA. Prialt is delivered intrathecally via continuous delivery from a surgically implanted pump or from an external microinfusion device and catheter [6, 22]. However, it has a half-life in cerebrospinal fluid (CSF) of only 5 h and must be administered continually since the CSF replenishes at over triple its total volume each day. Continuous dosing requires implantation of a delivery system, which must be titrated by a physician in a hospital setting to obtain the correct dose and drug delivery rate. In addition, excess ziconotide in the bloodstream could reduce blood pressure through inhibition of calcium channels in sympathetic neurons [23]. Extending the half-life of ziconotide could allow administration by single injection, thereby eliminating the need for surgery to implant a pump for continuous infusion and facilitating dose titration, thus making the drug available to more patients. It is unlikely that glycosylation or attachment of PEG groups would extend the half-life of ziconotide in the CSF given the relatively rapid turnover of fluid from the CSF. A potential improvement would be to restrict the peptide to the spinal cord and exclude it from the brain, thereby possibly eliminating central nervous system (CNS) side-effects [19].

Several strategies have been developed to improve the efficacy of therapeutic peptides, one of which is methylation of backbone amides (*N*-methylation) (Figure 4.1). Introduction of a backbone *N*-Me group has been shown to substantially improve a number of pharmacokinetically useful parameters, including membrane



**Figure 4.1** Schematic of a tetrapeptide unit consisting of Gly-Phe-(*N*-Me)Ala-Gly, with all peptide bonds in the *trans* configuration. Standard geometries were used in constructing this model, which has not been energy minimized. The *N*-Me group is highlighted in magenta.



permeability and proteolytic stability. Moreover, *N*-methylation results in a loss of hydrogen bonding potential at the affected site, reducing the role of main-chain hydrogen bonds at a binding interface and potentially altering binding properties [24]. Structurally, this modification largely restricts the affected residue and the amino acid preceding it to an extended conformation, as discussed in detail below. Modification of several biologically active peptides by the inclusion of *N*-Me amino acids into a sequence has been shown to enhance potency [25, 26], change receptor subtype selectivity [27, 28], and protect the peptide from proteolytic degradation [29]. In the pentapeptide ipamorelin, a highly potent and selective growth hormone-releasing peptide, several truncated and *N*-methylated analogs exhibited 10–20% oral bioavailability in animals [30]. This approach therefore offers the potential to overcome several potential limitations of peptides as therapeutics. *N*-Me substitutions have also proven useful in modulating the potency or selectivity of peptide ligands in the course of structure–function analyses.

## 4.2

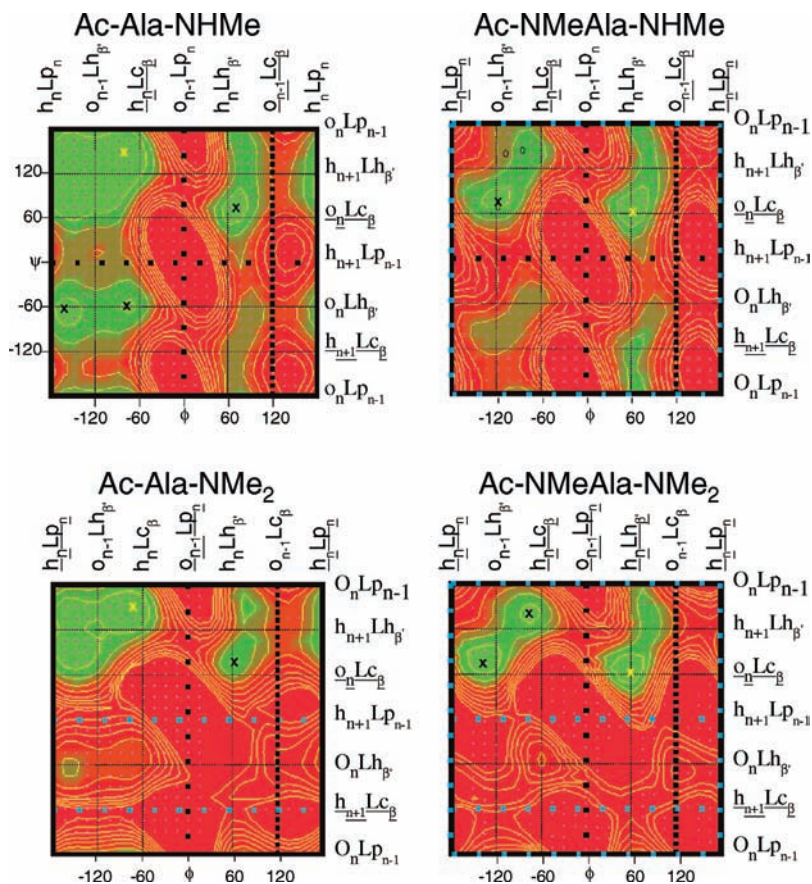
### Conformational Effects of *N*-Methylation

Manavalan and Momany [31] undertook empirical conformational energy calculations for *N*-Ac-*N*-Me-*N'*-Me-L-Ala-amide in both its *cis* and *trans* configurations. Introduction of the *N*-Me group in the *trans* configuration shifted the lowest energy position from  $\varphi -80^\circ/\psi 80^\circ$  to  $\varphi -120^\circ/\psi 70^\circ$ . The right-handed helical region  $\varphi -50^\circ/\psi -50^\circ$  was energetically forbidden, being 20 kcal/mol higher than the lowest energy state, and the second lowest energy state lay in the  $\alpha_1$  region ( $\varphi -50^\circ/\psi -50^\circ$ ), which is about 1 kcal/mol higher than the lowest energy position. The  $\varphi/\psi$  values obtained from the crystal structures of *N*-Me derivatives all fell within the 10 kcal/mol contours of the Ramachandran plots. In the *cis* configuration there were only two local minima, with the lowest energy state occurring near  $\varphi -140^\circ/\psi 70^\circ$ . Comparison of energy values between the *cis* and *trans* models at their minimum-energy positions showed that the peptide with a *cis* conformation was less stable than the *trans* by 4.5 kcal/mol. This energy difference was consistent with the experimentally observed preference for a *trans* configuration in poly(*N*-Me)-Ala [32].

In *N*-Ac-*N*-Me-*N'*,*N'*-diMe-L-Ala-amide, which serves as a model of *N*-methylation at both the *i* and *i* + 1 amide nitrogens, the low-energy regions were similar to those in *N*-Ac-*N*-Me-*N'*-Me-L-Ala-amide except that the areas were reduced within the 1 and 3 kcal/mol contours. The  $\alpha_1$  state was elevated by 3 kcal/mol relative to the lowest energy position, making the region around  $\varphi -140^\circ/\psi 80^\circ$  the most probable conformation. Similar conclusions were reached when the Ala side-chain was replaced with that of Phe. Energy calculations also showed that deviations from planarity for the peptide bond were more likely for *N*-methylated peptides than their unmethylated counterparts – a finding supported by crystal structures of cyclic peptides containing *N*-Me groups [33–35], which show deviations in  $\omega$  ranging from 5 to 19°.

In summary, *N*-methylation reduces the energy difference between the *cis* and *trans* isomers, thereby increasing the probability that a *cis* peptide bond will be found at the site of *N*-methylation [36–38], and favors an extended backbone conformation. In multiply *N*-methylated peptides, steric hindrance and the reduction in hydrogen bond donors make it difficult to predict the conformation.

More recently, Tran *et al.* [39] described simulations using a newer force field of the conformational effects of *N*-methylation and other peptide modifications. Their predictions for the effects of methylating both peptide bonds in Ac-Ala-NHMe (Figure 4.2) are very similar to those of Manavalan and Momany [31] except for the steeper energy minima in the previous maps. Side-chain atoms past the  $C^\beta$  position do not restrict the backbone conformation (although in proteins  $\beta$ -chain formation favors residues such as Leu, so the  $\varphi/\psi$  maps show Leu to be even more extended than Ala), so the calculations for Ala are likely to be representative of *N*-Me-Val and *N*-Me-Leu in peptides [40].



**Figure 4.2** Conformational energy maps for Ala peptide models modified by increasing degrees of *N*-methylation. (Adapted from Figure 2 of [39].)

The 11-residue cyclic peptide cyclosporin A is an example of a multiply N-methylated peptide, with seven N-methylated residues, including one N-Me-Val and four N-Me-Leu residues [1]. The  $\varphi/\psi$  angles for the four N-Me-Leu at positions 4, 6, 9, and 10 in the recently determined structure of cyclosporin A in complex with human cyclophilin G (Protein Data Bank (PDB ID: 2WFJ) [41, 42] are  $-114^\circ/94^\circ$ ,  $-111^\circ/177^\circ$ ,  $-129^\circ/65^\circ$ , and  $-111^\circ/167^\circ$ , respectively. Thus, the  $\varphi$  values were quite tightly clustered and close to the values predicted from energy calculations, whereas the  $\psi$  values were more dispersed, but nonetheless all were in the predicted quadrant of a Ramachandran plot. The corresponding values for cyclosporin A in complex with human cyclophilin D (PDB ID: 2Z6W) are  $-109^\circ/93^\circ$ ,  $-120^\circ/170^\circ$ ,  $-130^\circ/69^\circ$ , and  $-106^\circ/169^\circ$ , respectively.

### 4.3

#### Effects of N-Methylation on Bioactive Peptides

##### 4.3.1

##### Thyrotropin-Releasing Hormone

Manavalan and Momany [31] also described empirical conformational energy calculations for the tripeptide thyrotropin-releasing hormone (pGlu-His-Pro-NH<sub>2</sub>, where pGlu indicates pyroglutamate) and its (N-Me)His2 analog. The N-methylated analog is equipotent with the native peptide in both *in vitro* and *in vivo* assays [43]. Two distinct low-energy conformers were predicted for the N-methylated peptide, separated by an energy difference of only 1.1 kcal/mol. The first conformation had negative  $\varphi_2$  and positive  $\psi_2$  values, and the second had positive  $\varphi_2$  and  $\psi_2$  values. The conformational space available to this analog was less than that available to the unmodified peptide [44]. Nuclear magnetic resonance (NMR) analyses of the (N-Me)His2 analog [43] in both aqueous and nonaqueous solvents [45] yielded  $\varphi_2 -150^\circ$  and  $\varphi_2 155^\circ$ , which are in close agreement with the values calculated for the lowest energy conformation.

##### 4.3.2

##### Cyclic Peptides

Kessler *et al.* [25] investigated the influence of N-methylation of the selective  $\alpha_v\beta_3$  antagonist cyclo(RGDfV) (where lower case denotes a D-amino acid) on biological activity. Cyclo(RGDf-(N-Me)V) was found to be more active than the unmethylated peptide, and one of the most active and selective compounds in inhibiting vitronectin binding to  $\alpha_v\beta_3$ . Its structure in aqueous solution was determined from NMR data and molecular dynamics calculations [25]. The N-methylated peptide adopts a conformation characterized by a fast equilibrium between two inverse  $\gamma$  turns at Arg1 and Asp3 and a  $\gamma$  turn at Gly2. It was proposed that the N-Me group imposed steric repulsion via the peptide bonds Asp-D-Phe and Val-Arg, and blocked hydrogen bond formation between Arg1 HN and Asp3 CO, leading to a less-kinked orientation of the RGD pharmacophore.

Subsequently, an NMR study of the effect of *N*-methylation on a series of cyclic pentapeptides, cyclo(-D-Ala-L-Ala<sub>4</sub>-) [38], showed that only seven out of the 30 compounds adopted a single conformation (on the NMR time scale of chemical shift separation), whereas the others displayed two or more conformations in slow exchange. Among those analogs that adopted a single conformation, the incorporation of an *N*-Me moiety in place of the NH group introduced relatively minor conformational changes where no steric clashes would arise. In contrast, where introduction of the *N*-Me group was not sterically allowed, the conformation changed by introduction of a *cis* peptide bond between Ala4 and Ala5. Notably, of the seven well-defined structures, six had the D-residue *N*-methylated, implying that this combination represents a potentially valuable means of specifying a preferred conformation. It was also observed that further *N*-methylation of the conformationally homogeneous peptides did not guarantee the existence of a preferred conformer on the NMR timescale, suggesting that the most promising templates for future rational design were those with either mono- or di-*N*-methylation.

#### 4.3.3

##### Somatostatin Analogs

The tetradecapeptide somatostatin exerts potent inhibitory effects on secretory processes in tissues such as pituitary, pancreas, or gastrointestinal tract, as well as acting as a neuromodulator in the CNS. These biological effects are elicited by inhibition of a series of G-protein-coupled receptors, of which five different subtypes have been characterized; these subtypes have similar affinities for the ligand but different distribution in various tissues [46]. *N*-Methylation at Lys9 of a somatostatin analog based on octreotide (sandostatin; D-Phe5-cyclo[Cys6-Phe7-D-Trp8-Lys9-Thr10-Cys11]-Thr12) enhanced potency by around 4-fold and modified receptor subtype selectivity [27]; this was somewhat surprising given that Lys9 has been considered to constitute the active center of somatostatin, but may indicate that the introduced *N*-Me group makes favorable interactions with the receptor and/or stabilizes the peptide in its bound conformation.

Previously, Veber *et al.* [26] demonstrated that the mono-*N*-methylated cyclic somatostatin analog cyclo[(*N*-Me)Ala6-Tyr7-D-Trp8-Lys9-Val10-Phe11] had 50- to 100-fold greater potency than somatostatin in the inhibition of insulin, glucagon, and growth hormone release. This hexapeptide showed good metabolic stability, but only limited oral bioavailability.

Kessler *et al.* recently undertook a complete *N*-Me scan of the cyclopeptidic somatostatin analog cyclo[Pro6-Phe7-D-Trp8-Lys9-Thr10-Phe11], known as the Veber-Hirschmann peptide [47, 48]. They synthesized and characterized 30 *N*-methylated analogs. Screening of these analogs against the human receptor subtypes *hsst*<sub>1-5</sub> showed that seven had affinities similar to that of the parent peptide (i.e., nanomolar affinity for receptor subtypes *hsst*<sub>2</sub> and *hsst*<sub>5</sub>). In all seven of these active analogs, the βII' and the βVI turns were conserved (as determined by NMR spectroscopy and molecular dynamics calculations), confirming the importance of these two turns in maintaining the peptide in its bioactive conformation. There was

also no significant conformational perturbation associated with N-methylation, even at multiple sites.

In addition, no significant degradation was observed for any of the N-methylated peptides after 7 h incubation in rat serum. An analog triply N-methylated at D-Trp8, Lys9, and Phe11 retained receptor-binding activity, was not degraded by digestive enzymes isolated from the brush border, and showed the highest intestinal permeability in an *in vitro* model. Importantly, following administration of this analog by oral gavage at a dose one order of magnitude higher than the intravenous dose (i.e., 10 versus 1 mg/kg), oral bioavailability was found to be 10%, in contrast to the complete lack of oral bioavailability for the parent peptide [47].

#### 4.3.4

##### Antimalarial Peptide

The 20-residue peptide, R1, is a potent inhibitor of malaria parasite invasion of red blood cells [49]. R1 is an important lead compound for drug development, because its ability to block parasite growth indicates that it targets a site critical for apical membrane antigen-1 (AMA1) function. However, this peptide inhibitor is only effective against a limited subset of parasite isolates and does not exhibit broad strain specificity. To address this problem, Foley *et al.* [50] sought to improve the proteolytic stability and AMA1 binding properties of R1 by systematic methylation of backbone amides. The inclusion of a single N-Me group increased AMA1 affinity (Table 4.1), bioactivity, and proteolytic stability without introducing global structural alterations (as assessed by NMR). In addition, N-methylation of multiple R1 residues further improved these properties, as summarized in Table 4.1.

In this study, the deviations of backbone NH and C<sup>α</sup>H chemical shifts from random coil values were used as a proxy for conformational changes associated with N-methylation. Plots of the differences between these shifts for each peptide and corresponding values for native R1 showed that the differences caused by N-methylation were predominantly local, mostly within two to three residues either

**Table 4.1** Equilibrium constants determined by surface plasmon resonance for the interaction of N-methylated R1 derivatives with different strains of *P. falciparum* AMA1 [50].

AMA1 variant peptide	3D7 K <sub>D</sub> (nM) <sup>a</sup>	3D7 K <sub>D</sub> (nM) <sup>b</sup>	W2mef K <sub>D</sub> (μM) <sup>b</sup>	HB3 K <sub>D</sub> (μM) <sup>b</sup>
R1	77	80	17	69
[(N-Me)Leu8]-R1	23	ND <sup>c</sup>	15	55
[(N-Me)Leu8/(N-Me)Ser14]-R1	11	ND <sup>c</sup>	6	24
[(N-Me)Val1/(N-Me)Leu8/(N-Me)Ser14]-R1	13	ND <sup>c</sup>	6	25

- K<sub>D</sub> estimated using a kinetic analysis of Biacore data, as described by Harris *et al.* [50]. All K<sub>D</sub> values in this table have been rounded to the nearest integer; exact values and errors are given in Harris *et al.* [50].
- K<sub>D</sub> estimated using a steady-state analysis of Biacore data, as described by Harris *et al.* [50].
- ND = not determined.

side of the *N*-Me substitution, implying that *N*-methylation did not cause any long-range structural changes in R1. Calculated structures were consistent with the previously described solution structure of R1, which consists of two structured regions, both involving turns; the first of these, encompassing residues 5–10, is hydrophobic and the second, involving residues 13–17, is more polar [49]. Even in solution these turns are unlikely to represent the only conformations sampled by the peptide, because linear peptides lacking any covalent cross-links, as in the case of R1, are known to sample a range of rapidly interconverting conformations in aqueous solution.

Harris *et al.* [50] also addressed whether the *N*-Me groups nucleated local structure (as opposed to long-range structure, which is effectively ruled out by the lack of extensive chemical shift changes). A confounding factor here is that *N*-Me groups give sharp, strong NMR resonances, which are likely to detect interproton nuclear Overhauser effects (NOEs) over longer distances than the parent backbone amide proton. Structures were therefore calculated for various *N*-Me analogs in the presence and absence of NOEs to the *N*-Me groups. In essence, the structures of *N*-Me analogs calculated without *N*-Me NOEs were similar to the structure of R1, but inclusion of *N*-Me NOEs and two long-range NOEs in the structure calculations for the *N*-Me-Leu-8 analog caused an apparent stabilization of structure in the vicinity of the *N*-Me group, resulting in an apparently more compact global structure. These local effects have to be considered in NMR analyses of linear peptides such as R1, where there is a dearth of long-range NOEs and an absence of covalent constraints.

The enhanced affinity for AMA1 and broader strain specificity exhibited by a number of the *N*-Me R1 analogs implies that neither intra- nor intermolecular hydrogen bonding interactions play a critical role in R1 binding to AMA1. As the binding site for R1 on AMA1 is likely to be a hydrophobic groove (Richard *et al.*, unpublished results), the reduction in peptide polarity associated with *N*-methylation may contribute to the higher affinity. The combination of higher affinity, broader strain specificity, and resistance to proteolysis in plasma [50] makes the *N*-methylated analogs of R1 an attractive starting point for further development.

#### 4.4

#### Concluding Remarks

There are many ways in which *N*-methylation can affect a peptide's conformation, affinity, selectivity, stability and bioavailability. *N*-Methylation could interfere with amide group hydrogen bond formation, or the free energy of binding could be affected as a result of the introduction of a hydrophobic methyl group or the imposition of steric hindrance. Perturbation of the local conformation, although unlikely to be major, may be sufficient to modulate affinity. The overall effect of *N*-methylation on affinity will reflect a balance across all of these factors. For example, in the *N*-methylated R1 analogs, altered local conformation, loss of hydrogen bonding, and steric repulsion will contribute to the loss of binding affinity observed for some analogs. In contrast, for those analogs with higher affinity, *N*-methylation

presumably favored the bound conformation of the peptide, while hydrophobic interactions between introduced *N*-Me groups and the nonpolar binding groove of AMA1 may also contribute.

*N*-Methylated analogs with the desired affinity and target specificity identified by *N*-Me scanning are likely to have other beneficial properties: resistance to proteolysis in the gut and bloodstream, reduced polarity, and enhanced bioavailability and pharmacokinetics are all potential attributes of *N*-methylated peptides that may be expected to make them attractive analogs for further development of peptides as therapeutics. The immunosuppressant drug cyclosporine (cyclosporin A), which is widely used in transplantation to reduce the risk of organ rejection, provides an excellent example of a naturally occurring peptide with a high degree of *N*-methylation and good bioavailability.

### Acknowledgments

R.S.N. acknowledges fellowship support from the Australian National Health and Medical Research Council. I am very grateful to Brian Smith for assistance with Figure 4.1, Tony Burgess for assistance with Figure 4.2 and for helpful advice, Chris MacRaid for assistance with  $\varphi/\psi$  analyses, and Charles Galea and Jeff Babon for helpful comments on the text.

### References

- 1 Wenger, R.M. (1984) Synthesis of cyclosporine. Total syntheses of "cyclosporin A" and "cyclosporin H", two fungal metabolites isolated from the species *Tolypocladium inflamatam* GAMS. *Helvetica Chimica Acta*, **67**, 502–525.
- 2 Kahan, B.D. (2009) Forty years of publication of *Transplantation Proceedings* – the second decade: the cyclosporine revolution. *Transplantation Proceedings*, **41**, 1423–1437.
- 3 Öberg, K. (2009) Somatostatin analog octreotide LAR in gastroentero-pancreatic tumors. *Expert Review of Anticancer Therapy*, **9**, 557–566.
- 4 Aghi, M. and Blevins, L.S. Jr. (2009) Recent advances in the treatment of acromegaly. *Current Opinion in Endocrinology, Diabetes and Obesity*, **16**, 304–307.
- 5 Chia, C.W. and Egan, J.M. (2008) Incretin-based therapies in type 2 diabetes mellitus. *Journal of Clinical Endocrinology and Metabolism*, **93**, 3703–3716.
- 6 Miljanich, G.P. (2004) Ziconotide: neuronal calcium channel blocker for treating severe chronic pain. *Current Medicinal Chemistry*, **11**, 3029–3040.
- 7 Varkony, H., Weinstein, V., Klinger, E., Sterling, J., Cooperman, H., Komlos, T., Ladkani, D., and Schwartz, R. (2009) The glatiramoid class of immunomodulator drugs. *Expert Opinion on Pharmacotherapy*, **10**, 657–668.
- 8 Pennington, M.W., Beeton, C., Galea, C.A., Smith, B.J., Chi, V., Monaghan, K.P., Garcia, A., Rangaraju, S., Giuffrida, A., Plank, D., Crossley, G., Nugent, D., Khaytin, I., Lefievre, Y., Peshenko, I., Dixon, C., Chauhan, S., Orzel, A., Inoue, T., Hu, X., Moore, R.V., Norton, R.S., and Chandy, K.G. (2009) Engineering a stable and selective peptide blocker of the  $K_v1.3$  channel in T lymphocytes. *Molecular Pharmacology*, **75** 762–773.

- 9 Norton, R.S., Pennington, M.W., and Wulff, H. (2004) Potassium channel blockade by the sea anemone toxin ShK for the treatment of multiple sclerosis and other autoimmune diseases. *Current Medicinal Chemistry*, **11**, 3041–3052.
- 10 Tudor, J.E., Pallaghy, P.K., Pennington, M.W., and Norton, R.S. (1996) Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. *Nature Structural Biology*, **3**, 317–320.
- 11 Beeton, C., Smith, B.J., Sabo, J.K., Crossley, G., Nugent, D., Khaytin, I., Chi, V., Chandy, K.G., Pennington, M.W., and Norton, R.S. (2008) The D-diastereomer of ShK toxin selectively blocks voltage-gated K<sup>+</sup> channels and inhibits T lymphocyte proliferation. *Journal of Biological Chemistry*, **283**, 988–997.
- 12 Caruso, F., Caruso, R.A., and Mohwald, H. (1998) Nanoengineering of inorganic and hybrid hollow spheres by colloidal templating. *Science*, **282**, 1111–1114.
- 13 des Rieux, A., Fievez, V., Garinot, M., Schneider, Y.-J., and Preat, V. (2006) Nanoparticles as potential oral delivery systems of proteins and vaccines: a mechanistic approach. *Journal of Controlled Release*, **116**, 1–27.
- 14 Yap, H.P., Johnston, A.P.R., Such, G.K., Yan, Y., and Caruso, F. (2009) Click engineered bioresponsive, drug-loaded PEG spheres. *Advanced Materials*, **21**, 4348–4352.
- 15 Bailon, P. and Won, C.-Y. (2009) PEG-modified biopharmaceuticals. *Expert Opinion on Drug Delivery*, **6**, 1–16.
- 16 Yang, B.-B., Lum, P.K., Hayashi, M.M., and Roskos, L.K. (2004) Polyethylene glycol modification of filgrastim results in decreased renal clearance of the protein in rats. *Journal of Pharmaceutical Sciences*, **93**, 1367–1373.
- 17 Newton, H.B. (2006) Advances in strategies to improve drug delivery to brain tumors. *Expert Review of Neurotherapeutics*, **6**, 1495–1509.
- 18 Kratz, F. (2008) Albumin as a drug carrier: design of prodrugs, drug conjugates and nanoparticles. *Journal of Controlled Release*, **132**, 171–183.
- 19 Norton, R.S. and McDonough, S.I. (2008) Peptides targeting voltage-gated calcium channels. *Current Pharmaceutical Design*, **14**, 2480–2491.
- 20 Atanassoff, P.G., Hartmannsgruber, M.W.B., Thrasher, J., Wermeling, D., Longton, W., Gaeta, R., Singh, T., Mayo, M., McGuire, D., and Luther, R.R. (2000) Ziconotide, a new N-type calcium channel blocker, administered intrathecally for acute postoperative pain. *Regional Anesthesia and Pain Medicine*, **25**, 274–278.
- 21 Staats, P.S., Yearwood, T., Charapata, S.G., Presley, R.W., Wallace, M.S., Byas-Smith, M., Fisher, R., Bryce, D.A., Mangieri, E.A., Luther, R.R., Mayo, M., McGuire, D., and Ellis, D. (2004) Intrathecal ziconotide in the treatment of refractory pain in patients with cancer or AIDS: a randomized controlled trial. *Journal of the American Medical Association*, **291**, 63–70.
- 22 McGivern, J.G. (2006) Targeting N-type and T-type calcium channels for the treatment of pain. *Drug Discovery Today*, **11**, 245–253.
- 23 Norton, R.S., Baell, J.B., and Angus, J.A. (2004) Calcium channel blocking polypeptides: structure, function and molecular mimicry, in *Calcium Channel Pharmacology* (ed. S.I. McDonough), Kluwer, New York, pp. 143–181.
- 24 Bergseng, E., Xia, J., Kim, C.-Y., Khosla, C., and Sollid, L.M. (2005) Main chain hydrogen bond interactions in the binding of proline-rich gluten peptides to the celiac disease-associated HLA-DQ2 molecule. *Journal of Biological Chemistry*, **280**, 21791–21796.
- 25 Dechantsreiter, M.A., Planker, E., Matha, B., Lohof, E., Hölzemann, G., Jonczyk, A., Goodman, S.L., and Kessler, H. (1999) N-Methylated cyclic RGD peptides as highly active and selective  $\alpha_v\beta_3$  integrin antagonists. *Journal of Medicinal Chemistry*, **42**, 3033–3040.
- 26 Veber, D.F., Saperstein, R., Nutt, R.F., Freidinger, R.M., Brady, S.F., Curley, P., Perlow, D.S., Paleveda, W.J., Colton, C.D., Zacchei, A.G., Tocco, D.J., Hoff, D.R., Vandlen, R.L., Gerich, J.E., Hall, L., Mandarino, L., Cordes, E.H., Anderson, P.S., and Hirschmann, R. (1984) A super



- active cyclic hexapeptide analog of somatostatin. *Life Sciences*, **34**, 1371–1378.
- 27 Rajeswaran, W.G., Hocart, S.J., Murphy, W.A., Taylor, J.E., and Coy, D.H. (2001) Highly potent and subtype selective ligands derived by *N*-methyl scan of a somatostatin antagonist. *Journal of Medicinal Chemistry*, **44**, 1305–1311.
  - 28 Rajeswaran, W.G., Hocart, S.J., Murphy, W.A., Taylor, J.E., and Coy, D.H. (2001) *N*-Methyl scan of somatostatin octapeptide agonists produces interesting effects on receptor subtype specificity. *Journal of Medicinal Chemistry*, **44**, 1416–1421.
  - 29 Bruehlmeier, M., Garayoa, E.G., Blanc, A., Holzer, B., Gergely, S., Tourwe, D., Schubiger, P.A., and Bläuenstein, P. (2002) Stabilization of neurotensin analogues: effect on peptide catabolism, biodistribution and tumor binding. *Nuclear Medicine and Biology*, **29**, 321–327.
  - 30 Ankersen, M., Johansen, N.L., Madsen, K., Hansen, B.S., Raun, K., Nielsen, K.K., Thogersen, H., Hansen, T.K., Peschke, B., Lau, J., Lundt, B.F., and Andersen, P.H. (1998) A new series of highly potent growth hormone-releasing peptides derived from ipamorelin. *Journal of Medicinal Chemistry*, **41**, 3699–3704.
  - 31 Manavalan, P. and Momany, F.A. (1980) Conformational energy studies on *N*-methylated analogs of thyrotropin releasing hormone, enkephalin, and luteinizing hormone-releasing hormone. *Biopolymers*, **19**, 1943–1973.
  - 32 Conti, F. and De Santis, P. (1971) On the conformations of poly-*N*-methyl-L-alanine (PNMA) in solution. *Biopolymers*, **10**, 2581–2590.
  - 33 Groth, P. (1970) Crystal structure of cyclotetrasarcosyl. *Acta Chemica Scandinavica*, **24**, 780–790.
  - 34 Groth, P. (1976) Crystal conformation of cyclotriscarcosyl at  $-160^{\circ}\text{C}$ . *Acta Chemica Scandinavica*, **30**, 838–840.
  - 35 Benedetti, E., Marsh, R.E., and Goodman, M. (1976) Conformational studies on peptides. X-ray structure determinations of six *N*-methylated cyclic dipeptides derived from alanine, valine, and phenylalanine. *Journal of the American Chemical Society*, **98**, 6676–6684.
  - 36 Zimmerman, S.S. and Scheraga, H.A. (1976) Stability of *cis*, *trans*, and nonplanar peptide groups. *Macromolecules*, **9**, 408–416.
  - 37 Kessler, H., Anders, U., and Schudok, M. (1990) An unexpected *cis* peptide bond in the minor conformation of a cyclic hexapeptide containing only secondary amide bonds. *Journal of the American Chemical Society*, **112**, 5908–5916.
  - 38 Chatterjee, J., Mierke, D., and Kessler, H. (2006) *N*-Methylated cyclic pentaalanine peptides as template structures. *Journal of the American Chemical Society*, **128**, 15164–15172.
  - 39 Tran, T.T., Treutlein, H., and Burgess, A.W. (2006) Designing amino acid residues with single-conformations. *Protein Engineering, Design and Selection*, **19**, 401–408.
  - 40 Burgess, A.W., Ponnuswamy, P.K., and Scheraga, H.A. (1974) Analysis of conformations of amino acid residues and prediction of backbone topography in proteins. *Israel Journal of Chemistry*, **12**, 239–286.
  - 41 Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., and Zardecki, C. (2002) The Protein Data Bank. *Acta Crystallographica D*, **58**, 899–907.
  - 42 Altschuh, D. (2002) Cyclosporin A as a model antigen: immunochemical and structural studies. *Journal of Molecular Recognition*, **15**, 277–285.
  - 43 Donzel, B., Goodman, M., Rivier, J., Ling, N., and Vale, W. (1975) Synthesis and conformations of hypothalamic hormone releasing factors: two QRF-analogues containing backbone *N*-methyl groups. *Nature*, **256**, 750–751.
  - 44 Burgess, A.W., Momany, F.A., and Scheraga, H.A. (1975) On the structure of thyrotropin releasing factor. *Biopolymers*, **14**, 2645–2647.
  - 45 Donzel, B., Rivier, J., and Goodman, M. (1974) Conformational studies on the hypothalamic thyrotropin releasing factor and related compounds by  $^1\text{H}$  nuclear

- magnetic resonance spectroscopy. *Biopolymers*, **13**, 2631–2647.
- 46 Patel, Y.C. and Srikant, C.B. (1997) Somatostatin receptors. *Trends in Endocrinology and Metabolism*, **8**, 398–405.
- 47 Biron, E., Chatterjee, J., Ovadia, O., Langenegger, D., Brueggen, J., Hoyer, D., Schmid, H.A., Jelinek, R., Gilon, C., Hoffman, A., and Kessler, H. (2008) Improving oral bioavailability of peptides by multiple N-methylation: somatostatin analogues. *Angewandte Chemie (International Edition in English)*, **47**, 2595–2599.
- 48 Chatterjee, J., Gilon, C., and Hoffman, A., and Kessler, H. (2008) N-Methylation of peptides: a new perspective in medicinal chemistry. *Accounts of Chemical Research*, **41**, 1331–1342.
- 49 Harris, K.S., Casey, J.L., Coley, A.M., Masciantonio, R., Sabo, J.K., Keizer, D.W., Lee, E.F., McMahon, A., Norton, R.S., Anders, R.F., and Foley, M. (2005) Binding hot spot for invasion inhibitory molecules on *Plasmodium falciparum* apical membrane antigen 1. *Infection and Immunity*, **73**, 6981–6989.
- 50 Harris, K.S., Casey, J.L., Coley, A.M., Karas, J.A., Sabo, J.K., Tan, Y.Y., Dolezal, O., Norton, R.S., Hughes, A.B., Scanlon, D., and Foley, M. (2009) Rapid optimization of a peptide inhibitor of malaria parasite invasion by comprehensive N-methyl scanning. *Journal of Biological Chemistry*, **284**, 9361–9371.

## 5 High-Performance Liquid Chromatography of Peptides and Proteins

Reinhard I. Boysen and Milton T.W. Hearn

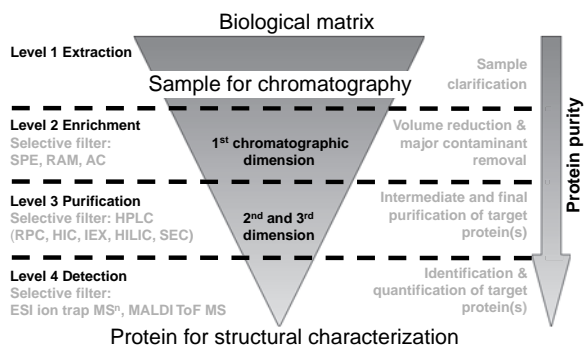
### 5.1 Introduction

High-performance liquid chromatography (HPLC) has become an essential tool for the separation, purification and characterization of biomacromolecules. Using HPLC for the isolation of naturally occurring peptides or proteins from complex biological matrices and for their unambiguous structural elucidation still presents considerable challenges. Several consecutive purification steps are usually required to obtain the target compound in sufficient yield and purity, which results in complex multidimensional LC (multidimensional liquid chromatography MDLC) purification schemes. Since proteins can occur as protein isoforms (as the result of chemical or biological post-translational modifications) or protein variants (from genetic modifications) their identification additionally requires mass spectrometry (MS) and bioinformatic analysis. Similar considerations are valid for the purification of synthetic peptides and recombinantly produced proteins.

The process of protein isolation and identification can be split into several subprocesses. After bulk extraction from the biological material, the crude protein extract is usually fractionated in a way that takes into account the chemical properties of the target protein and the chemical nature of the feedstock. These separations can be performed with a variety of chromatographic techniques, including open-column chromatography, and may include an initial protein precipitation with organic solvent or salt.

The workflow integration of high performance liquid chromatography into the extraction, enrichment, purification, and structural elucidation of peptides and proteins is illustrated in Figure 5.1.

After the extraction of a protein from a biological matrix (Level 1), the crude extract is clarified (i.e., by filtration, centrifugation) to be free from particulate matter to be suitable for chromatography. An appropriate sample buffer compatible with the mobile phase(s) of the particular chromatographic mode used in the next step is chosen. This is followed by an enrichment step (Level 2), preferably using solid-phase extraction (SPE) or restricted access materials (RAMs) in a step elution mode in order



**Figure 5.1** Example of workflow in protein isolation from a complex biological matrix using HPLC for the target compound purification and identification. The successive application of several chromatographic modes of different selectivity renders the chromatographic separation process multidimensional.

to eliminate the majority of low molecular weight materials and to drastically reduce the volume of the sample. The next step (Level 3) comprises the intermediate purification of target compound(s) and a final chromatographic purification using a variety of high-performance chromatographic modes of different selectivity (i.e., separating the analytes according to their molecular size, hydrophobicity/hydrophilicity, charge) which may encompass size-exclusion chromatography (HP-SEC), reversed-phase chromatography (HP-RPC), hydrophobic interaction chromatography (HIC), hydrophilic interaction chromatography (HP-HILIC), ion-exchange chromatography (HP-IEX) or affinity chromatography (HP-AC), to yield the desired compound(s) in the required amount and degree of purity. At the detection level (Level 4), besides diode array detection, electrospray ionization (ESI) MS can often be applied for the identification and – depending on the type of mass spectrometer – may allow quantification of the separated compound(s). Further structural elucidation can be performed with multiple-stage MS (multiple-stage mass spectrometry MS<sup>n</sup>), such as with an ion-trap mass spectrometer, quadruple or Fourier transform mass spectrometer, or with proton, carbon, or heteronuclear nuclear magnetic resonance (NMR) spectroscopy.

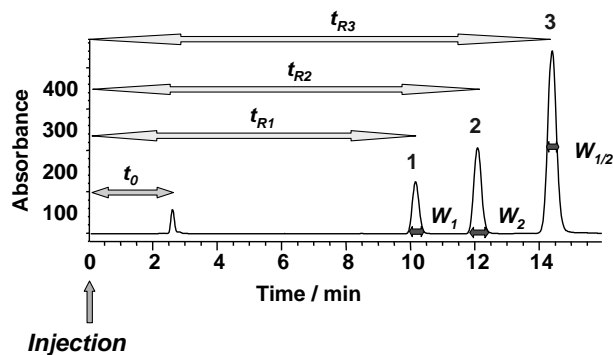
In order to develop a purification strategy tailored to a target peptide and protein, an intricate knowledge of the fundamental terms and concepts in chromatography and of the basic physicochemical characteristics of peptides and proteins is required, which are reviewed in Sections 5.2 and 5.3, respectively. The major chromatographic modes for the isolation and purification of peptides and proteins are described in Section 5.4. In order to take full advantage of a specific HPLC mode, and to effectively utilize time and resources, a comprehensive method development is always recommended. Guidelines for the method development in HP-RPC, the major chromatographic mode employed in peptide and protein analysis, are given in Section 5.5. Since peptides and proteins are generally isolated from complex biological matrices, often more than one chromatographic step is needed for their purification. The successive application of several suitable different chromatographic modes must

consider their applicability to the protein of interest, and their compatibility with each other and with the detection procedures. The concepts and implementations of two- or higher-dimensional separation schemes for peptide and protein purification is discussed in Section 5.6.

## 5.2

### Basic Terms and Concepts in Chromatography

A chromatographic separation begins with the injection of the sample, whereby its components are distributed between two phases, one of which is stationary while the other, the mobile phase, moves in a definite direction. The individual compounds carried by the mobile phase migrate through the chromatographic column with different velocity, depending on the elution methods employed, and are detected in the order in which they leave the column by, for example, UV absorbance. The resulting chromatogram plots the response of a detector to the concentration of the compound (peak) versus the length of time it takes to elute it from the column (retention time). The time an unretained compound needs to migrate through the column is the column void time  $t_0$ . It is related to the void volume  $V_0$ , which is the sum of the interstitial volume between the particles of the stationary phase and the available volume within the particle pores. The elution of sample components is delayed through their interaction with the stationary phase, which is described by the retention time  $t_R$  or by the retention volume  $V_R$ . The peak is characterized through the peak width  $w$ , the peak width at half height  $w_{1/2}$ , and the peak area  $A$  (Figure 5.2). Optimized conditions strive for narrow, symmetric peaks that are not fronting or tailing.



**Figure 5.2** Schematic depiction of a chromatogram of three retained compounds. Abbreviations:  $t_0$  = column void time (time from injection to detection of unretained compound),  $t_{R1}$ ,  $t_{R2}$ ,  $t_{R3}$  = retention times (time

from injection to detection of retained compounds),  $w$  = peak width (measured at baseline), and  $w_{1/2}$  = peak width at half height (measured at half the peak height).

The *retention factor*  $k$  is used to describe the retention independent from the column dimensions or flow rate:

$$k = \frac{t_R - t_0}{t_0} \quad (5.1)$$

where  $t_R$  is retention time and  $t_0$  is column void time.

Alternatively, the retention factor can be expressed in terms of elution volumes, since retention times,  $t_R$  and  $t_0$ , are related to the elution volumes and the flow rate  $F$  of the chromatographic system through the relationships:

$$V_R = t_R F \quad \text{and} \quad V_0 = t_0 F \quad (5.2)$$

Hence:

$$k = \frac{V_R - V_0}{V_0} \quad (5.3)$$

Thus, the retention factor relates to the number of additional column volumes beyond  $V_0$  required to elute a compound. The retention factor can have values between  $k = 0$  (no retention) and  $k = \infty$  (irreversible adsorption) with values of 1–20 most preferred. The retention factor can also be defined as the ratio  $n_s/n_m$ , where  $n_s$  is the total number of moles of the solute associated with the stationary phase and  $n_m$  is the total number of moles of the solute in the mobile phase:

$$k = \frac{n_s}{n_m} \quad (5.4)$$

In order to resolve two components, their retention factors must be different. The selectivity  $\alpha$  of a chromatographic system describes the ability of a chromatographic system to separate two compounds (1 and 2) based on their different retention factors:

$$\alpha = \frac{k_2}{k_1} \quad (5.5)$$

where  $k_1$  and  $k_2$  are the retention factors of the compounds. For a selectivity of  $\alpha = 1$  no separation is possible.

In order to evaluate the quality of a separation not only the peak distance of the two components must be considered but also their respective peak width. The resolution,  $R_S$ , of two adjacent peaks in a chromatogram is defined by the ratio of peak distance and their peak widths:

$$R_S = \frac{t_{R2} - t_{R1}}{(1/2)(w_1 + w_2)} \quad (5.6)$$

with retention times,  $t_{R1}$  and  $t_{R2}$ , of two adjacent peaks and the respective peak widths,  $w_1$  and  $w_2$ , of the two peaks. Two peaks can be either unseparated ( $R_S < 1$ ), partially overlapping ( $R_S = 1$ ), or baseline separated ( $R_S > 1.5$ ).

Dispersion effects of solutes in the chromatographic system are one cause of band broadening. The extent of band broadening is reflected in the column efficiency, which is usually expressed as the plate number,  $N$ , or as the plate height,  $H$  (also called the height equivalent to one theoretical plate (HETP)):

$$N = \frac{L}{H} \quad (5.7)$$

where  $L$  is the column length. The concept of “theoretical plates” goes back to the number of distillation plates during fractionating distillation. The higher the number of theoretical plates at a particular column length  $L$ , the better the quality of column and the narrower the peaks. The value of  $N$  is dependent on a variety of chromatographic and solute parameters including the column length,  $L$ , the chromatographic particle diameter,  $d_p$ , the linear flow velocity,  $u$ , and the solutes’ diffusivities ( $D_m$  and  $D_s$ ) in the bulk mobile phase and within the stationary phase, respectively. The plate number can be defined as:

$$N = \left( \frac{t_R}{\sigma_i^2} \right) \quad (5.8)$$

or:

$$N = 16 \left( \frac{t_R}{w} \right)^2 \quad (5.9)$$

where  $t_R$  is the retention time and  $\sigma_i^2$  the peak variance of the eluted zone in time units. For practical convenience,  $\sigma_i^2$  is often replaced with the peak width  $w$ . For Gaussian peaks  $w$  approximately corresponds to  $4\sigma$  ( $4 \times$  peak standard variation).

In order to permit a comparison of column efficiencies of columns with identical bed dimensions packed with sorbent particles of different physical or chemical characteristics (e.g., different average diameter, ligand type), the plate height  $H$  is defined through the reduced plate height,  $h$ , while the linear flow velocity  $u = L/t_0$ , can be defined in terms of reduced mobile phase velocity  $\nu$ :

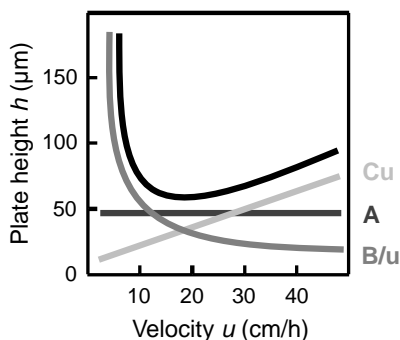
$$h = \frac{H}{d_p} \quad \text{and} \quad \nu = \frac{u d_p}{D_m} \quad (5.10)$$

where  $L$  is the column length,  $d_p$  is the particle diameter, and  $D_m$  is the diffusivity of the solute in the mobile phase.

The peak zone band broadening, which is caused by various mass transport effects in the column, is described through the dependency of the reduced plate height,  $h$  on the velocity  $u$  or the reduced velocity  $\nu$  through the van Deemter–Knox equation:

$$h = A + \frac{B}{u} + Cu \quad \text{or} \quad h = A\nu^{1/3} + \frac{B}{\nu} + C\nu \quad (5.11)$$

where the  $A$ -term expresses the eddy diffusion and mobile phase mass transfer effects, and is a measure of the packing quality of the chromatographic bed (and is constant for a given column); the  $B$ -term entails the longitudinal molecular diffusion



**Figure 5.3** Van Deemter–Knox plot (top curve) and plots of individual terms of the van Deemter–Knox equation: the A-term, which expresses the eddy diffusion and mobile phase mass transfer effects; the B-term, which

accounts for longitudinal molecular diffusion effects; and the C-term, which incorporates mass transfer resistances within the stationary phase microenvironment.

effects; while the C-term incorporates mass transfer resistances within the stationary phase microenvironment which describes the interaction of the solutes with the stationary phase. In order to achieve optimal separation performance, the reduced plate height needs to be as small as possible. The optimal flow velocity can be deduced from the minimum in  $h$  in the  $h$  versus  $v$  plots (Figure 5.3).

The van Deemter–Knox equation can be used to mathematically describe the chromatographic behavior of low-molecular-weight compounds like peptides with good approximation; for large molecules like proteins it has limitations, since the proteins vary considerably in shape and surface properties. As a consequence, the chromatographic effects can only be described with approximations and average values. It was established that (i) the behavior of small molecules is determined by their diffusion, the influence of the B-term (longitudinal diffusion) is negligible for large molecules particularly at higher flow velocities; (ii) the interaction with the stationary phase (C-term) of large molecules results in band broadening; and (iii) the optimal plate height for large molecules can be obtained at lower flow velocities.

Apart from zone band broadening within the column, the band broadening can also arise from extra-column band broadening due to instrument characteristics:

$$\sigma_t^2 = \sigma_{\text{column}}^2 + \sigma_{\text{extra}}^2 \quad (5.12)$$

where  $\sigma_{\text{column}}^2$  and  $\sigma_{\text{extra}}^2$  are the peak variances arising due to column and extra-column effects, respectively. Careful attention must be paid to possible sources of extra-column effects (e.g., the choice of tubing in terms of length and diameter, the type of fittings, the choice of detector cell volume) that can reduce the impact of  $\sigma_{\text{extra}}^2$  on the overall  $h$  value.

The zone band broadening or peak dispersion, expressed as reduced plate height,  $h$ , arises from kinetic, time-dependent phenomena. In the absence of secondary effects (e.g., slow chemical equilibria, pH effects, conformational



changes) that could influence the chromatographic process, the resolution  $R_S$  can be expressed as:

$$R_S = 1/4N^{1/2}(\alpha-1)(k/(1+k)) \quad (5.13)$$

This equation links three essential parameters that determine the quality of a chromatographic separation (the retention factor  $k$ , the selectivity  $\alpha$ , and the plate number  $N$ ) and therefore describes the extent to which the zone spreading may cause the loss of separation performance. As will be demonstrated in Section 5.5, this equation can be used to guide systematic method development for resolution optimization.

### 5.3

#### Chemical Structure of Peptides and Proteins

Peptides and proteins are a class of molecules containing amino acids as the basic units. The chemical organization (i.e., the primary structure or amino acid sequence) and the folded structure (i.e., the secondary, tertiary, and quaternary structure) are the essential features of a peptide or protein, around which a chromatographic separation is designed.

##### 5.3.1

#### Biophysical Properties of Peptides and Proteins

The 20 naturally occurring L- $\alpha$ -amino acids found in peptides and proteins vary dramatically with respect to the properties of their side-chains (or R-groups). This chemical diversity is further increased where some of these side-chains have been post-translationally modified with a variety of chemical and biological modifications (e.g., acetylation, deamidation, phosphorylation, glycosylation, and lipidation). The side-chains are generally classified according to their polarity (e.g., nonpolar or hydrophobic versus polar or hydrophilic). The polar side-chains are divided into three groups: uncharged, positively charged or basic, and negatively charged or acidic side-chains. In addition to the common L- $\alpha$ -amino acids (Table 5.1) that constitute the majority of proteins and most naturally occurring peptides, other amino acids such as ornithine, selenocysteine, hydroxyproline, allothreonine, and alloisoleucine also occur as part of some protein families, with much greater diversity available in terms of composition, molecular sequence, and chirality for peptides/polypeptides prepared by solution of solid-phase synthesis methods. Peptides and proteins generally contain several ionizable basic and acidic functionalities. They therefore typically exhibit characteristic isoelectric points with the overall net charges and polarities in aqueous solutions varying with pH, solvent composition, and temperature. Cyclic peptides without ionizable side-chains have a zero net charge and they represent an exceptional subgroup. The number and distribution of charged groups influences the polarizability and ionization status of a peptide or protein, as well as

**Table 5.1** Properties of the common L- $\alpha$ -amino acid residues including their monoisotopic and average mass, partial specific volume [1], accessible surface area [2], pK<sub>a</sub> of side-chain [3] and termini [4], and relative hydrophobicity [5].

Three-letter code	One-letter code	Monoisotopic mass (amu)	Average mass (amu)	Partial specific volume ( $\text{\AA}^3$ )	Accessible surface area ( $\text{\AA}^2$ )	pK <sub>a</sub> of side-chain	Relative hydrophobicity
Ala	A	71.03711	71.0788	88.6	115		0.06
Arg	R	156.10111	156.1876	173.4	225	12.48	-0.85
Asn	N	114.04293	114.1039	117.7	160		0.25
Asp	D	115.02694	115.0886	111.1	150	3.9	-0.20
Cys	C	103.00919	103.1448	108.5	135	8.37	0.49
Gln	Q	128.05858	128.1308	143.9	180		0.31
Glu	E	129.04259	129.1155	138.4	190	4.07	-0.10
Gly	G	57.02146	57.0520	60.1	75		0.21
His	H	137.05891	137.1412	153.2	195	6.04	-2.24
Ile	I	113.08406	113.1595	166.7	175		3.48
Leu	L	113.08406	113.1595	166.7	170		3.50
Lys	K	128.09496	128.1742	168.6	200	10.54	-1.62
Met	M	131.04049	131.1986	162.9	185		0.21
Phe	F	147.06841	147.1766	189.9	210		4.8
Pro	P	97.05276	97.1167	122.7	145		0.71
Ser	S	87.03203	87.0783	89	115		-0.62
Thr	T	101.04768	101.1051	116.1	140		0.65
Trp	W	186.07931	186.2133	227.8	255		2.29
Tyr	Y	163.06333	163.1760	193.6	230	10.46	1.89
Val	V	99.06841	99.1326	140	155		1.59

Terminal group	Composition	Mono-isotopic mass (amu)	Average mass (amu)	pK <sub>a</sub> of terminus
N-terminal groups				
Hydrogen	H	1.00782	1.0079	
N-Formyl	HCO	29.00274	29.0183	
N-Acetyl	CH <sub>3</sub> CO	43.01839	43.0452	
α-Amino				7.7–9.2
C-terminal groups				
Free acid	OH	17.00274	17.0073	
N-Acetyl	NH <sub>2</sub>	16.01872	16.0226	
α-Carboxyl				2.75–3.2

the hydrophobicity. Table 5.1 lists the characteristic data for the most common L- $\alpha$ -amino acids found in peptides and proteins, and gives a summary of the N- and C-terminal groups. These data are instrumental in determining the selection of the optimal separation conditions for the resolution of peptide and protein mixtures. They can be used to evaluate the impact of amino acid composition (which, for example, directs the choice of eluent composition or the gradient range in RP-HPLC) on retention behavior. They can also point to the impact of amino acid substitution or deletion with small peptides on retention, or alternatively can be used to guide the identification of peptide fragments derived from proteolytic (chemical or enzymatic) digestion of proteins.

### 5.3.2

#### Conformational Properties of Peptides and Proteins

In solution, a polypeptide or protein can, in principle, explore a relatively large array of conformational spaces. For small peptides (approximately up to 15 amino acid residues) a defined secondary structure ( $\alpha$ -helical,  $\beta$ -sheet, or  $\beta$ -turn motif) is generally absent. With increasing polypeptide chain length, depending on the nature of the amino acid sequence, specific regions/domains of a polypeptide or protein can adopt preferred secondary, tertiary, or quaternary structures. In aqueous solutions this folding, which internalizes the hydrophobic residues and thus stabilizes the polypeptide structure, becomes a significant feature of peptides and proteins for chromatographic separations. A critical factor in the selection of a HPLC procedure is that the experimental conditions will inevitably cause perturbations of the conformational status of these biomacromolecules. Polypeptide and protein conformational stability can be manipulated in a number of ways (mobile and stationary phase composition, temperature, etc.) in HPLC. However, in most cases, an integrated biophysical experimental strategy (including  $^1\text{H}$  two-dimensional NMR, Fourier transform IR (FTIR), ESI-MS, circular dichroism optical rotatory dispersion (CD-ORD) spectroscopy) is required in order to determine the secondary and higher-order structure of a polypeptide or protein in solution or in the presence of specific ligands or possible self-self aggregation effects with peptides or proteins which may have occurred during various HPLC separation conditions.

### 5.3.3

#### Optical Properties of Peptides and Proteins

The peptide bond absorbs strongly in the far-UV region of the spectrum ( $\lambda = 205\text{--}215$  nm). Hence, UV detection is the most widely used method for detection of peptides and proteins in HPLC. Apart from absorbing in the far-UV range, the aromatic amino acid residues also absorb light above 250 nm, due to the delocalized  $\pi$ -systems of the aromatic residues, as shown in Table 5.2. Knowledge of the UV spectra, in particular the extinction coefficients of the nonoverlapping absorption maxima of these amino acids, allows, in conjunction with UV-diode array detection (DAD) and second

**Table 5.2** Optical properties of aromatic amino acids in aqueous solution [6].

Three letter code		Absorption maxima		Fluorescence maxima	
		Wavelength (nm)	Extinction coefficient ( $M^{-1} cm^{-1}$ )	Wavelength (nm)	Fluorescence quantum efficiency
Phenylalanine	Phe	257	200	282	0.04
Tryptophan	Trp	280	5600	348	0.2
Tyrosine	Tyr	274	1400	302	0.14

derivative or difference UV spectroscopy, verification of peak purity and determination of the aromatic amino acid content of peptides and proteins. The knowledge of the relative UV/Vis absorbance of a peptide or protein is thus crucial, since the choice of detection wavelength of peptides and proteins in RP-HPLC (and in the other HPLC modes) depends on the different UV cutoffs of the eluents used. The common use of  $\lambda = 215$  nm as the preferred detection wavelength for most analytical reversed-phase applications (and for those of other HPLC modes) with peptides and proteins is a good compromise between detection sensitivity and potential detection interference due to buffer absorption. However, wavelengths between 230 and 280 nm are frequently employed in preparative applications, where the use of more sensitive detection wavelengths could result in overloading of the detector response (usually above an absorbance value of 2.0–2.5 AU).

The three aromatic amino acids phenylalanine, tryptophan, and tyrosine also show a small but nevertheless significant fluorescence (Table 5.2), whereby the fluorescence of the incorporated amino acid is usually smaller than that of the free acid. Frequently, these aromatic amino acid residues are the sole origin of the intrinsic fluorescence of proteins. Since the fluorescence of the folded/unfolded state of proteins can vary considerably, it can be used to monitor the change of their folding status. In addition, the intensity of the fluorescence of, for example, tryptophan is dependent on the solvational environment, whereby the intensity of the fluorescence decreases inversely proportional with the polarity of the solvent. The fluorescence can also be reduced by a protonated acidic residue (i.e., aspartic or glutamic acid) next to the tryptophan.

## 5.4

### HPLC Separation Modes in Peptide and Protein Analysis

There are several modes of HPLC currently in use for peptide and protein analysis, namely HP-SEC, HP-RPC, normal-phase chromatography (HP-NPC), HP-HILIC, aqueous normal-phase chromatography (HP-ANPC), HP-HIC, HP-IEX, and HP-AC, which includes immobilized metal ion affinity chromatography (HP-IMAC)

**Table 5.3** Chromatographic modes and exploited molecular properties of target compounds.

Chromatographic mode	Acronym	Exploited molecular properties
Size exclusion or gel permeation	SEC or GPC	molecular mass, hydrodynamic volume
Reversed phase	RPC	hydrophobicity
Normal phase	NPC	polarity
Hydrophilic interaction	HILIC	hydrophilicity
Aqueous normal phase	ANPC	hydrophilicity
Hydrophobic interaction	HIC	hydrophobicity
Anion exchange	AEX	net negative charge
Cation exchange	CEX	net positive charge
Affinity	AC	specific interaction
Immobilized metal ion affinity	IMAC	complexation

and biospecific/biomimetic affinity chromatography (HP-BAC). The principles of the major modes are explained below. These and also a number of less frequently used chromatographic modes, such as mixed-mode chromatography (HP-MMC), charge-transfer chromatography (HP-CTC) or ligand-exchange chromatography (HP-LEC), can be operated under isocratic (i.e., fixed eluent composition), step-gradient, or gradient elution conditions, which either change eluent conditions in variable steps or continuously, with the exception of HP-SEC (usually only performed under isocratic conditions). All modes can be used in analytical, semipreparative, or preparative [7–13] situations.

In order to achieve optimal selectivity and hence resolution of peptides and proteins in high-performance chromatographic separations, irrespective of whether the task at hand is of analytical or preparative nature, the choice of the chromatographic mode must be guided by the properties of the analytes (i.e., molecular size/shape hydrophobicity/hydrophilicity, net charge, isoelectric point, solubility, function, antigenicity, carbohydrate content, content of free SH, exposed histidine residues, exposed metal ions). A list of chromatographic modes and the molecular properties of the target compounds that form the basis of each separation mode is given in Table 5.3.

In addition to the abovementioned functional characteristics of these chromatographic systems, other chemical and physical parameters of the mobile and stationary phase impact on the resolution, mass recovery, and bioactivity preservation in separations of polypeptides or proteins during liquid chromatographic separations. These parameters are listed in Table 5.4.

#### 5.4.1

##### SEC

HP-SEC, also called gel-permeation chromatography (HP-GPC), is performed on porous stationary phases and separates analytes according to their molecular mass or, more precisely, their hydrodynamic volume. The separation of analytes is based on

**Table 5.4** Chemical and physical factors of the mobile and stationary phase that contribute to variation in the resolution, mass recovery, and bioactivity preservation of polypeptides, proteins and other biomacromolecules in HPLC systems [14].

Mobile phase contributions	Stationary phase contributions
Organic solvents	ligand composition
pH	ligand density
Metal ions	surface heterogeneity
Chaotropic reagents	surface area
Oxidizing or reducing reagents	pore diameter
Temperature	pore diameter distribution
Buffer composition	particle size
Ionic strength	particle size distribution
Loading concentration and volume	particle compressibility

the concept that molecules of different hydrodynamic volume (Stokes radius) permeate to different extents into porous HP-SEC separation media and thus exhibit different permeation coefficients according to differences in their molecular masses/hydrodynamic volumes. Analytes with a molecular weight larger than the exclusion limit (usually listed in the technical information provided by the column manufacturer) are excluded from the pores and elute in the void volume of the column. As a nonretentive separation mode, HP-SEC is usually operated with isocratic elution using aqueous low salt mobile phases.

HP-SEC can be used for group separation or high-resolution fractionation. In the group separation mode, HP-SEC removes small molecules from large molecules and is also suitable for buffer exchange or desalting. In the high-resolution fractionation mode, HP-SEC separates various components in a sample due to their different hydrodynamic volumes or can be employed to perform a molecular weight distribution analysis.

As HP-SEC columns have no adsorption capacity and dilute the sample upon elution, they are not normally used in the initial capture or for intermediate purification in multistep chromatographic processes, however, they are suitable for final polishing (e.g., for the final removal of unwanted aggregates or multimeric forms of the proteins or other impurities of significantly different molecular weight). HP-SEC can be performed directly after HP-AC, HP-HIC, or HP-IEX.

#### 5.4.2

##### RPC

HP-RPC is the major analytical mode for peptide and protein analysis. Historically, normal-phase LC and silica-based reversed-phase LC have represented the major analytical modes of adsorption chromatography. In reversed-phase chromatography, the polarity of the stationary and mobile phase is the reverse of that used in NPC. Peptides and proteins are loaded onto the column under aqueous conditions and eluted with a mobile phase containing an organic solvent. The column contains a

porous or nonporous stationary phase with immobilized nonpolar ligands. HP-RPC separates compounds according to their relative nonpolarity or their hydrophobicity. The most commonly accepted theory of the retention mechanism in HP-RPC is based on the solvophobic theory, which describes the hydrophobic interaction between the nonpolar surface regions of the analytes and the nonpolar ligands/surfaces of the stationary phase [15, 16].

Typically, the nonpolar ligands are immobilized onto the surface of spherical, porous or nonporous silica particles, although nonpolar polymeric sorbents (e.g., those derived from cross-linked polystyrene–divinylbenzene) can also be employed. Silica-based packing materials of 3–10  $\mu\text{m}$  average particle diameter and 70–1000  $\text{\AA}$  pore size with *n*-butyl, *n*-octyl, or *n*-octadecyl ligands are widely used for the separation of peptides and proteins. Silica particles of 1  $\mu\text{m}$  to more than 65  $\mu\text{m}$  have been developed in various size distributions and configurations (e.g., spherical, irregular, with various pore geometries and pore connectivities; and in pellicular, fully porous or monolithic structures) by a variety of routes of manufacture and with different silica types, and are grouped into type I, type II, or type III silica according to the classification of Unger [17]. For low-molecular-mass (below 4000 Da) polypeptides, silica materials of 70–80  $\text{\AA}$  pore size and 3–5  $\mu\text{m}$  average particle diameters are often used, which maximizes loading capacity and retention. For proteins in the mass range of 4000–500 000 Da, the pore size of 300  $\text{\AA}$  allows maintenance of high efficiency, whereby the loading capacity can be increased by increasing the column diameter. Macro-porous HP-RPC columns of 1000  $\text{\AA}$  pore size are increasingly used for the fractionation of very complex proteins samples.

In HP-RPC, an organic solvent (i.e., methanol, ethanol, acetonitrile, *n*-propanol, tetrahydrofuran) is used as a surface tension modifier in the chromatographic eluent, which has a particular elution strength, viscosity and UV cutoff. Mobile phase additives, such as acetic acid, formic acid, trifluoroacetic acid (TFA) and hepta-fluorobutyric acid (HBFA), are used to obtain a particular pH value, typically at low pH (e.g., around pH 2 for silica-based materials) with the exception of polymeric stationary phases, which have an extended pH range from pH 1 to 12. Some mobile phase additives may also function as ion-pair reagents, which interact with the ionized analytes to form overall neutral eluting species, and also suppress silanophilic interactions between free silanol groups on the silica surface and basic functional groups of the analytes. The properties of the additives determine their suitability for use with ESI-MS. Strong ion-pair interactions between analytes and mobile phase additives can suppress the ionization of the analytes in ESI-MS.

HP-RPC can be operated in the isocratic, step-gradient, or continuous gradient elution mode, and is frequently used as an intermediate or final polishing step in a multistep purification. It is ideally positioned after HP-IEC because it allows desalting and the separation of the sample in a single step. Due to its versatility and flexibility, HP-RPC techniques dominate the separation of peptides and proteins at the analytical, laboratory-scale, and semipreparative levels, since the majority of peptides and proteins possess some degree of hydrophobicity.



### 5.4.3

#### **NPC**

Chromatographic systems in which the stationary phase is more polar than the mobile phase had been developed at the beginning of the modern era of LC and were known under the acronym “NPC”. HP-NPC can be performed on unmodified silica and separates analytes according to their intrinsic polarity. HP-NPC can be operated in isocratic, step gradient, or gradient elution mode, where the retaining mobile phase contains less polar organic solvents and the eluting mobile phase consists of more polar organic solvents. Water, due to its extreme polarity, adsorbs to most “normal-phase” stationary phases and significantly affects the separation reproducibility. In contrast to HP-RPC with immobilized *n*-alkyl ligands, where the interaction of solute and stationary phase is based on solvophobic phenomena, the interaction in HP-NPC is based on adsorption. The retention behavior of peptides and proteins in HP-NPC is often described in terms of the classical concepts inherent to multisite displacement and site occupancy theory [18]. HP-NPC is mainly used for the separation of, for example, polyaromatic hydrocarbons, heteroaromatic compounds, nucleotides and nucleosides, and much less frequently for protected synthetic peptides, deprotected small peptides in the “flash chromatographic mode,” and protected amino acid derivatives used in peptide synthesis [19]. Originally, HP-NPC was limited to unmodified silica columns; however, recent work utilized polar-bonded phases such as amino ( $-\text{NH}_2$ )-, cyano ( $-\text{CN}$ )-, or diol ( $-\text{COHCOH}-$ )-coated sorbents. Such modified normal-phase packing materials were suitable for polar bonded phase chromatography (PBPC), which was used for the separation of peptides [20] and proteins [21]. Today, one of the main applications of modified normal-phase silica materials is their use in HPLC-integrated SPE procedures [22]. These types of sorbents, particularly when used as precolumn packing materials in LC-LC column switching settings, in conjunction with restricted access sorbents materials (RAM), allow multiple injections of untreated complex biological samples (e.g., hemolysed blood, plasma serum, fermentation broth, and cell tissue homogenates) for the isolation of bioactive peptides. Typically, with RAMs, hydrophilic, electroneutral diol groups are immobilized onto the outer surface of spherical particles. This layer prevents nonspecific interactions between the support matrix and protein(s) or other high-molecular-weight biomolecules, which are thus excluded from the interior regions of the particle and elute as nonretained components. The inner surfaces of the porous RAM particles are, however, chemically modified with *n*-alkyl ligands, which are only freely accessible for low-molecular-weight analytes, such as peptides. As a consequence, significant enrichment or partial resolution of peptide analytes can be achieved.

### 5.4.4

#### **HILIC**

HP-HILIC is performed on porous stationary phases with immobilized hydrophilic ligands and separates analytes according to their hydrophilicity. This variant of the HP-NPC mode was introduced by Alpert in 1990 [23], based on polyaspartic acid

immobilized onto silica, and is used for the separation of amino acids, small peptides, and simple maltoglycosides with mobile phases of high organic solvent content. A pseudo-HILIC separation of simple saccharides was performed on a BondaPak-NH<sub>2</sub> column as early as 1975 [24]. The HP-HILIC mode has since been applied for the separation of various analytes, including simple carbohydrates and amino acids [25] as well as peptides [26, 27].

In HILIC, polar sorbents with amide, aminopropyl, cyanopropyl, diol, cyclodextrin, poly(succinimide), and sulfoalkylbetaine phases are employed, and the non-aqueous mobile phases of NPC are replaced with high-organic, low-aqueous eluents [23, 28–30]. Elution of compounds from HP-HILIC columns is achieved by increasing the water content in the mobile phase. The elution order was initially thought to be more or less opposite to that seen in HP-RPC separation [23, 28, 30]. Although it intuitively would seem that retention in HILIC would simply be the “reverse” of that in HP-RPC, studies on the orthogonality of separations in two-dimensional LC have demonstrated that HP-HILIC (with a bare silica sorbent) and HP-RPC can be a suitable combination for proteomic analysis in two-dimensional systems [31]. Compared to the nonaqueous mobile phase in HP-NPC, the partly aqueous mobile phase used in HP-HILIC allows greater solubility of many polar and hydrophilic compounds, and fast separation of polar compounds can be achieved due to the low viscosity of the highly organic mobile phase. Moreover, the high content of organic solvent in the mobile phase favors ionization of polar compounds in subsequent ESI-MS and thus provides enhanced detection sensitivity for these compounds [30, 32, 33].

Despite the growing interest in this mode of LC, considerable scientific debate exists about the physical basis of the separation mechanism in HP-HILIC [34]. The roles of thermodynamic or kinetic effects in controlling resolution and separation efficiencies have yet to be fully explored. Although it was proposed by Alpert [23] that the retention of polar compounds in HILIC is through partitioning between the bulk of the mostly organic mobile phase and a stagnant water-enriched layer semi-immobilized on the surface of silica, so far the retention mechanism(s) in HP-HILIC are still a matter of discussion; processes of partitioning or adsorption, or combinations of, have been suggested to be responsible for generating retention in HP-HILIC [23, 30, 35, 36].

Hodges *et al.* have substantively enhanced the understanding of mixed-mode HILIC/cation-exchange chromatography (HILIC/CEX) by applying the contact region concept developed for HP-RPC [15, 37] to rationalize the retention of amphipathic  $\alpha$ -helical peptides [38–40]. The authors have compared a poly(2-sulfoethyl aspartamide)-silica (PolySulfoethyl A) strong cation-exchange column in the HILIC/CEX mode with a Zorbax SB300-C8 reversed-phase column for separating amphipathic  $\alpha$ -helical peptides. A substitution in the hydrophilic face of the peptide resulted in a substantive effect on retention in HP-HILIC/CEX, this was not observed in HP-RPC, whereas a substitution in the hydrophobic face of the peptide resulted in a distinct effect on retention in HP-RPC not observed in HP-HILIC.

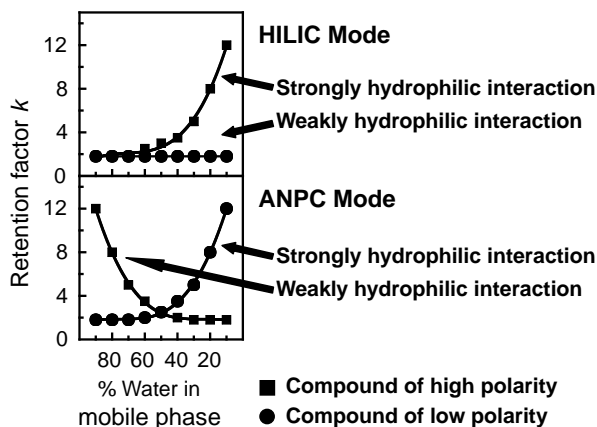
HP-HILIC can be operated in isocratic, step-gradient, or gradient elution mode, where the retaining mobile phase is organic and the eluting mobile phase is aqueous. Being more suited to the isolation of polar substances, HP-HILIC, when linked to

ESI-MS, has mainly found application for analysis of phosphopeptides [41–46] and glycopeptides [43–46].

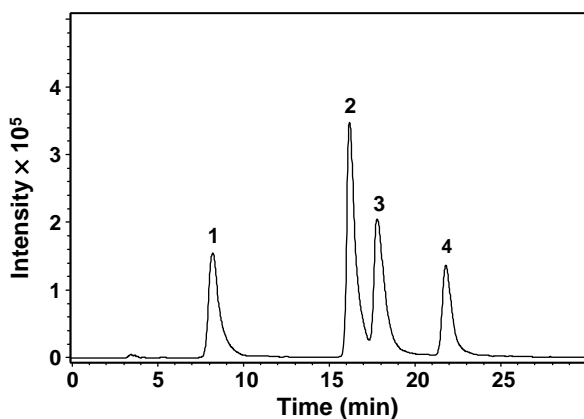
#### 5.4.5 ANPC

Recently, a new chromatographic mode, HP-ANPC, has been developed on stationary phases based on silica hydride surfaces [47–49]. One unique feature of silica hydride stationary phases is their ability to be employed over a broad range of mobile-phase compositions from 100% aqueous to pure organic solvents [50]. The retention principle in HP-ANPC is analogous to that found in HP-NPC, but the mobile phase contains water as part of the binary eluent. The difference between HP-ANPC and HP-HILIC is that in HP-HILIC retention is determined by an adsorbed water layer on the surface, which is either not present or substantially smaller on silica hydride surfaces employed in HP-ANPC.

A similar debate compared to that in HP-HILIC exists with respect to the precise separation mechanisms that operate in HP-ANPC. Since interchangeable use of the terms HP-HILIC and HP-ANPC has led to some confusion in the literature [30, 36], a recommendation has recently been put forward for separate descriptions for HILIC and ANPC stationary phases based on their specific separation capabilities and choice of mobile-phase compositions [51]. According to this rationale, the term HP-HILIC should describe the chromatographic mode that employs stationary phases capable of separating polar/ionic compounds with eluents of high organic solvent content, whereby compounds of low hydrophilicity are not significantly retained (Figure 5.4a). This behavior is in accord with theoretical predictions and modeling studies that were recently undertaken [52]. In contrast, the term HP-ANPC has been proposed to describe the chromatographic mode that allows the separation of compounds with a



**Figure 5.4** Scenarios for the HILIC retention mode and the ANPC retention mode for the separation of compounds of different hydrophilicities. The top plot (called Figure 5.4(a) in the text) refers to the HILIC mode, whilst the bottom plot (called Figure 5.4(b) in the text) refers to the ANPC mode.



**Figure 5.5** Aqueous normal-phase separation of polar peptides. Column: silica hydride Diamond from Higgins Analytical ( $0.30 \times 150$  mm, 4 mm); peptides: (1) Asp-Ala, (2) Arg-Gly-Asp, (3) Gly-Gly-His, (4) Gly-Arg-Ala-Asp-Ser-Pro-Ly s; eluent A: 0.1% formic acid in ACN, eluent B: 0.1% formic acid in water; gradient: 0.0–2.0 min 80% A, 2.0–15.0 min to 60% A, 15.0–15.1 min to 30% A, 15.1–30.0 min hold 30% A; flow rate: 4  $\mu$ L/min (Yang, Matyska, Pesek, Boysen and Hearn, unpublished results).

broad range of hydrophilicities in the same mixture using predominantly water-rich mobile-phase compositions (Figure 5.4b). An illustrative example of the separation of polar peptides with ANPC is seen in Figure 5.5.

#### 5.4.6

#### HIC

HP-HIC is used for protein separations involving hydrophobic sorbents. In HP-HIC the binding of proteins to the stationary phase occurs at high salt concentrations of the mobile phase and elution is performed through a decrease in salt concentration. HP-HIC separates proteins according to their hydrophobicity differences, and is based on the reversible interaction between a protein and a hydrophobic surface of a chromatographic support. Since protein–hydrophobic surface interactions are enhanced by high ion strength buffer solutions, HP-HIC is a suitable next step after ammonium sulfate precipitation or HP-IEX elution with high-salt buffer. Although the concept of HIC goes back to the work of Tiselius in 1948 [53, 54], it was not until the work of Shaltiel [55, 56], Hjertén [57] and Hofstee [58] in the early 1970s that HIC became important for the separation of peptides and proteins with stationary phases containing chemically immobilized hydrophobic ligands. The chromatographic sorbents in HP-RPC are typically octyl or octadecyl ligands immobilized onto, for example, nonporous or porous silica; however, the chromatographic sorbents in HP-HIC are more weakly hydrophobic due to the immobilized propyl, butyl, and phenyl groups [59–63]. In HP-HIC, similar to HP-RPC, the hydrophobic interaction between the peptide or protein and the nonpolar ligands immobilized onto the surface of the sorbent represents the dominant effect. The physicochemical

basis of separation selectivity for both HP-RPC and HP-HIC is associated with changes of microscopic surface tension involved with solute stationary-phase interactions [15, 64]. In both modes peptides and proteins are eluted by lowering the surface tension of the mobile phase. In HP-HIC, this is achieved with a decreasing salt concentration (i.e., by increasing the water content of the eluent), but in HP-RPC the decrease in surface tension of the eluent is achieved through an increase in the organic solvent content of the mobile phase. These differences between HP-RPC and HP-HIC have fundamental effects on the recovery of proteins in the bioactive state, as well as on the selectivity of the system.

The selectivity of protein separations in HP-HIC can be influenced through stationary-phase parameters (such as type and density of the immobilized hydrophobic ligand type of base matrix) and mobile phase parameters (such as type and concentration of salt [65], the addition of organic modifiers or surfactants [66], and the pH) as well as the temperature of the column.

In HP-HIC the two most widely used base matrices are hydrophilic carbohydrates (e.g., cross-linked agarose) or synthetic copolymer materials. The type of immobilized ligand (alkyl or aryl ligand) influences the selectivity of the stationary phase [55, 58]. In HP-HIC, nonpolar ligands with lower hydrophobicity and lower ligand density (about  $1/10^{\text{th}}$  of that of HP-RPC sorbents) are employed. The protein-binding capacities of HP-HIC sorbents increases with increasing *n*-alkyl chain length at a given ligand density [67] as well as with increasing ligand density, but will reach a plateau for very high ligand density levels. HP-HIC sorbents should be selected on the basis of the critical hydrophobicity concept [52, 63].

In HIC, the hydrophobic interactions between polypeptides or proteins and a sorbent are influenced by the use of salts in the mobile phase (various salts have differing molal surface tension increment values, see Table 5.5), by the use of different types or concentrations of salt. The surface tension of the mobile phase,  $\gamma$ , can be related to the molal surface tension increment,  $\sigma$ , and the molal concentration,  $m$ , of the salt by the equation  $\gamma = \gamma^{\circ} + \sigma m$ , where  $\gamma^{\circ} = 72$  dyn/cm for water.

Table 5.6 gives the corresponding parameters of the surface tension increment  $\sigma$ , the initial mobile phase concentration  $m$ , and the surface tension values  $\gamma$  of common aqueous salt buffers used in HP-HIC.

The minimum surface tension reached in the HP-HIC of polypeptides or proteins with binary water/salt systems corresponds to the surface tension of pure water (i.e., 72 dyn/cm). The effect of different salts on the hydrophobic interaction follows the Hofmeister (lyotropic) series for the precipitation of proteins from aqueous solution [68, 69], see Table 5.7.

The salts at the beginning of the series promote hydrophobic interactions and protein precipitation (salting-out effect) [70]. These are called antichaotropic (or kosmotropic) salts and are considered to be water structuring. The addition of kosmotropic salts to the equilibration buffer and sample buffer promote protein immobilized ligand interaction in HIC [71]. The salts at the end of the series (salting-in or chaotropic ions) randomize the structure of liquid water and tend to decrease the strength of hydrophobic interactions. The chloride anion is considered to be approximately neutral with respect to the water structure. In general, the effect of

**Table 5.5** Salts used in the mobile phase of HP-HIC.

Salt type	Molal surface tension increment $\sigma \times 10^3$ dyn-g/cm-mol
Calcium chloride	3.66
Magnesium chloride	3.16
Potassium citrate	3.12
Sodium sulfate	2.73
Potassium sulfate	2.58
Ammonium sulfate	2.16
Magnesium sulfate	2.10
Sodium dihydrogen phosphate	2.02
Potassium tartrate	1.96
Sodium chloride	1.64
Potassium perchloride	1.40
Ammonium chloride	1.39
Sodium bromide	1.32
Sodium nitrate	1.06
Sodium perchlorate	0.55
Potassium thiocyanate	0.45

**Table 5.6** Parameters of the surface tension of aqueous salt buffers.

Salt buffer type	$\sigma \times 10^3$ dyn-g/cm-mol	$m \times 10^3$ mol/g	$\gamma$ dyn/cm
Ammonium sulfate	2.16	2	77.31
Sodium chloride	1.64	2	76.29
Magnesium sulfate	2.1	1.4	75.95
Sodium sulfate	2.73	1	75.74
Sodium perchlorate	0.55	2	74.11
Sodium phosphate	2.02	0.05	73.01

**Table 5.7** Hofmeister series ranking the effect of anions and cations in promoting protein precipitation (ions with higher salting-out effect promote binding to hydrophobic interaction sorbents, whereas ions with higher salting in-effect promote elution from hydrophobic interaction sorbents).

Anions	← Increasing salting-out effect $\text{PO}_4^{3-}$ , $\text{SO}_4^{2-}$ , $\text{CH}_3\text{COO}^-$ , $\text{Cl}^-$ , $\text{Br}^-$ , $\text{NO}_3^-$ , $\text{ClO}_4^-$ , $\text{I}^-$ , $\text{SCN}^-$
Cations	$\text{NH}_4^+$ , $\text{K}^+$ , $\text{Na}^+$ , $\text{Li}^+$ , $\text{Mg}^{2+}$ , $\text{Ca}^{2+}$ , $\text{Ba}^{2+}$ Increasing salting-in effect →

the salt cations on the preferential interaction parameters is not as pronounced as the salt anions, in particular when the cation is monovalent, however divalent cations tend to bind to proteins [70]. Typically, kosmotropic (antichaotropic) salts (i.e., ammonium sulfate, sodium sulfate, magnesium chloride) of high molal surface tension increment are to be preferred in HP-HIC applications for polypeptides and proteins.

In combination with nondenaturing mobile phases, proteins can potentially be eluted in their native conformation from HP-HIC sorbents.

#### 5.4.7

#### IEX

HP-IEX is performed on stationary phases with immobilized charged ligands; separation occurs according to electrostatic interactions between the charged surface of the analyte(s) and the complementarily charged surface of the sorbent, and allows high-resolution, high-capacity separations. Peptides and proteins can be eluted in HP-IEX by either isocratic or by gradient elution [72–79]. In anion-exchange chromatography (HP-AEX), peptides and proteins are separated according to their net negative charge, whereby the retaining mobile phase is aqueous, of high pH and low salt concentration, and the eluting mobile phase is either aqueous, of high pH and high salt concentration, or aqueous, and of low pH. In contrast, HP-CEX separates analytes according to their net positive charge, whereby the retaining mobile phase is aqueous, of low pH and low salt concentration, and the eluting mobile phase is either aqueous, of low pH and high salt concentration, or aqueous, and of high pH.

The “net charge” concept has been widely used as a predictive basis to anticipate the retention behavior of proteins with both, anion and cation exchange stationary phases [80, 81]. According to this model, a protein will be retained on a cation-exchange column if the eluent pH is lower than the *pI* value of the protein, since under these conditions the protein will carry positive net charges. Conversely, a protein will be retained on an anion-exchange column when the eluent pH is above the *pI* of the protein. Finally, with a mobile phase of a pH that equals the *pI* of the protein, the surface of the protein can be considered as electrostatically neutral and the protein should not be retained on either cation- or anion-exchange columns.

Due to the insights into the relationships between protein structure, surface topography, and chromatographic retention made in the last two decades, this classical model is now considered as simplistic. Recent investigations into the effect of ionic strength, pH, buffer type and the concentration of counter- and co-ion on the chromatographic behavior on various proteins have revealed [73, 82–89], that the magnitude of electrostatic interactions between a protein and the stationary phase surface in HP-IEX is dependent on (i) the number and distribution of charged sites on the protein molecule defining its surface topography and electrostatic contact area with the stationary phase, (ii) the charge density of the stationary phase, and (iii) the mobile phase composition. As a consequence, a variation of chromatographic parameters can alter the affinity of the protein for the stationary phase through changes in the overall electrostatic surface charge or through specific electrostatic interactions of the displacer co- and counter-ions with surface charge groups on the

protein or with the immobilized charged ligand. In addition, possible changes of the three-dimensional structure of the proteins may have significant effects on protein retention. Studies on the influence of the experimental parameters on the number of charged interactive sites of the proteins involved in its binding to stationary phases have resulted in the development of the concept of an electrostatic interactive area (or ionotope) through which the protein is thought to bind to the stationary phase [90].

For peptide and protein separations, the use of a strong cation-exchange column has a considerable advantage over other ion-exchange separation modes [91], since this column can retain its negative charge character over a large pH range, from acidic to neutral. In peptides and proteins, at neutral pH, the side-chain carboxyl groups of the acidic amino acid residues (glutamic acid and aspartic acid) are completely ionized. Below pH 3 they are almost completely protonated. A change of pH therefore allows the retention of peptides and proteins to be varied according to the modified net charge of these biosolute(s). Both weak and strong cation exchangers (e.g., based on carboxymethyl or sulfonopropyl ligands), as well as weak and strong anion exchangers (e.g., dimethylamino or quaternary ammonium ligands) are commercially available, and are very suitable for the HP-IEX of peptides and proteins.

#### 5.4.8

##### AC

HP-AC is performed on stationary phases containing immobilized biomimetic or biospecific ligands and separates proteins according to principles of molecular recognition. HP-AC can be used for initial capture of proteins, as an intermediate step in a multiple-step purification procedure, or for the affinity removal of unwanted high abundance proteins, provided a suitable affinity ligand available for the target protein is available. HP-AC is highly selective and has usually a high capacity for the protein of interest. In HP-AC, analytes are eluted by step-gradient or gradient elution, where the capture (loading) mobile phase is aqueous and of low ionic strength, and the eluting mobile phase is aqueous and of higher ionic strength or of different pH value, or alternatively contains a mobile phase additive that competes with the target compound for binding to the immobilized biospecific ligand. HP-AC separations can be performed with immobilized chemical or biological ligands, or with molecular imprinted polymers. In achieving maximal selectivity and highest affinity in the interaction between the target substance(s) and the chromatographic sorbent, HP-AC excels all other modes, but each affinity sorbent must be tailored to the specific target compound. Nevertheless, HP-AC has found application in peptide and protein isolation.

IMAC exploits the affinities of the side-chain moieties of specific surface-accessible amino acids in peptides and proteins for the coordination sites of immobilized transition metal ions [92–94]. The majority of investigations employed tri- or tetradentate ligands, such as iminodiacetic acid (IDA), nitrilotriacetic acid (NTA), tris(carboxy methyl)ethylene-diamine (TED), *O*-phosphoserine (OPS), or carboxymethylaspartic acid (CMA) [93]. The retaining mobile phases are aqueous with neutral pH and high ionic strength, the eluting mobile phases are of low pH, contain competing ligands or EDTA. Novel immobilized chelate systems, such as 1,4,7-



**Table 5.8** Chromatographic mode employed in peptide and protein separation and their stationary and mobile phase characteristics.

Chromatographic mode	Stationary phase	Retaining mobile phase	Eluting mobile phase
SEC or GPC	porous	(nonretentive)	aqueous, low salt
RPC	hydrophobic	aqueous	organic solvents
NPC	polar	nonpolar organic solvents	polar organic solvents
HILIC	hydrophilic	nonpolar organic solvents	polar organic solvents, water
ANPC	polar	organic	aqueous
HIC	mildly hydrophobic	aqueous, high ionic strength	aqueous, low ionic strength
AEX	charged	aqueous, high pH, low ionic strength	aqueous, high pH, high ionic strength (or low pH), high selectivity counter ion
CEX	charged	aqueous, low pH, low ionic strength	aqueous, low pH, high ionic strength (or high pH)
AC	biomimetic, biospecific	low ionic strength	high ionic strength, competing ligand
IMAC	metal chelate	aqueous, neutral pH, high ionic strength	low pH, competing ligand, EDTA

triazolo-cyclononane (TACN), however, show different chromatographic properties compared to the IMAC behavior of traditional chelating ligands [94]. These AC techniques, in conjunction with soft gel matrices, have been applied in diverse analytical and preparative protein purifications. Novel procedures to immobilize an IMAC ligand at the surface of silica supports have provided guidelines for the design of very stable HP-IMAC systems for peptides and proteins [95]. Recently, various applications of different IMAC systems for the separation of peptides and proteins have been summarized [14, 38, 92, 96, 97].

Table 5.8 summarizes the nature of the retaining and eluting mobile phases for the above described chromatographic modes. Guidelines for how to choose a mode or a combination of modes for a specific separation task are given in Section 5.6.

## 5.5

### Method Development from Analytical to Preparative Scale Illustrated for HP-RPC

As noted above, HP-RPC is currently the most frequently used high-performance liquid chromatographic mode for the analysis and preparative purification of peptides and proteins, in particular for applications that involve off-line or on-line ESI-MS. The development of a method for preparative HP-RPC purification for the purpose of isolation of one or more component(s) from a peptide or protein product sample (or alternatively the purification of a synthesized product) is usually per-

formed in four steps: (i) development, optimization and validation of an analytical method, (ii) scaling up of this method to a preparative chromatographic system, (iii) application of the preparative method to the fractionation of the product, and, finally, (iv) analysis of the individual fractions.

### 5.5.1

#### Development of an Analytical Method

The development of an analytical method for the separation of a peptide or protein encompasses the selection of the stationary and mobile phase taking into consideration the analyte properties (hydrophobicity/hydrophilicity, acid–base properties, charge, temperature stability, molecular size), and is followed by a systematic optimization of the (isocratic or gradient) separations, using either aliquots of the crude extract or, if available, analytical standards.

In the selection of the stationary and mobile phase, a variety of chemical and physical factors of the chromatographic system that may contribute to variation in the resolution and recovery of peptides and proteins need to be considered. The stationary phase contributions relate to the ligand composition, ligand density, surface heterogeneity, surface area, particle size, particle size distribution, particle compressibility, pore diameter, and pore diameter distribution. The mobile phase contributions relate to the type of organic solvents, eluent composition, ionic strength, pH, temperature, loading concentration, and volume.

Typically, a particular HP-RPC material will be selected empirically as the starting point for the separation, taking into consideration its suitability for the separation task at hand, published procedures for similar types of peptides/proteins, availability of the stationary phase material for preparative chromatography, and if information is available on the analyte properties.

Since the quality of a separation is determined by resolution of individual peak zones, method development always aims at optimization of the resolution. The resolution of adjacent peak zones for a two analyte system can be defined as:

$$R_s = \frac{t_{R2} - t_{R1}}{(1/2)(w_1 + w_2)} \quad (5.6)$$

where,  $t_{R1}$  and  $t_{R2}$  are the retention times, while  $w_1$  and  $w_2$  are the peak widths, of two adjacent peaks corresponding to the analytes. To develop good resolution in the analytical separation of a complex mixture of peptides, method development always focuses on the least-well-resolved peak pair(s) of interest.

In isocratic elution, resolution depends on the column efficiency or plate number  $N$ , the selectivity  $\alpha$ , and the retention factor  $k$ , all of which can be experimentally influenced through systematic changes in individual chromatographic parameters. In the isocratic mode of separation, resolution is determined from:

$$R_s = (1/4)N^{1/2}(\alpha - 1)(k/(1 + k)) \quad (5.13)$$

As detailed above, the plate number  $N$  is the efficiency of the column and is a measure of the column performance. The selectivity  $\alpha$  describes the selectivity of a

chromatographic system for a defined peak pair and is the ratio of the  $k$  values of the second peak to the first peak. The retention factor  $k$  is a dimensionless parameter and is defined as  $k = (t_R - t_0)/t_0$ , where  $t_R$  is the retention time of a particular peak and  $t_0$  is the column void time. In this manner, normalization of the relative retention can be achieved for columns of different dimensions. While  $N$  and  $\alpha$  change only slightly during the solute migration through the column, the value of  $k$  can be readily manipulated through changes in the elutropicity of the mobile phase by a factor of 10 or more. The best chromatographic separations for low- or mid-molecular-weight analytes are generally achieved with mobile phase – stationary phase combinations that result in a  $k$  value between 1 and 20.

In gradient elution, in contrast to isocratic elution,  $\bar{N}$ ,  $\bar{\alpha}$ , and  $\bar{k}$  are the median values for  $N$ ,  $\alpha$ , and  $k$ , since they change during the separation as the shape and duration of the gradient changes. The “gradient” plate number  $\bar{N}$  has no influence on the selectivity or the retention (except for temperature change). The selectivity  $\bar{\alpha}$  and the retention factor  $\bar{k}$  usually have only a minor influence on  $\bar{N}$ . While  $\bar{N}$  and  $\bar{\alpha}$  change only slightly during the solute migration through the column, the  $\bar{k}$  value can change by a factor of 10 or more depending on the gradient steepness. Again, the best chromatographic separation is generally achieved with a  $\bar{k}$  value between 1 and 20. Although resolution in isocratic and gradient elution is mainly influenced by the mobile phase variables  $\alpha$  (or  $\bar{\alpha}$ ) and  $k$  (or  $\bar{k}$ ) and nearly independent of  $\bar{N}$ , for a given column, an optimization strategy should nevertheless start with appropriate selection of the stationary phase. This is because many initial choices (e.g., column dimensions, choice of ligand, etc.) are determined by the overall strategy (i.e., separation optimization for quantification of several analytes or separation optimization for planned scaling up to preparative purification of specific target compounds) and by the purification goals. A number of computer-assisted, expert systems can be used to guide this selection; for further insight into this field and the choice of different algorithmic expert systems approaches see I *et al.* [98].

Based on these considerations and in view of the implication of Eq. (5.13), the separation optimization for a peptide sample requires three steps to be performed: (a) the optimization of the column efficiency  $N$ , then (b) optimization of the selectivity  $\alpha$ , and, finally, (c) optimization of the retention factor  $\bar{k}$ -values.

**(a) Optimization of the column efficiency  $N$ .** The optimization of the peak efficiency, expressed as the theoretical plate number,  $N$ , requires an independent optimization of each of the contributing factors that influence the band-broadening of the peak zones due to column and the extracolumn effects. With a particular sorbent (ligand type, particle size, and pore size) and column configuration, this can be achieved through optimization of linear velocity (flow rate), temperature, detector time constant, and column packing characteristics, and by minimizing extracolumn effects (e.g., by using zero-dead volume tubing and connectors). The temperature of the column and the eluents should be thermostatically controlled in order to facilitate the reproducible determination of the various column parameters and to ensure resolution reproducibility. The flow rate (or, alternatively, the linear flow velocity) to achieve the minimum plate height,  $H$ , for a particular column can

be taken from the literature or experimentally determined according to published procedures.

**(b) Optimization of the selectivity  $\alpha$ .** Change in selectivity of the separation is the most effective way to influence resolution. This is mainly achieved by changing the chemical nature or concentration of the organic solvent modifier (acetonitrile, methanol, isopropanol, etc.) in conjunction with the appropriate choice of mobile phase additive(s). As noted above, this can be realized in both isocratic or gradient elution. Moreover, the interconversion of isocratic data to gradient data and vice versa can be achieved through the use of algorithms [98] based on linear and nonlinear solvent strength theory. However, if different organic solvents are used, different eluotropic strengths must be considered in order to allow elution of the analytes of the sample in the appropriate retention factor range [99, 100]. Once the selectivity parameter is fixed due to the initial choices of the mobile and stationary phase, further optimization should concentrate on resolution optimization via achieving the most appropriate retention factor for the different peptides in the mixture.

**(c) Optimization of the retention factor  $k$ -values.** In the isocratic elution mode of HP-RPC, resolution optimization can take advantage of the relationship between the retention time of an analyte (expressed as the retention factor  $k$ ) and the volume fraction of the organic solvent modifier,  $\varphi$ . Although typically these dependencies are curvilinear (i.e., not first order), for practical convenience they are often treated as linear relationships. Thus, the change in retention factor as a function of  $\varphi$  can be represented by:

$$\ln k = \ln k_0 - S\varphi \quad (5.14)$$

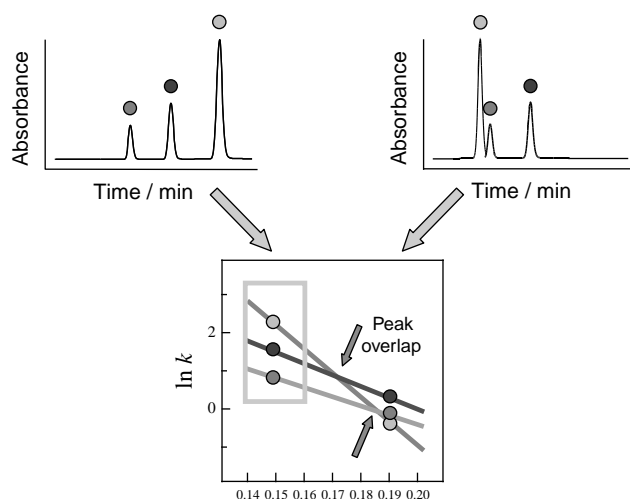
where  $k_0$  is the retention factor of the solute in the absence of the organic solvent modifier and  $S$  is the slope of the plot of  $\ln k$  versus  $\varphi$ . The values of  $\ln k_0$  and  $S$  can be calculated by linear regression analysis. Greater precision in the quality of fit of the experimental data, and thus improved reliability in the prediction of the retention behavior of analytes in HP-RPC systems for mobile phases of different solvent composition, can be achieved [14] through the use of an expanded form of Eq. (5.14):

$$\ln k = \ln k_0 - S\varphi + S'\varphi^2 - S''\varphi^3 + \dots \quad (5.15)$$

Similarly, in gradient elution HP-RPC, resolution optimization can take advantage of the relationship between the gradient retention time of an analyte (expressed as the median retention factor  $\bar{k}$ ) and the median volume fraction of the organic solvent modifier,  $\bar{\varphi}$ , in regular HP-RPC systems based on the concepts of the linear-solvent-strength theory [15, 101], such that:

$$\ln \bar{k} = \ln k_0 - S\bar{\varphi} \quad (5.16)$$

A mapping of the dependence of analyte retention (expressed as the natural logarithm of the retention factor,  $k$ ) on the mobile phase composition (expressed as the volume fraction of solvent in the mobile phase,  $\varphi$ ) in isocratic elution (or as  $\bar{k}$  versus  $\bar{\varphi}$  in gradient elution) with a minimum of two initial experiments, can be used to define the useful range of mobile phase conditions, and can



**Figure 5.6** Optimization of isocratic elution. Two chromatograms obtained for 19 and 14% (v/v) of organic solvent modifier in the mobile phase (corresponding to  $\varphi = 0.19$  and  $0.14$ , respectively) can be used to plot the resulting

retention factors versus the volume fraction of the organic solvent modifier in order to identify the mobile phase composition with optimal peak spacing (boxed).

indicate the mobile phase composition at which the band spacing is optimal, see Figure 5.6.

Irrespective of whether the data are obtained through isocratic or gradient elution techniques employing two initial experiments (differing only by their mobile phase composition or gradient run times, respectively) with tracking and assignment of the peaks, a relative resolution map (RRM) can be established, which plots resolution  $R_S$  against the separation time (or gradient run time  $t_G$ ). In the case of gradient elution, the RRM then allows determination of the optimal gradient run time (and gradient range). Such a procedure can be conveniently performed in any laboratory using Excel spreadsheets containing the relevant equations (see below) as macros or through software packages (e.g., DryLab, LabExpert, etc.). Such strategies greatly reduce the time to achieve an optimal separation, as well as saving on solvent, reagent and analyte consumption. Moreover, as shown in various studies, the more sophisticated of these methods permit [98] instrumentation to be operated in a nearly fully automated, unattended fashion 24 h/7 days per week.

In more advanced applications, optimization can be performed via computer simulation software (e.g., Simplex methods, multivariate factor analysis programs, DryLab G/plus, LabExpert, etc.). In such procedures, resolution of peak zones is optimized through systematic adjustment of mobile phase composition by successive change in the  $\varphi$ -value (or equivalent parameters, such as the concentration of the ion-pairing reagent employed). In gradient elution, advantage is taken of a strategy with the following eight steps: (i) performing of initial experiments; (ii) peak tracking and assignment of the peaks; (iii) calculation of  $\ln k_0$ - and  $S$ -values from initial

chromatograms; (iv) optimization of gradient run time  $t_G$  over the whole gradient range; (v) determination of new gradient range; (vi) calculation of new gradient retention times  $t_g$ ; (vii) change of gradient shape (optional); and, finally, (viii) verification of results.

**(i) Initial experiments.** In initial experiments, the peptide sample is separated using two linear gradients differing by a factor of 3 in their gradient run times (all other chromatographic parameters being held unchanged) to obtain the HP-RPC retention times for each of the peptide compounds [102]. Irrespective of what optimization strategy will be used, it is advisable to separate any sample with at least two different gradient run times, in order to identify overlapping peaks. For optimization of gradient shape and to achieve maximum resolution between adjacent peak zones, the ability to determine retention times of the peptides and to classify the parameters that reflect the contributions from the mobile phase composition and column dimensions is essential [103–105]. The determination of the volume,  $V_{\text{mix}}$ , is useful [106], determination of dead volume and gradient delay are crucial [101].

With various inputs regarding stationary and mobile phase parameters, algorithms such as that of DryLab G/plus can generate the RRM, based on calculation of corresponding  $S$ - and  $k_0$ -values for each component. If no computer program is available the resolution information can be plotted directly from the distances of the individual peak zones of adjacent peak pairs ( $R_s = t_{R2} - t_{R1}$ ) against the gradient run time  $t_G$ .

**(ii) Peak tracking and assignment of the peaks.** Complex chromatograms from reversed-phase gradient elution can often exhibit changes in peak order when the gradient steepness is changed. Before  $\ln k_0$ - and  $S$ -values are calculated, or computer simulation is used, the peaks from the two initial runs need to be correctly assigned. Several approaches to peak tracking have been described, using algorithms based on relative retention and peak areas [107], or alternatively, based on diode-array detection [108, 109].

**(iii) Calculation of  $\ln k_0$ - and  $S$ -values.** The retention times  $t_{g1}$  and  $t_{g2}$  for a solute separated under conditions of two different gradient run times ( $t_{G1}$  and  $t_{G2}$ , where  $t_{G1} < t_{G2}$ ) can be given by [103, 110]:

$$t_{g1} = \left(\frac{t_0}{b_1}\right) \log(2.3k_0 b_1) + t_0 + t_D \quad (5.17a)$$

and:

$$t_{g2} = \left(\frac{t_0}{b_2}\right) \log(2.3k_0 b_2) + t_0 + t_D \quad (5.17b)$$

with:

$$\frac{b_1}{b_2} = \frac{t_{G2}}{t_{G1}} = \beta. \quad (5.18)$$

where  $t_{G1}$  and  $t_{G2}$  are the gradient run time values of  $t_G$  for two different gradient runs, resulting in different values of  $b$  ( $b_1, b_2$ ), and  $t_g$  ( $t_{g1}, t_{g2}$ ) are the gradient retention times for a single solute in two different gradient runs;  $b_1$  and  $b_2$  are the gradient steepness parameters for a single solute over the two differing gradient run times;  $k_0$

$t_{G2}$  and  $t_{G1}$ , which is equivalent to the ratio of  $b_1$  and  $b_2$ ;  $t_0$  is the column dead time; and  $t_D$  is the gradient delay time. Steep gradients correspond to large  $b$ -values and small  $\bar{k}$ -values.

For small molecules there is an explicit solution [110] for  $b$  and  $k_0$ , namely:

$$b_1 = \frac{t_0 \log \beta}{\left[ t_{G1} - \left( \frac{t_{G2}}{\beta} \right) + (t_0 + t_D) \left( \frac{1-\beta}{\beta} \right) \right]} \quad (5.19)$$

and:

$$\log k_0 = \left( \frac{b_1}{t_0} \right) (t_{G1} - t_0 - t_D) - \log(2.3b_1) \quad (5.20)$$

From the knowledge of  $b$  and  $k_0$ , the values of  $\bar{k}$  and  $\bar{\varphi}$  can be calculated [101]:

$$\bar{k} = \frac{1}{1.15b_1} \quad (5.21)$$

$$\bar{\varphi} = \frac{\left[ t_{G1} - t_0 - t_D - \left( \frac{t_0}{\beta_1} \right) \log 2 \right]}{t_{G1}^0} \quad (5.22)$$

where  $\bar{k}$  is the value of  $k$  (retention factor) for a solute when it reaches the column midpoint during elution;  $\varphi$  is the volume fraction of solvent in the mobile phase;  $\Delta\varphi$  is the change in  $\varphi$  for the mobile phase during gradient elution ( $\Delta\varphi = 1$  for a 0–100% gradient);  $\bar{\varphi}$  is the effective value of  $\varphi$  during gradient elution and the value of  $\varphi$  at band center when the band is at the midpoint of column, and  $t_{G1}^0$  is the normalized gradient time with  $t_{G1}^0 = t_{G1}/\Delta\varphi$ .

By linear regression analysis, using  $\bar{k}$  and  $\bar{\varphi}$ , the  $S$ -value (empirically related to the hydrophobic contact area between solute and ligand) can be derived from the slope of the  $\log \bar{k}$  versus  $\bar{\varphi}$  plots, and  $\ln k_0$  (empirically related to the affinity of the solute towards the ligand) as the  $\gamma$ -intercept [15]:

$$S = (\ln k_0 - \ln \bar{k}) \bar{\varphi} \quad (5.23)$$

**(iv) Optimization of the gradient time,  $t_G$ , over the entire gradient range.** The retention factor  $\bar{k}$  is a linear function of the gradient run time  $t_G$  if  $\Delta\varphi$  is kept constant. Hence:

$$\frac{\bar{k}}{t_G} = \frac{0.87 F}{V_m \Delta\varphi S} = \text{const.} = C \quad (5.24)$$

The optimized gradient run time  $t_{GRRM}$  can be obtained from the RRM or alternatively, from the plot of  $R_S$  versus  $t_G$ , and yields for each analyte the new values of  $\bar{k}_{\text{new}}$  by  $t_{GRRM}$  being multiplied by  $C$ :

$$C t_{GRRM} = \bar{k}_{\text{new}} \quad (5.25)$$

**(v) Determination of the new gradient range.** If the gradient run time  $t_{GRRM}$  is changed in relation to  $\Delta\varphi$  with  $t_0 = \text{const.}$ , the  $k$ -values do not change, as can be seen is the solute retention factor at the initial mobile phase composition;  $\beta$  is the ratio of

from the following equation:

$$t_{G1}^0 = \frac{t_{GRRM}}{\Delta\varphi} = \frac{V_m \bar{S}k}{0.87F} \quad (5.26)$$

where:

$$\Delta\varphi_{opt} = \frac{t_{Gopt}}{t_{G1}^0} \quad (5.27)$$

and where the retention time  $t_g$  of the first peak is greater than  $(t_0 + t_D)$  and the retention time  $t_g$  of the last peak is less than  $\Delta t_{Gopt}$ .

**(vi) Calculation of the new gradient retention times  $t_g$ .** Based on the knowledge of the  $S$ - and  $\ln k_0$ -values, new gradient retention times can then be calculated.

**(vii) Change of gradient shape (optional).** Multisegmented gradients should only be used, once the gradient delay has been measured. With multisegmented gradients, an error in the gradient delay will reoccur at the beginning and end of each gradient step. In addition, the effect of  $V_{mix}$  (which can be determined according to the procedures described in [103]), which modifies the composition of the gradient at the start and end (rounding of the gradient shape), can lead to deviation of the experimentally determined retention times from the predicted “ideal” values as derived, for example, with DryLab G/plus simulations.

**(viii) Verification of the results.** After completion of the optimization process, the simulated chromatographic separation can now be verified experimentally using the predicted chromatographic conditions.

Examples where such systematic method development has been used for the analytical separation of peptides and proteins can be found in studies performed on ribosomal proteins [111].

## 5.5.2

### Scaling Up to Preparative Chromatography

While analytical HPLC aims at the quantification and/or identification of compounds (with the sample going from the detector to waste), preparative chromatography aims at the isolation of compounds (with the sample going to the fraction collector). For preparative separations, method development always focuses on the peaks of interest and the two adjacent eluting peaks. In many cases, all other peaks can be viewed as superfluous and are directed to the waste. Optimization of the resolution of the peak of interest from the adjacent peaks has to take into account the sample size and the relative abundances of the three components that form the basis of the separation task. Once an analytical method is established, it can be scaled up [12, 112] to a preparative separation by taking into consideration the operating ranges of the column (Table 5.9) or used for scaling up by deliberate column overloading.

The concept of parity in scaling up or down implies that the performance features, selectivity behavior, and recyclability of the stationary phase material used for the



**Table 5.9** Operating ranges of column types in HPLC with inner column diameter (ID), column lengths, flow rate range, and range of sample quantity.

Column type	Column ID (mm)	Column lengths (mm)	Flow rate range (ml/min)	Sample quantity range
Preparative	>4	15–250	5–20	mg–g
Analytical	2–4	15–250	0.2–1	μg–mg
Capillary	1	35–250	0.05–0.1	μg
Nano	<1	50–150	<0.05	ng–μg

analytical and the preparative separations are identical, with the exception of particle size. Both robust experimental methods as well as rules-of-thumb, acquired by experienced investigators, have been developed that enable such comparisons to be made. An extensive scientific literature is now available to indicate sound foundations for such scaling-up strategies, coupled with suitable experimental methods for their validation. Table 5.9 summarizes some of this information.

In order to obtain an equivalent elution profile, the flow rate needs to be adjusted for columns with different internal diameter, according to:

$$F_{\text{preparative}} = \left[ \frac{r_{\text{preparative}}}{r_{\text{analytical}}} \right]^2 \times F_{\text{analytical}} \quad (5.28)$$

where  $F$  is the flow rate and  $r$  is the column radius of the preparative or analytical column.

Estimates of the loading capacity of a particular column material can usually be obtained from the manufacturer. The mass loadability for a scaled-up separation can be calculated with:

$$M_{\text{preparative}} = \left[ \frac{r_{\text{preparative}}}{r_{\text{analytical}}} \right]^2 \times M_{\text{analytical}} \times C_L \quad (5.29)$$

where  $M$  is the mass,  $r$  is column radius of the preparative or analytical column, and  $C_L$  is the column length ratio.

In many cases, despite some loss of resolution, column overloading is an economic and viable method for compound purification. In analytical LC, the ideal peak shape is a Gaussian curve. If under analytical conditions a higher amount of sample is injected, peak height and area change, but not peak shape or the retention factor. However, if more than the recommended amount of sample is injected onto the column the adsorption isotherm becomes nonlinear. As a direct consequence, resolution decreases and peak retention times and peak shapes may change. There are two methods of column overloading – concentration overloading and volume overloading. In concentration overloading the volume of the injected sample is maintained, while the sample concentration is increased. The retention factors of the compound(s) decrease, and the peak shape may become triangular and tailing. The applicability of this method is limited by the solubility of the target compound(s) in the mobile phases employed. In volume overloading, the concentration of the sample

is maintained, but the sample volume is increased. The retention factor of the compound(s) increase(s) in the volume overload mode, with broadened peak shape. Once a suitable method is established, it can be applied to the preparative purification of the target compound(s).

### 5.5.3

#### Fractionation

There are four types of fraction collection [113]: (i) manual, with a manually pressed button to start and stop collection, (ii) time based, with a fraction collecting during fixed preprogrammed time intervals, (iii) peak based, based on a chosen threshold of the up- and down-slope of a detector signal, and (iv) mass based, with fraction collection occurring only if the specific mass of a trigger ion is detected by MS. In addition, a recovery collection can be performed, in which everything that is not collected as a fraction goes into a dedicated container where it can be easily recovered. Whatever the type of fraction collection, careful attention has to be given to the fraction collection delay times and a delay time measurement performed. For a peak with start time  $t_0$  and end time  $t_E$ , fraction collection needs to be started when the start of the peak arrives at the diverter valve ( $t_0 + t_{D1}$ ) and ended, when the end of the peak arrives at the needle tip ( $t_E + t_{D1} + t_{D2}$ ), where  $t_{D1}$  is the delay time between detector and valve and  $t_{D2}$  is the delay time between valve and needle tip.

### 5.5.4

#### Analysis of the Quality of the Fractionation

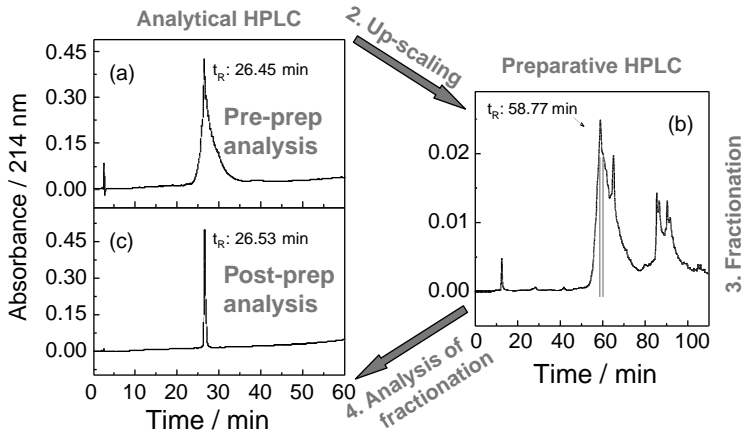
In the absence of on-line MS, fractionation is usually accompanied by an off-line mode of quality analysis. The typical workflow comprises (i) a preparative (e.g., prior to any fractionation) analysis of the unpurified material, (ii) the purification/fractionation of the compound, and (iii) a postpreparative analysis of the individual fractions. The pre- and postpreparative analysis can be performed with analytical HPLC, MS, and activity testing of the fractions, if an assay is available (Figure 5.7).

After the fractions have been collected, the solvent needs to be removed either by using a freeze dryer, rotary evaporator, or high-throughput parallel evaporator. Nonvolatile components can be removed with reversed-phase SPE procedures prior to solvent removal if the aqueous portion of the buffer is sufficiently large.

## 5.6

### Multidimensional HPLC

Although HPLC is a powerful separation technique for the fractionation of peptides from complex biological mixtures, very often more than one chromatographic step is necessary to achieve a required degree of purity of the target compounds. In practice, this is achieved through a series of purification steps. As there are material losses

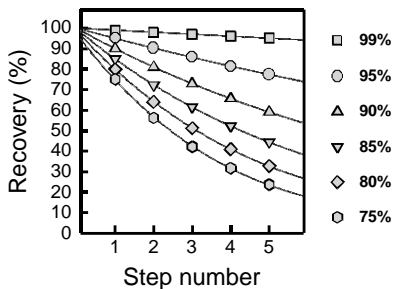


**Figure 5.7** Prepreparative analysis (a), preparative fractionation (b), and postpreparative analysis (c) with HP-RPC for the preparative isolation of synthetic polypeptides NH<sub>2</sub>-RIC(Acm)DKC(Acm)KVIRRHGRVYVIC(Acm)ENPKHKQ-COOH [114]. The preparative analysis of crude synthetic polypeptide (a) and postpreparative analysis of the purified product (c) was achieved with a TSK-ODS-120 T column (4.6 × 150 mm ID,

120 Å, 5 μm, endcapped) using a flow rate of 1 ml/min and a 60 min gradient from 100% A (0.1% TFA in water) to 100% B (90% acetonitrile 0.09% TFA), pH 2.1. The preparative fractionation (70 mg) of the crude mixture (b) was obtained with a TSK-ODS-120 T column (21.5 × 300 mm ID, 120 Å, 10 μm, endcapped) using a flow rate of 7.5 ml/min and a 90 min gradient.

associated with each purification step in these procedures (Figure 5.8) the overall recovery of the product has to be optimized. This can be achieved if the number of employed purification steps is minimized. Thus, strategies and techniques that reduce the number of unit operations are to be preferred since they will lead to minimization of product loss(es), and save on capital costs for equipment, reagents, and other consumables or operational costs for staff.

After initial extraction, the enrichment and purification of the target compound can be achieved typically in two to five chromatographic steps using a combination of different chromatographic modes. For peptides and proteins, various combinations



**Figure 5.8** Overall recovery at a fixed recovery per step value with an increased number of additional steps in multidimensional chromatography.

of chromatographic modes, including IEX-RPC, SEC-RPC, HILIC-RPC, RPC-HILIC, and AC-RPC combinations, are described in the literature.

Multidimensional (multistage, multicolumn) (MD)-HPLC offers the possibility of cutting the elution profiles into consecutive fractions, where these fractions can be treated independently from each other. One important consequence of this strategy is the gain in peak capacity, defined as the number of peaks that can be accommodated between the first and the last peak in a separation of defined resolution [115]. MDLC has developed from column switching and related techniques [116] for a specific target or class of targets [117]. MD-HPLC has the potential of independent optimization of the separation conditions for each fraction and allows a relative enrichment/depletion/peak compression of components. An advanced conceptual framework of MDLC has been developed for small-molecule separations [118–122]. MD-HPLC can be applied to the purification of a particular peptide or comprehensive fractionation of complex peptide mixtures.

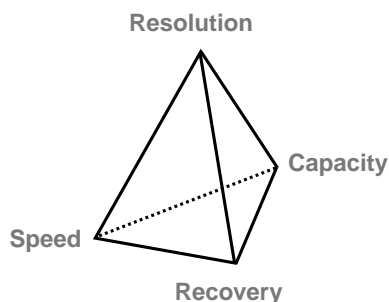
### 5.6.1

#### **Purification of Peptides and Proteins by MD-HPLC Methods**

To purify a particular peptide or protein, it is often possible to select complementary chromatographic modes that allow the target compound to be obtained in high purity with only a few separation steps, preferably three steps or less. In such noncomprehensive MD-HPLC, only a part of the analytes (as a single fraction) eluting from the first column is transferred to a second column for further purification (conventionally expressed by a hyphen, i.e., IEX-RPC). For such a “heart-cutting” technique, knowledge of the retention properties of the analyte mixture in the first column is needed in order to choose the segment(s). Heart-cutting techniques are fast, but not comprehensive, since the majority of analytes are not subjected to a separation in a second dimension. The main advantage of the technique is the improved resolution of compounds that coelute in the first dimension. A key requirement for such a purification scheme is that subsequent stages of the separation are orthogonal, with the two separation modes not correlated to each other in relation to their retention characteristics (i.e., selectivity).

For a single chromatographic dimension, the partly contradictory objectives of speed, resolution, capacity and recovery can usually not be maximized simultaneously (i.e., a high resolution can be achieved but at the expense of speed; a high speed separation can reduce resolution, etc.) (Figure 5.9).

A three-stage MD-HPLC protein purification process allows meeting the overall purification objectives by placing the emphasis for each purification stage on a different pair of objectives and choosing a chromatographic mode which is particularly suited for the task (Figure 5.10). At the enrichment or capture stage, the emphasis is on speed and capacity, employing HP-AC, HP-IMAC, HP-IEX, or HP-HIC, possibly in a SPE format as a low resolution step. This stage aims at the initial isolation of the target product from the crude sample, at its concentration and also at the removal of major or critical contaminants. At the intermediate purification stage, emphasis is placed on capacity and resolution, employing chromatographic modes



**Figure 5.9** Optimization goals for a chromatographic purification and their interrelationship.

with intermediate resolution (i.e., HP-IEX, HP-HIC, or HP-SEC). This stage has the objectives to remove the majority of impurities (i.e., other proteins, nucleic acids, viruses, and endotoxins). At the final chromatographic purification step, the emphasis is placed on resolution and recovery, employing high-resolution modes (i.e., HP-RPC). This stage strives to remove trace amounts of impurities or closely related compounds to obtain the final pure product. Therefore, HP-SEC may also be used as final polishing step (i.e., in order to remove unwanted multimeric forms of the target protein).

In some cases, the capture and intermediate purification, or for that matter intermediate purification and final polishing step may be achievable with a single separation step, resulting in a two-stage purification process. In other cases, such as for the purification of therapeutic proteins, four or more stages may be required to achieve the desired degree of protein purity. The selection of purification techniques for capture, intermediate purification, and final purification, and their intelligent combination is therefore vital for an efficient separation process.

Purification stage	Separation mode	Priority
<b>Enrichment (capture) stage:</b> rapid, high capacity, low resolution	AC, IMAC, IEX, HIC	
<b>Intermediate purification stage:</b> high capacity, high resolution, intermediate speed and recovery.	IEX, HIC, SEC	
<b>Final purification stage:</b> high resolution and recovery, low capacity and speed.	RPC, SEC	

**Figure 5.10** Optimization priorities at individual stages for a multidimensional, three-stage purification process after initial sample extraction exemplified for peptides and proteins and suitable chromatographic modes.

## 5.6.2

**Fractionation of Complex Peptide and Protein Mixtures by MD-HPLC**

If the objective of a purification scheme is the comprehensive fractionation of a complex, multicomponent peptide mixture, it is of advantage to use orthogonal chromatographic modes, but such extensive fraction collection requires additional, sometimes substantial, infrastructure. In comprehensive MD-HPLC, the entire analyte pool of the first column is transferred to the second column (expressed by a cross, i.e., IEX  $\times$  RPC) as sequential aliquots, either successively onto one column or alternating onto two parallel columns. The resulting data can be represented as three-dimensional contour plots, with retention times of the second dimension plotted against retention times of the first dimension. The information content of such comprehensive two-dimensional chromatograms is higher than the information content of individual one-dimensional chromatograms. The first comprehensive two-dimensional system was developed in the late 1970s by Erni and Frei [123]. In the subsequent decades, comprehensive MD-HPLC methods have been further developed, mainly for peptides and proteins [124, 125]. The theoretical aspects of MD-HPLC techniques have also been further developed [126–128].

## 5.6.3

**Operational Strategies for MD-HPLC Methods**

From an operational perspective, MDLC can be carried out off-line or on-line [129]. Regardless of which operational mode, off-line or on-line, is used, the compatibility of the mobile phases between successively employed chromatographic modes in a separation scheme needs to be considered (see Section 5.6.4.2). As a consequence, it may be necessary to process the fractions between two separation stages (e.g., through buffer exchange, concentration, or dilution) to enhance compatibility of eluent composition of fractions from the first chromatographic dimension with the retaining mobile phase of the second chromatographic dimension. If a nonretentive chromatographic mode such as SEC is employed in conjunction with a retentive chromatographic mode, such as RPC or IEX, it is usually performed first. This allows (i) relatively large eluent volumes stemming from isocratic elution in the nonretentive mode to be reduced through the capture of analytes under the retaining mobile phase conditions of the subsequent retentive chromatographic mode and (ii) reduction of extracolumn band broadening with resulting loss of resolution.

**5.6.3.1 Off-line Coupling Mode for MD-HPLC Methods**

The off-line coupling mode in HPLC is comparable with that employed in conventional (open) column chromatography in peptide and protein isolation. In the off-line mode, the eluent of the first column is collected as fractions, either manually or with an automated fraction collector, and reinjected onto the second column. Typical processing steps may include volume reduction by freeze-drying or automated high-

throughput parallel evaporation systems taking into account the boiling point(s) or volatility of the target analyte(s) and organic solvent if these are contained within the eluates. The use of volatile mobile phase additives then allows a buffer exchange.

#### 5.6.3.2 On-Line Coupling Mode for MD-HPLC Methods

The on-line mode uses high-pressure, multiposition, multiport switching valves, which allow selection of pathways for single fractions from the first chromatographic dimension to subsequent column(s) of the second chromatographic dimension. The fractions from the first dimension are either transferred directly, or through one (or more) intermediate trapping columns for the purpose of concentration and automated buffer exchange. This approach requires complex instrumentation, and results in increased optimization time and reduced system flexibility; however, it has numerous advantages in terms of reproducibility, recovery, speed, and automation.

#### 5.6.4

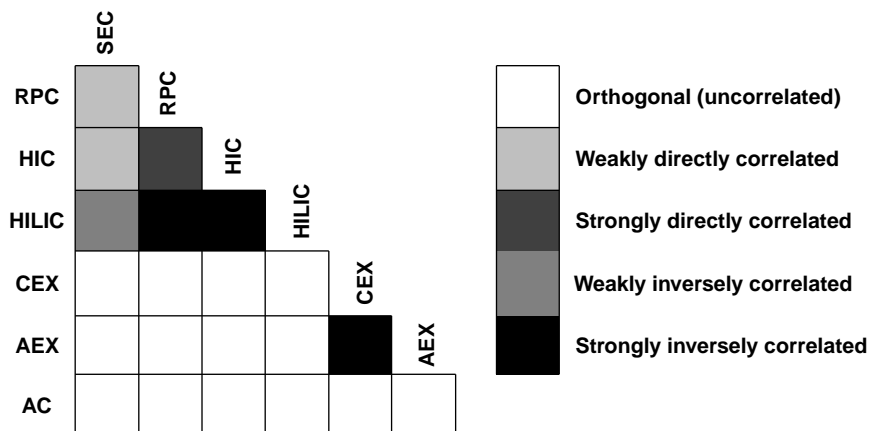
##### Design of an Effective MD-HPLC Scheme

MD-HPLC for peptides and proteins requires thoughtful selection of orthogonal and complementary separation modes, and of the order of their utilization and independent optimization in respect to the chromatographic goals (speed, resolution, capacity, and recovery). Furthermore, besides the mobile phase composition of the employed chromatographic modes, the elution mode (isocratic, step, or gradient elution), flow rates and, mobile phase temperatures need to be considered.

##### 5.6.4.1 Orthogonality of Chromatographic Modes

In order to exploit the full peak capacity of a two-dimensional system [122], it is advantageous if the applied chromatographic modes are orthogonal. It is generally accepted that the dimensions in a two-dimensional separation system are orthogonal, if the separation mechanism of the two dimensions are independent from each other causing the distribution of analytes in the first dimension to be uncorrelated to the distribution in the second dimension. An example of such orthogonality is LC-capillary electrophoresis (LC-CE), where totally different separation mechanisms are used (i.e., pressure-driven compared to electrically driven separation) [130]. In a similar manner different separation modes in HPLC can be viewed as being orthogonal (e.g., ion-exchange chromatography (CEX or AEX) and RPC are orthogonal as they separate according to net charge or hydrophobicity, respectively). A very coarse classification of chromatographic modes commonly applied in the MD-HPLC of peptides and proteins according to their separation principles is depicted in Figure 5.11.

In MD-HPLC systems, combinations of chromatographic modes are usually designed to achieve analyte separation according to different characteristic analyte properties [131].



**Figure 5.11** Degree of orthogonality of major chromatographic modes employed in the separation of peptides and proteins. Shading indicates the degree of correlation of the separation principles of paired modes.

For an ideal orthogonal, two-dimensional separation, the overall peak capacity,  $PC$ , is defined as the product of the peak capacities in each dimension:

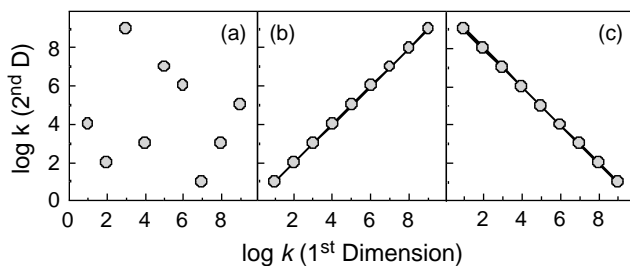
$$PC_{2D\text{-system}} = PC_{\text{first dimension}} \times PC_{\text{second dimension}} \quad (5.30)$$

However, if two nonidentical chromatographic modes with some degree of similarity are used in a two-dimensional system, the increase in the peak capacity and the total number of analytes that can be separated is much lower than the product of peak capacities of individual dimensions. The peak capacity also depends on the elution mode. Gradient elution provides a higher peak capacity than isocratic elution and is of advantage in two-dimensional LC.

It should be noted that since selectivity in chromatography depends not only on the stationary phase, but also on the mobile phase, orthogonal separations can be achieved through fine-tuning of the separation conditions, even if the principal separation mechanisms of both dimensions are similar. Such tuning removes the inaccessible area from the two-dimensional retention plane and ensures that the remaining retention space is used efficiently [126].

In addition, the structure of analytes has an effect on the peak capacity. In many separation systems, the contribution of structural units, especially the repeating units, to the Gibbs free energy of association of the analytes with the immobilized chromatographic ligands are additive [132]. Such structural repeating units can be hydrophobic or polar. If one chromatographic system in a two-dimensional LC has no selectivity for a structural element, then the first and the second dimension are noncorrelated (orthogonal) with respect to the repeating structural unit (Figure 5.12a). In a completely correlated separation system, with correlated retention factors in the two dimensions, the separations space is not utilized (Figure 5.12b). Such two-dimensional systems do not provide sufficient selectivity





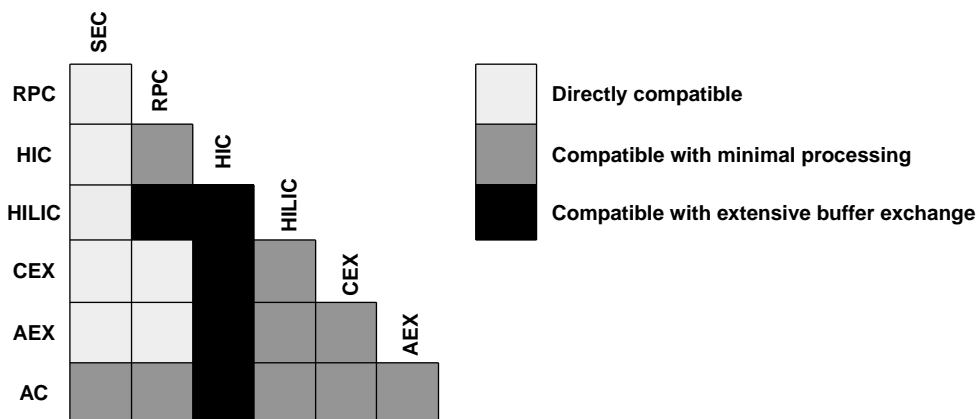
**Figure 5.12** Two-dimensional separation space for a set of peptides and proteins utilizing separation systems that are: (a) uncorrelated, (b) completely correlated, and (c) inversely correlated, where the retention factors obtained in the second dimension are plotted versus the retention factors obtained in the first dimension.

for the separation in respect to the structural property distribution of interest in either dimension and are generally not very useful in practice. In inversely correlated two-dimensional LC  $\times$  LC separation systems, the retention time increases in the first dimension, but decreases in the second dimension (Figure 5.12c). Neither correlated or inversely correlated two-dimensional LC  $\times$  LC increase the peak capacity significantly. The selectivity of a two-dimensional LC  $\times$  LC with respect to hydrophobic or polar repeating units can determine the suitability of chromatographic modes employed in two-dimensional separations and depends on the employed stationary as well as mobile phases. Orthogonal systems with noncorrelated selectivities provide the highest peak capacity and therefore the highest number of resolved peaks.

The peak capacity in two-dimensional LC  $\times$  LC decreases with increasing correlation of the selectivity between the first and the second chromatographic dimension. In practice, however, two-dimensional LC  $\times$  LC systems are rarely fully orthogonal with respect to each structure distribution type (i.e., hydrophobic, polar) [127]. Many partially orthogonal systems are using only part of the theoretically available two-dimensional separation space, but can be evaluated using analytes differing in the numbers of hydrophobic or polar structural units or by quantitative structure retention relationships.

#### 5.6.4.2 Compatibility Matrix of Chromatographic Modes

In the design of two-dimensional LC  $\times$  LC systems, the selection of the mobile phase for each chromatographic dimension is of fundamental importance in order to achieve maximal utilization of the two-dimensional separation space. In contrast to off-line two-dimensional LC procedures, where the collected fraction can be subjected to evaporation, dilution, or extraction, before injection onto the column of the second dimension, the compatibility of the mobile phases in on-line two-dimensional LC  $\times$  LC in terms of miscibility, solubility, viscosity, and elutropic strength is much more important. The mobile phases used in SEC  $\times$  RPC, SEC  $\times$  HILIC, RPC  $\times$  CEX, RPC  $\times$  AEX, RPC  $\times$  CEX, that are compatible, are shown in Figure 5.13.



**Figure 5.13** Pair-wise comparison of compatibility between common chromatographic modes. Compatibility is based on miscibility, solubility, and eluotropic strength for a particular class of peptides and proteins.

## 5.7

### Conclusions

Due to the enormous growth in capability and separation power that has occurred over the past two decades, the benefits of HPLC in peptide and protein chemistry may now seem obvious. However, as there is an immense choice of modes and procedures, further scope exists to improve the quality of such separations and achieve even higher resolutions based on even more efficient optimization procedures. For these reasons, a comprehensive overview of the principles and limitations of contemporary separation methods in various steps of purification and analysis of peptides and proteins has been presented at the beginning of this chapter.

In order to solve analytical problems for a particular compound or class of compounds, as well as to save time and resources, it is essential that systematic method development concepts are applied. Such methods then enable a successful scaling up to preparative purifications as well as the design and application of MD-HPLC purification schemes. Moreover, if used in conjunction with de-replication procedures, these advances in high-resolution chromatographic methods may lead to new discoveries that can be used to advance science or medicine, and at the same time respect the environment through reduced solvent and reagent usage.

To this end, it will also be the responsibility of future generations of analytical scientists to ensure that the development of new separation methods occurs responsibly and sustainably. It is therefore expected that increasingly the analysis of peptides and proteins will use the principles of green analytical chemistry, considering the issues of waste minimization and hazard reduction. Similar criteria will also apply to preparative and process developments. Thus, there is tremendous potential for investigators to pursue new aspects of method development, which hopefully has been encouraged by this chapter.

## References

- 1 Zamyatnin, A.A. (1972) *Progress in Biophysics and Molecular Biology*, **24**, 107–123.
- 2 Chothia, C. (1975) *Nature*, **254**, 304–308.
- 3 Dawson, R.M.C., Elliot, D.C., Elliot, W.H., and Jones, K.M. (1986) *Data for Biomedical Research*, 3rd edn, Clarendon Press, Oxford.
- 4 Rickard, E.C., Strohl, M.M., and Nielsen, R.G. (1991) *Analytical Biochemistry*, **197**, 197–207.
- 5 Wilce, M.C.J., Aguilar, M.-I., and Hearn, M.T.W. (1995) *Analytical Chemistry*, **67**, 1210–1219.
- 6 Walla, P.J. (2009) *Modern Biophysical Chemistry – Detection and Analysis of Biomolecules*, Wiley-VCH Verlag GmbH, Weinheim.
- 7 Hostettmann, K., Hostettmann, M., and Marston, A. (1986) *Preparative Chromatography Techniques: Applications in Natural Product Isolation*, 1st edn. Springer, London.
- 8 Bidlingmeyer, B.A. (ed.) (1987), *Preparative Liquid Chromatography*, vol. **38**, Journal of Chromatography Library, Elsevier, Amsterdam.
- 9 Grushka, E. (ed.) (1989) *Preparative-Scale Chromatography*, vol. **46**, Chromatographic Science Series, Marcel Dekker, New York, N.Y.
- 10 Unger, K.K. (ed.) (1994) *Handbook of HPLC, Part 2: Preparative Liquid Column Chromatography*, GIT-Verlag, Darmstadt.
- 11 Hostettmann, K., Marston, A., and Hostettmann, M. (1997) *Preparative Chromatography Techniques: Applications in Natural Product Isolation*, 2nd edn. Springer, London.
- 12 Rathore, A.S. and Velayudhan, A. (eds) (2003) *Scale-Up and Optimization in Preparative Chromatography: Principles and Biopharmaceutical Applications*, vol. **88**, Chromatographic Science Series, CRC Press, Boca Raton, FL.
- 13 Boysen, R.I. and Hearn, M.T.W. (2001) *Current Protocols in Protein Science* (eds J.E. Coligan, B.M. Dunn, H.L. Ploegh, D.W. Speicher and P.T. Wingfield), John Wiley & Sons, Inc., New York, pp. 1–40.
- 14 Hearn, M.T.W. (2000) *Handbook of Bioseparations*, vol. 2 (ed. S. Ahuja), Academic Press, San Diego, CA, pp. 71–235.
- 15 Horvath, C., Melander, W., and Molnar, I. (1976) *Journal of Chromatography*, **125**, 129–156.
- 16 Horvath, C., Melander, W., and Molnar, I. (1977) *Analytical Chemistry*, **49**, 142–154.
- 17 Unger, K.K. (1979) *Porous Silica: Its Properties and Use as Support in Column Liquid Chromatography*, vol. 16, Journal of Chromatography Library, Elsevier, Amsterdam.
- 18 Snyder, L.R. (1970) *Methods in Medical Research*, **12**, 11–36.
- 19 Ballschmiter, K. and Wössner, M. (1998) *Fresenius Journal of Analytical Chemistry*, **361**, 743–755.
- 20 Yoshida, T. (1998) *Journal of Chromatography*, **808**, 105–112.
- 21 Buchholz, K., Gödelmann, I., and Molnar, I.J. (1982) *Journal of Chromatography*, **238**, 193–202.
- 22 Papadoyannis, I.N., Zotou, A.C., and Samanidou, V.F. (1995) *Journal of Liquid Chromatography and Related Technologies*, **18**, 2593–2609.
- 23 Alpert, A.J. (1990) *Journal of Chromatography*, **499**, 177–196.
- 24 Linden, J.C. and Lawhead, C.L. (1975) *Journal of Chromatography*, **105**, 125–133.
- 25 Yang, M., Thompson, R., and Hall, G. (2009) *Journal of Liquid Chromatography and Related Technologies*, **32**, 628–646.
- 26 Yoshida, T. (2004) *Journal of Biochemical and Biophysical Methods*, **60**, 265–280.
- 27 Mant, C.T. and Hodges, R.S. (2008) *Journal of Separation Science*, **31**, 2754–2773.
- 28 Yoshida, T. (1997) *Analytical Chemistry*, **69**, 3038–3043.
- 29 Boutin, J.A., Ernould, A.P., Ferry, G., Genton, A., and Alpert, A.J. (1992) *Journal of Chromatography*, **583**, 137–143.
- 30 Hemström, P. and Irgum, K. (2006) *Journal of Separation Science*, **29**, 1784–1821.

- 31 Gilar, M., Olivova, P., Daly, A.E., and Gebler, J.C. (2005) *Analytical Chemistry*, **77**, 6426–6434.
- 32 Naidong, W. (2003) *Journal of Chromatography B*, **796**, 209–224.
- 33 Nguyen, H.P. and Schug, K.A. (2008) *Journal of Separation Science*, **31**, 1465–1480.
- 34 Hao, Z., Xiao, B., and Weng, N. (2008) *Journal of Separation Science*, **31**, 1449–1464.
- 35 Zhu, B.-Y., Mant, C.T., and Hodges, R.S. (1991) *Journal of Chromatography*, **548**, 13–24.
- 36 Wu, J., Bicker, W., and Lindner, W. (2008) *Journal of Separation Science*, **31**, 1492–1503.
- 37 Fausnaugh-Pollitt, J., Thevenon, G., Janis, L., and Regnier, F.E. (1988) *Journal of Chromatography*, **443**, 221–228.
- 38 Mant, C.T., Litowski, J.R., and Hodges, R.S. (1998) *Journal of Chromatography*, **816**, 65–78.
- 39 Mant, C.T., Kondejewski, L.H., and Hodges, R.S. (1998) *Journal of Chromatography*, **816**, 79–88.
- 40 Hodges, R.S., Chen, Y., Kopecky, E., and Mant, C.T. (2004) *Journal of Chromatography A*, **1053**, 161–172.
- 41 McNulty, D.E. and Annan, R.S. (2008) *Molecular and Cellular Proteomics*, **7**, 971–980.
- 42 McNulty, D.E. and Annan, R.S. (2009) *Methods in Molecular Biology*, **527**, 93–105.
- 43 Häggglund, P., Bunkenborg, J., Elortza, F., Jensen, O.N., and Roepstorff, P. (2004) *Journal of Proteome Research*, **3**, 556–566.
- 44 Picariello, G., Ferranti, P., Mamone, G., Roepstorff, P., and Addeo, F. (2008) *Proteomics*, **8**, 3833–3847.
- 45 Wang, C., Jiang, C., and Armstrong, D.W. (2008) *Journal of Separation Science*, **31**, 1980–1990.
- 46 Wührer, M., de Boer, A.R., and Deelder, A.M. (2009) *Mass Spectrometry Reviews*, **28**, 192–206.
- 47 Pesek, J.J., Matyska, M.T., Hearn, M.T.W., and Boysen, R.I. (2009) *Journal of Chromatography A*, **1216**, 1140–1146.
- 48 Pesek, J.J., Matyska, M.T., Loo, J.A., Fischer, S.M., and Sana, T.R. (2009) *Journal of Separation Science*, **32**, 2200–2208.
- 49 Pesek, J.J. and Matyska, M.T. (2010) *Advances in Chromatography*, **48**, 255–288.
- 50 Pesek, J.J. and Matyska, M.T. (2005) *Journal of Separation Science*, **28**, 1845–1854.
- 51 Pesek, J. and Matyska, M.T. (2007) *LCGC North America*, **25**, 480–490.
- 52 Hearn, M.T.W. (2002) *HPLC of Biological Macromolecules*, 2nd edn (ed. K.M. Gooding and F.E. Regnier), Dekker, New York, pp. 99–245.
- 53 Tiselius, A. (1948) *Arkiv för Kemi, Mineralogi, Geologi*, **26B**, 5.
- 54 Shepard, C.C. and Tiselius, A. (1949) *Discussions of the Faraday Society*, **7**, 275–285.
- 55 Shaltiel, S. and Er-El, Z. (1973) *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 778–781.
- 56 Shaltiel, S. (1974) *Methods in Enzymology*, **34**, 126–140.
- 57 Hjerten, S. (1973) *Journal of Chromatography*, **87**, 325–331.
- 58 Hofstee, B.H.J. (1973) *Analytical Biochemistry*, **52**, 430–448.
- 59 Antia, F.D., Fellegvari, I., and Horvath, C. (1995) *Industrial and Engineering Chemistry Research*, **34**, 2796–2804.
- 60 Fausnaugh, J.L., Pfannkoch, E., Gupta, S., and Regnier, F.E. (1984) *Analytical Biochemistry*, **137**, 464–472.
- 61 Gooding, D.L., Schmuck, M.N., Nowlan, M.P., and Gooding, K.M. (1986) *Journal of Chromatography*, **359**, 331–337.
- 62 Melander, W.R., El Rassi, Z., and Horvath, C. (1989) *Journal of Chromatography*, **469**, 3–27.
- 63 Wu, S.-L., Benedek, K., and Karger, B.L. (1986) *Journal of Chromatography*, **359**, 3–17.
- 64 Melander, W.R., Corradini, D., and Horvath, C. (1984) *Journal of Chromatography*, **317**, 67–85.
- 65 Fausnaugh, J.L., Kennedy, L.A., and Regnier, F.E. (1984) *Journal of Chromatography*, **317**, 141–155.

- 66 Wetlaufer, D.B. and Koenigbauer, M.R. (1986) *Journal of Chromatography*, **359**, 55–60.
- 67 Hofstee, B.H.J. (1979) *Pure and Applied Chemistry*, **51**, 1537–1548.
- 68 Pahlman, S., Rosengren, J., and Hjerten, S. (1977) *Journal of Chromatography*, **131**, 99–108.
- 69 Melander, W. and Horvath, C. (1977) *Archives of Biochemistry and Biophysics*, **183**, 200–215.
- 70 Roettger, B.F., Myers, J.A., Ladisch, M.R., and Regnier, F.E. (1989) *Biotechnology Progress*, **5**, 79–88.
- 71 Porath, J., Sundberg, L., Fornstedt, N., and Olsson, I. (1973) *Nature*, **245**, 465–466.
- 72 Chang, S.-H., Noel, R., and Regnier, F.E. (1976) *Analytical Chemistry*, **48**, 1839–1845.
- 73 Kopaciewicz, W. and Regnier, F.E. (1983) *Analytical Biochemistry*, **133**, 251–259.
- 74 Kopaciewicz, W., Rounds, M.A., and Regnier, F.E. (1985) *Journal of Chromatography*, **318**, 157–172.
- 75 Kopaciewicz, W. and Regnier, F.E. (1986) *Journal of Chromatography*, **358**, 107–117.
- 76 Hearn, M.T.W., Hodder, A.N., and Aguilar, M.I. (1988) *Journal of Chromatography*, **458**, 27–44.
- 77 Heinitz, M.L., Kennedy, L., Kopaciewicz, W., and Regnier, F.E. (1988) *Journal of Chromatography*, **443**, 173–182.
- 78 Hodder, A.N., Aguilar, M.I., and Hearn, M.T.W. (1990) *Journal of Chromatography*, **506**, 17–34.
- 79 Regnier, F.E. (1984) *Methods in Enzymology*, **104**, 170–189.
- 80 Boardman, N.K. and Partridge, S.M. (1955) *Biochemical Journal*, **59**, 543–552.
- 81 Himmelhoch, S.R. (1971) *Methods in Enzymology*, **22**, 273–286.
- 82 Kopaciewicz, W., Rounds, M.A., Fausnaugh, J., and Regnier, F.E. (1983) *Journal of Chromatography*, **266**, 3–21.
- 83 Rounds, M.A. and Regnier, F.E. (1984) *Journal of Chromatography*, **283**, 37–45.
- 84 Gooding, D.L., Schmuck, M.N., and Gooding, K.M. (1984) *Journal of Chromatography*, **296**, 107–114.
- 85 Gooding, K.M. and Schmuck, M.N. (1985) *Journal of Chromatography*, **327**, 139–146.
- 86 Stout, R.W., Sivakoff, S.I., Ricker, R.D., and Snyder, L.R. (1986) *Journal of Chromatography*, **353**, 439–463.
- 87 Hearn, M.T.W., Hodder, A.N., Stanton, P.G., and Aguilar, M.I. (1987) *Chromatographia*, **24**, 769–776.
- 88 Hearn, M.T.W., Hodder, A.N., and Aguilar, M.I. (1988) *Journal of Chromatography*, **443**, 97–118.
- 89 Hodder, A.N., Aguilar, M.I., and Hearn, M.T.W. (1989) *Journal of Chromatography*, **476**, 391–411.
- 90 Aguilar, M.I., Hodder, A.N., and Hearn, M.T.W. (1991) *HPLC Proteins, Peptides and Polynucleotides* (ed. M.T.W. Hearn), VCH, New York, pp. 199–245.
- 91 Mant, C.T. and Hodges, R.S. (1985) *Journal of Chromatography*, **327**, 147–155.
- 92 Porath, J., Carlsson, J., Olsson, I., and Belfrage, G. (1975) *Nature*, **258**, 598–599.
- 93 Zachariou, M., Traverso, I., and Hearn, M.T. (1993) *Journal of Chromatography*, **646**, 107–120.
- 94 Jiang, W., Graham, B., Spiccia, L., and Hearn, M.T.W. (1998) *Analytical Biochemistry*, **255**, 47–58.
- 95 Wirth, H.-J. and Hearn, M.T.W. (1993) *Journal of Chromatography*, **646**, 143–151.
- 96 Jenö, P., Scherer, P.E., Manningkrieg, U., and Horst, M. (1993) *Analytical Biochemistry*, **215**, 292–298.
- 97 Litowski, J.R., Semchuk, P.D., Mant, C.T., and Hodges, R.S. (1999) *Journal of Peptide Research*, **54**, 1–11.
- 98 I, T.-P., Smith, R., Guhan, S., Taksen, K., Vavra, M., Myers, D., and Hearn, M.T.W. (2002) *Journal of Chromatography A*, **972**, 27–43.
- 99 Schoenmakers, P.J., Biliot, H.A.H., and Galan, L.D. (1979) *Journal of Chromatography*, **185**, 179–195.
- 100 Patel, H.B. and Jefferies, T.M. (1987) *Journal of Chromatography*, **389**, 21–32.
- 101 Snyder, L.R. (1980) *HPLC – Advances and Perspectives*, vol. 1 (ed. C. Horvath), Academic Press, New York, pp. 208–316.

- 102 Dolan, J.W., Lommen, D.C., and Snyder, L.R. (1989) *Journal of Chromatography*, **485**, 91–112.
- 103 Ghrist, B.F.D., Coopermann, B.S., and Snyder, L.R. (1988) *Journal of Chromatography*, **459**, 1–23.
- 104 Ghrist, B.F.D. and Snyder, L.R. (1988) *Journal of Chromatography*, **459**, 25–41.
- 105 Ghrist, B.F.D. and Snyder, L.R.J. (1988) *Journal of Chromatography*, **459**, 43–63.
- 106 Stadalius, M.A., Gold, H.S., and Snyder, L.R. (1984) *Journal of Chromatography*, **296**, 31–59.
- 107 Lankmayr, E.P., Wegscheider, W., and Budna, K.W. (1989) *Journal of Liquid Chromatography and Related Technologies*, **12**, 35–58.
- 108 Berridge, J.C. (1986) *Techniques for the Automated Optimization of HPLC Separations*, Wiley Interscience, Chichester.
- 109 Strasters, J.K., Billiet, H.A.H., de Galan, L., Vandeginste, B.G.M., and Kateman, G. (1989) *Journal of Liquid Chromatography*, **12**, 3–22.
- 110 Quarry, M.A., Grob, R.L., and Snyder, L.R. (1986) *Analytical Chemistry*, **58**, 907–917.
- 111 Boysen, R.I., Erdmann, V.A., and Hearn, M.T.W. (1998) *Journal of Biochemical and Biophysical Methods*, **37**, 69–89.
- 112 Mazzei, J.L. and d'Avila, L.A. (2003) *Journal of Liquid Chromatography and Related Technologies*, **26**, 177–193.
- 113 Rosentreter, U. and Huber, U. (2004) *Journal of Combinatorial Chemistry*, **6**, 159–164.
- 114 Boysen, R.I. and Hearn, M.T.W. (2000) *Journal of Biochemical and Biophysical Methods*, **45**, 157–168.
- 115 Martin, M. (1995) *Fresenius Journal of Analytical Chemistry*, **352**, 625–632.
- 116 Huber, J.F.K., Van der Linden, R., Ecker, E., and Oreans, M. (1973) *Journal of Chromatography*, **83**, 267–277.
- 117 Mondello, L., Bartle, K., and Lewis, A. (eds) (2001) *Multidimensional Chromatography*. John Wiley & Sons, Ltd, Chichester.
- 118 Huber, J.F.K., Kenndler, E., and Reich, G. (1979) *Journal of Chromatography*, **172**, 15–30.
- 119 Giddings, J.C. (1984) *Analytical Chemistry*, **56**, 1258A–1260A, 1262A, 1264A, 1266A, 1268A, 1270A.
- 120 Davis, J.M. and Giddings, J.C. (1985) *Analytical Chemistry*, **57**, 2168–2177.
- 121 Davis, J.M. and Giddings, J.C. (1985) *Analytical Chemistry*, **57**, 2178–2182.
- 122 Giddings, J.C. (1995) *Journal of Chromatography A*, **703**, 3–15.
- 123 Erni, F. and Frei, R.W. (1978) *Journal of Chromatography*, **149**, 561–569.
- 124 Simpson, R.J. (ed.) (2004) *Purifying Proteins for Proteomics: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 1–15.
- 125 Unger, K.K., Racaiyte, K., Wagner, K., Miliotis, T., Edholm, L.E., Bischoff, R., and Marko-Varga, G. (2000) *Journal of High Resolution Chromatography*, **23**, 259–265.
- 126 Liu, Z. and Lee, M.L. (2000) *Journal of Microcolumn Separations*, **12**, 241–254.
- 127 Jandera, P. (2006) *Journal of Separation Science*, **29**, 1763–1783.
- 128 Dugo, P., Cacciola, F., Kumm, T., Dugo, G., and Mondello, L. (2008) *Journal of Chromatography A*, **1184**, 353–368.
- 129 Majors, R.E. (1980) *Journal of Chromatographic Science*, **18**, 571–579.
- 130 Hearn, M.T.W. (2001) *Biologicals*, **29**, 159–178.
- 131 Guiochon, G., Beaver, L.A., Gonnord, M.F., Siouffi, A.M., and Zakaria, M. (1983) *Journal of Chromatography*, **255**, 415–437.
- 132 Martin, A.J.P. (1949) *Annual Reports on the Progress of Chemistry*, **45**, 2 67–283.

## 6

# Local Surface Plasmon Resonance and Electrochemical Biosensing Systems for Analyzing Functional Peptides

Masato Saito and Eiichi Tamiya

### 6.1

#### Localized Surface Plasmon Resonance (LSPR)-Based Microfluidics Biosensor for the Detection of Insulin Peptide Hormone

##### 6.1.1

##### LSPR and Micro Total Analysis Systems

A deeper understanding of noble metal nanostructure phenomena is extremely important to develop innovative nanosensors for biomolecular interactions. It is well established that the metal nanostructures of particles, pores, rods, and rings have been investigated previously for plasmon properties as well as highly sensitive and specific sensors for biological targets [1, 2]. Localized surface plasmon resonance (LSPR), which results from the matching between the frequencies of incident photons and the collective oscillations of the conductive electrons in the metal nanostructures, has been presented in the recent literature [3–8]. Theoretically, the absorbance sensing property of LSPR is dependent upon the size, shape, and spacing of nanostructures as well as their local environment [9–14].

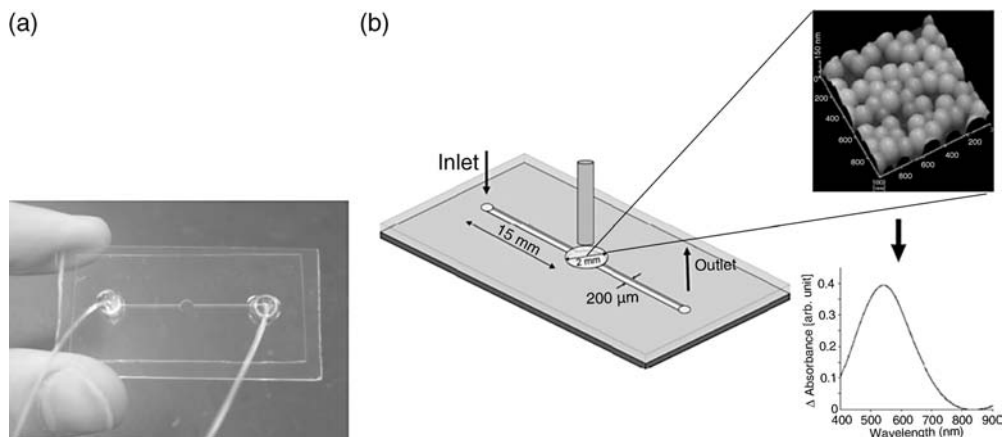
Given the current advances in biochip technology, micro total analysis systems ( $\mu$ TASs) [15] have received much attention for the highly efficient, simultaneously analysis of a number of important biomolecules from proteomics to genomics. If the LSPR-based chip, a surface detection tool for biomolecular interactions, is applied to a  $\mu$ TAS, this system will become more flexibly and widely used in analytical applications. Previously, molecular interactions would ordinarily have to be measured by surface plasmon resonance (SPR). However, due to the requirement of the Kretschmann configuration in total internal reflection mode, the planar SPR system has been faced with some drawbacks for lab-on-a-chip incorporation. Currently, a simple collinear optical system operated in transmission geometry without using the Kretschmann configuration has been reported for the label-free detection of antigen–antibody reactions and DNA–DNA hybridizations using LSPR-based nanochips [16–23]. Exploiting this advantage for  $\mu$ TASs, a microfluidic chip based on LSPR spectroscopy was proposed. On the basis of the characteristics of a previous chip, a polydimethylsiloxane (PDMS) microfluidic LSPR chip using the soft-lithog-

raphy technique has been fabricated. For evaluating the chip, the antibody–antigen reaction was performed to detect insulin – one of the most important indicators for diabetes diagnosis [23, 24]. This chip presents several advantages, such as real-time detection at low experimental cost with less reagent consumption, kinetic constant determination of antigen–antibody interaction, reduction of the total analysis time, and opening of the high potential for  $\mu$ TAS integration.

### 6.1.2

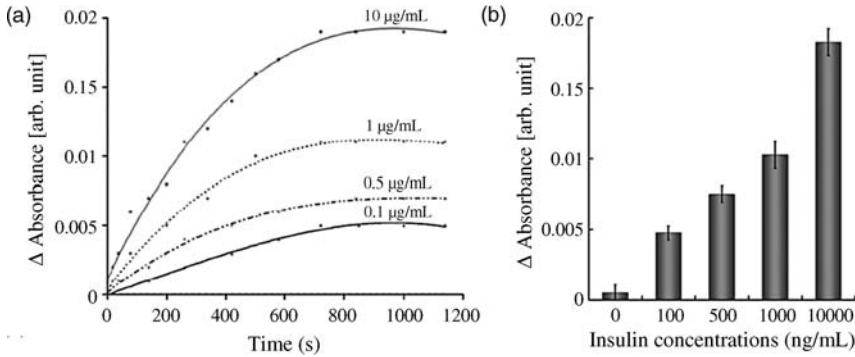
#### Microfluidic LSPR Chip Fabrication and LSPR Measurement

Microfluidic LSPR chips have a nano silica particle (100 nm diameter) monolayer on a glass substrate. A spot of LSPR substrate is created at the center of the glass slide with a diameter of 2 mm after deposition of a thin gold layer (30 nm) to cap the silica nanoparticles. Alternatively, a microfluidic device to cover the LSPR substrate has been fabricated using PDMS by a standard soft-lithography technique [25]. The microchannel has a width of 200  $\mu$ m, a height of 300  $\mu$ m and a length of 30 mm. LSPR measurements were performed using a set of instruments: spectrophotometer (USB-2000; wavelength range: 200–1100 nm), tungsten halogen light source (LS-1; wavelength range: 360–2000 nm), and optical fiber probe bundle (R-400-7 UV/V is: fiber core diameter: 200  $\mu$ m, wavelength range: 250–800 nm) (all from Ocean Optics). The microfluidic LSPR chip was placed proximally to the optical fiber probe bundle surface to satisfy that requirement that the incident light was reflected upon hitting the LSPR substrate surface, and coupled into a detection fiber probe and analyzed by the UV/Vis spectrophotometer over a wavelength range of 400–800 nm at room temperature as shown in Figure 6.1. Flow through the microfluidic LSPR chip was driven by the pressure from a microsyringe pump. A flow rate of 9  $\mu$ l/min was set up in all experiments. The absorbance data were plotted as functions of time and recorded by a PC using OOIBase32 software (Ocean Optics).



**Figure 6.1** Photograph of the fabricated microfluidic LSPR chip (a), and schematic illustration of the microfluidic LSPR chip and set-up (b).





**Figure 6.2** Real-time monitoring of insulin binding to anti-insulin antibody immobilized on the chip surface (a). Peak absorbance strength changes as a function of insulin concentrations (b).

### 6.1.3

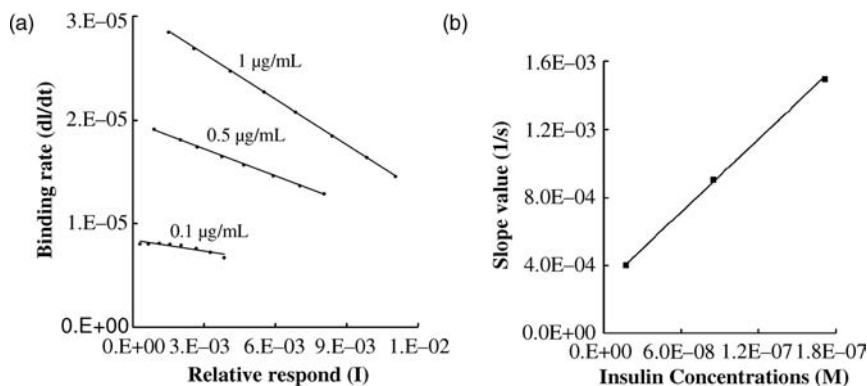
#### Detection of the Insulin–Anti-Insulin Antibody Reaction on a Chip

After immobilization of protein A on the LSPR spot surface through use of a chemical coupling agent [23], the anti-insulin antibody was captured by protein A. For label-free measurement of the antigen–antibody reaction, standard insulin solutions were introduced to the chip over 20 min and the peak absorbance intensity increases were recorded as a function of time. Figure 6.2(a) shows the representative plots of absorbance spectra response observed for the flow of insulin at different concentrations ranging from 0.1 to 10 μg/ml. In the absence of insulin, the absorbance spectrum was not changed due to the biomolecular binding event not occurring. In the presence of the insulin solution, insulin binds to the anti-insulin antibody on the microfluidic LSPR chip surface and thus the peak absorbance intensity was increased. These results also show that the absorbance intensity increases correspond to the increasing concentration of the insulin solution and the saturation value was attained after 12 min. The mean increases in peak absorbance intensity for binding of various insulin concentrations of 0.1, 0.5, 1, and 10 μg/ml were 0.0047, 0.0075, 0.0102, and 0.0175 ( $n = 4$ ) in succession, as shown in Figure 6.2(b). The results showed that the microfluidic LSPR chip could yield a limit of detection of 100 ng/ml insulin within the linear range from 0.1 to 10 μg/ml.

Moreover, the kinetic constants for the process of interaction of insulin and the anti-insulin antibody immobilized on the surface could be determined by linear transformation of sensograms [26]. The formation rate of the complex (antigen–antibody) at the time  $t$  based on the association rate constant ( $k_a$ ) and dissociation rate constant ( $k_d$ ) could be expressed by:

$$d[\text{Ag} - \text{Ab}]/dt = k_a[\text{Ag}][\text{Ab}] - k_d[\text{Ab} - \text{Ag}] \quad (6.1)$$

With the concentration of free antibody binding sites  $[\text{Ab}] = [\text{Ag} - \text{Ab}]_{\text{max}} - [\text{Ag} - \text{Ab}]$ , the following equation can be derived:



**Figure 6.3** Determination of individual rate constants in coordinates of Eq. (6.3) (a) and Eq. (6.4) (b).

$$d[\text{Ag}-\text{Ab}]/dt = k_a[\text{Ag}-\text{Ab}]_{\text{max}}[\text{Ag}] - (k_a[\text{Ag}] + k_d)[\text{Ab}-\text{Ag}] \quad (6.2)$$

The formation of the complex causes the increase in the absorbance intensity of LSPR in direct proportion to the refractive index of the variations in solute concentrations. The maximum absorbance intensity ( $I_{\text{max}}$ ) responds corresponding to the saturation of the available binding sites:

$$dI/dt = k_a[\text{Ag}]I_{\text{max}} - (k_a[\text{Ag}] + k_d)I \quad (6.3)$$

where:

$$k = k_a[\text{Ag}] + k_d \quad (6.4)$$

Figure 6.3(a and b) presents linearization data for the interaction process in coordinates of Eqs. (6.3) and (6.4). The kinetics constants calculated from these primary data were  $k_a = 7.18 \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$  and  $k_d = 3 \times 10^{-4} \text{ s}^{-1}$ . The overall affinity constant  $K$  ( $k_a/k_d$ ) was thus calculated to be  $2.39 \times 10^7 \text{ M}^{-1}$ .

These results suggest that the microfluidic LSPR chip could be used to transduce the biomolecular binding at the surface into an absorbance change with a required sensitivity for biosensor applications. Using the microfluidic LSPR chip, it is possible to measure the biomolecular interactions in real-time and calculate the kinetic constants using only a single optical fiber. Additionally, a possible advantage of the microfluidic LSPR chip is assumed to be the study of cell-transducing signals when this chip is connected to the cell culturing chamber. This characteristic is a great additional advantage compared to the previous LSPR systems, which were performed under air conditions. Although, the current detection limit is sufficient for many practical applications, a lower detection limit can be attained by optimization of the detection chemistry, such as surface density, competitive immunoassay, and so on.

## 6.2

### Electrochemical LSPR-Based Label-Free Detection of Melittin

#### 6.2.1

##### Melittin and E-LSPR

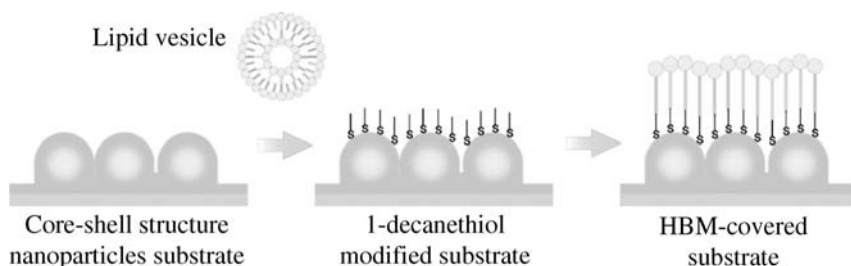
Pore-forming peptide toxins are cytolytic toxins that act on a plasma membrane for the purpose of permeabilizing the host cells [27, 28]. Melittin – a non-cell-selective lytic peptide from the venom of the honey bee – has a direct effect on human erythrocyte lysis with the perturbation of the membrane [29], leading to hemoglobin leakage [30]. Importantly, melittin can exhibit its pore-forming effect on a mimic of natural membranes, providing a feasibly alternative way to detect and study some of the desirable properties in artificial systems [31, 32]. A number of techniques to investigate the interactions of melittin toxin and the membrane have been detailed in the recent literature, including SPR [33, 34], electrochemical impedance spectroscopy [35], second harmonic generation [36], and the cantilever array sensor [37]. The combination of SPR and electrochemistry on a planar gold surface – electrochemical SPR – achieved a highly sensitive, reliable complementary analysis of the toxin–membrane interactions under fully identical experimental conditions [38, 39]. However, the requirement of the Kretschmann configuration in its total internal reflection mode limits the possibilities for massively parallel detection in a miniaturized package as well as lab-on-a-chip incorporation.

The surface plasmon band excitation of the core–shell nanostructure substrates in a simple collinear optical system has been previously demonstrated as the potential model to overcome these limitations [16–23]. Different from the fabrication of gold nanoparticles in solution, this nanostructure was suitable to process in various flexible formats – an important property to develop completed biochips in analytical and biosensor applications. In its construction, the silica nanoparticles were used as the “core,” and thin gold films were used as the “shell” coated at the bottom and the top of the “core”. The excitation mode of the plasmon absorption spectra of the core–shell nanoparticle structure could be controlled by changing the size of the silica nanoparticle and the shell thickness of the gold layer. Additionally, the LSPR microfluidic format was also developed to give the possibility to integrate LSPR measurement into a  $\mu$ TAS [23]. These previous achievements encouraged us to apply these core–shell nanoparticle structures to develop an electrochemical localized surface plasmon resonance E-LSPR sensor for the possible detection of peptide toxin.

#### 6.2.2

##### Fabrication of E-LSPR Substrate and Formation of the Hybrid Bilayer Membrane

The approach for fabricating the core–shell structure nanoparticle substrate has been well described [17, 20]. After cutting the silicon wafer, 5 nm of chromium and 30 nm of gold were deposited. Silica nanoparticles were formed on the gold substrate, producing the nanoparticle monolayer. After deposition of a thin gold-cap layer, this

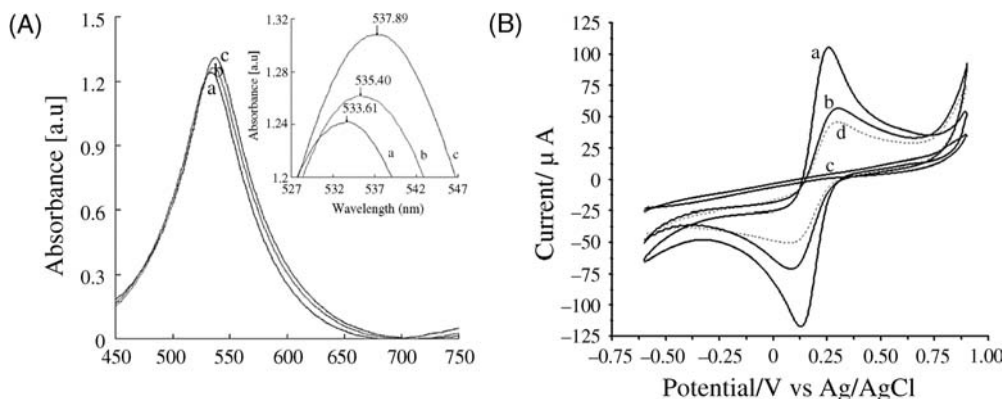


**Figure 6.4** Fabrication of the membrane-based sensor using the core-shell structure nanoparticle substrate.

core-shell structure nanoparticle substrate could be used simultaneously as a working gold electrode and LSPR-exciting device. All electrochemical measurements were performed by using a three-electrode system with a platinum wire as the counter electrode and an Ag/AgCl electrode as the reference electrode.

The lipid vesicles were prepared by dissolving dimyristoylphosphatidylcholine (DMPC) in pure chloroform. The organic solvent was then removed with a  $N_2$  stream to form a thin lipid layer and the samples were kept continuously in a vacuum desiccator for 12 h. An amount of a 0.01 M phosphate-buffered saline (PBS; pH 7.5) containing 0.1 M NaCl was added, giving a final lipid concentration of 0.5 mg/ml. The lipid vesicle solution was then sonicated for 1 h and used within 24 h. A self-assembled alkanethiol layer was achieved by introducing 1 mM decanethiol solution onto the substrate surface for 1 h. After fusing the lipid vesicles on the alkanethiol-modified surface for 2 h (Figure 6.4), the hybrid bilayer membrane (HBM)-immobilized surface was exposed to negative control bovine serum albumin (100  $\mu\text{g}/\text{ml}$  in PBS buffer) to confirm the complete coverage of the nonspecific binding sites. Then, aliquots of 20  $\mu\text{l}$  of melittin solutions were introduced for 20 min and the E-LSPR measurements were continuously performed.

The optical and electrochemical characteristics of the core-shell structure nanoparticle substrates were evaluated using a simple collinear optical system and the Autolab PGSTAT 100 system. The absorbance peak at 530 nm and the typical cyclic voltammogram of this substrate were clearly observed due to the rather regular nanoparticle surface, thus allowing performance of LSPR and electrochemistry analyses on the same surface. The LSPR behaviors of the core-shell structure nanoparticle substrate were investigated corresponding to the HBM deposition steps (Figure 6.5A). Compared to the bare substrates, the absorbance spectra of the alkanethiol-modified substrates were changed with an average peak shift of 2.03 nm and an absorbance strength increase of 0.02 AU due to the self-assembly formation of 1-decanethiol on the gold surface. Both the peak shift and the increase in absorbance of the core-shell structure nanoparticle substrates could be used as the optical signatures for studying biomolecular interactions. Owing to slightly higher sensitivity, monitoring the absorbance intensity changes in the LSPR response was focused on in this study. Dispersed with a lipid vesicle solution, the alkanethiol-modified hydrophobic surface could be contacted with acyl chains of polar lipids, orienting their polar head-groups toward the solution. As a result, the formation of HBM caused an intense



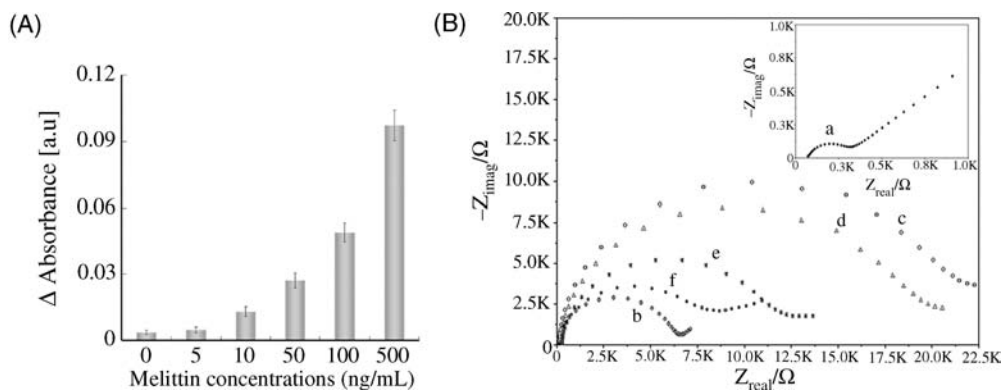
**Figure 6.5** (A) The peak absorbance intensity increases and the peak shift of the core-shell structure nanoparticle substrate due to the successive depositions steps: (a) bare substrate, (b) 1-decanethiol-modified surface, and (c) HBM-covered surface. (B) Cyclic voltammogram of the core-shell structure nanoparticle substrate (a), thiol-modified substrate (b), HBM-covered substrate (c), and after incubation with 100 ng/ml melittin (d) in 2 mM  $[\text{Fe}(\text{CN})_6]^{3-/4-}$ .

increase in the LSPR spectrum by about 0.039 AU in comparison to the decanethiol-modified surface. On the other hand, the presence of 1-decanethiol and successive DMPC layers on the bare substrate surface strongly suppressed the electrochemical reaction of the redox probes. As shown in Figure 6.5(B), a decrease in the magnitude and an insignificant change in the peak separation achieved with the 1-decanethiol-modified substrate surface indicated that the 1-decanethiol layer was not densely packed, thus maintaining the permeability for the electroactive species. Formation of HBM on the surface noticeably prevented the access of the redox probe and considerably restrained the faradaic current, resulting in a relatively flat-shaped curve. This result demonstrated that the formation of the DMPC layer on the core-shell structure nanoparticle surface mostly blocked the interfacial electron transfer between the redox probe and the gold surface. Consequently, the HBM had been successfully prepared on the gold surface, creating a simple membrane-based sensor from the core-shell structure nanoparticle substrate.

### 6.2.3

#### Measurements of Membrane-Based Sensors for Peptide Toxin

The absorbance strength increments in the LSPR response of the sensor were observed when various melittin concentrations of 0, 5, 10, 50, 100, and 500 ng/ml were independently introduced onto the membrane-based sensors ( $n = 4$ ). The slight absorbance peak increments in the buffer solution (without melittin) were a result of physical binding, but not by melittin peptide. With the higher concentrations of melittin, the peak absorbance intensities of the LSPR spectra were increased constantly, denoting the interactions of melittin with HBM, and the amount of bound melittin was directly related to the peptide toxin concentrations (Figure 6.6A).



**Figure 6.6** (A) Calibration curves for melittin on the membrane-based sensor using LSPR. (B) Impedance plots of the core-shell structure nanoparticle substrate (a), thiol-modified substrate (b), HBM-covered substrate before (c), and after interactions with 50 (d), 100 (e), and 500 ng/ml (f) melittin in 1 mM  $[\text{Fe}(\text{CN})_6]^{3-/4-}$ .

The LSPR measurement appears to be highly sensitive for detection of low toxin concentrations. Even at 10 ng/ml melittin, a distinct response in the peak absorbance intensity increase could be obtained. These results were in agreement with the previous study in which a few nanograms per milliliter of melittin could bind to an artificial membrane using the SPR method [33]. Impedance spectroscopy has been widely used to probe the electrode surface features because the interfacial electron transfer at the electrode surface could be changed by modifying various biomaterial layers on the surface. In this work, the interfacial electron transfer properties at the core-shell structure nanoparticles surface were altered by the adsorption of 1-decanethiol, HBM, and different melittin concentrations. From Figure 6.6(B), the electron transfer resistance at the gold surface increased upon the formation of alkanethiol and HBM layers due to the blocking of the redox probe to the electrode surface by densely arranged successive layers. After interaction with HBM, various melittin concentration solutions caused gradual decreases in the charge transfer resistance, resulting from the HBM permeability increases.

### 6.3

#### Label-Free Electrochemical Monitoring of $\beta$ -Amyloid ( $\text{A}\beta$ ) Peptide Aggregation

##### 6.3.1

##### Alzheimer's $\text{A}\beta$ Aggregation and Electrochemical Detection Method

One of the hallmarks of Alzheimer's disease is the formation of neuritic plaques in the brain of Alzheimer's disease individuals. The aggregation of  $\text{A}\beta$  peptides is central to the formation of the plaques.  $\text{A}\beta$  is a 4-kDa peptide present in the brain and cerebral spinal fluid. In its native form,  $\text{A}\beta$  is unfolded, but aggregates into a  $\beta$ -sheet structure of ordered fibrils under various conditions [40–42].  $\text{A}\beta(1-42)$  is more

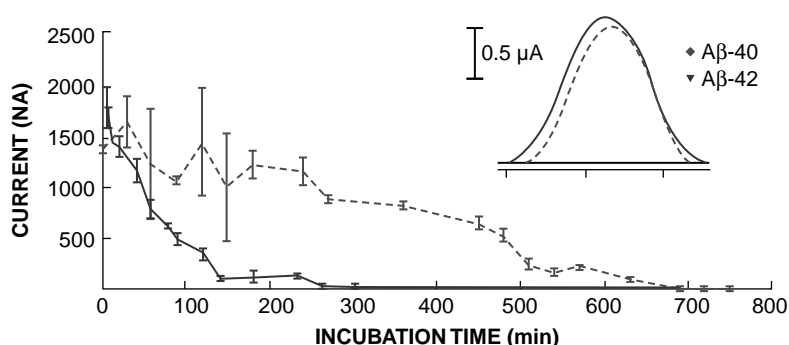
hydrophobic and aggregates more easily than  $A\beta(1-40)$ , and is predominant in the plaques of Alzheimer's disease individuals [43, 44]. The aggregation process starts with a nucleation step followed by a growth phase, which is dependent on the composition of the carboxyl end of  $A\beta$  [45].

The first introduction of the direct oxidation of Trp and Tyr residues on carbon electrodes was reported about two decades ago [46, 47]. Oxidation of Tyr and Trp at a wax-impregnated spectroscopic graphic electrode is reported to be a two-electron transfer process [48]. Although  $A\beta$  possesses only one redox-active residue – Tyr at position 10 – it was assumed that the changes in conformation and possibly charge(s) due to nucleation and later aggregation of the peptide might affect the adsorbability of the Tyr residue to the electrode surface, thus enabling detection of the aggregation process and possibly the initial stages. Moreover, since the method is based on conformational structure change, the study could provide information regarding the different structures that the peptides adopt prior to and during the aggregation process.

### 6.3.2

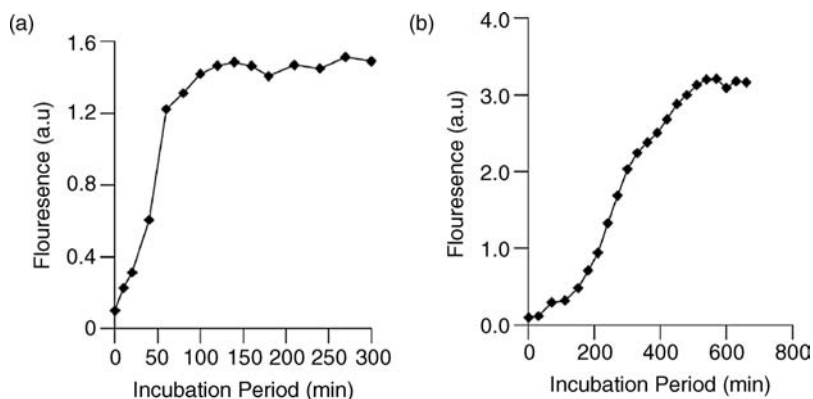
#### Label-Free Electrochemical Detection of $A\beta$ Aggregation

Initially, the electrochemical condition for  $A\beta$  detection was optimized using voltammetry and set to square-wave voltammetry (SWV) [49]. The peak potential of the two peptides was similar (inset of Figure 6.7). The dependence of detected current signal on the concentration of the analyte was studied and the results showed that an increase in concentration led to an increase in peak current, but the relationship was nonlinear. Concentration was increased until there was no increase in peak current (saturation) or until there was a change in peak shape (surface fouling). The detection limits, estimated from the relative standard deviation, corresponded to approximately 0.7  $\mu\text{g/ml}$  for  $A\beta(1-40)$  and  $A\beta(1-42)$  peptides. The aggregation kinetics were analyzed after incubation of the peptides at 80  $\mu\text{M}$  in a 20 mM Tris-HCl buffer, pH 7.0 (TBS) at  $37 \pm 1^\circ\text{C}$  using SWV (Figure 6.7). Samples



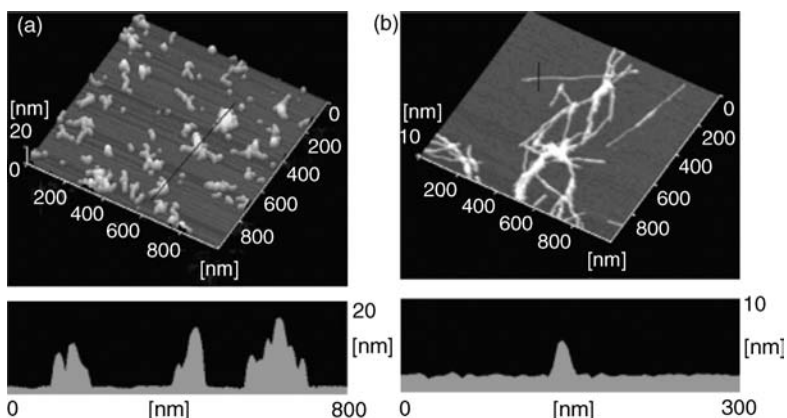
**Figure 6.7** Kinetic study of  $A\beta(1-42)$  (solid line) and  $A\beta(1-40)$  (dashed line) aggregation after incubation at 80  $\mu\text{M}$  in TBS at  $37 \pm 1^\circ\text{C}$ ; detected at 8 and 4  $\mu\text{M}$ , respectively, using SWV,

at room temperature. (Inset) Voltammograms of native  $A\beta(1-42)$  (solid line) and  $A\beta(1-40)$  (dashed line); detected at 8 and 4  $\mu\text{M}$ , respectively, using SWV, at room temperature.



**Figure 6.8** Kinetic study of A $\beta$ (1–42) (a) and A $\beta$ (1–40) (b) aggregation after incubation at 80  $\mu$ M in TBS at  $37 \pm 1^\circ\text{C}$ ; detected using Th-T fluorescence dye. The same arbitrary units (a.u.) are used throughout this report.

were analyzed until the current signals of the peptides were indistinguishable from background noise (i.e., after incubation periods of 300 and 750 min for A $\beta$ (1–42) and A $\beta$ (1–40), respectively). The peptides (80  $\mu$ M) were also analyzed using a spectrofluorometer in conjunction with Th-T as the indicating probe at excitation and emission wavelengths of 450 and 490 nm, respectively. The electrochemical data correlate highly ( $r = -0.9022$  and  $-0.9385$  for A $\beta$ (1–42) and A $\beta$ (1–40), respectively) with that obtained using Th-T fluorescence detection (Figure 6.8). After incubation for specified time periods, A $\beta$  peptides were deposited on a bare mica disk surface and on a 3-(aminopropyl)triethoxysilane-modified disk surface. After, the disk was rinsed with purified water and dried using nitrogen gas. Atomic force microscopy (AFM) images (Figure 6.9) were obtained in air in a dynamic force mode at optimal force. The AFM results for A $\beta$ (1–42) support the electrochemical results.



**Figure 6.9** AFM images of A $\beta$ (1–42) aggregates after incubation at 80  $\mu$ M in TBS at  $37 \pm 1^\circ\text{C}$  for 120 min on bare mica surface (a) and for 180 min on APTES-modified mica surface (b).



Further, a comparison of the kinetic data of A $\beta$ (1–40) and A $\beta$ (1–42) shows that A $\beta$ (1–40) adopts more varied conformational structures compared to A $\beta$ (1–42), as seen by the fluctuations of the Tyr signal displayed by A $\beta$ (1–40). The difference might be attributed to the distinct oligomerization pathways observed for the two peptides during the early stages of aggregation [50], thus making the electrochemical method slightly more informative than the labeled technique. The presence of metal ions, in particular, copper, zinc, and iron, has been reported to enhance A $\beta$  aggregation [51, 52]. Further, copper has been reported to mediate dityrosine cross-linking in A $\beta$  peptides [53]. None of these metals were detected in our buffer by graphite furnace atomic absorption spectroscopy. Trace amounts of iron (3.15 and 35.2  $\mu$ M) and zinc (0.26 and 3.9  $\mu$ M) were detected in A $\beta$ (1–40) and A $\beta$ (1–42) samples, respectively. Copper ions were not detected. The presence of iron and zinc would possibly influence the rate of peptide aggregation [51, 52], and probably have some effect on the Tyr oxidation signal. Their effect, if any, is being evaluated and will be reported. Bovine serum albumin, used as a control in these studies, displayed no changes in either the fluorescence signal (agreeing with a previous report [53]) or the Tyr oxidation signal. Also, the aggregation kinetics of A $\beta$ (1–42) incubated at  $0 \pm 1^\circ\text{C}$  showed no changes in the Tyr signal, which was not surprising since hydrophobic interactions are destabilized by low temperatures [54], further confirming the validity of the method for detecting the aggregation of A $\beta$ .

In summary, this study reported the first bioelectrochemical measurement of A $\beta$ -peptides using voltammetric techniques at a glassy carbon electrode. The kinetics of the peptide aggregation have been studied also for the first time using electrochemistry. Furthermore, it is a rapid and direct (label-free) technique, and the principle can be universally and readily extended to other protein aggregation studies. The method also has potential as a drug-screening tool and/or for assessing, *in vitro*, the effectiveness of Alzheimer's disease therapeutics that target A $\beta$  plaques.

## References

- 1 Link, S. and El-Sayed, M.A. (2003) *Annual Review of Physical Chemistry*, **54**, 331–366.
- 2 Prasad, P.N. (2004) *Nanophotonics*, Wiley-Interscience, New York.
- 3 Kelly, K.L., Coronado, E., Zhao, L.Z., and Schatz, G.C. (2003) *Journal of Physical Chemistry B*, **107**, 668–677.
- 4 Hutter, E. and Fendler, J.H. (2004) *Advanced Materials*, **16**, 1685–1706.
- 5 Prodan, E., Nordlander, P., and Halas, N.J. (2003) *Nano Letters*, **3**, 1411–1415.
- 6 Stuart, D.A., Haes, A.J., Yonzon, C.R., Hicks, E.M., and Van Duyne, R.P. (2005) *IEEE Proceedings on Nanobiotechnology*, **152**, 13–32.
- 7 Freeman, R.G., Grabar, K.C., Allison, K.J., Bright, R.M., Davis, J.A., Guthrie, A.P., Hommer, M.B., Jackson, M.A., Smith, P.C., Walter, D.G., and Natan, M.J. (1995) *Science*, **267**, 1629–1632.
- 8 Nath, N. and Chilkoti, A. (2004) *Analytical Chemistry*, **76**, 5370–5378.
- 9 Sherry, L.J., Jin, R., Mirkin, C.A., Schatz, G.C., and Van Duyne, R.P. (2006) *Nano Letters*, **6**, 2060–2065.
- 10 Sherry, L.J., Chang, S.-H., Schatz, G.C., Van Duyne, R.P., Wiley, B.J., and Xia, Y. (2005) *Nano Letters*, **5**, 2034–2038.
- 11 Huang, W., Qian, W., and El-Sayed, M.A. (2004) *Nano Letters*, **4**, 1741–1747.
- 12 Nath, N. and Chilkoti, A. (2002) *Analytical Chemistry*, **74**, 504–509.
- 13 Mock, J.J., Smith, D.R., and Schultz, S. (2003) *Nano Letters*, **3**, 485–491.

- 14 Himmelhaus, M. and Takei, H. (2000) *Journal Sensors and Actuators B*, **63**, 24–30.
- 15 Dittrich, P.S., Tachikawa, K., and Manz, A. (2006) *Analytical Chemistry*, **78**, 3887–3908.
- 16 Endo, T., Kerman, K., Nagatani, N., Takamura, Y., and Tamiya, E. (2005) *Analytical Chemistry*, **77**, 6976–6984.
- 17 Endo, T., Kerman, K., Nagatani, N., Hiepa, H.M., Kim, D.-K., Yonezawa, Y., Nakano, K., and Tamiya, E. (2006) *Analytical Chemistry*, **78**, 6465–6475.
- 18 Hiep, H.M., Kerman, K., Endo, T., Saito, M., and Tamiya, E. (2010) *Analytica Chimica Acta*, **661**, 111–116.
- 19 Hiep, H.M., Yoshikawa, H., Saito, M., and Tamiya, E. (2009) *ACS Nano*, **3**, 446–452.
- 20 Hiep, H.M., Endo, T., Saito, M., Chikae, M., Kim, D., Yamamura, S., Takamura, Y., and Tamiya, E. (2008) *Analytical Chemistry*, **80**, 1859–1864.
- 21 Endo, T., Yamamura, S., Kerman, K., and Tamiya, E. (2008) *Analytica Chimica Acta*, **614**, 182–189.
- 22 Kim, D.-K., Kerman, K., Saito, M., Sathuluri, R.R., Endo, T., Yamamura, S., Kwon, Y.-S., and Tamiya, E. (2007) *Analytical Chemistry*, **79**, 1855–1864.
- 23 Hiep, H.M., Nakayama, T., Saito, M., Yamamura, S., Takamura, Y., and Tamiya, E. (2008) *Japanese Journal of Applied Physics*, **47**, 1337–1347.
- 24 Henry, C. (1998) *Analytical Chemistry*, **70**, 594A–598A.
- 25 Duffy, D.C., McDonald, J.C., Schueller, O.J.A., and Whitesides, G.M. (1998) *Analytical Chemistry*, **70**, 4974–4984.
- 26 Karlsson, R., Michaelsson, A., and Mattsson, L. (1991) *Journal of Immunological Methods*, **145**, 229–240.
- 27 Parker, M.W. and Feil, S.C. (2005) *Progress in Biophysics and Molecular Biology*, **88**, 91–142.
- 28 Lehrer, R.I. and Ganz, T. (1999) *Current Opinion in Immunology*, **11**, 23–27.
- 29 Tosteson, M.T., Holmes, S.J., Razin, M., and Tosteson, D.C. (1985) *Journal of Membrane Biology*, **87**, 35–44.
- 30 Naito, A., Nagao, T., Norisada, K., Mizuno, T., Tuzi, S., and Saito, H. (2000) *Biophysical Journal*, **78**, 2405–2417.
- 31 Habermann, E. (1972) *Science*, **177**, 314–322.
- 32 Andersson, A., Biverstahl, H., Nordin, J., Danielsson, J., Lindahl, E., and Mäler, L. (2007) *Biochimica et Biophysica Acta*, **1768**, 115–121.
- 33 Papo, N. and Shai, Y. (2003) *Biochemistry*, **42**, 458–466.
- 34 Mozsolits, H., Wirth, H.-J., Werkmeister, J., and Aguilar, M.-I. (2001) *Biochimica et Biophysica Acta*, **1512**, 64–76.
- 35 Becucci, L., Leon, R.R., Moncelli, M.R., Rovero, P., and Guidelli, R. (2006) *Langmuir*, **22**, 6644–6650.
- 36 Kriech, M.A. and Conboy, J.C. (2003) *Journal of the American Chemical Society*, **125**, 1148–1149.
- 37 Pera, I. and Fritz, J. (2007) *Langmuir*, **23**, 1543–1547.
- 38 He, L., Robertson, J.W.F., Li, J., Kärcher, I., Schiller, S.M., Knoll, W., and Naumann, R. (2005) *Langmuir*, **21**, 11666–11672.
- 39 Bart, M., Van Os, P.J.H.J., Kamp, B., Bult, A., and Van Bennekom, W.P. (2002) *Sensors and Actuators B*, **84**, 129–135.
- 40 Stine, W.B. Jr., Dahlgren, K.N., Krafft, G.A., and LaDu, M.J. (2003) *Journal of Biological Chemistry*, **278**, 11612–11622.
- 41 Szabo, Z., Klement, E., Jost, K., Zarandi, M., Soos, K., and Penke, B. (1999) *Biochemical and Biophysical Research Communications*, **265**, 297–300.
- 42 Shen, C.L. and Murphy, R.M. (1995) *Biophysical Journal*, **69**, 640–651.
- 43 Lippa, C.F., Nee, L.E., Mori, H., and George-Hyslop, P. (1998) *Lancet*, **352**, 1117–1118.
- 44 Asami-Odaka, A., Ishibashi, Y., Kikuchi, T., Kitada, C., and Suzuki, N. (1995) *Biochemistry*, **34**, 10272–10278.
- 45 Jarrett, J.T., Berger, E.P., and Lansbury, P.T. (1993) *Biochemistry*, **32**, 4693–4697.
- 46 Reynaud, J.A., Malfoy, B., and Bere, A. (1980) *Electroanalytical Chemistry*, **116**, 595–606.
- 47 Brabec, V. and Schindlerova, I. (1981) *Bioelectrochemistry and Bioenergetics*, **8**, 451–458.
- 48 Brabec, V. and Mornstein, V. (1980) *Biophysical Chemistry*, **12**, 159–165.
- 49 Vestergaard, M., Kerman, K., Saito, M., Nagatani, N., Takamura, Y., and Tamiya, E. (2005) *Journal of the American Chemical Society*, **127**, 11892–11893.

- 50 Bitan, G., Kirkitadze, M.D., Lomakin, A., Vollers, S.S., Benedek, G.B., and Teplow, D.B. (2003) *Neuroscience*, **100**, 330–335.
- 51 Atwood, C.S., Perry, G., Zeng, H., Kato, Y., Jones, W.D., Ling, K.-Q., Huang, X., Moir, R.D., Wang, D., Sayre, L.M., Smith, M.A., Cheng, S.G., and Bush, A.I. (2004) *Biochemistry*, **43**, 560–568.
- 52 Huang, X., Atwood, C.S., Moir, R.D., Hartshorn, M.A., Tanzi, R.E., and Bush, A.I. (2004) *Journal of Biological Inorganic Chemistry*, **9**, 954–960.
- 53 Yoshiike, Y., Tanemura, K., Murayama, O., Akagi, T., Murayama, M., Sato, S., Sun, X., Tanaka, N., and Takashima, A. (2001) *Journal of Biological Chemistry*, **276**, 32293–32299.
- 54 De Felice, F.G., Houzel, J.-C., Garcia-Abreu, J., Louzada, P.R.F. Jr., Afonso, R.C., Meirelles, M.N.L., Lent, R., Neto, V.M., and Ferreira, S.T. (2001) *FASEB Journal*, **15**, 1297–1299.



## 7

# Surface Plasmon Resonance Spectroscopy in the Biosciences

*Jing Yuan, Yinqiu Wu, and Marie-Isabel Aguilar*

### 7.1

#### Introduction

Surface plasmon resonance (SPR) is a very powerful optical sensing technique and the capacity of SPR to monitor label-free binding events in real-time has made it a popular method to study macromolecular interactions [1–5]. SPR has now become a widely used technique to study antibody–antigen, DNA–DNA, DNA–protein, protein–protein, and receptor–ligand interactions [6, 7]. SPR spectroscopy has also been applied to the study of biomembrane-based systems that involve liposomes and planar mono- or bilayers [8–14]. This chapter provides an overview of SPR technology, and illustrates the power of SPR in studying biomolecular interactions with specific reference to the development of SPR-based immunosensors and the characterization of membrane interactions, and demonstrates the enormous potential of SPR to enhance our molecular understanding of membrane-mediated events.

### 7.2

#### SPR-Based Optical Biosensors

The SPR phenomenon is highly sensitive to small changes in refractive index occurring at the surface of a thin noble metal film [15]. This surface-sensitive optical technique can detect, monitor, and quantitatively measure binding events between ligands and an immobilized target in real-time without the use of radio, fluorescent, or enzyme labels. It also has excellent potential for studying surface-confined affinity interactions without removal of unreacted or excess reactants in the sample solutions. SPR biosensors provide rich information on the specificity, affinity, and kinetics of biomolecular interactions and/or the concentration levels of an analyte of interest from a complex sample [16, 17]. This contrasts sharply with traditional “end-point” immunoassays (e.g., enzyme-linked immunosorbent assays), which only provide information such as limited ranges of affinity or concentration and only at one binding characteristic per experiment. Since the initial characterization of

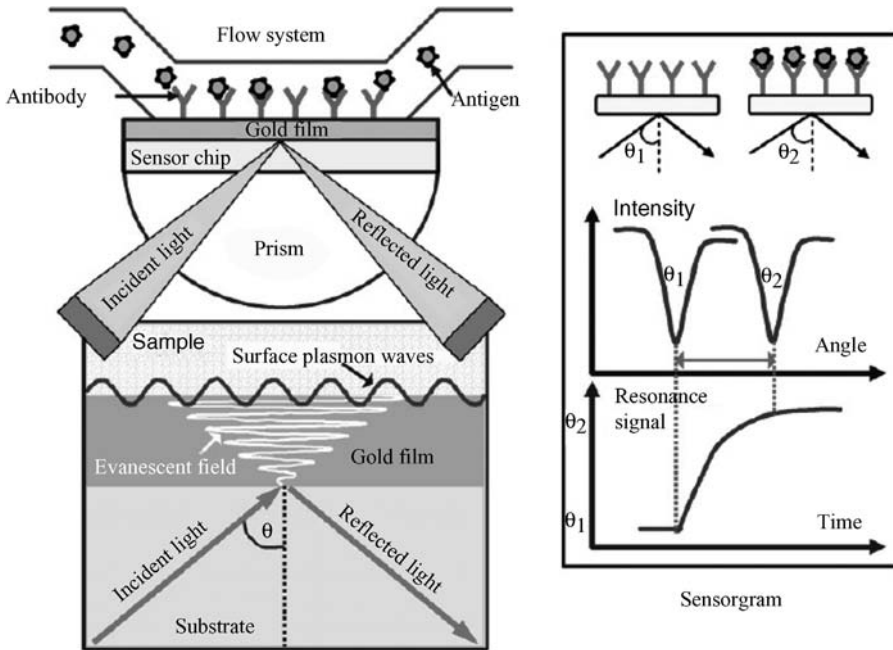
SPR in the late 1970s, surface plasmons have been intensively studied, and the technique has played a central role in the study of biomolecule characterization, kinetics of antibody–analyte interactions and ligand–fishing in drug discovery, and the detection of a variety of chemical and biological substances [18–21]. The remarkable progress in the development of new SPR biosensor technologies has made it a powerful tool in a variety of biomolecular interaction analyses, including antigen–antibody, ligand–receptor, protein–protein, DNA–DNA, biomembrane interactions, and so on, with enormous impact in wider application areas, such as environmental monitoring [22–24], biotechnology [25–29], medical diagnostics [30–32], drug screening [33–36], and food safety and security [37–39].

### 7.3

#### Principle of Operation of SPR Biosensors

SPR is a quantum optical/electrical phenomenon that occurs at the metal surface of a SPR-active substance when a photon of light is incident upon its surface [40]. It is best described as a charge density oscillation at the interface between two media with oppositely charged dielectric constants. Plasmons represent the “excited” free electron portion of the surface metal layer. This resonant excitation is provided by compatible light energy photons. Under appropriate conditions, the plasmons can be made to resonate with light, which results in the absorption of light [41]. The excitation of the surface plasmon is accompanied by the transfer of optical energy into the surface plasmon and its dissipation in the metal layer, which results in a narrow dip in the spectrum of reflected light. There are two kinds of configuration used for excitation of surface plasmons: Kretschmann [42] and Otto [43]. The Kretschmann configuration is most commonly used for SPR excitation.

In general, a SPR biosensor is comprised of the following components: a light source, a prism, a transduction surface (usually gold film), a biomolecule (antibody or antigen), a flow system, and a detector. Figure 7.1 shows a simple scheme of the principle and operation of an SPR immunoassay technique. With respect to stability and sensitivity considerations, thin gold films (50–100 nm) represent a better choice used to construct the transduction surface on a glass slide optically coupled to a glass prism through refractive index-matched oil. Plane polarized light is directed through the glass prism to the gold/solution dielectric interface over a wide range of incident angles; the intensity of the resulting reflected light is measured against the incident light angle with a detector. At selected incident light wavelength and angles, the photons of the light waves react with the free electron cloud in the metal film, causing a drop in the intensity of the reflected light. The angle at which the drop is maximum (minimum of reflectivity) is denoted as the “SPR angle.” This critical angle is extremely sensitive to the refractive index of the sample in contact (within around 200 nm) with the metal surface so that it is also highly influenced by the amount of biomolecules immobilized on the gold layer. Adsorption of biomolecules (antigen or antibody) on the metallic film as well as molecular interactions (antigen–antibody complex) induce a change in the refractive index near the surface, thus giving rise to a

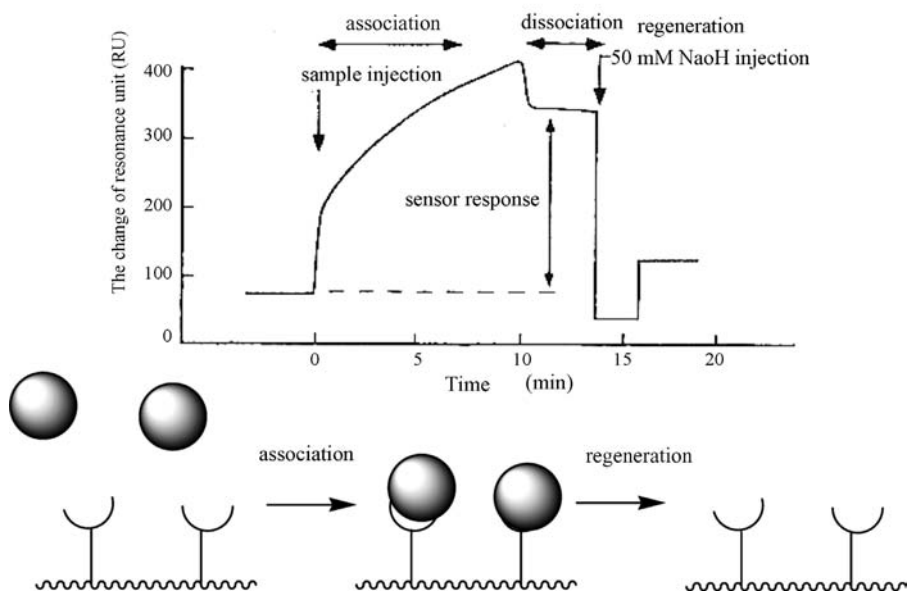


**Figure 7.1** Schematic view of the SPR immunoassay technique. Reprinted from [17], with permission from Elsevier.

shift of the resonance angle. The response (angular shift) is expressed in resonance units (RUs). This shift is directly proportional to the mass increase and concentration of the target analyte can thus be measured. Also, information on the affinity of analyte for the antibody and the association (or dissociation) kinetics between the antibody and analyte can be obtained [17, 19, 20].

It has been shown that the sensitivity of modern SPR sensing systems based on the Kretschmann configuration is such that they are effectively measurements of changes in the refractive index of the surface with 1 RU unit equivalent to  $5 \times 10^{-7}$  refractive index units, which corresponds to a  $1 \text{ pg/mm}^2$  surface coverage of biomolecules [44]. A computer plot can be made of the SPR response (RU) versus time and is known as a sensorgram. Figure 7.2 shows a SPR binding surface and a sensorgram highlighting the association and dissociation and regeneration phases.

There are several companies manufacturing SPR instruments for studying biomolecular interactions. Biacore was the first company to commercially develop SPR technology in 1990 [45]. During the last few years, Biacore technology has been used in approximately 90% of all work published in the optical biosensor field [36, 46–50]. The most commonly used applications of these systems are for the determination of affinity and kinetics of interaction between two or more biomolecules [51, 52], receptor–ligand interactions [53, 54], and nucleic acid hybridization [55, 56]. The systems have also been used for quantitative analysis using



**Figure 7.2** SPR schematic showing binding surface and the corresponding sensorgram.

antibodies as specific reagents [57, 58] and for the thermodynamic analysis of biomolecular interactions [59, 60]. Other commercially available SPR systems include the ProteOn XPR36 Protein Interaction Array System from BioRad, SensiQ from ICX Technologies, Spreeta from Texas instruments, and Nippon Laser and Electronics systems [61–63].

## 7.4

### Description of an SPR Instrument

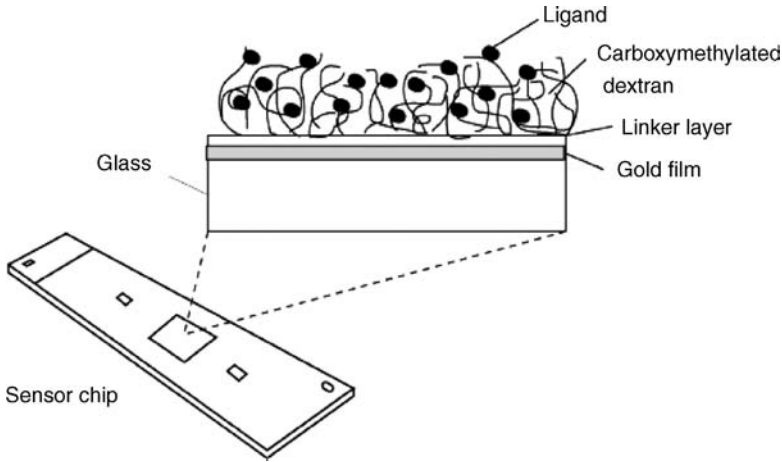
SPR biosensors are comprised of three key components that harness the SPR phenomenon for useful applications: a gold sensor chip, a microfluidic sample handling system, and an SPR detector. Aqueous samples are serially injected over the sensor chip surface immobilized with one of the interacting partners. Binding of the interactant is detected and quantified in real-time by the surface sensitive detector; we provide an overview of the Biacore system below.

#### 7.4.1

##### Sensor Surface

The sensor chip itself is comprised of a glass slide covered with a 50-nm thick gold film embedded in a plastic support platform as shown in Figure 7.3. This forms an





**Figure 7.3** Surface of a sensor chip consists of three layers (glass, a thin gold film, and a dextran layer) to which biomolecules can be immobilized.

interchangeable base from which a number of specialized surfaces may be created. Several different surface chemistries are now available for different types of biological applications. Current commercially available surfaces are CM5 (carboxymethylated dextran), SA (streptavidin), NTA (nickel chelation), HPA (hydrophobic monolayer), L1 (lipophilic dextran), B1 (low-charge carboxymethylated dextran), C1 (flat carboxymethylated), F1 (short dextran), J1 (unmodified gold surface) and SIA Kit (bare gold) [46]. The most common is the carboxymethylated dextran layer (CM5 Sensor Chip) coupled with different functional groups to make it suitable for immobilization of any ligand that is appropriate for many types of biological applications.

#### 7.4.2

##### Flow System

The SPR microfluidic systems are centered on integrated fluidics cartridges (IFCs). This allows a controlled flow of analyte in a continuous, pulse-free manner that ensures precise and consistent concentrations over the sensor chip surface. When a sensor chip is docked in the instrument, the IFC is pressed against the chip surface. The IFC forms three flow cell walls, while the sensor chip forms the fourth wall. There are four flow cells that can be fed and monitored separately. Flow cell volumes generally range from 20 to 60  $\mu\text{l}$  depending on the instrument model [62] and sample volumes of 5–450  $\mu\text{l}$  at flow rates of 1–100  $\mu\text{l}/\text{min}$  can be injected. For some models (e.g., Biacore 2000 and 3000), a sample can be passed over the four flow cells in sequence and response for all flow cells can be monitored in parallel, while for other models (e.g., Biacore 1000 and Biacore Q) only one flow cell can be monitored at any one time. All IFCs are specific to a defined SPR instrument series, but share common features, including low sample

consumption, the absence of an air/solution interface that could allow samples to evaporate or proteins to be denatured, stability mechanisms, and the ability to measure a range of surface ligand concentrations in one experiment to optimize kinetic and concentration analysis.

#### 7.4.3

##### **Detection System**

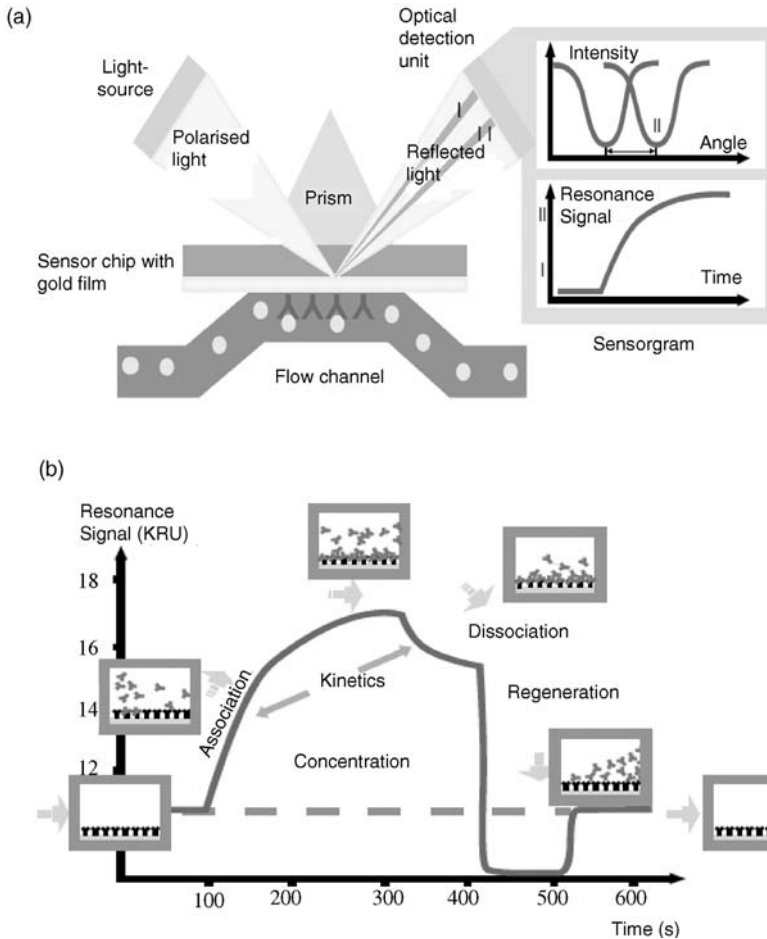
The SPR instruments detect changes in mass by measuring changes in the refractive index in the aqueous layer close to the sensor chip surface. When a prospective analyte binds to a target molecule attached to the surface of the chip, the mass at the surface increases. When the two dissociate, the mass returns to its original state. The change in mass concentration at the chip surface leads to a quantifiable change in the refractive index of the aqueous layer as depicted in Figure 7.4(a). The SPR detector measures this change and the biosensor's software then produces a sensorgram that shows the mass-concentration-dependent change in refractive index over time as shown in Figure 7.4(b) [64].

#### 7.5

##### **Application of SPR in Immunosensor Design**

The immunoassay, based on the specific recognition of an antigen by its antibody, has garnered widespread use for the determination of small and large analytes. The nature of the selectivity of antibody binding allows these reagents to be employed in the development of methods that are highly specific and that can often be used directly even in complex biological matrices. By combining the selectivity of antibody-analyte interactions with the vast array of antibodies that can be produced in nature and the availability of numerous readily detectable labels (e.g., radioisotopes, enzymatically or electrochemically induced absorbance, or fluorescence or chemiluminescence), immunoassays have been designed for a wide variety of analytes with extraordinarily low detection limits [65–75].

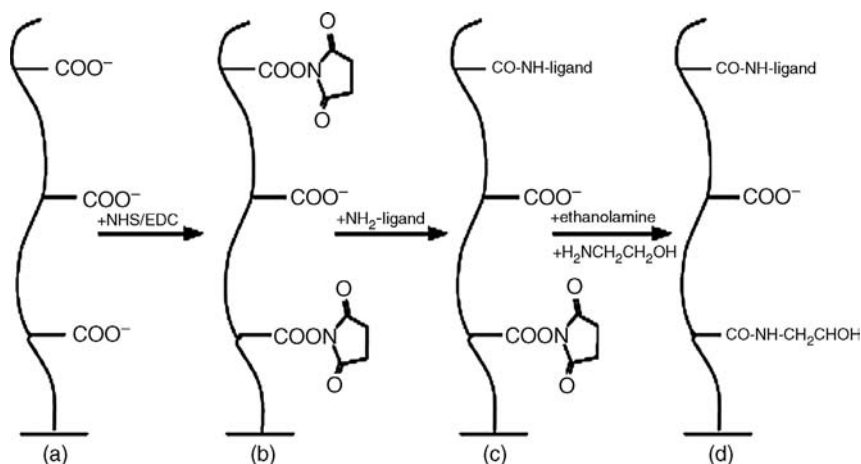
The primary attraction of SPR-based immunosensors is the high specific detection of small molecules with low detection limits for a wide variety of analytes in complex matrices [76–81]. In recent years, the need for simple and high-throughput analysis of low levels of small complex molecules, such as drug residues, pesticides, hormones, mycotoxins, and algal toxins, in biomedical, environmental, and food samples has been growing rapidly for consumer protection. However, until recently, most methods used for the detection of small molecules have used conventional instrumental analytical methods such as liquid chromatography [82–84], liquid chromatography-mass spectrometry [85, 86], capillary electrophoresis-mass spectrometry [87], and chemiluminescence detection [88]. These methods are generally quite costly, require highly skilled workers and time-consuming sample preparation steps, and are not suitable for real-time, *in situ*, or on-site detection applications. As a consequence, there is a growing demand



**Figure 7.4** (a) SPR phenomenon. Incident p-polarized light is focused into a wedge-shaped beam that is totally internally reflected at the interface of an exchangeable gold-coated glass slide. An increase in sample concentration in the surface coating of the sensor chip causes a corresponding increase in refractive index, which alters the angle of incidence at which the SPR phenomenon occurs (the SPR angle). This SPR angle is monitored as a change in the

detector position for the reflected intensity dip (from I to II). The kinetic events at the surface can then be displayed in a sensorgram by monitoring the SPR angle as a function of time. (b) Typical sensorgram. The association and dissociation phases of a compound: target interaction and surface regeneration are displayed in the sensorgram. SPR response is measured in RUs plotted against time.

for more rapid and more economical methods. It is now widely recognized that SPR immunosensors have the potential to fulfill this demand due to their highly desirable analytical characteristics, including sensitivity, selectivity, speed, and reliability in analyses with additional emphasis on portability, miniaturization, and on-site analysis.



**Figure 7.5** Amine coupling of a ligand to a sensor surface. (a) Nonactivated carboxyl groups on dextran. (b) The carboxyl groups are activated by addition of a mixture of succinimide (NHS) and carbodiimide (EDC). (c) The ligand

is covalently bound to the sensor surface.

(d) Remaining esters are deactivated by addition of ethanolamine. For some analytes the surface is amino modified by addition of ethylenediamine prior to immobilization.

### 7.5.1

#### Assay Development

Successful SPR analysis of a biomolecule requires optimization of several parameters, such as immobilization of the analyte to a specific chip surface, assay design, reduction of sample matrix effects, and surface regeneration.

##### 7.5.1.1 Immobilization of the Analyte to a Specific Chip Surface

Optimal immobilization is the most important parameter. There are several immobilization strategies, including amine coupling, thiol coupling, or aldehyde coupling [89–91]. However, the most commonly applicable immobilization strategy is amine coupling, whereby the ligand is coupled via primary amino groups. Amine coupling introduces *N*-hydroxysuccinimide (NHS) esters onto the surface matrix by activation of the carboxylic acid functions with a mixture of NHS and 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC). These esters then react spontaneously with amines and other nucleophilic groups on the ligand to form covalent links when they pass over the surface at a suitable pH [90]. Unreacted active esters are deactivated by the addition of ethanolamine hydrochloride (EAH). Amine coupling is shown in Figure 7.5.

The surface immobilized density is affected by a number of factors, such as percentage of activated carboxymethyl groups, pH, ionic strength, concentration of coupling buffer and ligand, and reaction time. The optimum pH for immobilization can be determined by “pH scouting”, or preconcentration, which involves injecting solutions at various pHs over the nonactivated surface and assessing the

preconcentration sensorgrams. The preconcentration is an important factor in enabling efficient immobilization of ligand from relatively dilute solution (20–100 µg/ml). Efficient preconcentration requires a low ionic strength in the ligand solution (maximum 10 mM monovalent cations). The pH of the coupling solution should be lower than the isoelectric point (pI) of the ligand to maximize electrostatic concentration of the ligand in the dextran layer. However, preconcentration is not possible with small ligands due to the mass sensitivity limitation of the instrument [91, 92].

#### 7.5.1.2 Assay Design

The assay design is the most important parameter for successful SPR analysis of biomolecules. Generally, the SPR-based immunoassay involves the immobilization of an antibody (or antigen) onto the sensor surface followed by the binding interaction with an analyte at the interface, which is monitored by detection of the sensor response. The most frequently used immunoassay formats are direct detection [93, 94], the sandwich assay [95, 96], the displacement assay [97, 98], and the competitive inhibition assay format [80, 81]. The most suitable assay format depends on the nature of the target analyte, the analytical sample, the sensitivity of the instrument, and the particular application.

**7.5.1.2.1 Direct Immunoassay** In direct immunoassays, antibodies are immobilized on the sensor surface and subjected to the binding interaction with the analyte of interest. The SPR change is directly proportional to the concentration of analyte. This method is simple, but only useful for the detection of large molecules with a molecular weight above 10 kDa because small molecules have insufficient mass to effect a measurable change in the refractive index. Several research groups have achieved a sensitive detection limit by using such simple formats (e.g., 100 pM for bovine serum albumin (BSA) [99] and 2.5 ppb for the cardiac marker protein troponin I [96]).

**7.5.1.2.2 Sandwich Immunoassay** The sandwich assay is more sensitive than the direct assay due to a larger mass increase in the detection step. The sandwich assay consists of two recognition steps. The first step is immobilization of an antibody onto the sensor surface that allows binding to the analyte of interest. The second step is the passing of a secondary antibody over the sensor surface to bind to the previously captured analyte. The two separate recognition steps result in a large mass increase, and consequently an enhanced sensitivity and specificity. This format is also suitable for the measurement of large molecules (above 5000 Da), such as proteins, bacteria and viruses, which have multiple epitopes for simultaneous binding of two antibodies. By using this method, the assay sensitivity can be greatly improved, as demonstrated by the reported limit of detection of 0.25 ppb for cardiac troponin I [96] and 0.5 ppb for staphylococcal enterotoxin B in milk [100].

**7.5.1.2.3 Indirect Competitive Inhibition Immunoassay** The sensitivity of all the above immunoassay formats in current instruments is insufficient for detection

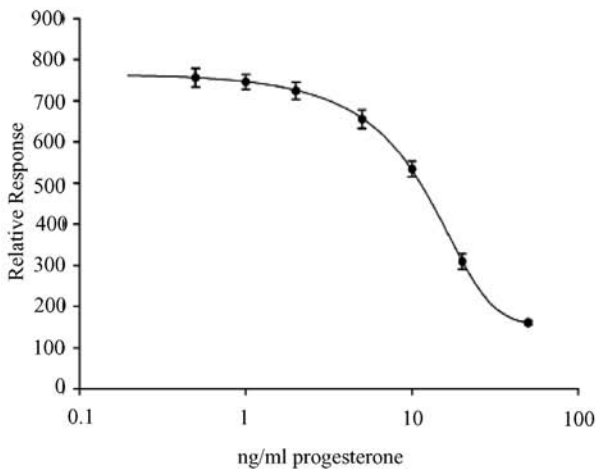
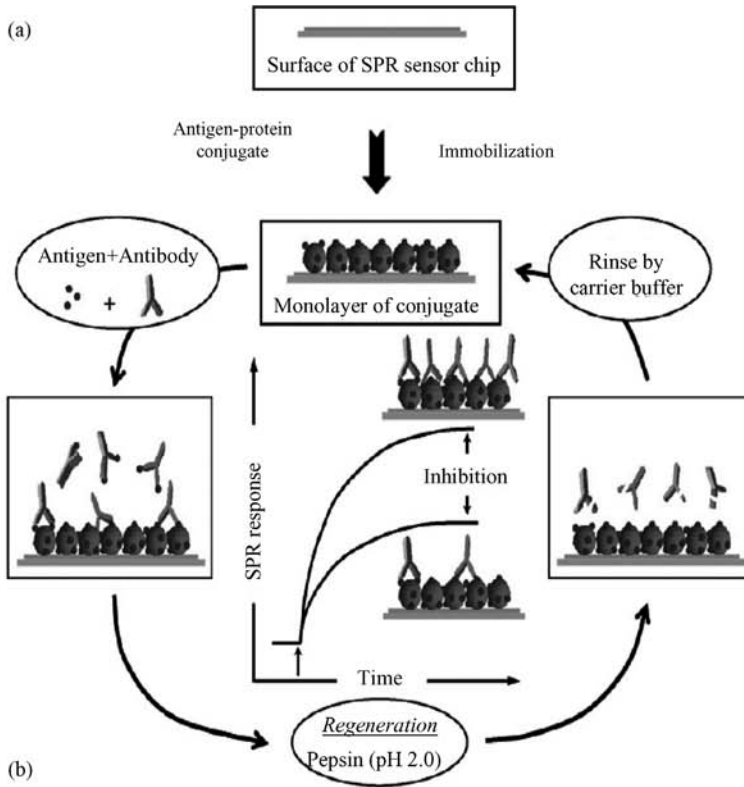
and quantification of low levels of low-molecular-weight analytes, such as hormones, antibiotics and residues. Instead, a competitive inhibition assay format is used, in which the analyte (usually analyte–protein conjugate) is immobilized on the surface. The analyte is then mixed with the respective antibody and introduced over the immobilized surface. The concentration of the antibody is kept constant so that the response variations are proportional to the amount of analyte mixed with the antibody. The concentration of antibody free to bind to the sensor surface is then measured. At a high analyte concentration most antibodies will be bound by the antigen in solution and so very little will be free to bind to the surface, resulting in a low response, while a blank sample will give a high response. Thus, the SPR change is inversely proportional to the analyte concentration in the solution in a competitive inhibition assay format. Since the majority of analytes of biomedical, food, and environmental interest are small in size, this competitive inhibition assay format has received widespread interest in the development of SPR immunosensors for a variety of applications [101–105]. Figure 7.4 depicts a schematic view of the indirect competitive immunoassay (Figure 7.4a) [17] and the SPR response observed for detection of progesterone (Figure 7.4b) [101].

## 7.6

### Application of SPR in Membrane Interactions

Biomolecular membrane interactions are central to many important biological processes and characterization of the molecular details of these interactions is important for understanding a wide range of cellular events, such as cellular signaling, ion channel formation, and protein trafficking. SPR is a widely used technique for investigating biophysical analysis of membrane-mediated events [14, 106–110].

While there are a variety of surfaces available for use in different SPR instruments for the study of biomolecular membrane interactions, there are two sensor chips available from Biacore commonly used to study membrane systems. The HPA sensor chip consists of self-assembled alkanethiol molecules covalently attached to the gold surface of the chip. This chip can be used to prepare hybrid bilayer lipid membranes by the fusion of liposomes onto the hydrophobic surface (Figure 7.7a) [111]. However, the pioneer L1 sensor chip is more widely used and can be used to prepare lipid bilayer membrane systems that mimic the fluid bilayer structure of the cell membrane more closely than those of the HPA chips. The L1 sensor chip is composed of a thin dextran matrix modified with lipophilic branches on a gold surface on which the lipid bilayer system is generated by the capture of liposomes by the lipophilic branches (Figure 7.7b) [8]. The immobilization of the biomimetic lipid surface onto the sensor chips is generally a fast and easily reproducible process. The HPA and L1 sensor chips can be applied to the study of membrane-based biomolecular interactions and to measure the binding affinity related to these interactions. There have been a number of



**Figure 7.6** (a) Schematic view of the indirect competitive inhibition SPR immunoassay. (b) RU response versus concentration of progesterone in bovine milk – a standard curve. Reprinted from [17, 101], with permission from Elsevier.

examples where these sensor chips have been applied to the study of peptide/protein–membrane interactions. These applications range from the analysis of protein–protein and protein–ligand interactions in a membrane environment to the study of the direct binding of peptides and proteins to a specific phospholipid surface [14, 54, 112].

### 7.6.1

#### General Protocols for Membrane Interaction Studies by SPR

##### 7.6.1.1 Liposome Preparation

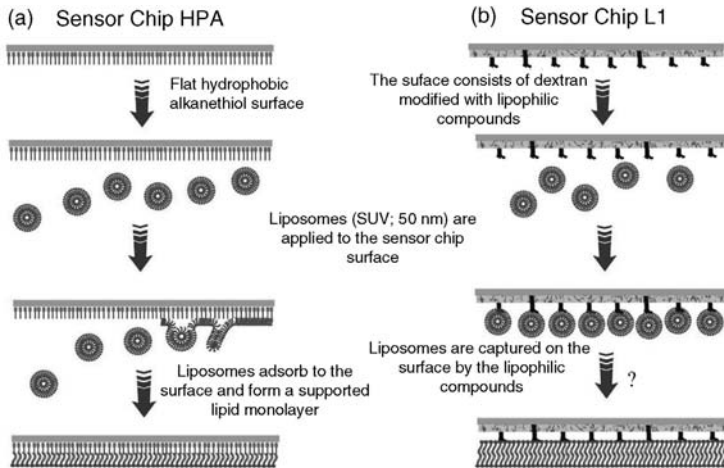
Small unilamellar vesicles (SUVs) composed of different phospholipids such as dimyristoylphosphatidylcholine (DMPC), dimyristoylphosphatidylglycerol (DMPG), 4 : 1 palmitoyl-oleoylphosphatidylcholine (POPC)/palmitoyl-oleoylphosphatidylcholine (POPG) or 4 : 1 POPE/POPG (around 50 nm) are used to prepare the membrane surface. The instrument manual provides clear instructions for SUV preparation, and in general they are prepared in 0.02 M phosphate buffer by sonication and/or extrusion. Dry phospholipid is dissolved in ethanol-free chloroform or in the case of DMPG, 4 : 1 POPC/POPG, and 4 : 1 POPE/POPG the dry lipid is dissolved in a  $\text{CHCl}_3/\text{MeOH}$  mixture (2 : 1, v/v). The solvent is then removed by evaporation under a gentle stream of nitrogen to form a dry lipid cake and residual solvents are removed under vacuum overnight. The dry lipid cakes are then resuspended in 0.02 M phosphate buffer via vortex mixing and the lipid suspension is allowed to stand at room temperature for 2 h in order to complete the swelling process. The lipid suspension is sonicated, and then extruded through polycarbonate filters 17–21 times to gain SUVs at around 50 nm size (LiposoFast, pore diameter 50 nm) and used to prepare a lipid bilayer system.

##### 7.6.1.2 Formation of Bilayer Systems

After cleaning as outlined in the instrument manual, the Biacore instrument is left running overnight using Milli-Q water as eluent to thoroughly wash all liquid handling parts of the instrument. The appropriate sensor chip is then installed into the machine, and the surfaces cleaned at a flow rate of 5  $\mu\text{l}/\text{min}$  by an injection of the nonionic detergent 40 mM octyl-glucoside (3000s) and 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS; 60 s), respectively. SUVs (30  $\mu\text{l}$  (HPA), 80  $\mu\text{l}$  (L1), 0.5 mM) are then immediately applied to the chip surface at a flow rate of 2  $\mu\text{l}/\text{min}$ . In the case of the HPA sensor chip, the liposomes adsorb spontaneously to form a supported lipid monolayer on the alkanethiol surface, providing a monolayer model membrane system as shown in Figure 7.7(a).

In the case of the L1 sensor chip, the liposomes are captured on the surface of the sensor chip by the lipophilic branches and this provides a bilayer model membrane system as illustrated in Figure 7.7(b). The anionic DMPG liposomes are thus deposited onto the L1 surface using a lower flow rate of 1  $\mu\text{l}/\text{min}$ , which allows more time for deposition and compensates for the electrostatic repulsions. To remove any multilamellar structures or weakly bound liposomes from the lipid surfaces,





**Figure 7.7** Preparation of a monolayer and a bilayer model membrane system on an (a) HPA and (b) L1 sensor chip, respectively.

sodium hydroxide (10 mM, 36 s) is injected at a higher flow rate (50  $\mu\text{l}/\text{min}$ ), which results in a stable baseline corresponding to the monolayer or bilayer systems as previously shown [8, 111]. In the case of the HPA chip, the negative control BSA is injected (0.1 mg/ml in phosphate buffer, flow rate 5  $\mu\text{l}/\text{min}$ , 120 s) to confirm the complete coverage of the chip surface with lipid by the absence of nonspecific binding. The prepared lipid monolayer or bilayer systems are then ready to use as a model cell membrane surface to perform the bioactive peptide–membrane binding studies.

#### 7.6.1.3 Analyte Binding to the Membrane System

Analyte solutions are prepared by dissolving each sample in 0.02 M phosphate buffer (pH 6.8). It is critical that the buffer used to suspend the analyte is the running buffer used in the assay and, where possible, the same buffer used for preparing the liposomes in order to avoid the presence of refractive index changes. The analyte concentrations are typically in the micromolar range, reflecting the more common affinities of these interactions. Solutions of the different analyte concentrations are injected over the lipid surfaces at a constant flow rate and for a contact time that is optimized for the particular assay.

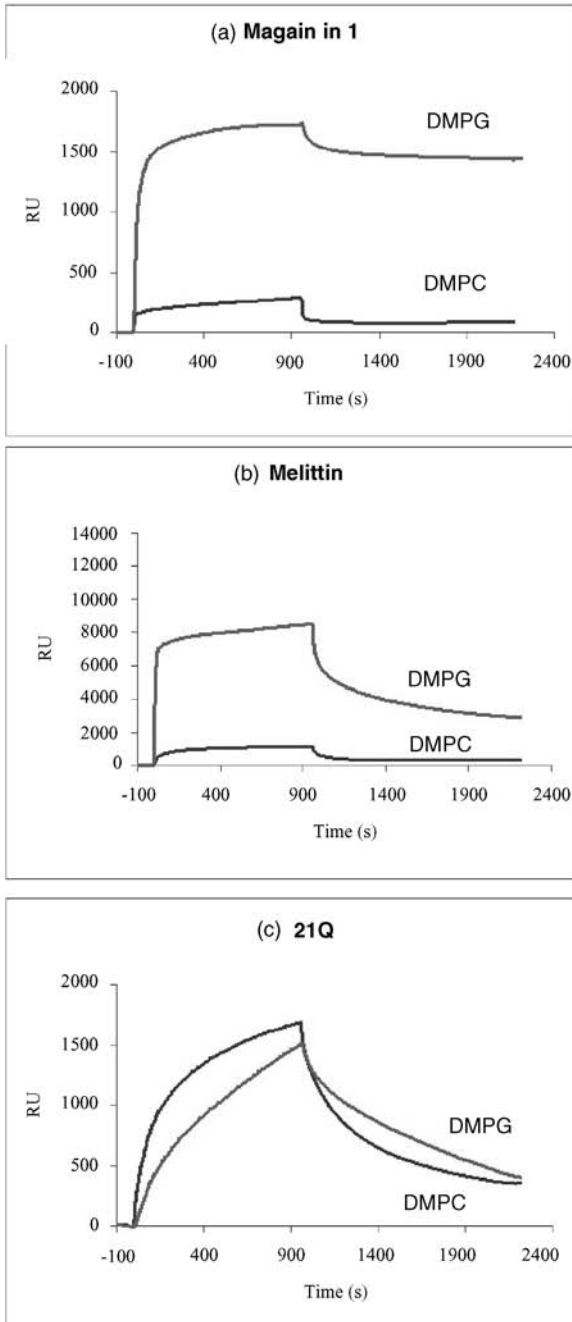
Since the long-term analyte interaction with the membrane is through hydrophobic interactions, the mild regeneration of the membrane surface by removal of the bound peptide is not generally possible without also removing the phospholipid layer. However, in some cases, very long dissociation times can allow weakly bound analyte to totally dissociate from the lipid surfaces before the next analyte concentration is injected. The phospholipid layer is therefore most typically removed completely with an injection of detergents such as octyl-glucoside (HPA) or CHAPS (L1), and each analyte injection is performed on a freshly generated

lipid surface. All binding experiments are carried out at a uniform temperature, typically close to ambient and above the lipid transition temperature (e.g., 25 °C or above is suitable for most membrane systems) and each analyte concentration performed in at least duplicate. The kinetics/affinity of the analyte–membrane binding event is then determined from analysis of a series of sensorgrams collected at different analyte concentrations over each different lipid surface studied.

#### 7.6.1.4 Membrane Binding of Antimicrobial Peptides by SPR

Antimicrobial peptides act via binding to, and disruption of, cell membranes and SPR has been used to study their membrane-binding properties. A number of different mechanisms have been proposed to describe the mode of action of these peptides. These models vary from general bilayer disruption following the binding of a critical concentration of peptide, the formation of pores, or through the binding with specific lipophilic components of the membrane such as lipopolysaccharides [113]. Since these models all involve the binding of peptides to the membrane, the determination of the relative affinity of the peptides for a particular target membrane is central to delineating the mechanism of action of these peptides. SPR therefore has a very important role in furthering our understanding of the molecular basis of action of this class of peptides.

We have used SPR to study the interaction of antimicrobial peptides with membranes of different composition [114–117]. The HPA chip was initially used to characterize the binding of magainin 1 (GIGKFLHSAGKFGKAFVGEIMKS), melittin (H<sub>2</sub>N-GIGAVLKVLTTGLPALISWIKRKRQQ-NH<sub>2</sub>), and its C-terminally truncated analog (21Q; H<sub>2</sub>N-GIGAVLKVLTTGLPALISWIQ-NH<sub>2</sub>) to liposomes comprising either DMPC or DMPG. Figure 7.8 shows the sensorgrams obtained for each peptide with both lipids. The results demonstrated that magainin 1 binds more strongly to negatively charged lipids (Figure 7.8a) and is known to have a high helical content in liposome solution containing anionic lipids as measured by circular dichroism [11]. In contrast, melittin interacts in relative terms with both zwitterionic and anionic lipids (Figure 7.8b), which correlates with observations that melittin binds and lyses both bacterial and eukaryotic cells. The role of the C-terminal positive residues of melittin was also examined through analysis of the membrane-binding properties of the C-terminally truncated analog (21Q) of melittin. As shown in Figure 7.8(c), 21Q exhibits lower binding affinity for both lipids, showing that the positively charged C-terminus of melittin greatly influences its membrane-binding properties. The results have also demonstrated that peptides such as melittin and magainin had higher affinities for both DMPC and DMPG membranes when using the L1 surface compared to the HPA surface [115]. This presumably reflects the ability of the peptides to penetrate the liposomes, whereas membrane insertion is restricted on the HPA surface. Indeed, it was shown that at sufficiently high concentrations, the cytolytic peptide melittin caused the lysis of the immobilized DMPG liposomes as evidenced by a drop in RU during the association phase.



**Figure 7.8** Comparative peptide sensorgrams for the binding of (a) magainin 1, (b) melittin and (c) 21Q (des 22–25-melittin) to DMPC and DMPG. Peptide concentration = 50  $\mu$ M. Reproduced from [11], with permission from Elsevier.

Kinetic analysis of the results generally revealed a poor fit to the simple 1:1 Langmuir binding model. However, analysis with more complex models including a two-state binding or a parallel binding mechanism resulted in a significantly improved fit [114, 115]. The results suggest that there is likely to be at least two steps involved in the interaction between the antimicrobial peptides and the model membrane systems when the peptides may first bind to the lipid head groups and localize themselves on the surface, and then insert further into the hydrocarbon region of the membrane. The absence of positively charged C-terminal residues in 21Q resulted in a loss of binding specificity with low binding affinities of this peptide with both DMPC and DMPG. Comparison of the results of melittin and 21Q (Figure 7.8b and c) thus indicates that the positive tail of melittin allows it to bind more rapidly and more strongly to the anionic lipids by electrostatic interactions, thereby enhancing subsequent hydrophobic binding. These results also demonstrate the role of electrostatic interactions in the initial orientation and binding of these peptides to the membrane, and the ability of SPR measurements to provide insight into membrane-mediated events.

## 7.7

### Data Analysis

The association ( $k_a$ ) and dissociation ( $k_d$ ) rate constants control the formation and breakdown of the complex AB, where A is the protein analyte and B is the surface-bound recognition partner, which bind to form the complex AB:



Using the resultant sensorgrams of each peptide, linearization analysis and curve fitting with numerical integration using standard plotting and statistical software such as Excel (Microsoft, Redmond, WA, USA), GraphPad Prism (GraphPad Software, San Diego, CA, USA) and BIAevaluation software (GE Healthcare, formerly Biacore, Uppsala, Sweden) can be performed to derive estimates for kinetic rate constants ( $k_a$  and  $k_d$ ) and equilibrium constants (steady-state approximations of  $K_A$  and  $K_D$ ).

#### 7.7.1

##### Linearization Analysis

The sensorgrams resulting from the biomolecular interactions can be analyzed by linearization analysis [118, 119]. This method involves a simple linear transformation of binding data to determine binding parameters. The limitation of this method is that it is only suitable for interpreting reactions that follow a simple bimolecular mechanism. The corresponding differential rate equation for this reaction model is:

$$dR/dt = k_a^*[C_P]^*R_{\max} - (k_a^*[C_P] + k_d)^*R \quad (7.2)$$

where  $k_a$  and  $k_d$  are the association and dissociation rate constants, respectively,  $R_{\max}$  is the maximum signal, which is proportional to the initial concentration of  $L$ ,  $[C_P]$  is the analyte concentration, and  $R$  is the signal from the biosensor which is proportional to the amount of complex AB.

To determine whether the bimolecular interaction can be described by a simple 1 : 1 bimolecular reaction, the association phase of the sensorgram of the particular analyte at all concentrations is plotted against the response ( $dR/dt$  versus  $R$ ). In the case of a simple bimolecular interaction, the association rate ( $k_a$ ) of each peptide can be easily determined from the slope of the  $dR/dt$  versus  $R$  curves against the peptide concentration according to Eq. (7.2). The dissociation rate of each peptide can also be determined by linearizing the dissociation phase of the highest concentration of the analyte by plotting  $\ln(R_0/R)$  versus time(s). A linear slope should also be observed in the case of a simple bimolecular interaction and the slope can be used to calculate the dissociation rate ( $k_d$ ) constant.  $R_0$  is the response at the start of the dissociation phase and  $R$  is the response at a selected time  $t$ . However, nonlinear slopes reflect complex interactions and the rate constants often cannot be accurately determined by linear analysis [119].

### 7.7.2

#### Numerical Integration Analysis

Where possible, the sensorgrams for each interaction should also be analyzed by curve fitting using numerical integration analysis [119, 120]. In order to distinguish between the possible binding models, the data is fitted globally by simultaneously fitting of the sensorgrams obtained over a range of different concentrations (minimally seven to 10 concentrations for accuracy). Where significant fitting problems occur, improved fitting is achieved through local (nonglobal) fitting or fitting the association and dissociation phases separately.

Initially, each biomolecular-membrane interaction is first analyzed using the simplest Langmuir (1 : 1) binding model (Eq. (7.1)). For SPR, like any kinetic analysis, the simplest model that accurately describes an interaction is always taken as correct unless there are valid reasons to believe otherwise, as the simplest model is always less susceptible to deviation from the experimental data when changes are applied to the parameters of the assay (such as concentration, flow rates, and temperature), and a complex model implies a very specific mechanistic series of events for interaction with significant consequences in terms of the behavior of the biomolecule and membrane.

Typically, the two-state reaction model and the parallel reaction model are applied to data sets that deviate from the 1 : 1 model. The two-state reaction model (Eq. (7.3)) describes a more complex 1 : 1 binding event. The initial binding of the biomolecule and membrane ( $A + B$ ) is followed by a conformational change or rearrangement in either (or both) the biomolecule or membrane which stabilizes the complex (shown below as  $AB^*$ ). The complex AB changes to  $AB^*$ , which cannot dissociate directly to

A + B and which may correspond to partial insertion of the peptide into the model membranes.

The two-state reaction therefore corresponds to:



The subsequent differential rate equations for this reaction model are represented by:

$$dR_1/dt = k_{a1} * C_A * (R_{max} - R_1 - R_2) - k_{d1} * R_1 - k_{a2} * R_1 + k_{d2} * R_2 \quad (7.4)$$

$$dR_2/dt = k_{a2} * R_1 - k_{d2} * R_2 \quad (7.5)$$

Another example is the parallel reaction model, which assumes that two simple interactions occur in parallel with different rate constants giving a complex overall interaction. In this scenario there will be two separate rates of reaction for the two distinct regions of the surface and the model is described by two parallel simple 1 : 1 models, where one set of processes is related to the surface B1 and the other set of processes is related to the discontinuous surface B2 thus:



The preinteracting biomolecule complex species AB<sub>1</sub> therefore provides a new surface B<sub>2</sub> for the binding of additional biomolecules. The corresponding differential rate equations for this reaction model are:

$$dR_1/dt = k_{a1} * C_p * (R_{max} - R_1) - R_1 * k_{d1} \quad (7.7)$$

$$dR_2/dt = k_{a2} * C_p * (R_{max} - R_2) - R_2 * k_{d2} \quad (7.8)$$

### 7.7.3

#### Steady-State Approximations

The inference of equilibrium constants from steady-states is useful in situations where binding is too complex or parameters too difficult to control when fitting to conventional models. If we again consider the biomolecular interaction at a surface (Eq. (7.1)); when the concentrations of A and AB are equal, and the rate at which the complex AB is formed is equal to the rate at which the complex dissociates, then the concentration of the free analyte and complex no longer change and the system is at equilibrium.

An important parameter to know for equilibrium measurements is the maximum binding capacity  $R_{max}$ . This is the value where 100% of all the available membrane B is saturated with the analyte A to form the AB complex.

We can calculate the maximum theoretical binding capacity ( $R_{\max}$ ) of a surface as follows:

$$R_{\max}(\text{RU}) = \text{biomolecule molecular weight} \times \text{immobilization level (RU)} \\ \times \text{stoichiometry bound partner molecular weight} \quad (7.9)$$

Caution should be applied though, as in practice the measured  $R_{\max}$  rarely reaches the theoretical  $R_{\max}$  for two reasons. (i) The assumption of an equilibrium state is that the analyte has infinite time to interact with the ligand to allow equilibrium to be reached; in biosensors there are often limitations as to how much biomolecule can be injected onto a surface due to liquid handling issues and so it is rarely feasible for sufficient biomolecule to be injected on the surface to reach the  $R_{\max}$ , especially at lower analyte concentrations and faster flow rates. (ii) Most biological reactions occur a long way from equilibrium and it is beyond most SPR instrument sensitivity to directly measure such equilibrium systems.

Despite these complications, equilibrium measurements can be inferred from what is termed the steady-state approximation. A steady-state can best be described as “a plateau” where the rate of change over a defined region is zero as shown in Figure 7.2. A point taken along one of the plateau regions termed  $R_{\text{eq}}$  is used for calculations.

It is important to understand that the steady-state is *not* equilibrium (the term equilibrium should be more correctly used for the ideal state). In a steady-state, the concentration of reactants and products are not necessarily equal, and other processes may be occurring at the same time; however, the rates of association and dissociation of the biomolecular complex AB are such that they yield a zero rate of overall change at the given time point. Despite not being equilibrium, the steady-state can be used to approximate or infer equilibrium constants because, analogously to equilibrium, the association and dissociation rates yield zero changes. A Scatchard analysis can be undertaken by plotting the steady-state report points  $R_{\text{eq}}/[A]$  versus  $R_{\text{eq}}$ , which has a slope equal to the  $K_A$ . Alternatively, at 50% of the  $R_{\max}$  from a plot of  $R_{\text{eq}}$  versus  $[A]$ , the biomolecule concentration is equal to the equilibrium dissociation constant  $K_D$ .

## 7.8

### Conclusions

The last decade has seen the development of a number of SPR-based biosensors capable of analyzing biomolecular interactions in real-time. These instruments have transformed the speed and ease with which these interactions can be studied, and are equally suited to both basic research and clinical and agricultural applications. Together with the appropriate control experiments to ensure specificity, SPR can provide extraordinary insight into biomolecular processes and also allow the detection of a wide range of molecules through the immunoassay format. There is no

doubt that SPR-based biosensors will continue to underpin new developments in biosciences both through improved sensitivity and the development of new hand-held formats for field applications.

## References

- 1 Karlsson, R., Katsamba, P.S., Nordin, H., Pol, E., and Myszka, D.G. (2006) *Analytical Biochemistry*, **349**, 136–147.
- 2 Nice, E.C., Rothacker, J., Weinstock, J., Lim, L., and Catimel, B. (2007) *Journal of Chromatography A*, **1168**, 190–210, editorial 189.
- 3 Rich, R.L., Cannon, M.J., Jenkins, J., Pandian, P., Sundaram, S., Magyar, R., Brockman, J., Lambert, J., and Myszka, D.G. (2008) *Analytical Biochemistry*, **373**, 112–120.
- 4 Rich, R.L. and Myszka, D.G. (2008) *Journal of Molecular Recognition*, **21**, 355–400.
- 5 Nice, E.C. and Catimel, B. (1999) *BioEssays*, **21**, 339–352.
- 6 Fivash, M., Towler, E.M., and Fisher, R.J. (1998) *Current Opinion in Biotechnology*, **9**, 97–101.
- 7 Malmqvist, M. and Karlsson, R. (1997) *Current Opinion in Chemical Biology*, **1**, 378–383.
- 8 Cooper, M.A., Hansson, A., Lofas, S., and Williams, D.H. (2000) *Analytical Biochemistry*, **277**, 196–205.
- 9 Heyse, S., Stora, T., Schmid, E., Lakey, J.H., and Vogel, H. (1998) *Biochimica et Biophysica Acta*, **1376**, 319–338.
- 10 Kuziemko, G.M., Stroh, M., and Stevens, R.C. (1996) *Biochemistry*, **35**, 6375–6384.
- 11 Mozsolits, H., Wirth, H.-J., Werkmeister, J., and Aguilar, M.-I. (2001) *Biochimica et Biophysica Acta*, **1512**, 64–76.
- 12 Saenko, E., Sarafanov, A., Ananyeva, N., Behre, E., Shima, M., Schwinn, H., and Josic, D. (2001) *Journal of Chromatography A*, **921**, 49–56.
- 13 Wang, W., Smith, D.K., Moulding, K., and Chen, H.M. (1998) *Journal of Biological Chemistry*, **273**, 27438–27448.
- 14 Mozsolits, H. and Aguilar, M.-I. (2002) *Biopolymers*, **66**, 3–18.
- 15 Green, R.J., Frazier, R.A., Shakesheff, K.M., Davies, M.C., Roberts, C.J., and Tendler, S.J.B. (2000) *Biomaterials*, **21**, 1823–1835.
- 16 Jung, L.S., Campbell, C.T., Chinowsky, T.M., Mar, M.N., and Yee, S.S. (1998) *Langmuir*, **14**, 5636–5648.
- 17 Shankaran, D.R., Gobi, K.V., and Miura, N. (2007) *Sensors and Actuators B*, **121**, 158–177.
- 18 Englebienne, P., Van Hoonacker, A., and Verhas, M. (2003) *Spectroscopy*, **17**, 255–273.
- 19 Karlsson, R. (2004) *Journal of Molecular Recognition*, **17**, 151–161.
- 20 Mullett, W.M., Lai, E.P.C., and Yeung, J.M. (2000) *Methods*, **22**, 77–91.
- 21 Winzor, D.J. (2003) *Analytical Biochemistry*, **318**, 1–12.
- 22 Dillon, P.P., Daly, S.J., Killard, A.J., and O’Kennedy, R. (2003) *International Journal of Environmental Analytical Chemistry*, **83**, 525–543.
- 23 Mallat, E., Barceló, D., Barzen, C., Gauglitz, G., and Abuknesha, R. (2001) *Trends in Analytical Chemistry*, **20**, 124–132.
- 24 Mauriz, E., Calle, A., Montoya, A., and Lechuga, L.M. (2006) *Talanta*, **69**, 359–364.
- 25 Boltovets, P.M., Snopok, B.A., Boyko, V.R., Shevchenko, T.P., Dyachenko, N.S., and Shirshov, Y.M. (2004) *Journal of Virological Methods*, **121**, 101–106.
- 26 Kanoh, N., Kyo, M., Inamori, K., Ando, A., Asami, A., Nakao, A., and Osada, H. (2006) *Analytical Chemistry*, **78**, 2226–2230.
- 27 Kim, M., Jung, S.O., Park, K., Jeong, E.-J., Joung, H.-A., Kim, T.-H., Seol, D.-W., and Chung, B.H. (2005) *Biochemical and Biophysical Research Communications*, **338**, 1834–1838.
- 28 Oli, M.W., McArthur, W.P., and Brady, L.J. (2006) *Journal of*



- Microbiological Methods*, **65**, 503–511.
- 29 Zhang, W., Krishnan, N., and Becker, D.F. (2006) *Archives of Biochemistry and Biophysics*, **445**, 174–183.
  - 30 Chung, J.W., Kim, S.D., Bernhardt, R., and Pyun, J.C. (2005) *Sensors and Actuators B*, **111/112**, 416–422.
  - 31 Haes, A.J., Chang, L., Klein, W.L., and Van Duynne, R.P. (2005) *Journal of the American Chemical Society*, **127**, 2264–2271.
  - 32 Kumbhat, S., Shankaran, D.R., Kim, S.J., Gobi, K.V., Joshi, V., and Miura, N. (2007) *Biosensors and Bioelectronics*, **23**, 421–427.
  - 33 Abdiche, Y.N. and Myszka, D.G. (2004) *Analytical Biochemistry*, **328**, 233–243.
  - 34 Myszka, D.G. and Rich, R.L. (2000) *Pharmaceutical Science and Technology Today*, **3**, 310–317.
  - 35 Reddy, T.R.K., Mutter, R., Heal, W., Guo, K., Gillet, V.J., Pratt, S., and Chen, B. (2006) *Journal of Medicinal Chemistry*, **49**, 607–615.
  - 36 Rich, R.L. and Myszka, D.G. (2001) *Journal of Molecular Recognition*, **14**, 273–294.
  - 37 Ivnitski, D., Abdel-Hamid, I., Atanasov, P., and Wilkins, E. (1999) *Biosensors and Bioelectronics*, **14**, 599–624.
  - 38 Pattnaik, P. and Srivastav, A. (2006) *Journal of Food Science and Technology*, **43**, 329–336.
  - 39 Spadavecchia, J., Manera, M.G., Quaranta, F., Siciliano, P., and Rella, R. (2005) *Biosensors and Bioelectronics*, **21**, 894–900.
  - 40 Homola, J., Yee, S.S., and Gauglitz, G. (1999) *Sensors and Actuators B*, **54**, 3–15.
  - 41 Hoa, X.D., Kirk, A.G., and Tabrizian, M. (2007) *Biosensors and Bioelectronics*, **23**, 151–160.
  - 42 Kretschmann, E. (1971) *Zeitschrift fur Physik*, **241**, 313–324.
  - 43 Otto, A. (1968) *Zeitschrift fur Physik*, **216**, 398–410.
  - 44 Stenberg, E., Persson, B., Roos, H., and Urbaniczky, C. (1991) *Journal of Colloid and Interface Science*, **143**, 513–526.
  - 45 Liedberg, B., Nylander, C., and Lundström, I. (1995) *Biosensors and Bioelectronics*, **10**, i–ix.
  - 46 Rich, R.L. and Myszka, D.G. (2000) *Current Opinion in Biotechnology*, **11**, 54–61.
  - 47 Rich, R.L. and Myszka, D.G. (2000) *Journal of Molecular Recognition*, **13**, 388–407.
  - 48 Rich, R.L. and Myszka, D.G. (2002) *Journal of Molecular Recognition*, **15**, 352–376.
  - 49 Rich, R.L. and Myszka, D.G. (2005) *Journal of Molecular Recognition*, **18**, 1–39.
  - 50 Rich, R.L. and Myszka, D.G. (2006) *Journal of Molecular Recognition*, **19**, 478–534.
  - 51 McCormick, A.N., Leach, M.E., Savidge, G., and Alhaq, A. (2004) *Clinical and Laboratory Haematology*, **26**, 57–64.
  - 52 Myszka, D.G. (1997) *Current Opinion in Biotechnology*, **8**, 50–57.
  - 53 Mistrík, P., Moreau, F., and Allen, J.M. (2004) *Analytical Biochemistry*, **327**, 271–277.
  - 54 Mozsolits, H., Thomas, W.G., and Aguilar, M.I. (2003) *Journal of Peptide Science*, **9**, 77–89.
  - 55 Kai, E., Sawata, S., Ikebukuro, K., Iida, T., Honda, T., and Karube, I. (1999) *Analytical Chemistry*, **71**, 796–800.
  - 56 Persson, B., Stenhagen, K., Nilsson, P., Larsson, A., Uhlén, M., and Nygren, P.-Å. (1997) *Analytical Biochemistry*, **246**, 34–44.
  - 57 Cheng, Y., Dubovoy, N., Hayes-Rogers, M.E., Stewart, J., and Shah, D. (1999) *Journal of Immunological Methods*, **230**, 29–35.
  - 58 Kyprianou, D., Guerreiro, A.R., Chianella, I., Piletska, E.V., Fowler, S.A., Karim, K., Whitcombe, M.J., Turner, A.P.F., and Piletsky, S.A. (2009) *Biosensors and Bioelectronics*, **24**, 1365–1371.
  - 59 Navratilova, I., Papalia, G.A., Rich, R.L., Bedinger, D., Brophy, S., Condon, B., Deng, T., Emerick, A.W., Guan, H.-W., Hayden, T., Heutmekers, T., Hoorelbeke, B., McCroskey, M.C., Murphy, M.M., Nakagawa, T., Parmeggiani, F., Qin, X., Rebe, S., Tomasevic, N., Tsang, T., Waddell, M.B., Zhang, F.F., Leavitt, S., and Myszka, D.G. (2007) *Analytical Biochemistry*, **364**, 67–77.

- 60 Roos, H., Karlsson, R., Nilshans, H., and Persson, A. (1998) *Journal of Molecular Recognition*, **11**, 204–210.
- 61 Baird, C.L. and Myszka, D.G. (2001) *Journal of Molecular Recognition*, **14**, 261–268.
- 62 Brawman, T., Bronner, V., Lavie, K., Notcovich, A., Papalia, G.A., and Myszka, D.G. (2006) *Analytical Biochemistry*, **358**, 281–288.
- 63 Marchesini, G.R., Koopal, K., Meulenberg, E., Haasnoot, W., and Irth, H. (2007) *Biosensors and Bioelectronics*, **22**, 1908–1915.
- 64 Löfås, S. (2004) *Assay and Drug Development Technologies*, **2**, 407–416.
- 65 Gonzalez-Techera, A., Kim, H.J., Gee, S.J., Last, J.A., Hammock, B.D., and Gonzalez-Sapienza, G. (2007) *Analytical Chemistry*, **79**, 9191–9196.
- 66 Hage, D.S. (1999) *Analytical Chemistry*, **71**, 294R–304R.
- 67 Kim, H.-J., Gonzalez-Techera, A., Gonzalez-Sapienza, G.G., Ahn, K.C., Gee, S.J., and Hammock, B.D. (2008) *Environmental Science and Technology*, **42**, 2047–2053.
- 68 Kobayashi, N., Iwakami, K., Kotoshiba, S., Niwa, T., Kato, Y., Mano, N., and Goto, J. (2006) *Analytical Chemistry*, **78**, 2244–2253.
- 69 Ngo, T.T. (2000) *Methods*, **22**, 1–3.
- 70 Strachan, G., Whyte, J.A., Molloy, P.M., Paton, G.I., and Porter, A.J.R. (2000) *Environmental Science and Technology*, **34**, 1603–1608.
- 71 Wellman, A.D. and Sepaniak, M.J. (2006) *Analytical Chemistry*, **78**, 4450–4456.
- 72 Yuan, J., Addo, J., Aguilar, M.-I., and Wu, Y. (2009) *Analytical Biochemistry*, **390**, 97–99.
- 73 Yuan, J., Deng, D., Lauren, D.R., Aguilar, M.-I., and Wu, Y. (2009) *Analytica Chimica Acta*, **656**, 63–71.
- 74 Yuan, J., Oliver, R., Aguilar, M.-I., and Wu, Y. (2008) *Analytical Chemistry*, **80**, 8329–8333.
- 75 Yuan, J., Oliver, R., Li, J., Lee, J., Aguilar, M., and Wu, Y. (2007) *Biosensors and Bioelectronics*, **23**, 144–148.
- 76 Daly, S.J., Keating, G.J., Dillon, P.P., Manning, B.M., O’Kennedy, R., Lee, H.A., and Morgan, M.R.A. (2000) *Journal of Agricultural and Food Chemistry*, **48**, 5097–5104.
- 77 Gobi, K.V., Tanaka, H., Shoyama, Y., and Miura, N. (2004) *Biosensors and Bioelectronics*, **20**, 350–357.
- 78 Kim, S.J., Gobi, K.V., Harada, R., Shankaran, D.R., and Miura, N. (2006) *Sensors and Actuators B*, **115**, 349–356.
- 79 Miura, N., Ogata, K., Sakai, G., Uda, T., and Yamazoe, N. (1997) *Chemistry Letters*, 713–714.
- 80 Shankaran, D.R., Matsumoto, K., Toko, K., and Miura, N. (2006) *Sensors and Actuators B*, **114**, 71–79.
- 81 Yu, Q., Chen, S., Taylor, A.D., Homola, J., Hock, B., and Jiang, S. (2005) *Sensors and Actuators B*, **107**, 193–201.
- 82 Gratacos-Cubarsi, M., Castellari, M., Valero, A., and Garcia-Regueiro, J.A. (2006) *Analytical and Bioanalytical Chemistry*, **385**, 1218–1224.
- 83 Wen, Y., Zhang, M., Zhao, Q., and Feng, Y.-Q. (2005) *Journal of Agricultural and Food Chemistry*, **53**, 8468–8473.
- 84 Yang, S., Cha, J., and Carlson, K. (2005) *Journal of Chromatography A*, **1097**, 40–53.
- 85 Maudens, K.E., Zhang, G.-F., and Lambert, W.E. (2004) *Journal of Chromatography A*, **1047**, 85–92.
- 86 Pang, G.-F., Cao, Y.-Z., Zhang, J.-J., Jia, G.-Q., Fan, C.-L., Li, X.-M., Liu, Y.-M., Li, Z.-Y., and Sin, Y.-Q. (2005) *Journal of AOAC International*, **88**, 1304–1311.
- 87 Santos, B., Lista, A., Simonet, B.M., Rios, A., and Valcarcel, M. (2005) *Electrophoresis*, **26**, 1567–1575.
- 88 Paseková, H., Polásek, M., Cigarro, J.F., and Dolejšová, J. (2001) *Analytica Chimica Acta*, **438**, 165–173.
- 89 Johnsson, B., Löfås, S., and Lindquist, G. (1991) *Analytical Biochemistry*, **198**, 268–277.
- 90 Lofas, S., Johnsson, B., Edström, A., Hansson, A., Lindquist, G., Hillgren, R.-M.M., and Stigh, L. (1995) *Biosensors and Bioelectronics*, **10**, 813–822.
- 91 O’Shannessy, D.J., Brigham-Burke, M., and Peck, K. (1992) *Analytical Biochemistry*, **205**, 132–136.
- 92 Nagata, K. and Handa, H. (eds) (2000) *Real-Time Analysis of Biomolecular*

- Interactions: Applications of BIACORE*, Springer, Tokyo.
- 93 Nakamura, C., Hasegawa, M., Nakamura, N., and Miyake, J. (2003) *Biosensors and Bioelectronics*, **18**, 599–603.
- 94 Sonezaki, S., Yagi, S., Ogawa, E., and Kondo, A. (2000) *Journal of Immunological Methods*, **238**, 99–106.
- 95 Severs, A.H., Schasfoort, R.B.M., and Salden, M.H.L. (1993) *Biosensors and Bioelectronics*, **8**, 185–189.
- 96 Wei, J., Mu, Y., Song, D., Fang, X., Liu, X., Bu, L., Zhang, H., Zhang, G., Ding, J., Wang, W., Jin, Q., and Luo, G. (2003) *Analytical Biochemistry*, **321**, 209–216.
- 97 Charles, P.T., Rangasammy, J.G., Anderson, G.P., Romanoski, T.C., and Kusterbeck, A.W. (2004) *Analytica Chimica Acta*, **525**, 199–204.
- 98 Rabbany, S.Y., Lane, W.J., Marganski, W.A., Kusterbeck, A.W., and Ligler, F.S. (2000) *Journal of Immunological Methods*, **246**, 69–77.
- 99 Lee, W., Lee, D.-B., Oh, B.-K., Lee, W.H., and Choi, J.-W. (2004) *Enzyme and Microbial Technology*, **35**, 678–682.
- 100 Homola, J., Dostálek, J., Chen, S., Rasooly, A., Jiang, S., and Yee, S.S. (2002) *International Journal of Food Microbiology*, **75**, 61–69.
- 101 Gillis, E.H., Gosling, J.P., Sreenan, J.M., and Kane, M. (2002) *Journal of Immunological Methods*, **267**, 131–138.
- 102 Gobi, K.V., Tanaka, H., Shoyama, Y., and Miura, N. (2005) *Sensors and Actuators B*, **111/112**, 562–571.
- 103 Nguyen, B., Tanius, F.A., and Wilson, W.D. (2007) *Methods*, **42**, 150–161.
- 104 Sakai, G., Nakata, S., Uda, T., Miura, N., and Yamazoe, N. (1999) *Electrochimica Acta*, **44**, 3849–3854.
- 105 Shankaran, D.R., Gobi, K.V., Sakai, T., Matsumoto, K., Toko, K., and Miura, N. (2005) *Biosensors and Bioelectronics*, **20**, 1750–1756.
- 106 Henriques, S.T., Pattenden, L.K., Aguilar, M.-I., and Castanho, M.A.R.B. (2008) *Biophysical Journal*, **95**, 1877–1889.
- 107 Mozsolits, H., Unabia, S., Ahmad, A., Morton, C.J., Thomas, W.G., and Aguilar, M.I. (2002) *Biochemistry*, **41**, 7830–7840.
- 108 Karlsson, O.P. and Lofas, S. (2002) *Analytical Biochemistry*, **300**, 132–138.
- 109 Papo, N. and Shai, Y. (2003) *Biochemistry*, **42**, 458–466.
- 110 Veiga, A.S., Pattenden, L.K., Fletcher, J.M., Castanho, M.A.R.B., and Aguilar, M.I. (2009) *ChemBioChem*, **10**, 1032–1044.
- 111 Cooper, M.A., Try, A.C., Carroll, J., Ellar, D.J., and Williams, D.H. (1998) *Biochimica et Biophysica Acta*, **1373**, 101–111.
- 112 Hou, X., Richardson, S.J., Aguilar, M.-I., and Small, D.H. (2005) *Biochemistry*, **44**, 11618–11627.
- 113 Shai, Y. (1999) *Biochimica et Biophysica Acta*, **1462**, 55–70.
- 114 Jin, Y., Mozsolits, H., Hammer, J., Zmuda, E., Zhu, F., Zhang, Y., Aguilar, M.I., and Blazyk, J. (2003) *Biochemistry*, **42**, 9395–9405.
- 115 Lee, T.H., Mozsolits, H., and Aguilar, M.-I. (2001) *Journal of Peptide Research*, **58**, 464–476.
- 116 Hall, K. and Aguilar, M.-I. (2009) *Biopolymers*, **92**, 554–564.
- 117 Hall, K., Lee, T., and Aguilar, M. (2011) *Journal of Molecular Recognition*, **24**, 108–118.
- 118 Morton, T.A. and Myszka, D.G. (1998) *Methods in Enzymology*, **295**, 268–282.
- 119 Morton, T.A., Myszka, D.G., and Chaiken, I.M. (1995) *Analytical Biochemistry*, **227**, 176–185.
- 120 Karlsson, R. and Falt, A. (1997) *Journal of Immunological Methods*, **200**, 121–133.



## 8

# Atomic Force Microscopy of Proteins

Adam Mechler

### 8.1

#### Foreword

“Seeing is believing,” says the proverb. Yet, when it comes to the molecular world, our present knowledge is almost exclusively derived from indirect evidence. Even our perception of the “visual appearance” of a molecule is based on models and simplifications. This applies even more pronouncedly to our view of chemical reactions, especially the spatial aspects thereof. In the case of small molecules, sufficient details of a reaction mechanism might be obtained with spectroscopic measurements. For larger biomolecules, however, the steric factors inherent in the complexity of their structures prohibit characterization with such simple means; studying the interactions of proteins poses thus a disproportionately larger problem. The common solution is the painstaking process of testing interaction affinities of a large number of mutants, aiming at identifying key residues and thus the preferential binding sites. An alternative possibility is performing molecular dynamic simulations. Both are time- and resource-consuming; it would be much more desirable to simply observe the process, by using microscopic techniques.

Fluorescent confocal microscopy is frequently used to image protein distributions in live cells and by using Förster resonance energy transfer measurements even interactions of proteins can be identified. In spatial resolution, however, all optical techniques are constrained by the optical diffraction limit. While methods to “cheat” the diffraction limit do exist, resolving nanometer-scale structures is still out of reach. Electron microscopy is another frequently used method to study biomolecules with high resolution. However, electron microscopy requires a dehydrated, fixed or frozen sample; it is thus limited to image, at best, a snapshot of an interaction. Only atomic force microscopy (AFM) [1] offers the means of imaging individual biomolecules in their native, physiological environment *in vitro* and, in the case of unicellular organisms, *in vivo*. Thus, this chapter will explore the means, the advantages, and the disadvantages of AFM imaging.

### 8.1.1

#### Importance of Asking the Right Question

AFM creates a three-dimensional map of a surface morphology. Thus, AFM imaging might be performed on surface-confined systems. That includes natural surface processes, such as membrane–protein interactions, ion channel activity, and so on, as well as surface-immobilized processes where biomolecules are attached to the surface artificially. Sample preparation is a key issue and successful imaging often requires a long, iterative process to optimize the immobilization method. Thus, AFM imaging should be used in cases when it is able to provide a new perspective or to solve an outstanding problem; basically, to answer a question that is not possible with any other method. Given the ambiguity of the definition of the surface of a molecule, it is likely that simply “having a look” at a protein does not do any of the above – most monomeric proteins appear roundish, distorted more or less due to surface adhesion, dehydration, or compression by the imaging probe. Asking the right question is central to the AFM of proteins.

As AFM maps morphology, it might be able to identify any distinct *geometric* feature predicted from theory or indirect measurements. Such geometric features might be the oligomerization of proteins, specific binding of protein(s) to DNA, membrane insertion/disruption, pore formation, pore opening and closing, and so on. An alternative possibility is to use the force-sensing capabilities of AFM to perform mechanical measurements of, for example, elasticity or to identify/map antibody–antigen interactions. The feasibility of using AFM imaging must be evaluated on a case-by-case basis and it presumes substantial knowledge of the operation principles as well as the physics of the probe–sample interaction.

## 8.2

### AFM

#### 8.2.1

##### Principle and Basic Modes of Operation

AFM uses a mechanical microprobe to map the sample surface in three dimensions. While the principle is old – it resembles the way turntables read music encoded as topography – the AFM stands out due to a number of clever technical solutions. The first is the machining of the probe – a near atomically sharp tip mounted on a cantilever spring made entirely from silica – by using standard clean-room technology making it possible to use the probe as a disposable unit, retaining its sharpness and sensitivity by regular replacements. Another such solution is the use of laser lever position sensing. Optical resolution is limited by diffraction: two objects the separation distance of which is shorter than half the wavelength of the imaging light will appear as a single object, since light cannot be focused to an arbitrary small spot. No such limitations exist, however, to the angle of specular reflection, provided that the area of the mirroring surface remains large compared to the wavelength of

the light. Thus, even if one corner of a mirror is lifted/lowered by only a small fraction of the wavelength of the reflected light, the resulting tilt is converted into a corresponding angular direction change of the reflected laser beam and at a long enough distance, the displacement of the laser beam is measurable with a photodetector. In the case of the AFM, the mirror is the polished back of the probe cantilever, with a position sensing detector at a distance of a few centimeters. Thus, images could be recorded by simply mapping the bending of the cantilever. However, the range would be seriously limited and the sample distorted or even destroyed due to the variations in the pushing force. For this reason a feedback circuit is implemented to maintain constant force: the signal read from the photodetector is compared to a reference value, the *setpoint*, and, while raster scanning the surface, a piezo actuator (*Z scanner*) adjusts the height of the clamped end of the cantilever to return the bending to the set value. This mode of operation is known as *contact mode*. The trajectory traced by the Z scanner is then recorded as topography, while the torsion of the cantilever (*lateral force*) reveals the tribological properties of the surface.

A major disadvantage of the contact mode is the emergence of high shear forces. Hard surfaces might be imaged this way; however, the working mode is unsuitable for soft, poorly bound biological samples. The problem was solved successfully with the invention of further working modes where the cantilever is not static, but is oscillated at its resonance frequency. If this oscillating probe is moved into the vicinity of the surface, the tip would periodically hit the surface and this “tapping” restrains the amplitude of oscillation. By maintaining constant amplitude, the feedback can control the distance from the surface similarly to the constant force mode. This kind of operation is known by many names (e.g., intermittent contact mode, semicontact mode, AC mode), due to trademark and patent issues; generally it is referred to as “*tapping mode*” (TappingMode™, Veeco Instruments; [www.veeco.com](http://www.veeco.com)).

### 8.2.2

#### How Does a Tip Tap?

A well-known paper published under this title describes the dynamical properties of the probe in fine detail by introducing a realistic continuum-mechanical treatment of the probe–surface interaction [2]. Since all the relevant physical details are described in the cited article, here we omit the discussion of the full, rather complex interaction model and embark on a semiphenomenological treatment of the probe mechanics, which is nevertheless sufficient to demonstrate a number of important issues related to the control and accuracy of imaging.

First, look at the physics of the probe oscillation far from the surface. The probe is commonly treated as a massless spring (spring constant:  $D$ ) with an effective mass ( $m$ ) at the end. This is known in physics as a linear harmonic oscillator, with its resonance (circular) frequency given by  $\omega_0 = \sqrt{D/m}$ . Considering some environmental damping  $\eta$  proportional to the velocity of the probe and drive the base of the cantilever by an actuator, the equation of motion takes the form:

$$m \frac{d^2x}{dt^2} = -Dx - \eta \frac{dx}{dt} + F_0 \sin \omega t \quad (8.1)$$

where  $x$  is the distance along the trajectory of the oscillation and  $F_0$  is the periodic drive force with  $\omega$  (circular) frequency. The solution of this differential equation is a sinusoidal periodic motion:

$$x(t) = A \sin(\omega t + \delta) \quad (8.2)$$

with an amplitude given by:

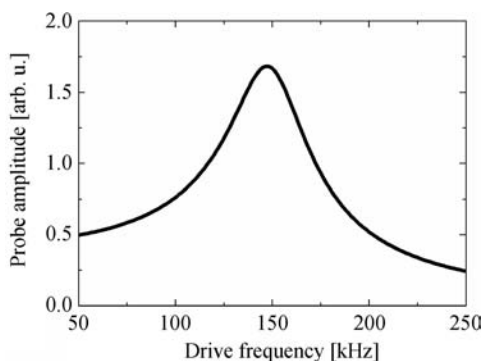
$$A = \frac{a_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2 \omega^2}} \quad (8.3)$$

where  $a_0 = F_0/m$  and  $\beta = \eta/2m$ , and the phase lag between the drive and the oscillator is:

$$\text{tg} \delta = \frac{2\beta\omega}{\omega_0^2 - \omega^2} \quad (8.4)$$

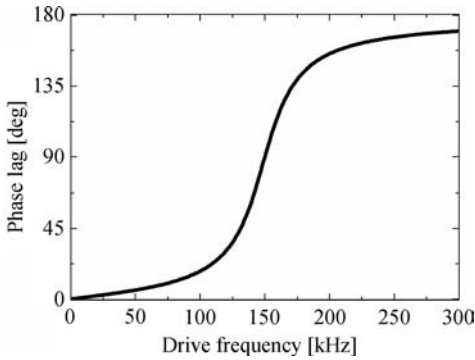
Equation (8.3), when plotted against the drive frequency, is near zero except for a peak around the natural frequency  $\omega_0$ , called the resonance curve (Figure 8.1). Accordingly, the tapping mode AFM probe has to be driven close to the natural frequency – or resonance frequency – to have a measurable amplitude. The phase lag  $\delta$  (Eq. (8.4)) also exhibits resonance properties: it undergoes a  $180^\circ$  change in the proximity of the resonance frequency, with a steep slope (Figure 8.2). Both the phase and the amplitude of the oscillation are sensitive to small changes of the mass or the spring constant of the oscillator; both are also sensitive to the introduction of an external force field, such as the tip–surface interaction. Thus either might be used to establish a feedback to maintain constant probe–surface distance.

Tapping, while facilitating the imaging of soft material, introduces a number of other problems, which we will discuss in due course. First, however, consider the capabilities of AFM through a few imaging highlights.



**Figure 8.1** Example of an amplitude resonance curve, based on Eq. (8.3).





**Figure 8.2** Example of a phase resonance curve, based on Eq. (8.4).

### 8.3

#### Bioimaging Highlights

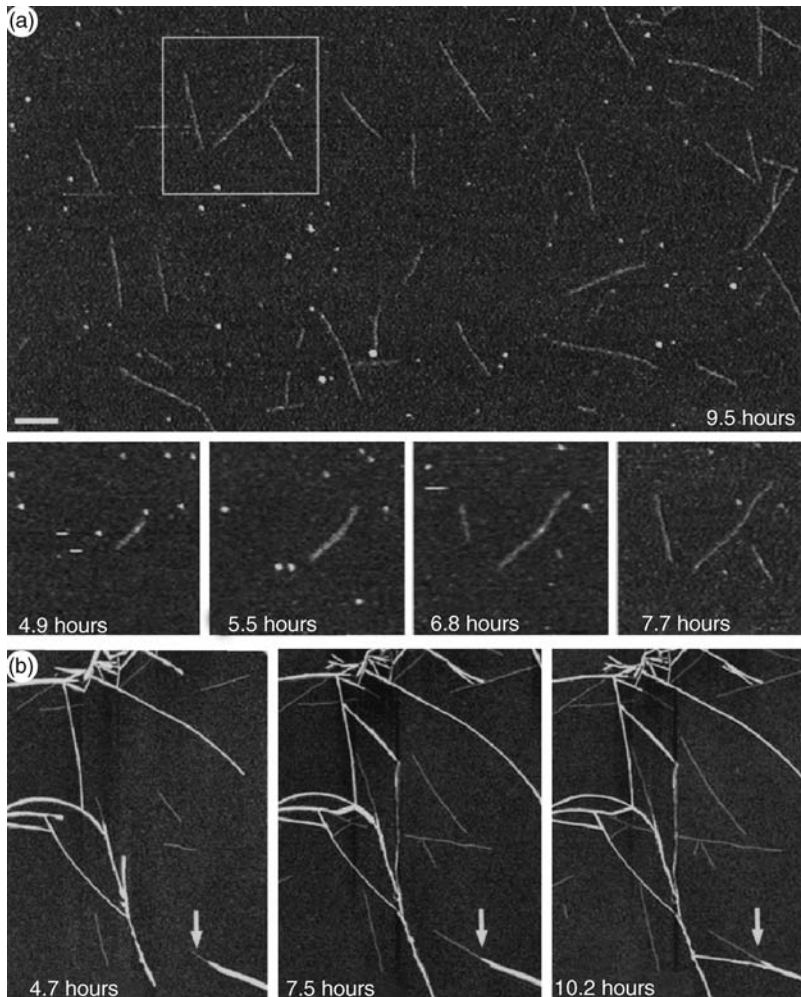
AFM is inherently a surface method. Many biological systems of interest, while being spatially confined (e.g., membrane proteins, ion channels), are not particularly surface processes. The geometry therefore often prohibits *in vivo* studies. However, using the principle of biomimetics, when a sufficient complexity of physiological environment is recreated to provide biological-identical conditions, protein function and interactions might be successfully studied. Several reviews have been written about AFM bioimaging (e.g., [3–5]). Here, we limit discussion to a few highlights. These by no means constitute a review of the field or even of the potential applications; the intention is to demonstrate, through a few select examples, how to use AFM in biomolecule research to obtain information that is not possible or not convenient with other methods.

#### 8.3.1

##### Protein Oligomerization, Aggregation, and Fibers

The simplest imaging problem is measuring the size of an object, such as a protein. Due to ambiguities of size measurements on the nanometer scale, however, it is more practical if the question to be answered is the presence or absence of multiple populations, such as monomeric and oligomeric forms of a protein (e.g., [6–8]). Asking the question this way allows for a quantitative answer, as the number of objects in each of the populations might be easily counted with manual or automated methods.

An important form of protein aggregation is fiber growth (biopolymerization), which is implicated in a number of diseases including diabetes and Alzheimer's. In the first example, the growth of amylin fibers – constituents of pancreatic deposits in diabetes – was tracked with AFM over time [9]. Imaging was performed under liquid, using the tapping mode. Figure 8.3 shows the formation of the smaller protofibrils as well as the assembly and growth of the larger-order amylin fibers. Unique to AFM



**Figure 8.3** Watching amylin fibrils grow on a piece of mica in the AFM. (a) Selected area on the mica revealing exclusively the 2.4-nm protofibrils. The gallery displays previous scans of the boxed area in the larger overview picture at the times indicated. Bidirectional fibril growth is evident. (b) Gallery of three scans of an area selected to show both the 2.4-nm protofibrils and higher-order fibrils growing over time. Arrows point to a protofibril growing from the end of a higher-order fibril. Scale bar = 200 nm. The galleries shown in (a) and (b) show selected scans from two different experiments. (Reproduced with permission from [9].)

imaging was the information obtained on the (i) growth rate of individual fibrils, (ii) direction of growth (it was bidirectional), and (iii) morphological changes of individual fibers with growth.

The second example is of the oligomerization of transthyretin – a protein implicated in a particular neurodegenerative disease, familial amyloidotic polyneuropathy

(FAP) [10]. In this study, aggregation properties of the wild-type and an amyloidogenic mutant, typical to those mutants found in the body of FAP sufferers, were compared in time-lapse studies to structurally identify the neurotoxic species. AFM imaging complemented dynamical light scattering studies, which give an average size without revealing the shape, and tissue culture toxicity studies, which identified the age of the neurotoxic species. The advantage of using AFM is to distinguish between globular and fibrillar aggregates (Figure 8.4).

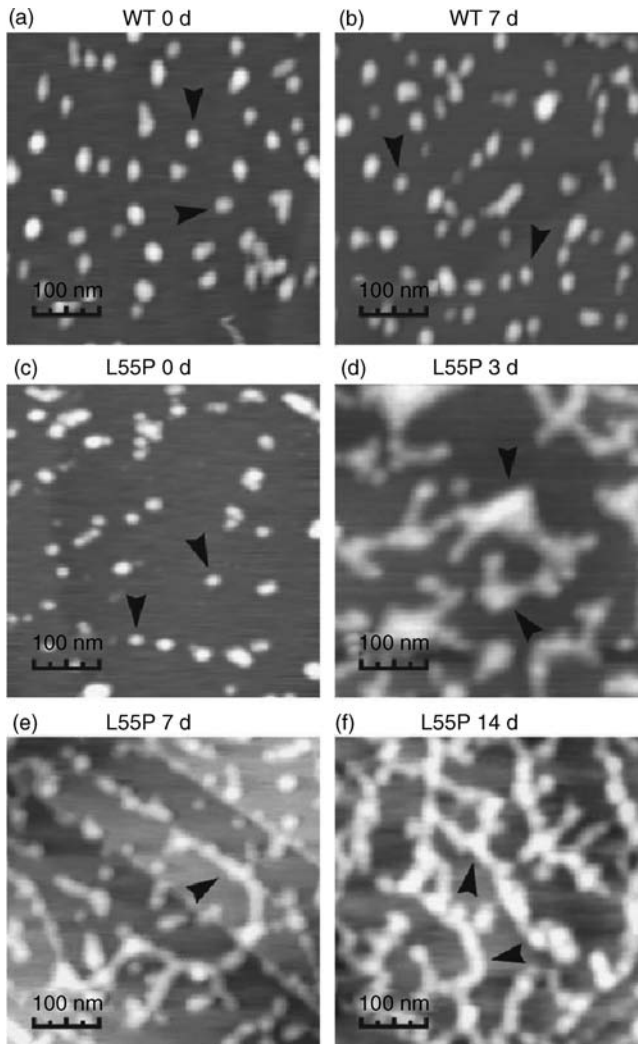
### 8.3.2

#### **Membrane Binding and Lysis**

Membrane interaction of proteins is a broad research area, ranging from the study of integral membrane proteins to lytic antimicrobial peptides. AFM might be used to image proteins on the surface of a live cell; what is more common, however, is to image protein interactions with a supported biomimetic membrane, in a physiological buffer solution. In a stable system, time-lapse measurements might be possible, recording a slow movie of, for example, a membrane disruption process [11, 12].

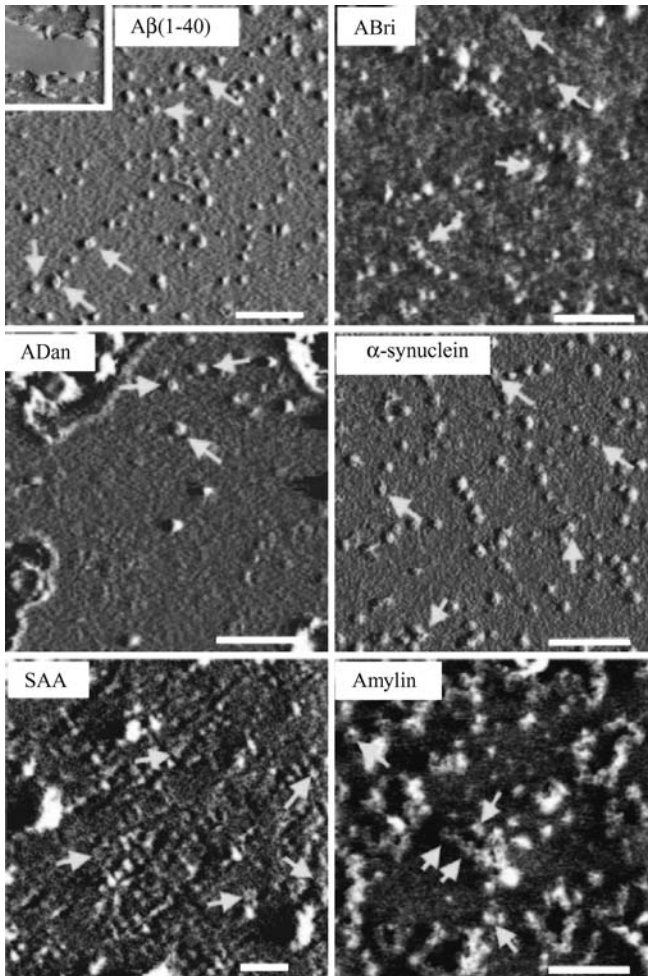
The first example is of the membrane binding of small oligomers of amyloid proteins, implicated in neurodegenerative diseases such as Alzheimer's disease. It was hypothesized that these small oligomers form pores on the membrane surface, and exhibit neurotoxicity by creating an uncontrolled "leak" between the extracellular fluid and the cytosol. Here, AFM was used to confirm the membrane binding of the small oligomers of a range of amyloidogenic proteins (Figure 8.5), while electrophysiology confirmed the presence of transmembrane pores. The unique information provided by the AFM was the identification of the size and geometry of the oligomers [13].

Membrane interactions of antimicrobial lytic peptides constitute a second example. In this field, there is ongoing debate about the mechanism by which the peptides disrupt bacterial membranes. The fact of disruption was established with dye leakage and electrophysiological measurements, membrane binding and lysis with surface plasmon resonance and quartz crystal microbalance studies, and the antimicrobial effect with live cell experiments. Hypothetically, two main mechanisms were identified: transmembrane pore formation and carpet-like disruption, when the membrane suddenly falls apart after a "carpet" of peptides binds to its surface. No direct proof of either mechanism could be obtained, however, without direct imaging by solution AFM of the process of membrane interaction of these peptides. When imaging model peptides incorporating into membranes [14], it was found that mixed domains may form where the peptides modulate the membrane morphology into striated domains (Figure 8.6). This morphology was unanticipated as no other method suggested the formation of such structures; however, it has been found since for a range of supposedly pore-forming antimicrobial peptides. Thus, in this study, AFM has given a unique insight into the peptide–membrane interaction.



**Figure 8.4** Analysis of transthyretin aggregation by AFM. Wild type (WT) transthyretin was examined at 0 (a) and 7 days (b) of aging, whereas L55P (mutant) transthyretin was examined at 0 (c), 3 (d), 7 (e), and 14 days (f) of aging. For wild-type transthyretin, no significant change in morphology was seen after aging for 7 days; species of around 20 nm in diameter were predominant (arrowheads in a and b). However, although L55P transthyretin at 0 days had

similar morphology (arrowheads in c) to wild-type transthyretin, larger irregular aggregates (50–100 nm in diameter) appeared after aging for 3 days (arrowheads in d). Aging for 7 days led to the appearance of short isolated fibrillar structures with an apparent diameter of around 25 nm (arrowheads in e). After aging for 14 days, L55P transthyretin was predominantly in the form of interweaving fibrils (arrowheads in f). (Reproduced with permission from [10].)



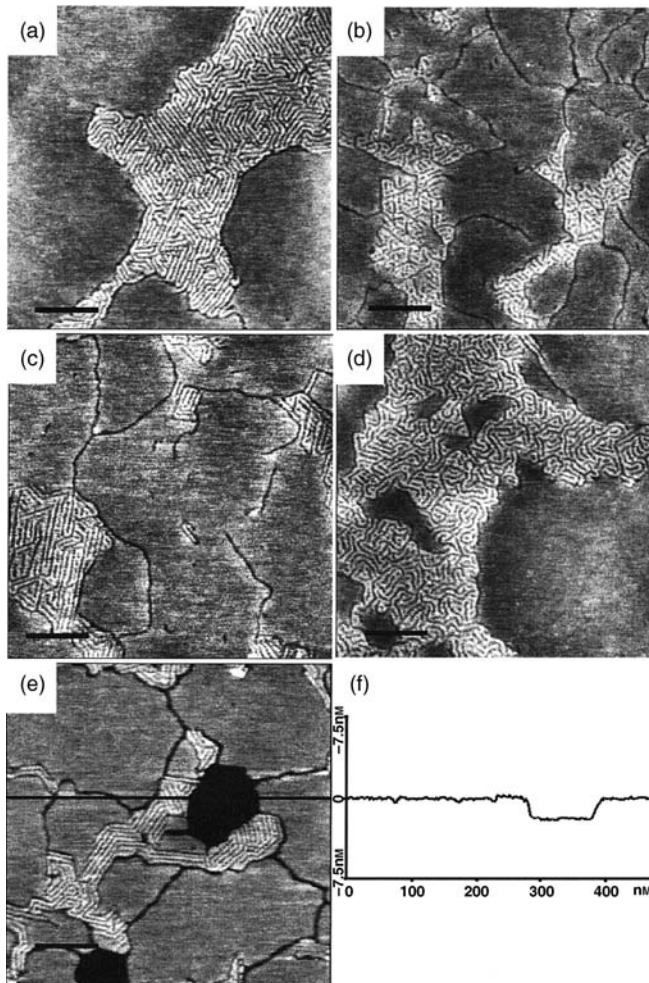
**Figure 8.5** AFM images of amyloid peptides reconstituted in membrane bilayers. Inset shows lipid bilayer with thickness of around 5 nm. For amyloid  $\beta(1-40)$ , ADan, and  $\alpha$ -synuclein, channel-like structures with a central pore can be easily resolved. For ABri,

SAA, and amylin, the central pore is only resolved on some multimer structures. Arrows indicate locations where annular structures can be observed clearly. Scale bars = 100 nm. (Reproduced with permission from [13].)

### 8.3.3

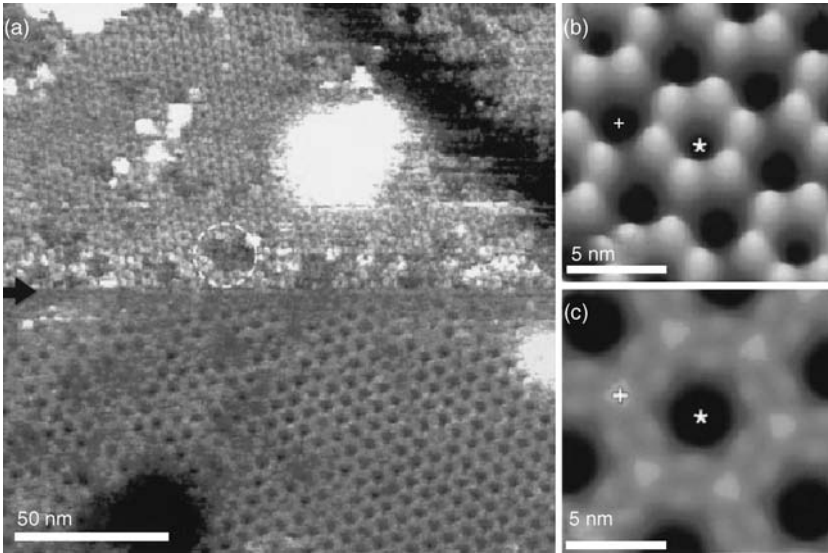
#### Ion Channel Activity

As has been seen in the previous example, peptides forming transmembrane pores might be imaged with AFM. There are also transmembrane pores formed by multimers of large proteins. Gating of such pores, often very specific to a small ionic species, is an important area of research. In particular, as the gating is



**Figure 8.6** AFM images of striated domains induced in supported dipalmitoylphosphatidylcholine (DPPC) bilayers by uncharged model peptides. (a) DPPC with 2 mol% Ac-GWWL(AL)8WWA-Etn; (b) DPPC with 2 mol% Ac-GWWL(AL)8WWA-Etn; (c) DPPC with 2 mol% Ac-GYYL(AL)7YYA-amide; (d) DPPC with 2 mol% Ac-GFFL(AL)7FFA-amide; (e) occasionally lower areas are visible; (f) cross-section of the line drawn in (e). Imaged in 20 mM NaCl. All scale bars = 100 nm. (Reproduced with permission from [14].)

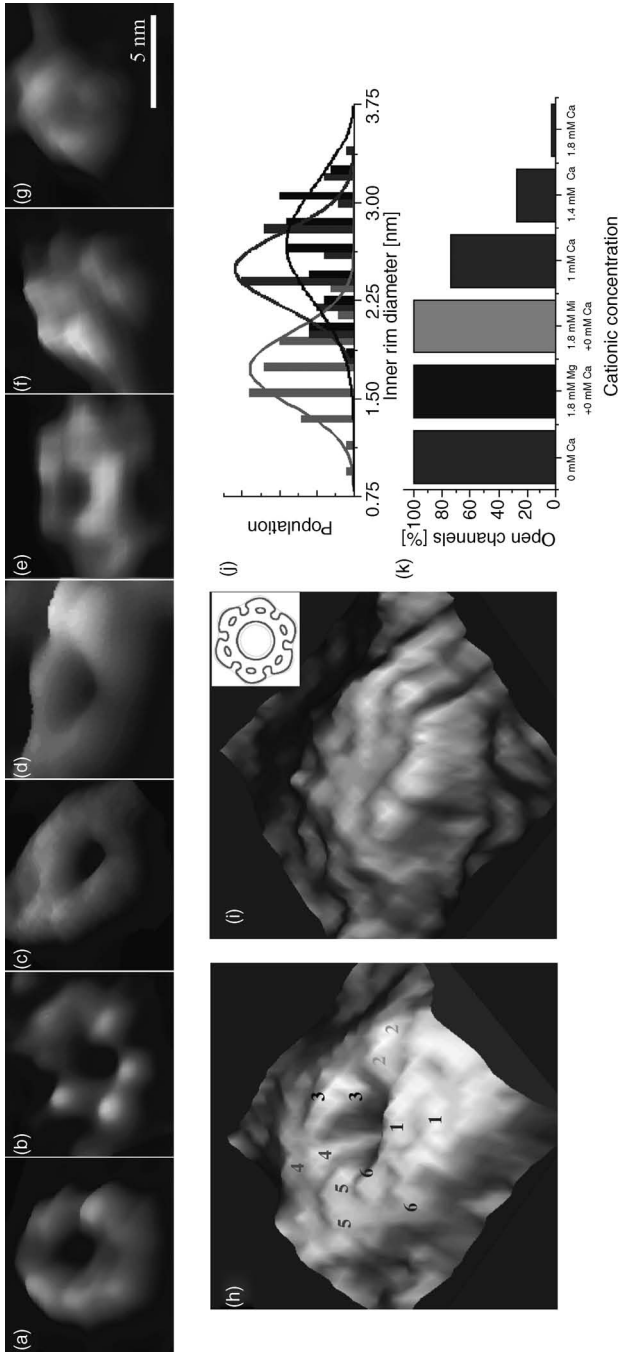
accompanied by a substantial conformational change, questions include whether all channels open and close at a threshold concentration, or is it a statistical process; and how is the gating affected by the ionic environment? These pores are trimers to hexamers of large units that are several tens of kiloDaltons each, with an internal diameter on the nanometer scale; thus the conformational change indicative of the gating might be visible with AFM. The first example as seen in Figure 8.7 is the imaging of the extracellular face of connexon 26, a hexamer of 26-kDa connexin units.



**Figure 8.7** Images of the cytoplasmic gap junction surface of connexon 26 hemichannels. Contact mode imaging. (a) AFM topography demonstrating the variability of cytoplasmic gap junction domains. Individual gap junction domains appear disordered (circle). The initial applied force of 50 pN (top to center of image) was enhanced at the center of the topograph (black arrow) to 70 pN (center to bottom of image). A conformational change is distinct: pore-forming gap junction hexamers collapse onto the membrane surface, thereby

transforming into pores with larger channel diameters. (b) Average of the extended conformation of gap junction. The central pore is visible (asterisks) and the protrusion from the membrane surface is measured to be  $1.7 \pm 0.2$  nm ( $n = 30$ ). (c) Average of gap junction domains collapsed onto the membrane surface. Here, the cytoplasmic domains protrude only by  $0.2 \pm 0.2$  nm ( $n = 30$ ) above the lipid bilayer (+). The vertical level scale is 3 (a), 2 (b) and 1 nm (c). (Reproduced with permission from [15].)

Crystalline arrays of connexons were isolated from a cell membrane surface and imaged in liquid. The channels are open; the central pore is visible on AFM. Importantly, a change in the imaging force results in a dramatic change in the morphology of the surface (arrow), forcing the channels into a different conformation [15]. This highlights the importance of using the correct imaging parameters for sensitive specimens. In another example, the effect of ionic perturbations on the gating of a larger channel of the connexon family, connexon 43, was imaged. Open and closed conformations of the channel were seen coexisting at salt concentrations approaching extracellular levels, thus suggesting that the transition between open and closed conformations is not a sudden change, and might be affected by other parameters as well (Figure 8.8). It was shown that the gating is  $\text{Ca}^{2+}$  specific as similar concentrations of divalent cations  $\text{Mg}^{2+}$  and  $\text{Ni}^{2+}$  did not induce the same conformational change [16]. With AFM imaging, statistics of open versus closed channels could be measured, which was not possible with any other method.



**Figure 8.8** Ultrastructure and size distribution of connexon pores in buffers containing different divalent cations. (a–g) Three-dimensional height images of individual connexons imaged in buffer with no calcium (a); no calcium but with 1.8 mM  $[\text{Mg}^{2+}]$  (b); no calcium but with 1.8 mM  $[\text{Ni}^{2+}]$  (c); with 1.0 mM  $[\text{Ca}^{2+}]$  (d); with 1.4 mM  $[\text{Ca}^{2+}]$ , an open channel (e); with 1.4 mM  $[\text{Ca}^{2+}]$ , a closed channel (f); with 1.8 mM  $[\text{Ca}^{2+}]$  (g). (h and i) High-resolution three-dimensional height images of the extracellular face of an open connexon, imaged in nominally  $\text{Ca}^{2+}$ -free buffer (h) and a closed connexon imaged in the presence of 1.8 mM  $[\text{Ca}^{2+}]$  (i), respectively.

Putative extracellular loops E1 and E2 for each of the six connexin subunits forming a hemichannel are denoted by the numbers 1–6. (j) Distribution of the inner diameter of the pore, measured at one-third height from the topmost point in the absence of calcium (black) and in 1.8 mM  $[\text{Ca}^{2+}]$  (dark gray). In the absence of calcium, other divalent cations (e.g., 1.8 mM  $[\text{Ni}^{2+}]$ , light gray) have no significant effect on the pore diameter. (k) Proportion of the open channels at different cationic concentrations. (Reproduced with permission from [16].)



### 8.3.4

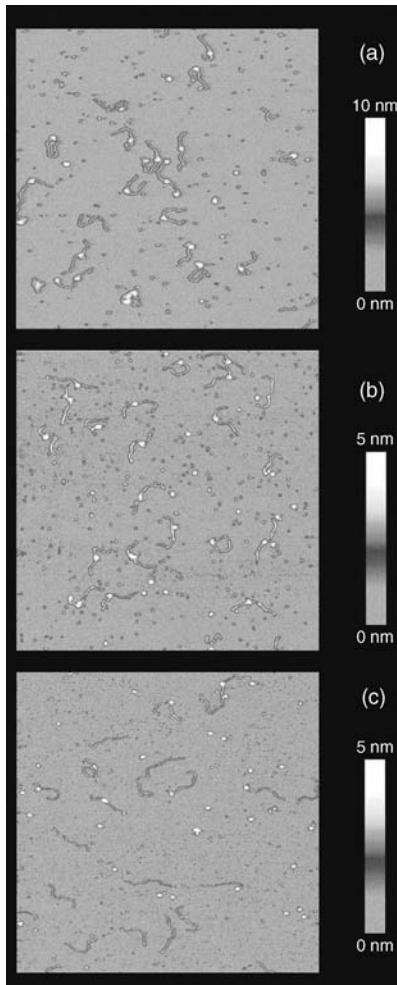
#### Protein–DNA-Specific Binding

Another important area of protein studies where microscopy might prove useful is protein–oligonucleotide binding. To determine a binding site, molecular biology relies on systematically mutating the oligonucleotide and then screening for the protein complex with gel electrophoresis. AFM imaging can significantly shorten the process making mutations unnecessary. After imaging a DNA strand that contains a suggested binding site, the presence and exact location of proteins can be identified, and the existence of the binding site can be confirmed. In Figure 8.9, the protein is RNA polymerase (RNAP) binding to a short DNA sequence [17]. The specific binding of the protein to a site at around one-third of the DNA is visible, with unbound DNA and protein also present. In Figure 8.7(c), nonspecific binding of protein to the end of the DNA can be observed, indicating an inappropriate sample preparation protocol. Importantly, when unbound DNA is present within the imaged area, it might be used as an *in situ* reference to identify any length changes occurring during binding. Length change might be the result of DNA binding in a loop around the protein and thus it has a high significance in identifying the conformation of the DNA–protein complex.

## 8.4

### Issues

As has been seen, AFM is a powerful tool to image single biomolecules. The interpretation of high-resolution imaging data is, however, not without controversy. Published imaging protocols are various and inconsistent, the reproducibility is poor, and the resolution is often much lower than expected. These problems persist in spite of the extensive hardware development of the past few years, suggesting that the problems stem from misperceptions of the imaging process itself. The AFM is not free of imaging artifacts, especially when imaging on the nanometer range. It is understood that determining the accuracy of the morphological measurements and distinguishing the real surface features from the imaging artifacts needs intensive image processing [18], and often numerical simulation of the imaging mechanism (e.g., [16]). Most imaging problems are related to the dynamical properties of the AFM probe as it is influenced in a nonlinear fashion by the interaction with the surface [2]. Of particular interest are the resolution, the magnitude of the force exerted on the sample, the speed of imaging, and the accuracy of surface tracking, all inter-related (e.g., [2, 24, 25]). Several articles deal with the numerical simulation of the AFM imaging process (e.g., [19–23]), the discussion of which is beyond the scope of this chapter. Here, the findings of those works are briefly summarized: the properties of the tapping mode AFM imaging as they are known so far.



**Figure 8.9** Complexes of RNAP  $\sigma^{54}$  with DNA:  $2\ \mu\text{m} \times 2\ \mu\text{m}$  images of a sample (a) in buffer and (b, c) in air prepared with two different protocols. Apart from the RNAP  $\sigma^{54}$ -DNA complexes, the pictures display some free DNA

fragments and free RNAP  $\sigma^{54}$ . In (c), nonspecific binding of the RNAP  $\sigma^{54}$  may be also observed. (Reproduced with permission from [17].)

#### 8.4.1

##### Resolution

In routine operation, in a physiological solution (aqueous buffer of low concentration) AFMs can easily capture images of 500-nm lateral scan size with  $512 \times 512$ -pixel resolution. Evidently, this means that the size of a data pixel is about 1 nm. The objects of interest (i.e., peptides, proteins) are a few nanometers in size; with this

resolution, these can be seen, but cannot be analyzed. The scan size can be easily reduced; however, an actual improvement of resolution might not be achieved as high-resolution imaging is influenced and limited by numerous factors, the first and most obvious of which is the tip size. An average AFM tip has an apex radius of 10–50 nm. This is significantly larger than the structures intended to be resolved. Consistently, if the sample surface is considered a rigid boundary (i.e., no compression can happen nor any morphology change whatsoever), the contour of the tip is convoluted from the contour of the sample. In practical terms, while the height of an object stays correct, the lateral dimensions extend significantly and the geometry becomes rounded. While deconvolution algorithms are often used to improve such images, it is not possible to tell the exact shape of the sample where the morphology prohibits direct tip access, such as at vertical steps and narrow holes. A simple solution of this problem would be the application of sharper probes with high aspect ratio, such as ion-beam-etched silica tips or electron-beam-deposited amorphous carbon extra-tips. Some initial assumptions are, however, misleading. The sample surface, in the majority of cases, cannot be treated as a rigid envelope. The tip does deform the surface upon contact, in both an elastic and nonelastic manner. When the maximal force exerted by the probe upon imaging – the “imaging force” – exceeds what the toughness of the substrate can tolerate, imaging damage occurs. The description of the force interaction needs some deeper insight to the physics of AFM operation, which will be briefly discussed below. Nevertheless, the damage caused by the tip can also be simply related to the contact area – the same force acting on smaller area can cause more damage. Accordingly, sharper tips pose a higher risk of surface destruction. It is also obvious that, if the surface is indeed deformed upon contact with the tip, the contour measured by the AFM is a result of a complex interaction where the softer surface areas might appear lower in morphology. However, it also means that a relatively blunt tip can provide a surprisingly accurate surface tracking, when sensing a step in the change of the surface properties of the sample. The probe shape is therefore an important parameter and it is indeed desired to use sharper tips; however, one needs to keep in mind that the selection is also a tradeoff where the improvement of the resolution can easily end in surface damage.

#### 8.4.2

##### **Imaging Force**

A central issue is thus the amount of force exerted on the sample during imaging, the minimization of which is of key importance for biological samples. The impact of one “tap” of the tip – the *imaging* force – is the force exerted on the sample upon scanning over a featureless surface. When the probe encounters a sudden change in the morphology or the physicochemical properties of the surface, the perturbation of the oscillation causes a *transient* force that is often much higher than the imaging force.

First, consider the variables that might have an influence on the strength of the probe–surface interaction. These are the working parameters of the probe: the oscillating amplitude of the tip far from the surface (free amplitude), the setpoint

amplitude (that is, the amplitude of the probe during operation), and the frequency of the oscillation, as well as the properties of the sample: the surface energy, elasticity, charge distribution, and radii of morphology. Contrary to a very common assumption, imaging force is not *directly* related to the spring constant. The momentum of the tip determines the force, similarly to the way a folded sheet of paper can break a pencil if impacted with high enough velocity. If the probe is barely touching the surface at the end of its trajectory, the force remains largely attractive. However, when it is moved closer to the surface, the momentum at impact will increase rapidly, as determined by

$$I = mA \omega \cos(\omega t + \delta) \quad (8.5)$$

where  $m$  is the (effective) mass of the probe,  $A$  is the maximal oscillating amplitude,  $\omega$  is the circular frequency of the oscillation,  $t$  is time, and  $\delta$  is the phase angle between the drive signal and the oscillation of the end of the probe. This formula gives the momentum of the probe at any point of its oscillation. It must be noted that  $I$  is linearly proportional to the mass and the maximal (free) amplitude of the probe, and its oscillating frequency. Given that the probe is forced to an oscillation very close to its eigenfrequency, it might be assumed that the two are the same and thus through the latter the spring constant is indirectly included in the formula. However, there are more significant contributions from the cosine term. Employing a simple approximation, consider that bringing the probe to the proximity of the surface, to  $z$  distance, where  $z < A$ , the momentum of the probe hitting the surface equals the momentum of the probe at that point of the oscillation in the absence of the surface. In this case the tip–surface distance  $z$  might be included through the argument of the cosine which is equal to  $\arcsin(A/z)$ . From this it follows that the free amplitude  $A$  and the setpoint  $A/z$  determine the force upon impact, and are the two key parameters (and, incidentally, the easiest to change) that need to be optimized. To minimize the imaging force, the smallest possible amplitude is desired, with a setpoint amplitude as close to the free amplitude as possible.

It must be mentioned that Eq. (8.5) describes the momentum of the *first* impact of the probe and for a periodic case the tapping of the surface becomes what is known in physics as an impact oscillator, requiring complex treatment. However, the conclusions drawn from the above simple model have been verified both experimentally and by numerical simulations, and so we can avoid discussing mathematical details.

#### 8.4.3

##### Repetitive Stress

As has been seen, the tapping mode, while reducing shear damage, also introduces higher and less controllable vertical forces. However, even reducing the impact of each “tap” might not be enough to save the sample from damage. Consider a tip repeatedly hitting the surface with a high frequency. On large scale, recording a picture in a reasonable time (i.e., 1–10 min) means that the tip is scanned relatively fast over the sample and thus the individual “taps” by the tip happen somewhat apart. Using the same frame capture rate, for higher-resolution pictures the range is soon

reached where the contact areas of the consequent impacts substantially overlap. Eventually, at the highest resolution where the nondestructive treatment of a delicate sample would be the most important, the tip will end up hammering the same area at an ultrasonic frequency. Thus, the destructive effect of this local sonication needs to be understood and minimized for imaging at small scan scales.

From the application point of view, it seems plausible to reduce the effect of local sonication by simply increasing the distance between, and decreasing the force of, the individual impacts (i.e., using faster scanning with smaller free amplitudes and/or higher setpoint). However, there are inherent limitations to the scan speed – the bandwidths of the hardware and the probe. Furthermore, reduction of the probe amplitude also leads to the rise of imaging artifacts, as will be seen below.

#### 8.4.4

##### **Artifacts Related to too Low Free Amplitude**

It is well known that various user manuals recommend working with a large, several hundred nanometers free amplitude, as this provides good surface tracking for hard inorganic samples. Not only does it eliminate the influence of weak attractive forces, but also provides a large dynamic range: stepping up and down surface steps incites a substantial change in the probe amplitude, up or down, and thus feedback can react quickly. However, large amplitude leads to large force, as was seen in Section 8.4.2. When working on soft, compliant, and poorly bound structures, these might be destroyed, moved, or compressed during imaging, leading to damage and artifacts. Thus, for bioimaging, conditions other than the ones outlined in the manuals are needed. The use of a small free amplitude is the most important difference. However, there are limitations.

When working with small amplitudes, the probe is largely within the range of the attractive noncontact forces. This allows for imaging conditions when the probe does not physically touch the surface. In this “zero force” case, however, spatial changes in the strength of the attractive interaction appear as height differences [20, 26]. Even if the probe does touch the surface, the contribution of the attractive forces remains strong; thus the magnitude of the height artifact is smaller, but it persists [21]. If the attractive forces are too high, another artifact might occur – a sawtooth-like periodic signal is superposed on the topography, with a periodicity sensitive to the value of the feedback gain. In this case, the energy of the low amplitude oscillation is insufficient to break the tip away from a sample; the amplitude drops to zero, to which the feedback reacts by suddenly lifting up the probe, after which the amplitude recovers and the feedback attempts to bring the probe back to the surface, where it will be trapped again, and so on. Thus, if the sawtooth pattern appears in the topography signal, the only way to get good surface tracking is by increasing the free amplitude.

There is a further limitation to the smallest working amplitude. The sensitivity of the A/D converter is a non-negotiable characteristic of the AFM hardware; too small free amplitudes, especially in combination with too high feedback ampli-

tudes, would fall into the range where the discrete steps are reached, leading to unstable operation.

#### 8.4.5

##### Transient Force and Bandwidth

Above, we discussed two kind of forces exerted during AFM operation, the imaging force and the transient force. Now the discussion returns to the latter one. So far only a harmonic solution of Eq. (8.1) has been discussed. The complete solution, however, includes the term:

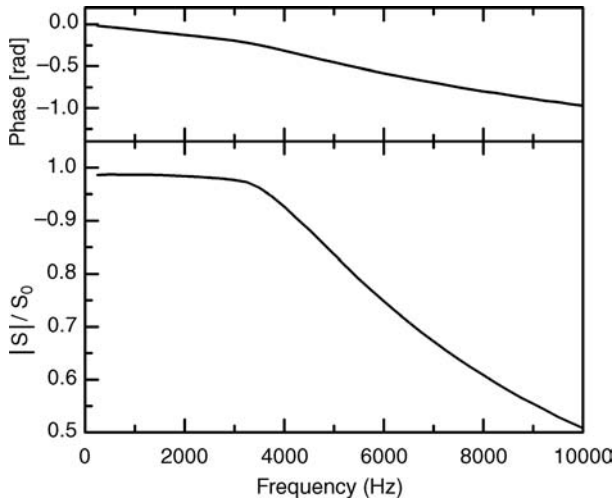
$$\gamma = Ae^{-\beta t} \sin(\omega_1 t - \varphi_0) \quad (8.6)$$

where  $\gamma$  is the trajectory of the tip,  $\varphi_0$  is the initial phase lag, and  $\omega_1 = \sqrt{\omega_0^2 - \beta^2}$ , and  $A$  and  $\beta$  were defined before as the maximal amplitude and the damping coefficient, respectively. This equation describes an exponentially decaying sinusoidal motion. Thus, any perturbation to the system that changes the amplitude and/or the phase lag would generate a transient term Eq. (8.6) in addition to the periodic term Eq. (8.2). The probe reaches a stable amplitude under the new conditions that perturbed the system after the decay of the transient term. It is obvious that the inclusion of the transient amplitude into the feedback mechanism would lead to improper response to the surface structures, and, eventually, instability and imaging artifacts. Thus, the sampling frequency of the feedback cannot be higher than the reciprocal of the time constant of a typical transient (or, rather, it is desired to reduce the transient length). This transient thus limits the effective response time of the AFM to surface structures and thus the accuracy of surface tracking at a given scan speed. The accuracy of surface tracking is often discussed in terms of *bandwidth* – the frequency of a periodic modulation that is still visible to the system. It is also obvious that during the transient the force exerted on the surface will be higher than the imaging force, as the feedback is virtually inactive.

#### 8.4.6

##### Accuracy of Surface Tracking

After the exerted force, accuracy of imaging is the second most important characteristic of an AFM. Here, the discussion will focus on poorly understood phenomena – the inherent nonlinearities and artifacts of the probe. As it was briefly mentioned in Section 8.4.4, the dynamical range of the amplitude response is key to generate sufficient control signal for the feedback to approach or retract the probe. It was also briefly mentioned that probe response time to amplitude perturbations is related to the ability of the AFM to sense small morphology changes. There is a more accurate way of looking at this phenomenon – the amplitude transfer efficiency. Basically, this is the measure of the ability of the



**Figure 8.10** Calculated describing function (transfer function for nonlinear systems) of the tapping mode AFM probe in air for 40 N/m and 10 nm tip radius (for other parameters, see [27]). The free and setpoint amplitudes were 50 and 35 nm, respectively. (Reproduced with permission from [27].)

probe amplitude to follow a sinusoidal modulation of a certain frequency. A plot of this parameter against the frequency of the modulation is called the transfer function of the probe (Figure 8.10). The frequency value at which the amplitude transfer drops below  $-3$  dB is typically interpreted as the bandwidth of the system; however, for AFM, since only the 1:1 amplitude transfer is acceptable, the corner frequency is used as the bandwidth. To relate the bandwidth defined in terms of modulation frequency to topography, it is easy to see that periodic surface features appear to the probe more or less frequently depending on the scanning velocity. The shape of the structures matters as well, since the transfer function is defined for sinusoidal modulation. It is known that an arbitrary shape can be reconstructed from a series of sinusoidal waves – a process also known as Fourier analysis. The steeper a structure is, the higher frequency Fourier components need to be included; for a vertical step, it is an infinite series. What this means in practical terms is that in the case where the probe encounters a vertical step on the surface, it is inherently unable to perfectly follow the topography. For lesser extremes, a morphology that has a spatial frequency (from the combination of scanning speed and the Fourier analysis) beyond the bandwidth causes the probe amplitude to lag behind the topography change, leading to inaccurate surface tracking and thus inaccurate height measurements. Accordingly, a tell-tale sign of bandwidth-limited behavior is if the measured height of a periodic object, such as a test grid, is a function of the scan velocity. We have to compromise with a scan speed that provides consistent measurements; this is often much lower than the linear velocity range achievable with the AFM scanner.

It was shown with numerical simulation of the probe transfer properties upon interaction with surfaces of varied physical properties [27] that the bandwidth can be tuned over a wide range by changing the free amplitude and setpoint. In general, higher free amplitudes and lower setpoints increase the bandwidth; however, as was discussed above, this is accompanied by an increased imaging force, and thus there is a tradeoff between imaging velocity and the force exerted on the sample.

There are two important implications of the sensitivity of the bandwidth to the setpoint. The first is that a situation is possible when the AFM is accurately tracking the *substrate* (low Fourier frequencies), but fails to sense the *sample*, such as small monomeric proteins. The second is that a small change in the setpoint up or down can make such a sample become invisible or visible, respectively, while the substrate is correctly imaged all the time. Considering that the system maintains the setpoint typed into the software with the help of the feedback circuit, it is possible that, on a sloping surface and with low feedback gains, the *actual* ratio of the free and operational amplitude can change slightly between scanning up and down; a result of which is that the proteins are visible on one picture but “disappear” from the consecutive one. Cited as “black magic,” this phenomenon is clearly explained with bandwidth limited behavior.

#### 8.4.7

##### Step Artifacts

Finally, mention should be made of the phenomenon of contrast inversion. If the oscillating probe is pushed against the surface, the amplitude usually decays linearly with the distance (after a brief “rounded” range) and reaches zero once the energy of the drive dissipates into the surface altogether. In case of compliant, elastic samples and strong attractive interaction, this approach curve might become nonlinear. Visually, a shoulder or a peak is formed on the approach curve. If this peak appears at the far end, after engaging the surface, the probe can work at higher setpoints than the free amplitude. After a large surface perturbation, however, it is possible for the probe to find the same amplitude on the other side of the peak; this being a positive slope, the feedback would generate an inverted image – higher structures would appear lower. This is an unstable working condition and usually ends in losing the surface. If the peak appears on the slope, the inverted range is usually too short for even temporary stable operation, thus the other possibility is for the probe to fall on the same amplitude on a negative slope but a certain distance farther away from (or closer to) the surface, thus recording a step artifact parallel to the scan direction.

Here is the end of the discussion of probe behavior and imaging artifacts. The key control parameters were identified: the free amplitude, setpoint, and the scan velocity, the careful and sensible control of which is sufficient to optimize the imaging of an arbitrary sample most of the time. Even more importantly, this reasoning highlighted the importance of careful analysis in the interpretation of the images to avoid premature conclusions.



## 8.5

### Force Measurements

It is often convenient to use the AFM as a force measurement tool [28]. After recording a picture and selecting the points of interest, AFM can record a simple approach and retract curve, pushing the probe against and then withdrawing from the sample. The bending of the cantilever, which is directly proportional to the normal load, is plotted against the vertical displacement, thus providing a force–distance curve. This working mode is used to measure mechanical surface properties of a wide range of materials including lipid membranes [29–34]. However, it can be used to measure the strength of protein–protein and protein–antibody interactions, and protein unfolding as well, if one of the interacting proteins or one end of the stretched protein is chemically attached to the probe. Such force measurements might be performed in a mapping fashion, when at each raster point of an image a statistical number of force curves is being collected. After averaging, each raster point might be represented by, for example, its elasticity or maximal adhesion. A comparison of this map to the surface morphology helps to perform necessary corrections for the geometry (the maximal adhesion force is the function of the curvature of the surface), and thus the force interaction map might be used to distinguish different materials on the related topography.

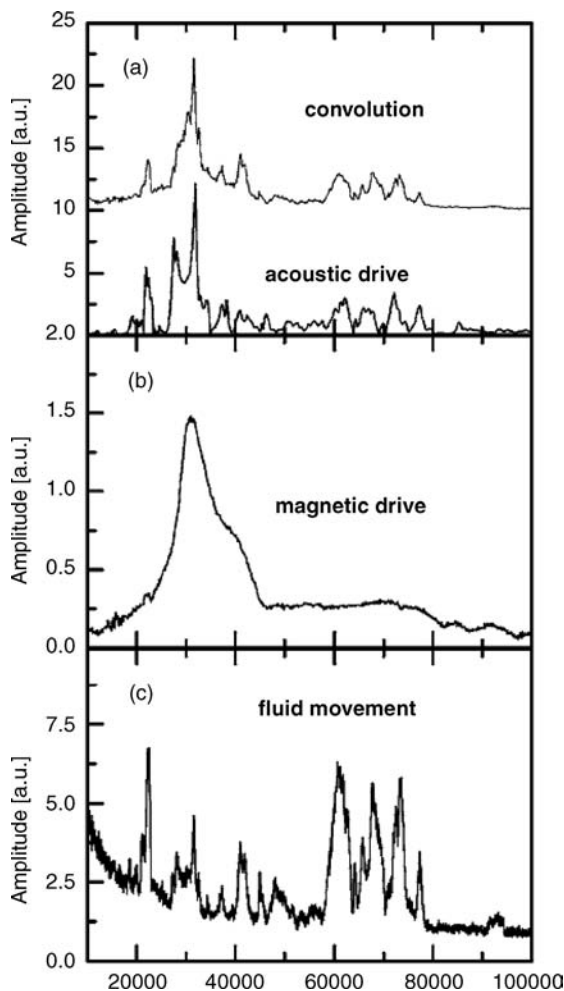
The accuracy of these measurements is a function of the accuracy of the probe parameters used for data analysis. Whereas most AFM probes come with a nominal value and an error range for the radius of curvature and the spring constant, the actual values of the individual probe properties can be way off compared to the nominated range. It is important therefore to determine these parameters with independent measurements. The tip shape can be determined by using standard calibration samples. These are usually test grids with factory-guaranteed geometry. Step grids can be used to measure the radius of curvature in one direction; pin grids provide a convolution picture of the whole tip apex. Measuring the spring constant is far more complicated. Whereas more than a dozen different methods exist, based on comparative studies [35, 36] four are considered the most accurate. These are the added mass [37], the reference cantilever [38], the thermal fluctuation [39, 40], and the Sader model [41]. As most force measurements are performed comparatively, the calibration of the cantilever is less important for practical use; we omit the detailed discussion of the methods, which, if needed, the reader can find in the cited articles.

## 8.6

### Liquid Imaging

The majority of questions about proteins seeking an answer from AFM methods require the imaging to be performed in a physiological solution. While AFM is certainly capable of doing that, there are special considerations to be made. Imaging in a liquid cell is usually done by using the common “acoustic drive” (i.e., a dither piezo to force the cantilever oscillation). Unlike the case of imaging in air, however,

the dither piezo cannot be placed directly behind the cantilever, as it has to be insulated from the liquid while the cantilever has to be immersed in it. Most liquid cell designs would thus fasten the cantilever to a watertight holder, called the liquid cell (i.e., the term does not refer to a complete enclosure for the liquid) and drive the entire “cell” to oscillation. The result of this design is that a frequency sweep searching for the natural frequency of the probe would produce several apparent resonance peaks instead of the single real one (Figure 8.11a). It is often noted that the



**Figure 8.11** Resonance curves of an NSC18 cantilever (MikroMasch; [www.spmtips.com](http://www.spmtips.com)) in liquid. Frequency sweep showing (a) forest of peaks using acoustic excitation and (b) a single peak using magnetic excitation. (c) The forest of peaks can also be observed in the oscillations of the liquid droplet enclosed into the liquid cell.

The spectrum was recorded by an external laser reflected from the exposed meniscus of the liquid, registered on a segmented photodiode. Convolution of the magnetic excitation and fluid frequency spectrum is also shown in (a). (Reproduced with permission from [43].)

position and height of the peaks in this “forest” are often changing upon approach to the surface, even at very long distances, of the order of 100  $\mu\text{m}$ . This would imply that the thickness of the liquid in the cell might be related to the appearance of the peaks and thus the “forest of peaks” is usually explained with resonances of the liquid [42]. The same spectrum might be measured in the oscillations of the liquid itself (Figure 8.11c). However, the assumption is not true. The peaks remain if the water is removed from the cell, indicating that what is seen on the frequency sweep is the spectrum of the resonances of various components of the probe holder [43].

An easy way to deal with the “forest” is to use cantilevers of known eigenfrequencies. However, measuring the resonance frequency of the cantilever under liquid requires instrumentation that the average AFM operator does not have. Furthermore, the measured probe eigenfrequency often does not coincide with any peaks on the frequency sweep recorded by the AFM. If these resonances do not overlap with the eigenfrequency of the probe, it is driven off-resonance. In practical terms this means that the ratio of the probe amplitude and the drive amplitude is less than 1. This situation was often explained with overdamping of the cantilever by the viscous liquid environment. Indeed, damping does play a role in reducing the apparent amplitude ratio below 1; however, for a meaningful resonance it should remain larger than 1. Interestingly, even under these conditions tapping the surface leads to amplitude loss, and thus imaging appears to work. However, the interaction force can be up to 10 times higher than in the case of real probe resonances and the sensitivity is seriously compromised [43]. This is the reason why AFM manufacturers recommend soft probes for tapping under liquid. However, a simple protocol might be used to determine which peak is the best suited to imaging, for an arbitrary cantilever. First, an amplitude–distance curve, similar to a force–distance curve, has to be recorded. In air, and in liquid as well if the probe is driven at its natural frequency, the amplitude is near constant until the probe hits the surface; there a linear slope begins which reaches zero amplitude at a distance corresponding to the probe amplitude. If the probe is driven above the eigenfrequency, however, a positive peak appears before the linear decay part of the curve. Below the natural frequency the peak does not reappear, however the slope of the linear section is becoming less steep. Thus, starting from higher frequencies, recording amplitude–distance curves gives a clear indication which peak is the closest to the resonance frequency of the probe. It might be still slightly off-resonance; however, choosing a drive frequency this way is typically sufficient to achieve nondestructive imaging.

There is a way to circumvent this process by eliminating the need to oscillate the entire cell. This might be done by using magnetic actuation of the cantilever [44]. As it is depicted in Figure 8.11(b), a frequency sweep graph recorded with magnetic drive displays a single, broad resonance peak. However, magnetic actuation requires magnetically coated probes, which are unstable, deteriorate fast, and have a larger apex radius due to the coating. Accordingly, magnetic drive is used only occasionally for high-resolution bioimaging and other methods are being sought.

Another issue of liquid imaging that requires special consideration is the bandwidth of the probe under liquid. As seen above, bandwidth determines the achievable speed and sensitivity. For linear systems, bandwidth is inversely proportional to the quality

factor. Consistently, the lower  $Q$  under liquid should lead to higher bandwidth and thus faster operation. Not so – even if operated at the natural frequency, the nonlinearity of the probe–surface interaction counteracts any bandwidth increase [27]. Thus, liquid imaging is not faster than ambient imaging; however, due to the lower  $Q$ , it is less sensitive. Thus, a *slower* scan rate is recommended for imaging under liquid.

Finally, a discussion of the multitude of tip artifacts frequently observed on liquid images. These typically appear as multiplications or shadows of the surface features, occasionally escalating to bizarre cauliflower-like structures. The effect is seldom seen in dry images, as it usually originates from an uneven, broken tip, where smaller satellite tips surround the main apex. Each of these satellites will sense the surface structures. If the surface structure is round, such as a globular protein, the recorded image shows the morphology of the tip “imaged” – convoluted – on the protein. Relating the artifact to the case of dry imaging, poor quality probes are often blamed for this effect also when the imaging is performed under liquid. In reality, it might be just a poor choice of substrate. When the probe and the sample are submerged in water, the interaction between the two is influenced by the surface tension of the liquid. In this case, attachment of the sample, such as proteins, to a surface is driven by surface energy, coulombic interactions play a lesser role. Surface energy of a substance determines if it is hydrophobic or hydrophilic, and, in water, a protein will adhere to the surface that has a similar relation to water. In effect, a hydrophobic protein would adhere to the more hydrophobic surface, independent of whether it is the substrate or the probe. Thus, proteins might be picked up by the probe and this might happen at the first opportunity (i.e., the first time the probe scans the surface). The attachment of one or more proteins to the probe apex will lead to the multiple asperities artifact, even if the tip was originally of excellent quality. Accordingly, sample preparation is a key issue for liquid imaging.

## 8.7

### Sample Preparation for Bioimaging

AFM is a surface technique. It can image only surface morphology, thus any sample of interest that is not naturally a surface structure must be immobilized on a surface. It is relatively straightforward for dry imaging; more problematic for imaging under liquid. The following section will examine the most effective methods. The aim of this section is to give a practical guide to sample preparation, the information about which is largely missing from the literature.

#### 8.7.1

##### Adhesion

The simplest way to prepare a sample for imaging in air is by drying down a dilute solution of the protein of interest. Upon removal of the water, surface adhesion is usually strong enough to keep the protein attached to the surface, independent of whether the surface is hydrophilic or hydrophobic. In physical terms, a hydrophilic

surface has higher and a hydrophobic surface has lower surface energy in respect to the surface tension of the water. In the absence of water, the strength of surface energy only determines the strength of the adhesion. If the surface energy is very low, however, it might promote the aggregation of hydrophilic proteins. Hydrophilic areas on a hydrophobic surface, such as edge planes on basal plane graphite, act as molecule traps and might even promote protein–protein interactions such as amyloid fiber formation [45]. Accordingly, even in this simple case, the potential of the surface to alter the state of the imaged molecules must be assessed.

Another, often overlooked issue is the presence of buffer molecules, salts, or other stabilizing agents such as glycerol. The evaporation of the water leaves crystals of salt, droplets of glycerol, or a blanket of large molecules such as Tris behind. These structures can be mistaken for the protein or they just simply obscure the information sought with AFM imaging. Repeated gentle rinsing with distilled water might be used to remove highly soluble salts; larger organic molecules, however, usually remain on the surface. Indeed, the repeated rinse is more likely to remove some of the proteins of interest. It is thus recommended that the proteins are deposited from a physiological buffer, such as phosphate-buffered saline, that contains only inorganic salts.

It should be remembered that even a few microliters of nanomolar concentrations can contain enough molecules to coat the substrate surface with several layers of proteins. It is better to incubate a droplet of a low-concentration sample solution on the surface (typically for 5–15 min), and drying it only after removing excess liquid and rinsing the surface with a few drops of distilled water. This method gives a better control of surface coverage; however it is very sensitive to the adhesion of the protein to the surface. If varying hydrophilic and hydrophobic surfaces, such as mica and highly oriented pyrolytic graphite, respectively, does not lead to a workable deposition procedure, pretreatment of the surface might become necessary. Coulombic immobilization might be possible for amphipathic molecules. The substrate is either functionalized with a ligand that becomes charged at certain pH, like a carboxylic acid, or simply by rinsing the substrate with a solution of an appropriate ion. Typically,  $\text{Ni}^{2+}$  is used.

If imaging is performed in liquid, in general, there is no need to avoid organic molecules in the solution, provided these do not preferentially attach to the substrate or tip surface or coat the proteins under investigation. In this case surface adhesion is typically weaker; however, amphiphilic or slightly hydrophobic molecules might readily, and strongly, attach to hydrophobic surfaces, in the absence of which they often compensate for the solvation effects by aggregation. Surface pretreatment with ionic solutions can be also used to improve immobilization; however, here typically chelating ions are used, which coordinate to the surface and the proteins as well. If these simple immobilization methods fail, there are further possibilities to consider: physical entrapment and chemical attachment.

### 8.7.2

#### **Physical Entrapment**

Physical entrapment requires the modification of the surface in a way that it can specifically enclose part of the protein, while still retaining a physiologically relevant

conformation. A thin mesh of polyelectrolyte or actin fibers might serve this purpose; however, a more practical solution is using a biomimetic phospholipid bilayer [12]. Given that many proteins of interest are membrane-associated, using a biomimetic membrane platform does not only solve the issue of the immobilization but also improves the bioidentity of the imaging conditions. These methods are seldom used in ambient imaging. In an aqueous environment, proteins cannot only preferentially and bioidentically bind to a membrane, but also retain full functionality, thus their interactions might be observed *in situ* and, under special circumstances, in *real-time* [11, 46].

### 8.7.3

#### Chemical Binding

There are special circumstances when none of the above methods would provide sufficient results. For very large or easily aggregating proteins, fibers that tend to loop away from the surface, large functional complexes, or for cases when electronic access to the protein is required, such as for *in situ* electrochemical characterization, chemical immobilization is the best suited. This might be coordination binding, where, for example, a histidine-tagged protein is immobilized via a metal held by a ligand such as nitrilotriacetic acid. Such a system is used for protein purification and thus it is relatively easily adapted to protein immobilization. Alternatively, the protein might be chemically attached with a linking reaction such as *N*-ethyl-*N'*-(3-dimethylaminopropyl) carbodiimide/*N*-hydroxy succinimide. In both cases, the surface first has to be functionalized with the desired ligand. While functionalization of mica and graphite is possible, in practice a gold substrate is preferred for the ease of thiol chemistry. Gold, however, is typically not smooth enough and thus it has to be annealed in a reductive flame to create crystalline facets of sufficient size [47]. The smoothness of the resulting layer determines the success of the AFM imaging, as surface asperities must be much smaller – or much larger – than the feature size to be imaged. Furthermore, chemical attachment can easily modify the protein tertiary structure, changing its physicochemical and interaction characteristics. As the protein is attached to the end of a “blade of molecular grass,” as self-assembled monolayers are often described, it is more mobile, creating more opportunity for interactions, but reducing the ability of the AFM probe to image the proteins without displacing them. The result is a blurred image. In general, chemical immobilization should be avoided, and only attempted as the last resort.

### 8.8

#### Outlook

The sample preparation methods finish the discussion of AFM imaging of proteins. The discussion has not, so far, addressed the future – in what direction will AFM as a technique evolve and how will that enhance the usefulness for protein imaging? There are two obvious fields where any improvement would be welcomed: in the

imaging speed and the resolution. Solutions for improved imaging velocity include the use of small cantilevers [48], designing scanners with lower inertia [49, 50], and experiments with advanced control methods [51]; few methods combine these solutions [52], and, as yet, there is no information in the public domain about the accuracy and imaging characteristics of such fast AFMs. If a commercial system does become available, the benefits would be enormous; not only would it be possible to record pictures in a blink of an eye and thus map large surface areas, but also to record movies of protein–protein interaction. The resolution is partially addressed with the use of extremely sharp diamond-like carbon probes; however, without also improving the sensitivity of the probe, the enhancement in resolution is not substantial. Higher-frequency probes, more accurate position detection could provide the answer. Finally, liquid imaging would largely benefit from the elimination of the “forest of peaks.” In summary, AFMs of the near future will be much faster, more sensitive, and hopefully provide the means of imaging protein interactions *in situ*, in *real-time*: as they happen in nature.

## References

- Binnig, G., Quate, C., and Gerber, Ch. (1986) *Physical Review Letters*, **56**, 930–933.
- Burnham, N.A., Behrend, O.P., Oulevey, F., Gremaud, G., Gallo, P.-J., Gourdon, D., Dupas, E., Kulik, A.J., Pollock, H.M., and Briggs, G.A.D. (1997) *Nanotechnology*, **8**, 67–75.
- Lal, R. and John, S.A. (1994) *American Journal of Physiology*, **266**, C1–C21.
- Santos, N.C. and Castanho, M.A.R.B. (2004) *Biophysical Chemistry*, **107**, 133–149.
- Alessandrini, A. and Facci, P. (2005) *Measurement Science Technology*, **16**, R65–R92.
- Mou, J., Sheng, S., Ho, R., and Shao, Z. (1996) *Biophysical Journal*, **71**, 2213–2221.
- Gruber, C.W., Čemažar, M., Mechler, A., Martin, L.L., and Craik, D.J. (2009) *Biopolymers*, **92**, 35–43.
- Praporski, S., Ng, S.M., Nguyen, A.D., Corbin, C.J., Mechler, A., Zheng, J., Conley, A.J., and Martin, L.L. (2009) *Journal of Biological Chemistry*, **284**, 33224–33232.
- Goldsbury, C., Kistler, J., Aebi, U., Arvinte, T., and Cooper, G.J.S. (1999) *Journal of Molecular Biology*, **285**, 33–39.
- Hou, X., Parkington, H.C., Coleman, H.A., Mechler, A., Martin, L.L., Aguilar, M.-I., and Small, D.H. (2007) *Journal of Neurochemistry*, **100**, 446–457.
- Mechler, A., Praporski, S., Atmuri, K., Boland, M., Separovic, F., and Martin, L.L. (2007) *Biophysical Journal*, **93**, 3907–3916.
- Mechler, A., Praporski, S., Piantavigna, S., Heaton, S.M., Hall, K.N., Aguilar, M.-I., and Martin, L.L. (2009) *Biomaterials*, **30**, 682–689.
- Quist, A., Doudevski, I., Lin, H., Azimova, R., Ng, D., Frangione, B., Kagan, B., Ghiso, J., and Lal, R. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10427–10432.
- Rinia, H.A., Boots, J.-W.P., Rijkers, D.T.S., Kik, R.A., Snel, M.M.E., Demel, R.A., Killian, J.A., van der Eerden, J.P.J.M., and de Kruijff, B. (2002) *Biochemistry*, **41**, 2814–2824.
- Müller, D.J., Hand, G.M., Engel, A., and Sosinsky, G.E. (2002) *EMBO Journal*, **21**, 3598–3607.
- Thimm, J., Mechler, A., Lin, H., Rhee, S., and Lal, R. (2005) *Journal Biological Chemistry*, **280**, 10646–10654.
- Schulz, A., Mücke, N., Langowski, J., and Rippe, K. (1998) *Journal of Molecular Biology*, **283**, 821–836.
- Philippson, A., Im, W., Engel, A., Schirmer, T., Roux, B., and Müller, D.J. (2002) *Biophysical Journal*, **82**, 1667–1676.
- Tamayo, J., Humphris, A.D.L., and Miles, M.J. (2000) *Applied Physics Letters*, **77**, 582–584.

- 20 Mechler, Á., Kokavecz, J., Heszler, P., and Lal, R. (2003) *Applied Physics Letters*, **82**, 3740–3742.
- 21 Mechler, Á., Kopniczky, J., Kokavecz, J., Hoel, A., Granqvist, C.-G., and Heszler, P. (2005) *Physical Review B*, **72**, 125407/1–125407/6.
- 22 Kokavecz, J., Heszler, P., Toth, Z., and Mechler, Á. (2003) *Applied Surface Science*, **210**, 123–127.
- 23 Mechler, Á., Kokavecz, J., and Heszler, P. (2001) *Materials Science and Engineering C*, **15**, 29–32.
- 24 Kokavecz, J., Horváth, Z.L., and Mechler, Á. (2004) *Applied Physics Letters*, **85**, 3232–3234.
- 25 Anczykowski, B., Cleveland, J.P., Krüger, D., Elings, V., and Fuchs, H. (1998) *Applied Physics A*, **66**, S885–S889.
- 26 Heszler, P., Révész, K., Reimann, C.T., Mechler, Á., and Bor, Z. (2000) *Nanotechnology*, **11**, 37–46.
- 27 Kokavecz, J., Marti, O., Heszler, P., and Mechler, Á. (2006) *Physical Review B*, **73**, 155403/1–155403/8.
- 28 Benz, M., Gutschmann, T., Chen, N., Tadmor, R., and Israelachvili, J. (2004) *Biophysical Journal*, **86**, 870–879.
- 29 Dufrene, Y.F., Boland, T., Schneider, J.W., Barger, W.R., and Lee, G.U. (1998) *Faraday Discussions*, **111**, 79–94.
- 30 Zhang, L., Vidu, R., Waring, A.J., Lehrer, R.I., Longo, M.L., and Stroeve, P. (2002) *Langmuir*, **18**, 1318–1331.
- 31 Kaasgaard, T., Mouritsen, O.G., and Jorgensen, K. (2002) *FEBS Letters*, **515**, 29–34.
- 32 Desmeules, P., Grandbois, M., Bondarenko, V.A., Yamazaki, A., and Salesse, C. (2002) *Biophysical Journal*, **82**, 3343–3350.
- 33 Maeda, N., Senden, T.J., and di Meglio, J.M. (2002) *Biochimica Biophysica Acta*, **1564**, 165–172.
- 34 Schneider, J., Barger, W., and Lee, G.U. (2003) *Langmuir*, **19**, 1899–1907.
- 35 Levy, R. and Maaloum, M. (2002) *Nanotechnology*, **13**, 33–37.
- 36 Burnham, N.A., Chen, X., Hodges, C.S., Matei, G.A., Thoreson, E.J., Roberts, C.J., Davies, M.C., and Tendler, S.J.B. (2003) *Nanotechnology*, **14**, 1–6.
- 37 Cleveland, J.P., Manne, S., Bocek, D., and Hansma, P.K. (1993) *Review of Scientific Instruments*, **64**, 403–405.
- 38 Gibson, C.T., Watson, G.S., and Myhra, S. (1996) *Nanotechnology*, **7**, 259–262.
- 39 Hutter, J.L. and Bechhoefer, J. (1993) *Review of Scientific Instruments*, **64**, 1868–1873.
- 40 Butt, H.-J. and Jaschke, M. (1995) *Nanotechnology*, **6**, 1–7.
- 41 Sader, J.E. (1998) *Journal of Applied Physiology*, **84**, 64–77.
- 42 Schäffer, T.E., Cleveland, J.P., Ohnesorge, F., Walters, D.A., and Hansma, P.K. (1996) *Journal of Applied Physiology*, **80**, 3622–3627.
- 43 Kokavecz, J. and Mechler, A. (2007) *Applied Physics Letters*, **91**, 023113–023115.
- 44 Han, W., Lindsay, S.M., and Jing, T. (1996) *Applied Physics Letters*, **69**, 4111–4113.
- 45 Klug, G.M.J.A., Losic, D., Subasinghe, S., Aguilar, M.-I., Martin, L.L., and Small, D.H. (2003) *European Journal of Biochemistry*, **270**, 4282–4293.
- 46 Kienberger, F., Mueller, H., Pastushenko, V., and Hinterdorfer, P. (2004) *EMBO Reports*, **5**, 579–583.
- 47 Golan, Y., Margulis, L., and Rubinstein, I. (1992) *Surface Science*, **264**, 312–326.
- 48 Viani, M.B., Schäffer, T.E., Paloczi, G.T., Pietrasanta, L.I., Smith, B.L., Thompson, J.B., Richter, M., Rief, M., Gaub, H.E., Plaxco, K.W., Cleland, A.N., Hansma, H.G., and Hansma, P.K. (1999) *Review of Scientific Instruments*, **70**, 4300–4303.
- 49 Ando, T., Kodera, N., Takai, E., Maruyama, D., Saito, K., and Toda, A. (2001) *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12468–12472.
- 50 Schitter, G., Åström, K.J., DeMartini, B.E., Thurner, P.J., Turner, K.L., and Hansma, P.K. (2007) *IEEE Transactions on Control Systems Technology*, **15**, 906–915.
- 51 Schitter, G., Stark, R.W., and Stemmer, A. (2004) *Ultramicroscopy*, **100**, 253–257.
- 52 Ando, T., Uchihashi, T., Kodera, N., Yamamoto, D., Miyagi, A., Taniguchi, M., and Yamashita, H. (2008) *Pflügers Archiv*, **456**, 211–225.



## 9

# Solvent Interactions with Proteins and Other Macromolecules

Satoshi Ohtake, Yoshiko Kita, Kouhei Tsumoto, and Tsutomu Arakawa

### 9.1

#### Introduction

Proteins *in vivo* are surrounded by a large number of solutes present at wide range of concentrations, leading to an extremely high total solute concentration [1–6]. It has been reported that proteins are only marginally stable when isolated and purified in dilute solutions [7–10]. Even when proteins are conformationally stable, they tend to aggregate and lose their functional structure [11]. Furthermore, it is widely accepted that some proteins are intrinsically unstable or are present in a natively unfolded state [12–14]. In any of these circumstances, proteins have to be stabilized against unfolding or aggregation. As mentioned above, in the natural environment, proteins are present in a crowded solution with many macromolecules. Such a crowded condition can stabilize proteins [1, 15, 16]. In addition, proteins are bound by their specific substrates or other ligands, which increase their thermodynamic stability [17, 18] through linkage function [19–22]; ligand binding that is specific to the native structure is linked to the stabilizing energy of the ligand-bound protein structure. When proteins are isolated from their natural environments, these stabilizing factors are lost. Various solvent additives (cosolvent) have been developed for such critical applications. Note that the term “cosolvent” indicates that they exert their effects only at high concentrations, thus becoming part of the solvent. We will review cosolvent applications in a wide variety of fields, including: (i) research, (ii) purification, (iii) protein refolding and expression, (iv) chromatography, (v) formulation, (vi) disinfection, and (vii) freezing and freeze-drying.

Biological functions of protein molecules are regularly studied in pure solution. The solution conditions can be made close to their natural environment by adding salts and buffers. However, such a native-like condition is often insufficient to maintain protein stability, and there are numerous examples of proteins losing their functional structure in dilute solutions. The addition of cosolvents, however, has been demonstrated to restore their structural stability to that observed in their natural environments. For example, Weisenberg and Deery [23] and Weisenberg and Timasheff [24] made a pioneering observation during the isolation of brain microtubules.

The microtubules were readily purified from brain tissue homogenate by a simple temperature-controlled assembly and disassembly process [25, 26]. However, purification required the microtubule proteins to remain in their functional structure, which was maintained *in vivo* by microtubule-associated proteins (MAPs). Previously, it had been assumed that the binding of MAPs is essential for the observed microtubule assembly [25, 26]. This was shown to be incorrect when Weisenberg *et al.*, as well as others, purified MAP-free tubulins, which are subunit proteins of microtubules. What made this purification possible was the presence of cosolvents, glycerol or sucrose, at high concentration: around 30% glycerol and around 1 M sucrose [27–31]. Without these cosolvents, tubulin lost the ability to polymerize into microtubules during purification and storage. More importantly, the pure tubulins were able to polymerize to microtubules in the presence of the cosolvents; that is, these cosolvents were capable of replacing the activity of MAPs, demonstrating that MAPs are not essential for microtubule assembly. There are many other examples demonstrating the use of cosolvents at high concentrations to examine the function of proteins and other macromolecules in *in vitro* systems.

Cosolvents not only stabilize, but also serve to precipitate proteins [32, 33]. A classical example is cold ethanol precipitation of plasma proteins [34–37]. An aqueous solution of organic solvents has been used to induce polymerization of microtubules [38] and precipitate and fractionate proteins and other macromolecules [33, 39–41]. In addition to differential precipitation, protein purification is conducted by various chromatography methods, some of which use cosolvents at fairly high concentrations, such as salting-out salts for hydrophobic interaction chromatography (HIC) and organic solvents for reversed-phase chromatography [42–47]. These cosolvents modulate binding and elution of proteins.

Recombinant proteins are the basis of biotechnology. As will be described in Section 9.2.1, proteins that are developed for therapeutic purposes must meet a rigorous regulatory guideline. Recombinant protein expression is also assisted by cosolvents. One such example is periplasmic expression of foreign proteins in *Escherichia coli*. It has been shown that the addition of folding-enhancing osmolyte [48] or aggregation-suppressing arginine [49] in culture media increased the periplasmic expression of recombinant proteins. When recombinant proteins are expressed as insoluble inclusion bodies, they need to be solubilized and refolded for function. Refolding of recombinant proteins usually requires assistance of either folding-enhancing or aggregation-suppressing cosolvents in an appropriate buffer system [50–55]. Various solubilization and refolding technologies are often combined with the use of cosolvents. The Refold database suggested that the dilution refolding method is more frequently combined with folding-enhancing cosolvents, while arginine is combined with a majority of refolding technologies [56, 57] (see also [refold.med.monash.edu.au](http://refold.med.monash.edu.au)). Nevertheless, how these cosolvents are used effectively for protein refolding is still determined case-by-case, depending greatly on the protein, while arginine appears to be a universal cosolvent, applicable for a wide variety of proteins, independent of the refolding technology employed.

Pharmaceutical proteins require high-quality products possessing reasonable storage stability. While it would be ideal to make the protein solution for injection

match the environment found in physiological condition, this is often not possible and requires the presence of cosolvents for preparing a stable protein formulation. Even with this aid, proteins are often too unstable to be stored in liquid form and may need to be stored frozen or be lyophilized. In either case, the proteins are exposed to various stresses, including ice and salt crystal formation, freeze-concentration of the protein to extremely high concentration, and removal of hydrating water. The addition of cryo- or lyo-protectants is usually essential to preserve the native conformation of the proteins and their long-term storage stability. In the frozen and dried state, the cosolvent interactions with macromolecules are expected to differ greatly from those in solution.

The last application relates to our recent invention of using a concentrated arginine solution to inactivate viruses [58–61]. Low pH or alcohol solution has been used primarily as a disinfectant of virus- or microbe-contaminated surfaces [62–67]. Aqueous acidic arginine solution appears to provide a more effective and safer alternative to the current disinfectant formulation. These cosolvent applications will be reviewed in detail later in this chapter.

Although cosolvent systems have been and may be used as a routine procedure in a wide variety of applications described above, it is clear that they will be better utilized if the mechanism of cosolvent effects in each application is fully understood. There have been various attempts to explain the mechanism of cosolvent effects. To our knowledge, the very first mechanistic explanation was made by Traube [68], who demonstrated the effects of salts on the surface tension of water. Two important observations were noted. First, is the effect of different salts on gas solubility, which follows their effects on protein solubility. The salt effect on protein solubility was initially observed by Hofmeister in 1888 [69]; there is a universal order of salts in changing the solubility of solutes, whether they are protein or gas, in water. The second observation is that salts that increase the surface tension of water to a greater extent are more effective in decreasing the solubility of solutes. This surface tension principle was further expanded by Sinanoglu and Abdunur [70, 71] and later by Melander and Horvath [72]. The surface tension principle is clearly defined by the property of the salt solution alone and has no involvement of proteins or other macromolecular solutes. In other words, this principle is independent of the solute surface; the term “solute” may represent macromolecules or viruses, with which water and cosolvents interact. Furthermore, the surface tension principle is related to the hydration property of salt ions. It can thus be derived that the hydration potential of salt ions is closely correlated to the effects of salts on macromolecules [73–75].

Another fundamental principle of the cosolvent effect arises from its excluded volume [1, 76–81]. Proteins and other macromolecules are surrounded by other macromolecules at high concentrations in the natural environment, as described earlier. In such a case, the surrounding macromolecules restrict the motion of the macromolecule of interest and increase its stability. A similar mechanism operates in solutions containing cosolvents at high concentrations. Solvent molecules, including water, are restricted in rotational and translation motion near the solute surface (e.g., protein, DNA, or virus), relative to their motions in the bulk phase, which makes the system thermodynamically less stable [82, 83]. As cosolvents typically have a larger hydrodynamic radius than a water molecule, they are excluded from the protein

surface to a greater extent than the latter. As a result, the addition of cosolvents destabilizes the protein in water; more so for the protein with greater surface area. This mechanism is independent of the chemical nature of the solute species.

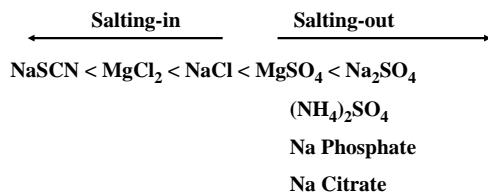
While these two principles certainly play a role in explaining the effects of cosolvents on proteins, the surface of proteins and other solute species cannot be completely inert. In other words, cosolvents can interact with the protein surface directly. If this occurs, however, the effects that are expected from the above principles no longer comply with the observed effects. Direct measurements of protein–solvent interaction should reflect both cosolvent binding and the solute-independent mechanism. In fact, equilibrium dialysis measurements demonstrated that a fine balance between binding and exclusion of cosolvents with the solutes (i.e., macromolecules) determines the effects of the solvent, in which exclusion arises from the above two principles. As such equilibrium measurements relate to the differential interaction of proteins with water and cosolvents, they are termed “preferential” interactions. Preferential interaction measurements revealed that proteins are highly hydrated in cosolvent solutions that enhance protein stability. This excess hydration can be described by either the surface tension or excluded volume principle. However, it could also be due to the direct binding of water, as has been demonstrated previously [84–86]. While preferential interaction measurements cannot distinguish water binding from cosolvent exclusion, they are independent of the mechanism involved and can explain the cosolvent effects without understanding the hydration mechanism. It is an established concept that proteins and other macromolecules bind water with varying degrees of affinity. Thus, preferential interaction reflects all different types of binding or exclusion mechanisms of cosolvents with the solute species. More importantly, the thermodynamic interactions can explain the solvent effects, through thermodynamic linkage of the interactions, on the reaction of the solutes, such as unfolding or aggregation reactions. As is evident, a better understanding of these thermodynamic interactions will aid in the effective application of cosolvents for proteins and other macromolecules.

Hydrated water becomes critical when water activity is reduced, such as upon freezing or freeze-drying. Protein–water interactions must be replaced by the cosolvents in the dried state for maintenance of protein stability/activity. This will be covered later as a unique case of the interaction between solutes and cosolvents. First, we will review the interaction mechanisms of proteins with cosolvents and how such interactions can explain the effects of solvents, through thermodynamic linked function, on various cosolvent applications for proteins, viruses, and nucleic acids.

## 9.2 Solvent Applications

### 9.2.1 Research

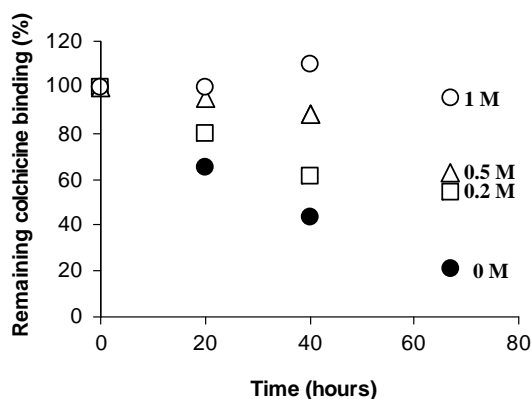
Among the various known effects of salts, the Hofmeister series [69] may be the most popular and useful (even to this day) in purification and processing of proteins,



**Figure 9.1** Order of salts in altering protein solubility. Buffers that are often used to enhance protein binding are also included.

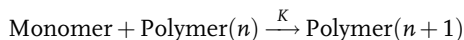
nucleic acids, and viruses. An abbreviated Hofmeister series is summarized in Figure 9.1. The series shows, for example, that NaSCN and MgCl<sub>2</sub> belong to the salting-in class while MgSO<sub>4</sub> and Na<sub>2</sub>SO<sub>4</sub> belong to the salting-out class. Even buffers, such as phosphate and citrate at near-neutral pH, were observed to salt-out proteins and other biological molecules. Any research on biological samples requires, more or less, an unnatural environment. Sample proteins, peptides, nucleic acids, or viruses in *in vitro* experiments are most likely present at concentrations far above their physiological concentration and in purer form. As stated, many proteins are often unstable under such conditions, and must be stabilized against conformational changes and/or aggregation. There are numerous examples demonstrating instability under unnatural environments and how the cosolvents were incorporated to circumvent the stability issues. Cosolvents are also used to characterize proteins. Due to the sheer number of publications, it is not possible to cover all of the references that have employed cosolvents in research applications. This chapter will, however, highlight some examples of their applications in basic research.

In research, glycerol and sucrose have been the compounds/stabilizers of choice, through several decades of experience in biochemical research on proteins and other macromolecules. As stated in Section 9.1, tubulin rapidly loses its structure in solution and in the frozen state [30, 87, 88]. Figure 9.2 demonstrates the ability of



**Figure 9.2** Plot of colchicine binding of tubulin as a function of sucrose concentration during storage at 4 °C in PMG buffer (10 mM phosphate, 5 mM MgCl<sub>2</sub>, 0.1 mM GTP, pH 7.0): 0 (●), 0.2 (□), 0.5 (△), and 1 M sucrose (○). (Data adapted from [88].)

tubulin to bind an anticancer drug, colchicine, when stored in PMG buffer (10 mM phosphate, 5 mM MgCl<sub>2</sub>, 0.1 mM GTP, pH 7.0) at 4 °C [88]. Drug binding indicates that the protein is in the native structure. The drug-binding activity of tubulin is rapidly lost with time in the PMG buffer (closed circles) even when stored at low temperatures. However, the colchicine-binding activity can be maintained upon the addition of sucrose (Figure 9.2). At 1 M sucrose, there is very little activity loss even after 60 h of incubation (open circles). The observed activity loss is most likely due to the conformational change or aggregation of tubulin upon storage, as evidenced by sedimentation velocity analysis, in which the tubulin stored frozen without stabilizers resulted in extensive aggregation (nonspecific polymerization). In the presence of 1 M sucrose, the native structure of tubulin was maintained [30]. Such stabilization enhanced the ability of tubulin to undergo specific polymerization reactions to form microtubules. The stabilization effects of cosolvents have been manifested in their ability to enhance polymerization of tubulin and other structures [89–93]. A simplistic model of protein polymerization (e.g., tubulin, actin, virus, and fragellin) can be described by the equilibrium between the monomer and polymer states:



The equilibrium constant  $K$  is independent of the size of the polymer and hence equivalent to the inverse of solute solubility ( $S$ ), or the critical micelle concentration (CMC) of micelles [94–96] (i.e., the activity (concentration) of polymer does not depend on the size of the polymer). Thus, in this case, the critical monomer concentration,  $C_m$ , which is in equilibrium with the polymer, is defined by  $C_m = 1/K$ . Table 9.1 shows the  $C_m$  value as a function of glycerol concentration [97]. In the absence of glycerol,  $C_m$  is 8 mg/ml, indicating that there will be no microtubule formation below 8 mg/ml. As with CMC, a protein concentration higher than 8 mg/ml, at minimum, is required for microtubule formation to occur. This is the reason why tubulin does not assemble following purification. MAPs (later called tau) lower the  $C_m$  to a level that allows *in vitro* polymerization [91, 98–100]. Glycerol not only stabilizes tubulin, but also lowers the  $C_m$  to 0.75 mg/ml at a concentration of 4.11 M and allows *in vitro* polymerization feasible. Other cosolvents, including sodium

**Table 9.1** Critical monomer concentration of microtubule assembly as a function of tubulin concentration.

Glycerol concentration (mol/l)	$C_m$ (mg/ml)
0	8.0
2.06	2.0
2.74	1.45
3.43	1.10
4.11	0.75

Data adapted from [97].

piperazine-*N,N'*-bis(2-ethanesulfonic acid), sodium glutamate, sodium 2-morpholinoethanesulfonic acid, and creatine phosphate, were also found to enhance tubulin polymerization at high concentrations (e.g., 0.1–0.8 M) [101–104]. However, these cosolvents also caused formation of other assembled structures. Since these cosolvents are electrolytes, the increased ionic strength could have affected the specific electrostatic interactions essential for microtubule formation, leading to an alteration in protein–protein interactions.

The addition of  $\text{Na}_2\text{SO}_4$ , a strong salting-out salt, resulted in the reversal of urea-induced actin depolymerization [105]. More specifically, actin depolymerization, which resulted from the addition of 4 M urea, was reversed upon the addition of 0.6 M  $\text{Na}_2\text{SO}_4$ . A similar effect of  $\text{Na}_2\text{SO}_4$  has also been observed for flagella assembly [106]; while the subunit flagellin cannot polymerize readily in dilute buffer, the addition of 0.6 M  $\text{Na}_2\text{SO}_4$  resulted in rapid assembly, while 1.2 M addition of  $\text{Na}_2\text{SO}_4$  caused aggregation of assembled short flagella filaments. Concentrated phosphate, in particular at higher pH, was also effective in enhancing assembly, consistent with the mechanism of a strong salting-out salt [107, 108]. While NaF, KF, and sodium citrate were also effective, NaCl,  $\text{MgCl}_2$ , and  $\text{CaCl}_2$  were either ineffective or enhanced the depolymerization of flagella, again consistent with the weaker salting-out or salting-in effects of these salts [106] (Figure 9.1). Although a detailed description of preferential interactions of salts is given later, Table 9.2 introduces the various cosolvents in relation to their interactions with proteins: salting-in salts and salting-out salts are clearly distinguished by their tendency to bind to proteins.

Tobacco mosaic virus (TMV) comprises a single protein species and a RNA. In the absence of RNA, the protein assembles into a helical polymer and other assembled structures, including oligomers, disks, and “Lock-washer” [109]. The helical polymer dominates below pH 6.5, therefore the assembly reaction can be described as a helix-condensation mechanism. Such assembly reactions are enhanced in the presence of salts, such as KCl and  $(\text{NH}_4)_2\text{SO}_4$  [110, 111].

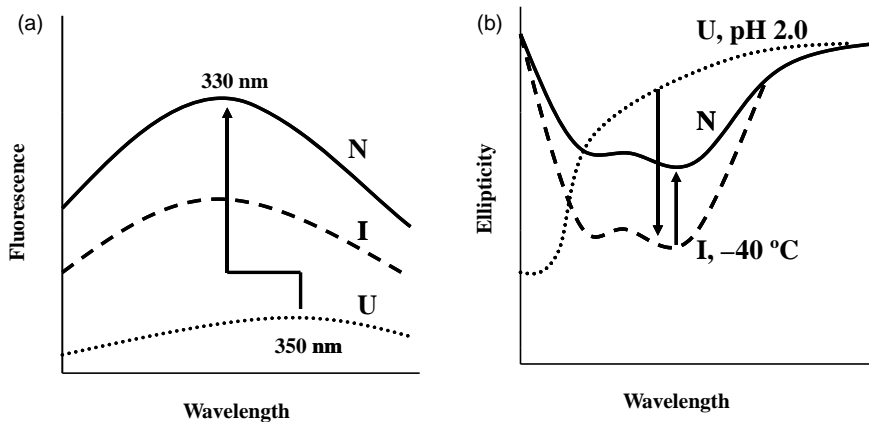
**Table 9.2** Classification of various cosolvents.

Cosolvents	Preferential interaction	
	Preferential binding	Preferential exclusion
Salts	salting-in	salting-out
Osmolytes		folding enhancer structure stabilizer aggregation suppressor (arginine)
Organic solvents	destabilizer denaturant (helix inducer)	precipitant
Polymers		precipitant stabilizer (inert polymer) destabilizer (PEG)
Denaturants	denaturant solubilizer	

Organic solvents have been extensively used as cosolvents, even as high as around 100%, in various research applications. For example, organic solvents are versatile precipitation reagents for proteins, which find application in their crystallization and purification [39–41]. Other applications include, but are not limited to, enzymology, kinetic analysis at subzero temperatures, and nuclear magnetic resonance (NMR) analysis, as a solubilizing agent for hydrogen/deuterium (H/D) exchange. Water-soluble enzymes function naturally in aqueous solution and use water-soluble substrates. It has been demonstrated that as long as the native functional structure is maintained, water-soluble enzymes are active in organic solvents, enabling them to act on poorly water-soluble substrates [112–116]; enzymes in organic solvents exhibit unique substrate specificity. Surprisingly, such conditions can even be achieved by suspending the dried enzymes in organic solvents. In fact, water-immiscible organic solvents make the suspended enzymes more stable; it should be noted that enzymes under such conditions are not dissolved, but are dispersed as aggregated particles.

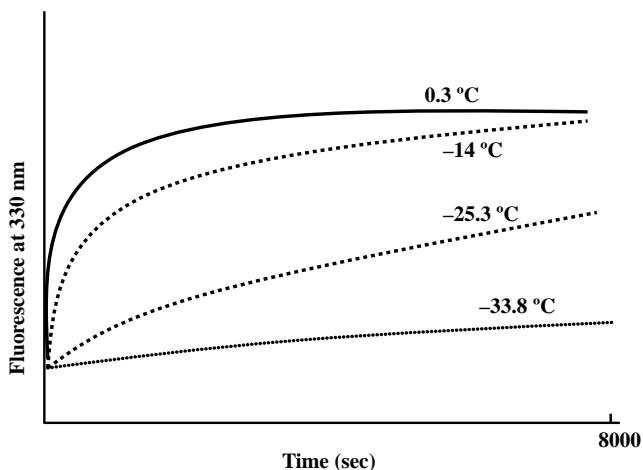
Protein folding or enzyme reaction occurs too fast to analyze the structure and/or activity of kinetic intermediates. Reduction of the experimental temperature is one way to slow the reaction kinetics, as has been demonstrated by the pioneering work of Douzou and Fink; water-miscible organic solvents have been used to study the reaction kinetics at subzero temperatures [117–122]. At these low temperatures, denaturing solvents often have no deleterious effects on protein structure, allowing protein folding or enzyme reaction to be unaltered. The only effect is to slow the kinetics of reaction. For example, staphylococcal nuclease (SNase) was demonstrated to retain the native structure at pH 7.0, even in 50% methanol, although the melting temperature at neutral pH decreased from around 55 °C in the absence of methanol to around 15 °C in 50% methanol [122]. While the stability is significantly decreased (as is evidenced by the lowered melting temperature), the native structure is still retained below 10 °C in 50% methanol at pH 7.0. Upon lowering the pH to 2.0, however, SNase denatured, even near 0 °C. Figure 9.3 shows the structure of SNase in 50% methanol at pH 2.0 (unfolded, U) and at pH 7.0 (native, N). A fluorescence peak is observed at 350 nm, when unfolded (see the spectrum U in Figure 9.3), and at 330 nm, when folded (see the spectrum N in Figure 9.3). Circular dichroism (CD) analysis, conducted at 0 °C, demonstrated that the protein was helical in the native state and disordered at pH 2.0 (U) (Figure 9.3b). To initiate folding, the pH of the unfolded sample, present at pH 2.0, was adjusted to 7.0. The timecourse of folding (Figure 9.4) was followed by monitoring the fluorescence intensity at 330 nm, as the fluorescence intensity at this wavelength is expected to increase upon folding (Figure 9.3). It is evident from Figure 9.4 that folding is slowed significantly at low temperatures (e.g., compare the kinetics at 0.3 and –33.8 °C), suggesting that the folding intermediate may be observed under these experimental conditions. Figure 9.3 also shows the fluorescence and CD spectra of the intermediate structure (I) at –40 °C; the fluorescence peak has already shifted to 330 nm, indicating that the native-like structure is formed. The lower intensity suggests that SNase must undergo further structural rearrangement. The CD spectrum is consistent with the observation that folding has occurred, although not fully native, as demonstrated by the formation of an  $\alpha$ -helix structure. It appears that more helices are formed in the



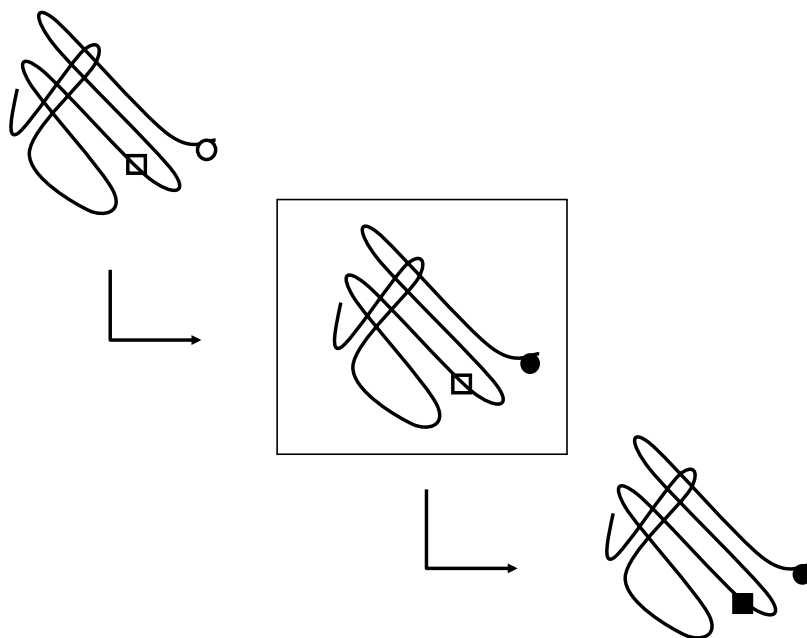


**Figure 9.3** Fluorescence and CD spectra of SNase in 50% methanol. The native structure (N) was obtained at pH 7.0 and the unfolded structure (U) was obtained at pH 2.0. The intermediate structure (I) was obtained by raising the pH 2.0 sample to pH 7.0 at  $-40^{\circ}\text{C}$ . (Data reformatted from [122].)

intermediate state (compare the spectrum I with N). Such overshooting of the helix formation has been observed from the rapid kinetic experiment using the stopped-flow technique [123, 124]. It is evident from these observations that while organic solvents enhance unfolding at elevated temperatures, they do not affect protein structure at low temperatures. The ability of organic solvents to denature proteins is due to their weakening effect on hydrophobic interactions, which becomes much



**Figure 9.4** Folding of SNase in 50% methanol at pH 7.0. Unfolded SNase at pH 2.0 was titrated to pH 7.0 to initiate folding at different temperatures, ranging from  $-33.8$  to  $0.3^{\circ}\text{C}$ . Folding was followed by fluorescence intensity at 330 nm. (Data reformatted from [122].)



**Figure 9.5** Timecourse of H/D exchange reaction. Solvent-exposed proton is shown by the open circle (○) and the less exposed proton is shown by the open square (□). Exchanged proton is shown by closed symbols (●, ■).

weaker at subzero temperatures: hydrophobic interaction is highly temperature-dependent.

H/D exchange methodology measures the solvent-accessible, exchangeable protons by H/D exchange. Figure 9.5 illustrates an example of a protein with two exchangeable protons (open circle and square). The open circle represents a more accessible proton than the one indicated by the open square (top left). A brief exposure of the protein to  $D_2O$  leads to a complete exchange of H by D for the solvent-accessible proton (D, expressed as a closed circle in the middle panel). If the measurement time is faster than the exchange rate for the less accessible proton (i.e., open square), then the timecourse of the exchange can be followed. A prolonged incubation should lead to the H/D exchange of the less accessible proton (closed square; bottom left). A recently developed NMR technique [125] has allowed the rapid acquisition of two-dimensional NMR data feasible. Two-dimensional  $^1H$ - $^{15}N$  correlation spectra can now be obtained within 2–3 s using this pulse-sequence mode technique. While this measurement mode is fast enough to follow the timecourse of proton exchange, the time requirement for sample preparation (i.e., sample mounting and optimizing measurement parameters) may deter the use of this NMR technique. Alternatively, reaction kinetics can be followed if the H/D exchange can be halted mid-reaction. The best approach involves the exchange of water with organic solvents, which contain no exchangeable protons. To stop the H/D exchange

instantaneously, the protein solution after the exchange reaction was quench-frozen in liquid nitrogen and then extensively dried to remove as much water as possible. The lyophilized protein was not soluble in benzene, chloroform, diethyl ether, or dioxane, but was slightly soluble in acetone and more so in acetonitrile and several alcohols. Strikingly, the lyophilized apomyoglobin was readily soluble in dimethyl sulfoxide (DMSO), which resulted in a reasonable NMR spectrum [126]. This technology has been successfully used to follow the folding kinetics of several proteins [127].

### 9.2.2

#### Precipitation

Precipitation and phase separation are conventional techniques used for the purification of biological macromolecules. Plasma fractionation by aqueous ethanol solution, developed by Cohn [34, 128], is still used to this day to produce important human therapeutics from blood [129–131]. Since the discovery of the Hofmeister series [69], differential precipitation by ammonium sulfate (salting-out salt) has been the simplest and most effective means to reduce the sample volume and also to fractionate proteins. While this process is useful for protein purification, it has not been as practical for plasma fractionation, which is one of the most important life-saving technologies available today. A critical parameter for plasma-derived products is the viral load, or product safety. While ammonium sulfate may be effective for plasma fractionation, it is not effective in virus inactivation; rather, it may even stabilize viruses. Based on Mellanby's observation of cold ethanol-induced fractionation of horse serum [132], Cohn developed a method for differential precipitation of human plasma proteins using an aqueous ethanol solution at low temperatures [34, 128], meeting the critical need for blood-derived products during World War II. This process consisted of the successive addition of 5–40% ethanol at  $-3$  to  $-6$  °C, and modulation of pH and ionic strength, resulting in fractionation of fibrinogen, albumin, antibodies, coagulation factors, and protease inhibitors simultaneously with a certain degree of virus inactivation [133]. Although other precipitating agents described above may be used, ethanol is safe to humans and is not strongly denaturing to the proteins at or below moderate temperature. Nevertheless, many alternative procedures using different cosolvents have been developed [36]. It should be noted again that organic solvents, including ethanol, can denature proteins at elevated temperatures and must be used with caution.

Organic solvents are also used to induce protein crystallization. Obtaining high-quality crystals is crucial for successful structure determination of proteins using X-ray crystallography. Salting-out salts, poly(ethylene glycol) (PEG), and organic solvents have all been used for this purpose. Among them, 2-methyl-2,4-pentanediol (MPD) is an extremely strong protein precipitant [134, 135], although it is a protein destabilizer at elevated temperatures, which is true of other organic solvents [136]. As described earlier, and also later in greater detail, such effectiveness comes from the strong repulsion of MPD from the charged protein surface, reflecting the observed phase separation of MPD by the addition of ions [39].

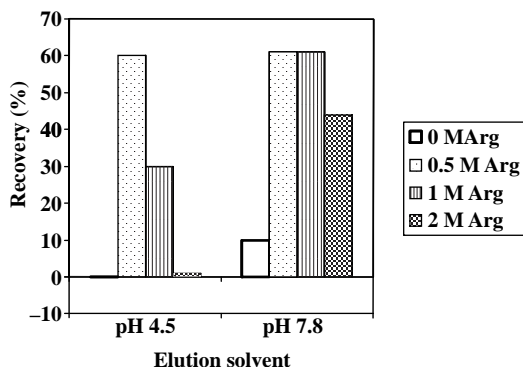
## 9.2.3

**Chromatography**

Separation of biological macromolecules employing chromatography is normally manipulated by solvent pH and ionic strength – the concentration of ions is normally within the range close to that resembling the physiological concentration. However, higher concentrations, at which the effects of cosolvents can no longer be explained from simple electrostatic theory, are sometimes required for optimal chromatographic performance [137–145]. Examples include HIC in the presence of concentrated salt solutions, reversed-phase high-performance liquid chromatography (RP-HPLC) in the presence of organic solvents, and size-exclusion chromatography (SEC) that is often carried out at high salt concentrations or in the presence of organic solvents, which results in the suppression of nonspecific protein interaction with the column matrix. When cosolvents at high concentrations are used, the results are often those not expected from simple electrostatic or hydrophobic effects. In addition, these cosolvents may have an unexpected impact on the samples to be purified.

Both HIC and RP-HPLC use hydrophobic ligands attached to the column matrix, to which macromolecules bind. In the former case, high concentrations of salting-out salts such as  $(\text{NH}_4)_2\text{SO}_4$ , citrate, or phosphate (Figure 9.1) are used to enhance binding [43, 108, 146–150], while in the latter case, high organic solvent concentrations are used for elution [151–153]. HIC is one of the fundamental analytical and separation tools used for protein purification [154, 155], and is now widely used as a platform technology for large-scale manufacturing of therapeutic proteins [156]. Various salts have been used for binding, including  $(\text{NH}_4)_2\text{SO}_4$ , NaCl, phosphate, and citrate [42, 157] (see Figure 9.1 for the order of salting-out strength). The mechanism of enhanced protein binding in HIC, by the salting-out salts, is due to the surface tension effect, or “attraction pressure,” as described earlier [68, 72]. Amino acids (e.g., glycine), polyols (ethylene glycol and glycerol), and sugars (sucrose) have all been shown to modulate protein binding and elution [158, 159]. Thermodynamic preferential interaction was used successfully to explain their effects in HIC. Effective elution of bound proteins can be obtained in the presence of arginine at 0.5–1 M range and in dilute ethanol solution [160]. For example, Figure 9.6 shows the enhanced elution of activin, a sticky protein, from an HIC column. At the two pH values examined, elution of the protein by buffer alone is extremely low: near-zero at pH 4.5 and 10% at pH 7.8. The recovery, however, is increased to over 60% with 0.5–1 M arginine. At both pH values, the recovery was lower using 2 M arginine for an unknown reason. As described below, elution of antibodies in affinity chromatography is monotonically increased with increasing arginine concentration. The difference observed using arginine, between HIC and affinity chromatography, is most likely due to the mode of protein binding: HIC by hydrophobic interaction alone and affinity by mixed-mode interactions (i.e., hydrophobic, electrostatic, and hydrogen bonding). Although not shown here, the addition of ethanol to 0.5 M arginine further enhanced the elution [160].

HIC can also be conducted using a polar ligand, such as ion-exchange and polysaccharide columns, when extremely strong salting-out conditions are



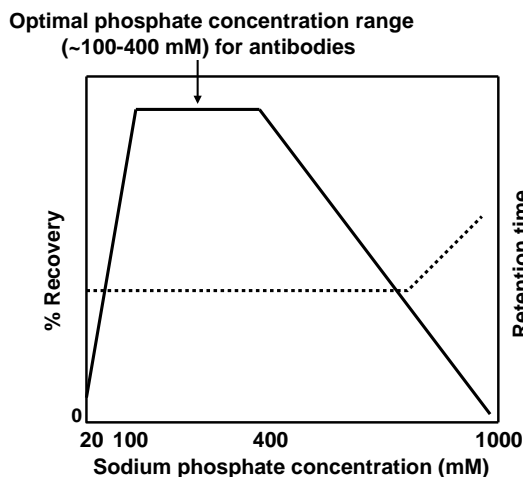
**Figure 9.6** Recovery of bound activin from HIC column using arginine at pH 4.5 or 7.8. Arginine concentration ranged from 0 to 2 M, as indicated. (Data reformatted from [160].)

employed [161, 162]. For example, nonhydrophobic columns, including DEAE, CM, and P-gel ion-exchange resins as well as nonsubstituted hydrophilic gels, combined with extremely strong salting-out conditions, have all been employed for this purpose [162–166]. Fujita *et al.* [165] showed protein binding to nonsubstituted Sepharose 4B in 3 M  $(\text{NH}_4)_2\text{SO}_4$ , and protein elution by descending concentration of  $(\text{NH}_4)_2\text{SO}_4$  in a similar manner to that employed for normal HIC operation. Although the data described above prove that a strong salting-out condition can confer protein binding, even on hydrophilic ligands [46], the application of HIC on hydrophilic columns can be found in the purification of halophilic enzymes. Enzymes from halophilic organisms require high salt concentrations for stability and function [167], which makes many conventional chromatography methods that are conducted under low salt conditions impractical (e.g., ion-exchange chromatography (IEC) and SEC). Normal-mode HIC would cause the binding to be too strong to dissociate under high salt concentrations, while the use of low salt concentration (required to dissociate the bound proteins) would cause inactivation. HIC using hydrophilic ligands has been successfully used, because a high salt concentration was required for sufficient binding. Furthermore, as the bound proteins elute at relatively high salt concentrations from the hydrophilic ligand, the eluted fractions contained a high enough salt concentration for direct evaluation of the enzyme activity [162, 168, 169].

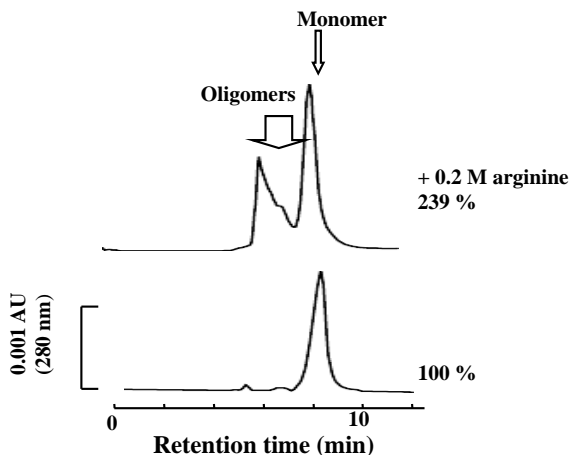
RP-HPLC typically employs stronger hydrophobic ligands and elution conditions under acidic pH, which enhance hydrophobic interactions between the protein and ligand: under such conditions, the protein is highly unfolded, exposing more hydrophobic amino acids to the solvent. While protein binding does not require salting-out salts, as in HIC, elution of the bound protein requires the use of organic solvents. Such elution effects of organic solvents can be readily explained from their effects on the solubility of nonpolar compounds, as described later.

SEC, also termed gel-filtration or gel-permeation chromatography, is among the most frequently used techniques for analysis, quality control, and purification of proteins [170–173]. The recovery and resolution of proteins are often compromised

by the nonspecific interaction of the proteins with the column matrix. Resolution between peaks is greater with longer columns and lower protein load, which both increase the surface area for a given amount of protein, and thus protein-binding sites on the matrix. Both electrostatic and hydrophobic interactions can occur between the protein and column matrix [174]. To reduce such nonspecific binding to the stationary phase, various approaches have been taken [175–177]. For example, the electrostatic interactions can be suppressed by the addition of electrolytes (e.g., high phosphate concentration or NaCl at moderate concentrations (0.2–0.5 M)) [175]. High salt concentrations, however, may enhance protein adsorption, occurring through hydrophobic interactions. Gagnon [178] demonstrated that the addition of both ammonium sulfate and phosphate delayed the elution of monoclonal antibodies. Due to the salting-out effects of ammonium sulfate and phosphate (Figure 9.1), the salts enhanced the association–dissociation equilibrium of the antibodies with the column matrix, thus retarding their elution. Figure 9.7 shows a schematic presentation of the effects of increasing phosphate concentration on the retention time and recovery of antibodies using SEC [49, 179, 180]. Although the retention time was not affected under low phosphate concentrations (up to around 800 mM), recovery increased significantly with increasing phosphate concentration (0–100 mM), indicating that phosphate suppressed the electrostatic binding of the antibodies to the column. However, both recovery and retention are strongly affected at higher phosphate concentration, indicating that extremely high salt concentration causes protein binding; there is almost no recovery at 1 M phosphate, with the retention time being delayed considerably. Thus, at such high phosphate concentrations, the loaded proteins are mostly retained by the column, and any proteins that dissociate no longer elute at their expected positions. When hydrophobic binding occurs in SEC, it can be suppressed by the addition of organic solvents. In fact,



**Figure 9.7** Schematic illustration of protein recovery (solid line) and retention (dotted line) by a SEC column as a function of phosphate concentration. (Data reformatted from [180].)



**Figure 9.8** Effect of 0.2 M arginine on the elution of an antibody sample containing soluble oligomers. Lower panel shows the elution with 0.1 M phosphate, pH 6.8, while upper panel shows the elution with the addition of 0.2 M arginine. (Data reformatted from [183].)

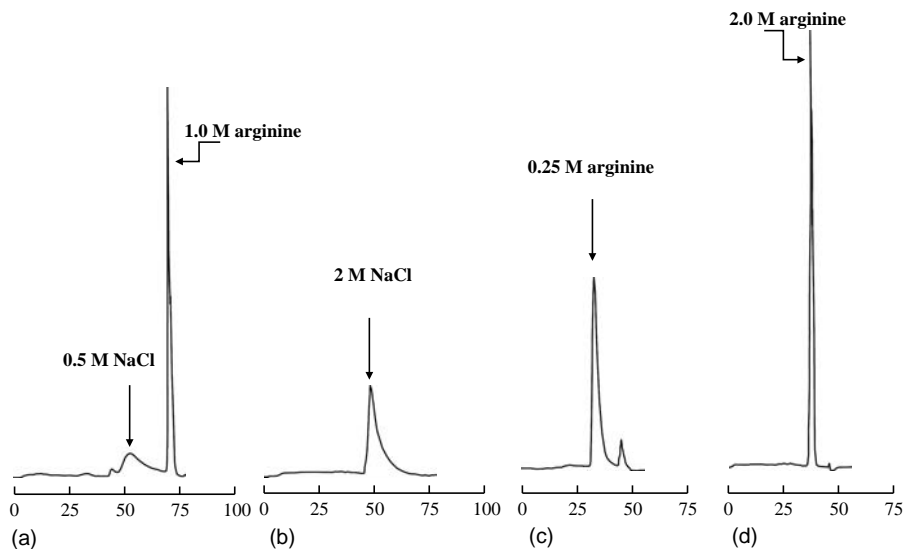
acetonitrile at high concentrations is often used for that purpose [181, 182]. Thus, while salts can be used to suppress electrostatic interactions and enhance hydrophobic interactions, organic solvents can be used to weaken hydrophobic binding. Arginine has been shown to effectively suppress nonspecific binding of proteins to both silica and polysaccharide-based columns [183]. Figure 9.8 demonstrates such an example, in which a highly aggregated antibody preparation was subjected to SEC analysis. It appears that the sample is fairly homogenous when run using 0.1 M phosphate buffer, demonstrating small amounts of oligomer peaks (lower panel). However, when analyzed in the presence of 0.2 M arginine, the elution profile resembles that of an extremely heterogeneous preparation. The addition of 0.2 M arginine to the phosphate-based elution buffer increased the overall recovery 2.4-fold. More importantly, the soluble oligomers, which were recovered only marginally using the phosphate mobile phase, were eluted as a large peak in the presence of 0.2 M arginine. This indicates that SEC run under these conditions, in which protein binding occurs, leads to a misleading conclusion concerning sample purity.

Electrostatic interactions determine the binding of proteins to the column matrix in IEC and hydroxyapatite chromatography, and thus salts play a major role in modulating the binding and elution of proteins [184]. In IEC, NaCl has been the choice of salt for elution. Ammonium sulfate at high concentration can either induce or retain binding of bovine serum albumin (BSA) to a cation-exchange resin, Q-Sepharose, under the salt concentrations at which electrostatic interactions are reduced and cause elution of proteins [185]. Ammonium sulfate serves as an elution salt, as does NaCl, at low to moderate concentration, but can cause protein binding at extremely high concentration to an IEC column, as observed in HIC of halophilic enzymes. In the current example, BSA bound to Q-Sepharose in the presence of 4 M ammonium sulfate and the bound BSA was eluted at lower ammonium sulfate

concentrations. BSA, bound at low salt concentration, remained bound upon increasing the ammonium sulfate concentration step-wise to 3.5 M or above. However, lowering the ammonium sulfate concentration caused the elution of BSA, resulting from the weakening of the salting-out effect and electrostatic interaction.

PEG enhances protein–protein interactions due to its large excluded volume effect, as will be described in detail in Section 9.5. Such an effect can be used to enhance the binding of protein to IEC. Although PEG delays the elution of all species (e.g., monomers, oligomers, and large aggregates), the higher-molecular-weight species are delayed further, leading to greater separation of oligomers from the monomer [186, 187].

Ligand-affinity chromatography is a convenient way to purify proteins, as the protein of interest can be purified in a single step. Ligands can constitute any one of the following: dyes, substrates, inhibitors, and many other small molecules, which specifically bind to target proteins. Among them, dye-affinity column chromatography offers a robust and rapid purification of proteins [188]. For dye-affinity chromatography, covalently attached, group-specific ligands are used for purification of enzymes, including dehydrogenases and kinases [189]. Blue-dye columns bind enzymes requiring adenylyl-containing cofactors, such as NAD and NADP. Lactate dehydrogenase (LDH) binds to blue-dye columns and the bound enzymes can be eluted by salts, but with low recovery and broad elution peaks [190, 191], perhaps due to the different types of interactions present between the protein and the dye. Another study examined the ability of aqueous arginine solution to elute LDH from Blue-Sepharose [192]. As shown in Figure 9.9, 0.5 and 2 M NaCl (Figure 9.9a and b) were



**Figure 9.9** Elution of LDH from a dye-affinity column by NaCl or arginine at neutral pH. The concentrations of NaCl and arginine are indicated: (a) 0.5 M NaCl and 1.0 M arginine, (b) 2 M NaCl, (c) 0.25 M arginine, and (d) 2 M arginine. (Data reformatted from [192].)



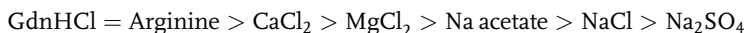
ineffective in eluting the protein, resulting in a skewed elution peak with less than 60% recovery. On the contrary, the use of 0.25 M arginine (Figure 9.9c) resulted in a sharp elution profile, with increased recovery at approximately 65%. Increasing arginine concentration to 2 M (Figure 9.9d) resulted in a nearly quantitative recovery of the bound LDH. It is evident that arginine is more effective in eluting the bound LDH than NaCl. Furthermore, 1 M arginine resulted in a sharp elution peak even following elution with 0.5 M NaCl, as shown in Figure 9.9(a).

Biological affinity chromatography is a convenient way to purify proteins and study protein–protein or protein–ligand interactions [142, 193–196], and it is a simple laboratory tool for clinical sample analysis [197]. The technique relies on specific interactions present between two molecules (e.g., between a specific inhibitor/substrate and enzyme, between ligand and receptor, between antigen and antibody, etc.). These interactions can be disrupted by the addition of protein denaturants or chaotropic salts [198, 199], which may denature or destabilize the proteins. However, such high-affinity interactions have a great advantage in that sample proteins can typically be applied to affinity columns as obtained, for example, in cell culture media, serum, tissue, or cell lysates, as the protein–protein or protein–ligand interactions normally occur under physiological conditions. The bound proteins are then eluted by raising the elution buffer salt concentration [190, 191], changing pH [191], or competition with high-affinity ligands. The last elution technique is conventionally used in tag-affinity purification. More specifically, target proteins are expressed as either the N- or C-terminal tag sequence, and then the expressed proteins are trapped by anti-tag antibodies. Dissociation can be conducted by tag peptides, which compete with the bound target proteins. However, regeneration of the antibody column often requires a harsh dissociating solvent to remove the bound free tags on the antibody resin. In this regard, elution by a cosolvent solution has a great advantage in terms of ease of use and cost.

Polyclonal antibodies are versatile reagents and are also used for the detection of therapeutic markers in different diagnostic assay formats. The polyclonal antibodies used for such applications must be specific for target antigens. Specific antibodies can readily be purified by antigen-affinity chromatography [142]. The bound antibodies are eluted by acid, but often with low recovery [194]. This is because the affinity of antibodies for an antigen is often too high, arising from a multitude of interaction forces [195, 200]. While there are various elution conditions for antigen chromatography, such as different pH (acid or alkaline elution), high salt concentrations, and different types of buffer, the most significant development in solvent manipulation for the dissociation of antibody–antigen interactions is the use of  $\text{MgCl}_2$  at high concentrations. Potts and Vogt [137] have applied various chaotropic agents to dissociate sarcoma virus kinase (which was used to generate the antibodies) from antibody columns developed against this specific kinase. Of the seven solvents tested (urea, GdnHCl, ethylene glycol,  $\text{MgCl}_2$ , NaSCN, formic acid, and  $\text{NH}_4\text{OH}$ ), urea,  $\text{MgCl}_2$ , NaSCN, and formic acid resulted in a substantial elution of the kinase. Among these four reagents, urea and formic acid destroyed the kinase activity. On the other hand, the kinase was effectively eluted with 1.9–2.5 M  $\text{MgCl}_2$ ; however, higher  $\text{MgCl}_2$  concentrations caused partial loss of activity. Durkee *et al.* [141] showed that

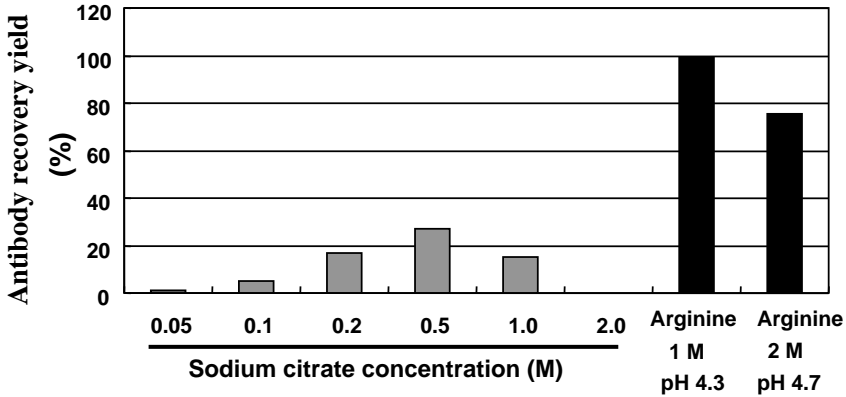
Russell's viper venom factor X can be purified by antibody-conjugated columns using 3.5 M  $\text{MgCl}_2$  at pH 7.0. In another example, Narhi *et al.* [143, 144, 201] used ligand columns to purify the different monoclonal antibodies raised against the ligand and showed that low pH elution was generally ineffective while denaturing cosolvents resulted in greater elution. However, eluted antibodies under such denaturing conditions required refolding to restore the antigen binding activity.

Protein A/G-affinity chromatography is the most convenient purification process for monoclonal antibody production [202, 203]. This affinity chromatography is a platform technology for commercial-scale production of therapeutic antibodies. Binding of Protein A with the Fc region of antibodies is so selective that the majority of contaminating proteins in the conditioned media does not bind to the column, resulting in increased capacity of the column for the antibodies. However, the interaction between the Fc region and Protein A or G is so strong that a harsh elution condition is often required. While the elution of antibodies bound to Protein A or G resin by an acidic solution is a common practice, the use of strong acid can cause antibody denaturation [204]. Cosolvents, such as arginine and arginine derivatives, at 0.3–2 M, were found to be effective in eluting several monoclonal antibodies [205, 206]. It has been shown that aqueous arginine solution at or above pH 4.0 can effectively elute antibodies from both Protein A and G columns: pH below 4.0 is normally required using acid alone. Various salts were tested for their effectiveness in weakening antibody binding to Protein A at pH 4.0 and 5.0, at which many antibodies have reduced affinity for the resin [207]. The order of their ability to reduce Protein A binding was found to be:



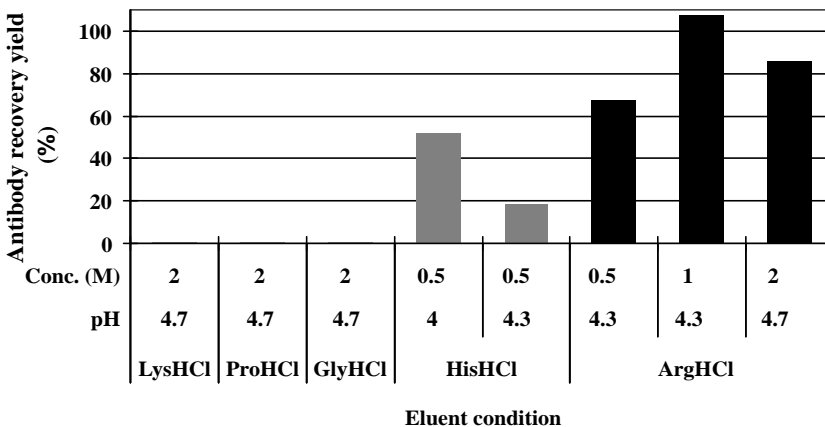
In fact,  $\text{Na}_2\text{SO}_4$  enhanced binding, consistent with its salting-out effect (Figure 9.1), while arginine caused elution of bound antibodies from the Protein A column. Figure 9.10 shows the elution of antibody from Protein A column at or above pH 4.3, which is higher than the normal operating pH value. At pH 4.3, the conventional elution buffer (citrate) is not effective; the recovery is at most around 30% at 0.5 M citrate. Interestingly, there is no recovery at 2 M citrate, reflecting its salting-out property (see Figure 9.1 for the order of salting-out salts). At the same pH, 1 M arginine resulted in near full recovery from the column. Increasing arginine concentration to 2 M resulted in approximately 75% elution, even at pH 4.7. It appears that among the amino acids, only arginine is effective. Figure 9.11 shows the effects of 2 M lysine (shown as  $\text{LysHCl}$  to indicate that HCl was used to adjust the pH), proline ( $\text{ProHCl}$ ), and glycine ( $\text{GlyHCl}$ ) at pH 4.7 with little elution of the antibody being observed. Histidine ( $\text{HisHCl}$ ) was slightly effective at 0.5 M and pH 4.0 or 4.3. Compared to these amino acids, 0.5 and 1 M arginine ( $\text{ArgHCl}$ ) at pH 4.3 and 2 M arginine at pH 4.7 were highly effective.

While biological affinity chromatography provides robust one-step purification, it does have a number of disadvantages. (i) Elution is difficult, as described above. (ii) The resin is expensive and tends to bleach, contaminating the eluted sample with the protein ligand (Protein A or G in the above case): Protein A/G, when present in



**Figure 9.10** Elution of antibody from a Protein A column by citrate at pH 4.3 or arginine at the indicated concentrations and pH values. Sodium citrate concentration ranged from 0.05 to 2 M and arginine was used at either 1 or 2 M. (Reformatted from [205].)

pharmaceutical products, causes serious side-effects due to its binding to circulating antibodies. (iii) The column cannot be regenerated with a strong solvent, including sodium hydroxide. Such washing would destroy the binding activity of the protein ligand. To overcome these shortcomings, so-called mixed-mode resins have been developed [186]. They are based on synthetic ligands and are designed to mimic the high-affinity binding afforded by biological ligands. Various organic solvents, and arginine, turned out to be effective for both washing and elution of the bound proteins in these mixed-mode resins, as was the case for biological affinity chromatography [208, 209].



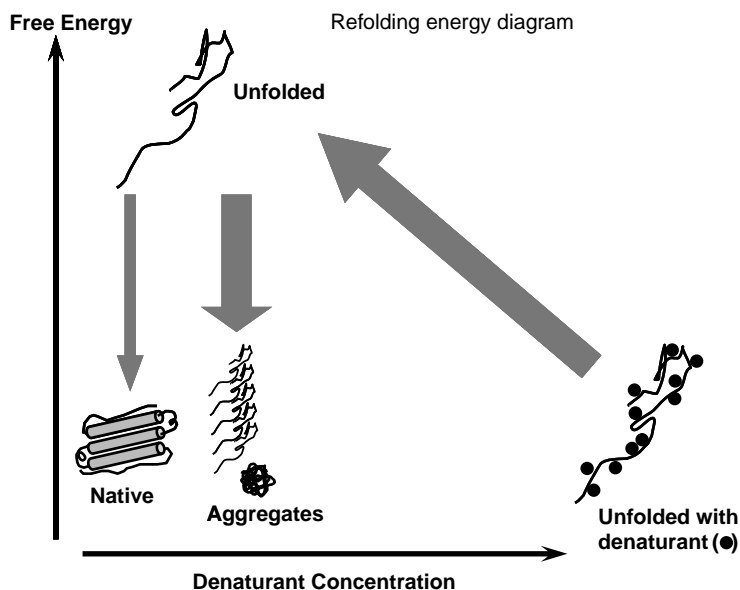
**Figure 9.11** Elution of antibody from a Protein A column by different amino acids at the indicated concentration and solution pH. The examined amino acids include lysine, proline, glycine, histidine, and arginine. (Reformatted from [205].)

## 9.2.4

**Protein Refolding**

Expression of recombinant proteins is a fundamental and critical technology, not only for basic science, but also for industrial applications. There are a number of heterologous recombinant expression systems. Among them, *E. coli* expression is the most convenient and frequently used. Heterologous expression of foreign genes in *E. coli*, however, often leads to the production of expressed proteins in an insoluble form called inclusion bodies. Extraction, solubilization, and refolding, when necessary, require the use of cosolvents. While most inclusion bodies consist of aggregates of largely unfolded proteins, certain inclusion bodies consist of native or native-like structures. In the latter case, solubilization of insoluble proteins by mild solvents may be able to generate an active protein, avoiding the time-consuming refolding procedure. Aqueous arginine solution at 0.5–2 M and neutral pH has been shown to effectively solubilize green fluorescent protein and  $\beta_2$ -microglobulin in the native state. Simple dialysis removal of arginine, without refolding, was sufficient to generate the final product [210, 211]. However, most inclusion bodies require refolding, which is not a straightforward process and often requires an extensive trial-and-error approach. Inclusion bodies are typically first solubilized using cosolvents such as urea and GdnHCl. These solvents denature proteins inherently, even when inclusion bodies were formed from native-like structures, which necessitate refolding.

Refolding is a process that leads to a change in protein conformation from the unfolded to the folded (native) state. At high denaturant concentrations, proteins are unfolded (disordered), well solvated, and flexible. As schematically depicted in Figure 9.12, the structure in the denaturant solution is in the low-energy state, due to the binding of denaturant molecules (e.g., urea and GdnHCl) (closed circles). As will be described in Section 9.5, binding of denaturant is the driving force for protein unfolding. In an aqueous buffer, proteins are folded, rigid, and compact. The low-energy state of the native protein is conferred by many intramolecular interactions that stabilize the protein structure (Figure 9.12). Ideally, the transfer of protein molecules from high denaturant concentration to aqueous buffer should lead to refolding (i.e., transfer of protein molecules from denaturant solution to aqueous solvent will force them to collapse into a compact structure). However, such a drastic process usually does not work, as it will lead to a misfolded and/or aggregated structure, which also happens to be in the low-energy state. Folding occurs through many intermediates, which are in the high-energy state, that are not stabilized by the bound denaturants, as shown in Figure 9.12 (shown as unfolded without bound denaturants). Such intermediates, if not sufficiently stabilized, tend to misfold and aggregate. Once misfolded or aggregated, protein molecules have no flexibility to disaggregate and refold into the native structure. Thus, two factors play a key role in successful refolding: methods to reduce the denaturant concentration and the assistance of refolding by cosolvents. There are a number of reports describing protocols to reduce denaturant concentration (e.g., dilution, dialysis, buffer exchange through gel filtration, and binding of denatured protein to solid support) [55, 212–216].



**Figure 9.12** Free energy diagram of various states of a protein during solubilization and refolding. Denaturant molecules are illustrated by filled circles. The unfolded structure is converted to either an amorphous, regularly structured aggregate or native structure.

The focus of this chapter is the use of small-molecule cosolvents to increase the recovery of active proteins and the efficiency of protein folding.

Cosolvents may be classified into two groups: folding enhancers or aggregation suppressors, as summarized in Table 9.1. Folding enhancers accelerate collapse of unfolded structure by salting-out or a preferential exclusion mechanism, and consist of a variety of cosolvents, including many osmolytes. As these folding enhancers also increase inter- as well as intramolecular associations, they can enhance aggregation, although they may prevent misfolding. Aggregation suppressors reduce the tendency for folding intermediates to aggregate and stabilize the intermediates, allowing sufficient time for them to proceed to the productive folding pathway. Weak binding of aggregation suppressors indicates that they can readily dissociate from the intermediates once they are stabilized by intramolecular interactions of the native structure. While there are protein-dependent aggregation suppressors, the most universal cosolvent appears to be arginine, which can weakly bind to the folding intermediates.

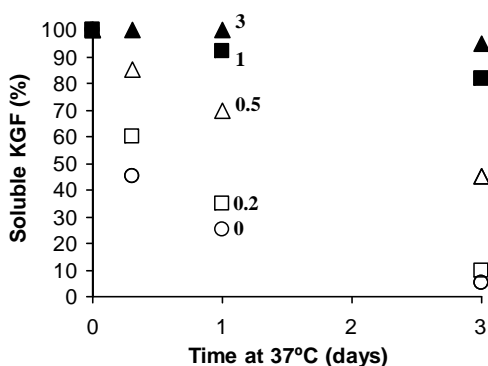
### 9.2.5

#### Formulation

Formulation of biological products, such as peptides, proteins, nucleic acids, and viruses, is designed to enhance their stability against chemical and physical

degradation during long-term storage. Chemical degradation can normally be controlled by solution pH or low concentration of different solvent components. While these conditions affect the physical stability of biological macromolecules, cosolvents at high concentrations also play a major role on their physical stability. Through an indirect mechanism, the cosolvents can affect the chemical stability as well. The most relevant case is protein unfolding, which can lead to enhanced chemical degradation, as conformational changes will alter the amount of solvent exposed, chemically labile groups. Thus, cosolvents that enhance the conformational stability should reduce chemical degradation. This section deals primarily with the cosolvent effects on physical stability of biological products. There are numerous examples examining the effects of solvent additives on the stability of biological products and some of these will be discussed here.

Fibroblast growth factors (FGFs) consist of many members, and have a common stability problem in that they are unstable and readily aggregate upon storage, in particular at elevated temperatures [217–219]. FGF members are also known as heparin-binding proteins, and thus heparin enhances their stability and thereby prevents aggregation [219]. Keratinocyte growth factor (KGF or FGF7) also has a short shelf-life during storage due to its unstable structure [220]. The amount of native monomer of KGF becomes about half in less than 2 days, if stored at 0.5 mg/ml at 37 °C [221]. Sodium citrate showed a concentration-dependent stabilization of KGF: 0.5 M citrate extended the half-life from 1.8 to 88 days during storage at 37 °C [221]. Other cosolvents were also tested for their effectiveness in stabilizing KGF [222]. Figure 9.13 plots the amount of soluble KGF with time after storing 0.5 mg/ml protein in 10 mM phosphate at pH 7.0, at the indicated concentration of NaCl at 37 °C. With this buffer system, the half-life is only 0.35 days without the salt (open circles). The half-life increased significantly with salt concentration, as almost no decay was observed in 3 M NaCl (closed triangles). Various salts and protein stabilizing osmolytes were examined for their effects on KGF stability.



**Figure 9.13** Plot of monomeric KGF after storage at 37 °C as a function of NaCl concentration, ranging from 0 to 3 M. (Reformatted from [221].)

**Table 9.3** Shelf-life and unfolding temperature of KGF in the presence of 0.5 and 1 M cosolvents.

Additive	0.5 M		1 M	
	Unfolding temperature (°C)	Half-life (days)	Unfolding temperature (°C)	Half-life (days)
None	41	0.35	41	0.35
Sucrose	+ 4	+ 0.46	+ 8	+ 4.37
Trehalose	+ 4	+ 0.25	+ 8	+ 3.1
Proline	+ 6	+ 0.01	+ 12	+ 0.04
Betaine	+ 4	+ 0.19	+ 5	+ 0.48
Sorbitol	+ 2	+ 0.08	+ 6	+ 0.66
Sodium citrate	+ 14	∞	+ 29	∞
Ammonium sulfate	+ 12	+ 48.5	+ 16	+ 120
Sodium phosphate	+ 14	+ 55.5	+ 20	∞

Data adapted from [332].

Table 9.3 shows the temperature at which 0.5 mg/ml KGF begins to unfold in a thermal unfolding measurement, and the half-life at which 50% of the native KGF is lost upon storage at 37 °C. It is evident that the salts are much more effective than neutral cosolvents in enhancing the stability of the protein. While proline is somewhat more stabilizing against thermal unfolding, it has little impact on the storage stability. It appears that while osmolytes were effective in increasing the melting temperature, they were not as effective as salts in enhancing the stability of KGF. It appears that electrostatic effects add to the stabilizing effects of the salts on the storage stability of KGF.

FGF20 has an extremely low solubility in aqueous solution [223]. Figure 9.14 plots the solubility of FGF20 as a function of pH between 5.0 and 8.5 – a pH range suitable for formulation. It is evident that the solubility in 50 mM phosphate buffer is below 0.5 mg/ml at any pH, and nearly zero between pH 6.0 and 6.5 (circles). Obviously, this makes protein formulation and purification very difficult. Fortunately, the addition of 0.2 M arginine in the same buffer was shown to greatly increase the solubility of the protein (squares). The solubility is above 1 mg/ml between pH 7.0 and 8.5 and sharply increases with decreasing pH, reaching 8 mg/ml at pH 5.5. Thus, it appears that the inclusion of arginine assisted in both the purification and formulation of the protein. The ability of arginine to suppress aggregation was also observed for monoclonal antibodies [224]. Table 9.4 shows the degradation of IgG1 upon exposure to intense UV light. It is clear that 0.2 M arginine protects the IgG1 from light-induced aggregation, which is the major degradation product. Consistent with other proteins and antibodies, arginine showed no stabilization effects against thermal denaturation: it actually resulted in a lower melting temperature. Thus, arginine reduced aggregation by its aggregation-suppressive effects and not by stabilization effects.

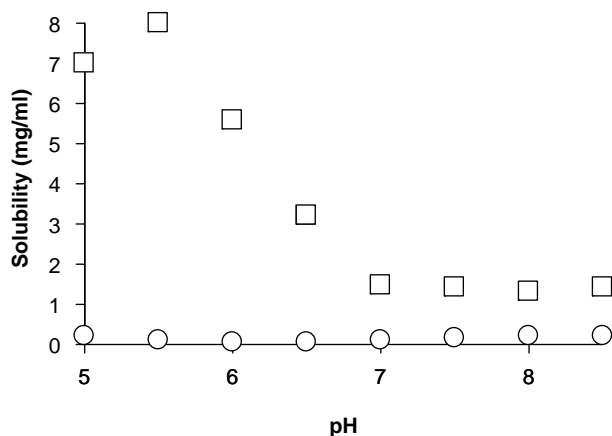
**Table 9.4** Protection of antibody from light-induced aggregation by arginine.

Degradation products	Storage in PBS		Storage in PBA	
	No light	Light	No light	Light
Aggregates (%)	1.3 ± 0.7	14 ± 9	0.9 ± 0.5	5 ± 0.2
Low-molecular-weight species (%)	2.9 ± 0.4	5.2 ± 0.6	1.8 ± 0.2	3.5 ± 0.3

The presented data are an average of the SEC data analyzed from four different IgG1 antibodies.

PBS: 10 mM sodium phosphate, 145 mM NaCl, pH 7.2.

PBA: 10 mM sodium phosphate, 200 mM arginine hydrochloride, pH 7. Data adapted from [224].



**Figure 9.14** Plot of FGF20 solubility as a function of pH with and without 0.2 M arginine: (○) 50 mM phosphate and (□) 0.2 M arginine. (Data reformatted from [223].)

### 9.3

#### Solvent Application for Viruses

The effects of solvents, particularly salts, on virus structure have been briefly described in an earlier section. Based on the effects of solvents on proteins and macromolecules, it is expected that solvents will affect the structure, stability, and aggregation propensity of viruses and their interactions with surface structures. In fact, cosolvents have been widely used to purify, process, and formulate viruses, as described here. Solvent application for viruses can be formally divided into several subcategories: the use of a cosolvent system for isolation and purification of viruses, the stabilization of viruses for biotherapeutic purposes, and the inactivation of viruses for anti-infective development. Each of these topics will be examined in detail below.



## 9.3.1

**Isolation and Purification of Viruses**

Isolation and purification of viruses are required for determining the severity of viral contamination and for developing viruses as therapeutic reagents (vaccines). To prevent the spread of virus infection, isolation, identification, and quantification of the virus species in the contaminated sources are required (e.g., water supply and solid surfaces with which humans come into contact). Isolation of viruses from contaminated sources often requires labor-intensive processes, including concentration and purification, due to their minute quantity or strong binding to solid surfaces [225–227]. Various salts and organic compounds at high concentrations have been employed to improve the recovery and concentration of virus preparations [228, 229]; however, the effects of cosolvents on viruses during processing are far more complex than those observed for simpler biological macromolecules, reflecting their complexity in structure and their tendency to aggregate or bind to cell debris and to other surfaces [228, 230]. For example, Downing *et al.* [228] have reported on the successful purification of respiratory syncytial virus and vesicular stomatitis virus using 1 M Na<sub>2</sub>SO<sub>4</sub> as eluent. However, the simple change of elution buffer counterion to Mg<sup>2+</sup> (i.e., 1.44 M MgSO<sub>4</sub>) rendered the purification ineffective. As both salts are salting-out for proteins, the quantitative difference between these salts suggests that their effects on viruses are much more complex than on proteins. To date, several other procedures have been developed to isolate, purify, and identify viruses [231–239].

In addition to understanding the physical properties of viruses to generate novel anti-infective agents, the development of an effective isolation procedure will aid in the production of purified, highly concentrated stock of viruses, which can serve as pharmaceutical vaccines, research reagents, drug targets, and delivery vectors. These reagents require a high degree of purity without compromised biological activity or structural integrity. Another issue to be considered during purification is the aggregated state of the virus particles, as purification efficiency is reduced significantly in the presence of aggregates [240, 241]. Several techniques have been developed to purify virus particles; however, the methods differ quite considerably both in terms of efficacy and time requirement. Although density gradient centrifugation has been the method of choice, the method suffers from low processing capacity [232, 236, 238]. Virus concentration has also been achieved through the use of charged membrane filters, followed by elution, employing media with appropriate buffer salt, pH, and ionic strength [226, 227, 231, 242]. Chromatography methods, including affinity and ion exchange, have also been developed to purify virus particles. However, the purification procedure is much more complex than that used for proteins, again reflecting the structural complexity of virus preparations [233, 234, 239]. Gao *et al.* [233] demonstrated the benefit of purifying Adeno-associated virus (AAV) vectors by employing a column chromatography method, in which both the purity and the potency of AAV were demonstrated to be higher than those prepared by CsCl gradient centrifugation. Magnesium salts and

sulfate salts in the concentration range of 0.24–1.44 M have both been used for elution from chromatographic columns [228, 237, 238]. It is interesting that a salting-in salt,  $\text{MgCl}_2$ , and a salting-out salt,  $\text{Na}_2\text{SO}_4$ , are often similarly effective in the elution of viruses from an affinity column [228] – an observation that does not occur on proteins [243], again indicating that virus behavior during purification or separation is qualitatively different from that encountered with proteins. Nevertheless, cosolvents, when used properly, facilitate purification of viruses and serve to reduce aggregation.

### 9.3.2

#### Stabilization and Formulation of Viruses

Widely variable effects of salts on virus stability were observed depending on the type and concentration of salt and on the physical properties of the virus. For certain viruses (e.g., herpes simplex virus (HSV) and infectious bronchitis virus (IBV)), strong salting-out and protein-stabilizing salts (e.g.,  $\text{Na}_2\text{HPO}_4$  and  $\text{Na}_2\text{SO}_4$ ) were found to enhance their stability against heat treatment [244]. For example, type I Inoue-Melnick virus (IMV) heated at 50 °C for 5 min in the presence of 1 M  $\text{Na}_2\text{SO}_4$  resulted in 0.2  $\log_{10}$  decrease, whereas in distilled water, the titer decreased by 1.7  $\log_{10}$  [245]. In the presence of other salts, specifically 1 M  $\text{MgCl}_2$  and 1 M  $\text{MgSO}_4$ , the virus decreased in titer by more than 4  $\log_{10}$  (i.e., salt addition inactivated the virus further than that observed in distilled water). Similarly to  $\text{Na}_2\text{SO}_4$ , HSV was stabilized by 1 M  $\text{Na}_2\text{SO}_4$  and demonstrated a small decrease in titer (0.1  $\log_{10}$ ) following incubation at 50 °C for 30 min, whereas in distilled water, it decreased in titer by 2  $\log_{10}$  [246]. As these salts also stabilize proteins, the mechanism of their effects on virus is expected to be similar to that operating on proteins. Weaker stabilizing salts, such as NaCl and KCl, demonstrated destabilizing effects on type I IMV and HSV. Even for phosphate salts, differences in their ability to stabilize viruses were observed depending on their counterion and their ionization state (e.g., while  $\text{Na}_2\text{HPO}_4$  enhanced the stability of certain viruses,  $\text{KH}_2\text{PO}_4$  destabilized them). The former salt has a much higher ionic strength than the latter and hence greater effects on water structure; the importance of salt ions on water structure is given later in Section 9.5.

A summary of literature data demonstrating the effects of several salts on a wide variety of viruses is given in Table 9.5. What stands out in this table is the variability of the two magnesium salts,  $\text{MgCl}_2$  and  $\text{MgSO}_4$ , in stabilizing viruses; the virus was stabilized by one salt and destabilized by the other. In the case of type 2 poliovirus, 1 M  $\text{MgCl}_2$  stabilized the virus against heat treatment, while  $\text{MgSO}_4$  at the same concentration destabilized the virus [247]. Measles virus, on the other hand, demonstrated the opposite trend, in which 1 M  $\text{MgSO}_4$  stabilized the virus to heat treatment, while 1 M  $\text{MgCl}_2$  further destabilized the virus [248]. Comparing the effects of the two magnesium salts on the physical properties of several viruses, it appears that, typically,  $\text{MgCl}_2$  increases the thermal stability of nonenveloped viruses while it destabilizes enveloped viruses (Table 9.5).  $\text{MgSO}_4$ , on the other hand,

**Table 9.5** Stabilization and destabilization of various viruses by salts.

<b>Virus</b>	<b>Stress</b>	<b>Stabilizing salt</b>	<b>Destabilizing salt</b>	<b>Remarks</b>	<b>References</b>	
<b>Nonenveloped viruses</b>						
Enterovirus (poliovirus, coxsackievirus, echovirus)	50°C, 1 h	1 M MgCl <sub>2</sub>		NaCl was not always stabilizing	[281]	
Type 1 poliovirus	50°C, 15 min	1 M CaCl <sub>2</sub>		effect greater than NaCl	[247]	
		2 M NaCl				
Type 2 and 3 poliovirus	50°C, 15 min	1 M MgCl <sub>2</sub>	1 M MgSO <sub>4</sub>		[247]	
		1 M MgSO <sub>4</sub>				
Type 1 and 3 poliovirus	50°C with formalin	2 M NaCl		antigenic activity preserved in viruses inactivated with MgCl <sub>2</sub>	[251]	
		1 M Na <sub>2</sub> SO <sub>4</sub>				
Infectious bronchitis virus	50°C, 15 min	1 M Na <sub>2</sub> SO <sub>4</sub>	1 M MgCl <sub>2</sub>	similar observation for poliovirus, echovirus and Japanese B encephalitis virus	[244]	
		1 M MgSO <sub>4</sub>	1 M KCl			[252]
		1 M Na <sub>2</sub> HPO <sub>4</sub>	1 M NaCl			[253]
Measles virus (Edmonston virulent strain)	50°C, 15 min	1 M K <sub>2</sub> SO <sub>4</sub>	1 M CaCl <sub>2</sub>	1 M KCl and 1 M K <sub>2</sub> SO <sub>4</sub> moderately stabilizing	[248]	
		1 M MgSO <sub>4</sub>	1 M MgCl <sub>2</sub>			
		1 M Na <sub>2</sub> SO <sub>4</sub>	1 M CaCl <sub>2</sub> 1 M Na <sub>2</sub> HPO <sub>4</sub>			

(Continued)

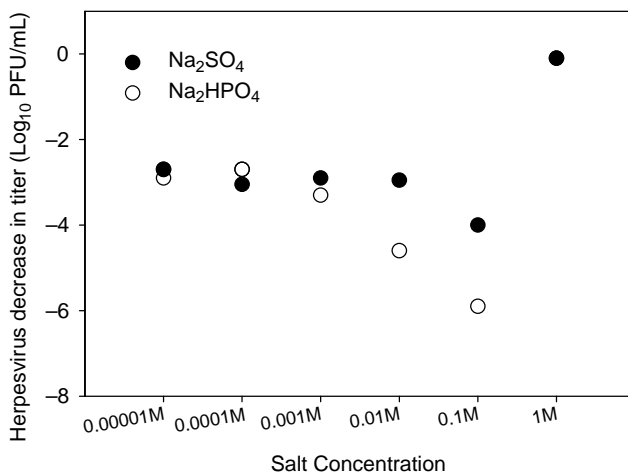
Table 9.5 (Continued)

Virus	Stress	Stabilizing salt	Destabilizing salt	Remarks	References
Influenza, parainfluenza, vesicular stomatitis, vaccinia, rubella	50 °C, 15 min	1 M MgSO <sub>4</sub>	1 M MgCl <sub>2</sub>		[247]
HSV	50 °C	1 M Na <sub>2</sub> HPO <sub>4</sub> 1 M Na <sub>2</sub> SO <sub>4</sub>	1 M MgCl <sub>2</sub> 2 M NaCl 1 M KH <sub>2</sub> PO <sub>4</sub> 2 M KCl		[246]
Type I IMV	50 °C, 5 min	1 M Na <sub>2</sub> SO <sub>4</sub>	1 M MgCl <sub>2</sub> 1 M MgSO <sub>4</sub> 2 M NaCl 2 M KCl		[245]

demonstrates no apparent trend, as it has been demonstrated to stabilize and destabilize, respectively, measles virus and type I IMV, both of which are enveloped viruses.  $\text{MgCl}_2$  is generally classified as a salting-in salt and  $\text{MgSO}_4$  as a salting-out salt [249, 250]. Their virus-dependent effects may be due to the difference in the mechanism by which heat treatment inactivates viruses; in other words, multiple mechanisms may exist in heat-induced virus inactivation. For example,  $\text{MgCl}_2$  may prevent virus aggregation, while  $\text{MgSO}_4$  may enhance the structural stability of viral proteins or viral particles, resembling their effects on proteins. Thus, the effects of the two salts may depend on how the viruses are inactivated, as well as on the physical characteristics of the viruses themselves. In this regard, it is interesting to note that  $\text{MgCl}_2$  not only stabilized poliovirus against formalin-induced inactivation at  $50^\circ\text{C}$ , but also helped maintain its antigenicity [251]. This may be due to the salting-in effect of  $\text{MgCl}_2$ , preventing viral aggregation, although no such data appear to be available. Overall, the observed effects of magnesium salts on virus stability clearly demonstrate that their effects cannot be explained solely from their effects on viral proteins alone; rather, these effects may also involve lipids, and nucleic acids.

It is also interesting to note the concentration dependence of the salts on their ability to stabilize the virus at increasing temperatures; a small amount of salt (i.e., 60 mM  $\text{MgSO}_4$  or  $\text{Na}_2\text{HPO}_4$ ) was sufficient to stabilize the virus at  $37^\circ\text{C}$ , but higher concentrations (i.e., above 0.5 M) were typically required to achieve similar stability at  $50^\circ\text{C}$  [244]. The salts were very effective in conferring thermostability to the virus such that the inactivation rate of IBV-42 at  $60^\circ\text{C}$  in 1 M  $\text{MgSO}_4$  was approximately the same as that observed in water at  $25^\circ\text{C}$ . Hopkins [244] suggested that the anion is the active component in the stabilization of IBV and that the activity is related to the valence of the anion. Similar observations have been reported by Wallis *et al.* [252] in their work on poliovirus and echoviruses, and by Nakamura and Ueno [253] on Japanese B encephalitis virus.

All of the experiments represented in Table 9.5 were carried out at high salt concentrations, in which salt-specific effects dominate. Careful examination of the salt effects (both at low and high concentrations) demonstrates a complex salt-dependent virus stabilization mechanism against heat treatment. Wallis and Melnick [246] reported that herpesvirus in 1 M  $\text{Na}_2\text{HPO}_4$  decreased in titer by  $0.5 \log_{10}$  following exposure to  $50^\circ\text{C}$  for 10 min, and more than  $1.7 \log_{10}$  decrease in titer after 60 min. In distilled water, the virus decreased in titer by  $2 \log_{10}$  following 15 min of heating and was completely inactivated after 60 min. The thermal stability of herpesvirus was improved further upon replacing  $\text{Na}_2\text{HPO}_4$  with  $\text{Na}_2\text{SO}_4$ . Although sufficient stabilization was observed at 1 M concentration of either salt, at lower concentrations, the inactivation rate of the virus became greater than that observed in distilled water (Figure 9.15). Interestingly, the highest amount of inactivation was observed at 0.1 M concentration, and the stability improved with further dilution of each salt solution. At the low concentrations examined (0.1 M and below), the primary effect of salts on the virus particles is ionic strength modification. Thus, it appears that these viruses become unstable in the presence of salts at low concentrations, in which the primary effect is electrostatic in nature. The electrostatic stabilization of virus



**Figure 9.15** Decrease in herpesvirus titer following 15 min of incubation at 50 °C in the presence of various amounts of (●) Na<sub>2</sub>SO<sub>4</sub> and (○) Na<sub>2</sub>HPO<sub>4</sub> salts. The concentrations ranged from 10<sup>-5</sup> to 1 M. (Data adapted from [246].)

structures may be compromised in the presence of small amounts of salt ions. On the other hand, these viruses were found to be much more stable in 1 M solution than in water, indicating that they exert salt-specific stabilizing effects, consistent with their effects on other viruses (Figure 9.15). Thus, not only the salt type, but also its concentration plays an important role in determining the stability of viruses.

Viruses are also destabilized by freezing and drying. Carpenter and Crowe [254–256] have shown that cosolvents that enhance the stability of proteins in solution also enhance their stability during freezing and thawing. Savithri *et al.* [257] reported a stabilizing effect of Tris–HCl on the structural integrity of belladonna mottle virus (BDMV) following freezing and thawing. Although the highest concentration tested was 100 mM (which demonstrated 80% protection of capsid structure), it should be noted that the Tris–HCl salt is concentrated during freezing; as water crystallizes, a freeze-concentrate of salt and virus is generated. Such high Tris–HCl concentration may exert its effect through specific interactions, as observed for certain buffers in solution, at a much higher concentration [107, 258, 259]. These authors also demonstrated the effects of other salts, including MgCl<sub>2</sub> at 40 mM, which resulted in 85% protection, demonstrating consistency with its effects on certain viruses observed against heat treatment. Although the examined concentration was low (40 mM), this simply represents an initial concentration prior to freezing and the effective concentration is much higher during freezing. Similar observations can be made about the formulation compositions utilized for lyophilized viruses (Table 9.6).

Lyophilization is a technique commonly utilized to improve the storage stability of a variety of biopharmaceutical products, including proteins, antibodies, and vaccines. The labile biopharmaceuticals are stabilized by restricting molecular mobility, thereby inhibiting degradation, whether physical or chemical in nature, and by

**Table 9.6** Formulation composition for freeze-dried viruses.

Virus	Stabilizers	Remark	References
Foot-and-mouth disease virus	skim milk, horse serum, horse meat bouillon sucrose, glutamate, gelatin, dextrin, and their mixtures	4.5 years at 4 and 23 °C 1 year at 4 °C; less effective than skim milk	[260, 261] [262]
Pseudorabies virus	SPG		[263]
Rickettsiae	SPG		[264]
Marek's disease virus	SPG		[265]
Varicella zoster virus	SPG		[264]
Yellow fever virus	SPG		[266]
Respiratory syncytial virus	SPGA (SPG + albumin): prefreezing concentration, 218 mM sucrose, 7.1 mM K <sub>2</sub> HPO <sub>4</sub> , 3.76 mM KH <sub>2</sub> PO <sub>4</sub> , 4.9 mM glutamate, 1% BSA	stable over 2 years at 25 °C	[267]
Alfalfa mosaic virus	1–7.5% sucrose  1% inositol 0.5% peptone 0.5% lysine	no stabilizing effect of 0.5% glutamate or 0.5% cysteine	[425]
Infectious pancreatic necrosis virus	10% lactalbumin hydrolysate 10% lactose	4 years at 4 °C	[269]
Live attenuated rinderpest virus	5% lactalbumin hydrolysate/10% sucrose	99.49 years at 4 °C >30 days at 37 °C	[270] [271]
Smallpox vaccine	5% peptone/5% sodium glutamate	potential side-reaction	[272]
BCG vaccine	sodium glutamate		[274, 275]
Calf dermal pulp vaccine	1% sodium glutamate	5% glutamate less effective	[272]
Vaccinia virus	sodium glutamate peptone	moisture content important	[426]
Inactivated influenza virus	1 : 500 ratio of virus to trehalose		[276]

structurally stabilizing the biomolecule through the use of excipients, including cosolvents. Vaccines, which may consist of a live attenuated virus, have been successfully freeze-dried and there are several lyophilized vaccines currently available. In fact, with the exception of the oral polio vaccine, all live viral vaccines and whole-cell bacterial vaccines are prepared in a lyophilized format and reconstituted prior to administration. The lack of liquid live virus and whole-cell bacterial vaccines is due to difficulties in maintaining adequate therapeutic titers during storage.

Initial efforts in stabilizing lyophilized viruses included complex formulation components such as skim milk, horse serum, and horse meat bouillon, all of which were used for stabilizing the foot-and-mouth disease (FMD) virus [260, 261]. Although the virus was successfully stabilized (complete retention of infectivity following 4.5 years of storage at both 4 and 23 °C), a simpler stabilizer or a mixture of stabilizers was desired. Fellows [262] lyophilized the FMD virus using a variety of stabilizers, including sucrose, glutamate, gelatin, dextrin, and their mixtures, and achieved moderate stability following 1 year of storage at 4 °C. However, none was as successful as skim milk in conferring storage stability. Scott and Woodside [263] successfully stabilized pseudorabies virus by lyophilizing it in a medium containing sucrose, potassium phosphate, and glutamate (SPG). This combination was first used by Bovarnick *et al.* [264] to increase the stability of rickettsiae and later by Calneck *et al.* [265] for freeze-drying Marek's disease virus. The authors suggested that glutamate was successful in conferring stability to the virus due to its ability to react preferentially with carbonyl groups in the medium, thus protecting viral proteins, which are sensitive to denaturation by carbonyl groups in the lyophilized state. Sucrose has been shown to be essential for preserving the virion structure and infectivity of varicella zoster virus, and has been shown to be necessary for the preservation of infectivity of yellow fever virus and rickettsiae [264, 266]. The stabilizing capability of the SPG formulation was further improved upon the incorporation of 1% BSA. The loss of virus infectivity due to inactivation at the gas/liquid interface (as well as liquid/enclosure interface) could be prevented by saturating these interfaces with another protein (i.e., BSA), effectively preventing virus access to the surface [267, 268]. Some of the other excipients screened and determined to be effective included lactalbumin hydrolysate, lactose, peptone, sodium glutamate, and trehalose (Table 9.6). Lactalbumin hydrolysate (10%) and lactose (10%) were found to stabilize infectious pancreatic necrosis virus following lyophilization, demonstrating negligible loss after 4 years of storage at 4 °C, similar to the stabilization offered by skim milk [269]. Lactalbumin hydrolysate (5%) was also reported to be effective in stabilizing lyophilized live attenuated rinderpest virus in the presence of 10% sucrose [270, 271]. The predicted half-life of the lyophilized rinderpest virus was 99.5 years and more than 30 days at 4 and 37 °C, respectively. For an unpurified smallpox vaccine, the best storage stability at 45 °C was obtained with 5% peptone and 5% sodium glutamate [272]. Peptone has been reported to act protectively in freeze-drying and preservation of smallpox vaccine [273], and sodium glutamate was shown to exert a powerful protective influence on the viability of dried BCG vaccine [274, 275]. Trehalose, another disaccharide similar to sucrose, has been shown to be an effective stabilizer of live



viruses upon lyophilization. Huang *et al.* [276] successfully lyophilized a whole inactivated influenza virus in trehalose at a ratio of 1: 500 from sterile saline.

The popular formulation composition for a wide variety of viruses, including pseudorabies, rickettsiae, and varicella zoster is SPG or SPGA (SPG + albumin) ([263–267], Table 9.6). The exact composition varied, but typically, the formulation contained 218 mM sucrose, 7.1 mM  $K_2HPO_4$ , 3.76 mM  $KH_2PO_4$ , 4.9 mM sodium glutamate, and 1% (w/v) BSA. Unlike the concentrations of cosolvents required for stabilization in solution (1 M and above, Table 9.5), the concentration of the formulation is quite low (below 0.25 M). During lyophilization, however, the concentration of the formulation components increases considerably. Thus, although the initial solution concentration of sucrose may be low, the amount necessary for effective stabilization may be reached during the drying process. Glass-forming excipients, such as sucrose, can also confer stability to viruses through the formation of an amorphous matrix, in which the molecular mobility, and thus degradation rate, is considerably retarded. This will be further expanded upon in Section 9.5.

### 9.3.3

#### Inactivation of Viruses

Removal of contaminant products, such as viruses, from cell culture media containing recombinant proteins represents an integral step in the production and purification of therapeutic proteins. Typical processing conditions employ high-temperature incubation and low pH treatment [277, 278], but these conditions can also be detrimental to the stability of freshly prepared proteins [204, 279, 280]. Arginine, employed at high concentrations (i.e., around 1 M), has been demonstrated to lower the temperature, as well as raise the pH, at which virus inactivation takes place, thus reducing the risk of degrading the protein being purified concurrently [58, 60]. HSV-1 in the presence of 1.2 M arginine has been reported to be inactivated at approximately 40 °C, while in its absence, complete inactivation was accomplished only at higher temperatures (above 50 °C) [60]. Significant reduction in viral titer (i.e., above 5.7  $\log_{10}$  reduction) has been obtained in the presence of 1 M arginine at pH 4.3, while in its absence, similar reduction was obtained only under more acidic conditions (pH 3.5) [58]. Thus, through the addition of arginine at a relatively high concentration, the virus inactivation procedure can be conducted at a milder processing condition (i.e., less acidic pH and lower temperature).

The addition of salts has also been reported to be effective in inactivating viruses, as described above in Table 9.5. The JES strain of herpesvirus was inactivated by heat treatment in distilled water (15 min at 50 °C), resulting in 2  $\log_{10}$  decrease in titer. Inactivation was further enhanced to more than 6  $\log_{10}$  decrease in titer upon the addition of 1 M  $MgCl_2$  [248]. On the other hand, the addition of 1 M  $Na_2SO_4$  improved the stability of viruses to heat inactivation (0.1  $\log_{10}$  decrease in titer), thus the choice of salt must be made carefully. In general, enveloped viruses are inactivated by the addition of molar quantities of  $MgCl_2$  (salting-in salt), while the same salt can stabilize a wide variety of nonenveloped viruses, such as poliovirus and coxsackievirus [247,

281] (Table 9.5). This suggests a difference in the interaction of the salt with the viral membrane components (i.e., lipids, proteins, carbohydrates, etc.) and with proteins alone, as is the case for nonenveloped viruses. Furthermore, enveloped viruses that are susceptible to  $\text{MgCl}_2$  are also susceptible to inactivation by KCl and NaCl, but at twice the effective concentration of  $\text{MgCl}_2$  [245, 248], suggesting the importance of salt solution osmolality (or ionic strength) on the stability of these viruses.

Common cleaning reagents used to disinfect surfaces typically contain alcohol. They are the product of years of extensive research conducted in examining the efficacy of alcohol in inactivating a wide variety of viruses. Moorer [282] reported a short inactivation time (20 s) required for disinfecting enveloped viruses, including HIV, bovine viral diarrhea virus, and pseudorabies virus, all with very high clearance factors of at least 6.5 (clearance factor =  $\log_{10}$  (total amount of virus added/total amount recovered from the treated sample)) with 80% ethanol containing 5% isopropanol. Similarly, 80% isopropyl alcohol was demonstrated to be effective in reducing the titer of both type 1 and type 2 HSV to less than  $10^1$  from  $10^6$  and  $10^{5.5}$ , respectively, upon mixing equal volumes of alcohol with the virus [283]. The efficacy of alcohol as a disinfectant is highly dependent on the physical properties of the virus, as has been described above for salts. For example, 70% ethanol was demonstrated to be efficacious in inactivating a wide variety of enveloped viruses, including Sindbis, HSV-1, and vaccinia, but was less effective for poliovirus type 1 – a nonenveloped virus [284]. Ethanol was also demonstrated to be weakly effective against bacteriophage MS2 and feline calicivirus – both nonenveloped viruses [285–287]. There are exceptions, however, as 70% ethanol was demonstrated to be effective against human rotavirus and adenovirus type 5 [288, 289]. These results suggest the potent, destabilizing effects of alcohols on lipid membranes. Although the effects on protein structure may also occur, alcohol-induced denaturation is generally reversible and may be incapable of irreversibly inactivating viruses simply through its interaction with the viral proteins [290]. Another possibility also exists that the addition of ethanol causes virus aggregation, whether enveloped or nonenveloped, due to the charges on the virus surface, as will be discussed below for DNA and proteins. Aggregation of viruses, if irreversible, should lower the effective concentration.

## 9.4

### Solvent Application for DNA

#### 9.4.1

##### Isolation and Purification of DNA

Similarly to viruses, which can be used as a vehicle for targeted therapy, plasmids, which have been studied for their use in gene delivery, are also frequently purified by chromatography with the aid of cosolvents. A typical process of plasmid production begins with the culture of transformed *E. coli* cells followed by alkaline lysis of the harvested bacteria. Following selective precipitation and concentration of the genetic material of the cell lysate, plasmids can be purified by several chromatography

methods, including ion-exchange [291, 292], size-exclusion [293, 294], affinity [295, 296], and hydrophobic chromatography [297, 298].

IEC, the most widely used and economical method for plasmid purification, separates DNA from RNA and proteins using cosolvents to modulate the binding affinity (i.e., charge–charge interaction) of the various components with the resins. Tseng and Ho [299] demonstrated that an elution buffer containing higher than 45% isopropanol was effective in separating plasmids from RNA and proteins, while at lower concentrations, separation through anion exchange Q-Sepharose was ineffective. The authors suggested that the role of the cosolvent, isopropanol, was to modify the dielectric constant of the system, which had a profound impact on the binding strength between both RNA and plasmid DNA to the resins, but to varying degrees: the addition of isopropanol to the column buffer decreased the dielectric constant and thus enhanced the overall electrostatic interactions, whether attractive or repulsive. The interaction strength between RNA and the resins increased less than that between DNA and the resins, resulting in efficient separation of the two components.

Another alcohol, ethanol, at high concentrations (i.e., above 80%) has been reported to lead to aggregation of DNA [300–302], thus care must be taken when using the alcohol cosolvent system for DNA purification: whether such aggregation is due to the repulsion between the surface charges of DNA and alcohol, as is observed for proteins, is an intriguing question. As will be described later in Section 9.5, unfavorable interaction between alcohol and polar compounds or salts leads to phase separation of the solute molecules. It has been postulated that the aliphatic chains of alcohols in solution may function like the hydrophobic ligands on the resins to create hydrophobic interactions with DNA [299]. The different hydrophobicities, resulting from the aliphatic chains of alcohols, may alter the conformations of plasmid and RNA in varying ways to affect the separation process. The base pairs of DNA are buried within the double helix, whereas the bases of RNA are exposed to the solution. Thus, the aliphatic chains of alcohols might interact with the exposed bases of RNA to alter its conformation, whereas the minimal interactions with the unexposed bases of DNA may be incapable of altering the DNA conformation. This might lead to differential interactions of alcohols with DNA and RNA; unfavorable, repulsive interactions between alcohol and DNA can strengthen the binding to the resin, while favorable hydrophobic interaction of alcohol with RNA may enhance its dissociation. Furthermore, the presence of certain metal salts has been reported to modulate the ethanol content at which aggregation occurs [303], suggesting the strong interplay between the compositions present in the cosolvent system on purification efficiency. This cooperative effect may be explained by the increased effective concentration of metal salts in the vicinity of nucleic acids, upon the addition of ethanol (as a result of mutual repulsion between metal salts and alcohol), which may increase the binding constant of the metal ions to nucleic acids.

Another solvent that demonstrates efficacy in the purification of plasmid DNA is PEG. Humphreys *et al.* [304] demonstrated that 10% PEG6000 can precipitate DNA present in lysates of *E. coli* carrying plasmids of a wide range of molecular weight (i.e.,  $6\text{--}123 \times 10^6$ ). DNA precipitation was rapid (90% completion within 2 h at 4 °C), and in comparison to other purification procedures, PEG precipitation was fairly gentle

and caused no changes to the biological activity of the purified plasmids; *E. coli* K12 strain JC7623 was transformed at similar frequencies by plasmid DNA isolated with or without a PEG precipitation step. The PEG precipitation method can also be used to separate the DNA fragments based on the size of the base pairs by modifying the amount of PEG present as a cosolvent. Lis [305] reported that the minimally required concentration of PEG, in general, is less for larger and/or more anisometric structures. Addition of 5% PEG caused precipitation of DNA fragments of *Drosophila melanogaster* larger than 1650 bp, and smaller fractions were collected by increasing the PEG concentration further. In another application, Sauer *et al.* [306] used 10% PEG8000 in the presence of 250 mM NaCl to selectively precipitate plasmid DNA (5369 bp) maintained in *E. coli* DH5 $\alpha$ , which was previously processed by selective precipitation of high-molecular-weight RNA from the cleared lysate with 1.4 M CaCl<sub>2</sub>. The main advantage of the PEG precipitation method is that it does not require the use of organic solvents nor does it require the use of RNase or spermidine, which can bind RNA and DNA indiscriminately. The reversible effects of PEG are expected from its exclusion mechanism (i.e., PEG does not bind).

Acid-phenol extraction system is another example of a cosolvent system that can be used to purify closed circular DNA. Phenol (present at 50%) has been demonstrated to selectively extract DNA species, other than the covalently closed DNA, at acidic pH and low ionic strength. Zasloff *et al.* [307] have reported on a number of systems in which more than 95% of the nicked circular species and linear DNA molecules (greater than about 1500 bp) are cleared from the water phase (containing 50 mM sodium acetate, pH 4.0). After three extractions, more than 99% of the contaminants have been removed, while closed circular DNA remained in the water phase.

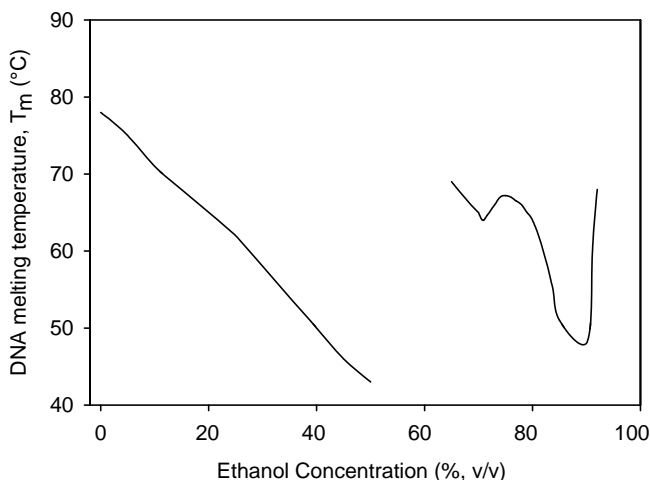
#### 9.4.2

##### Stability of DNA in a Cosolvent System

The presence of cosolvents, such as ethanol, and certain metal salts can have a profound impact on the structural state of DNA and may impact the purification procedure significantly. The presence of high ethanol content has been demonstrated to result in the aggregation of DNA and the structural stability of DNA, as manifested by the B  $\rightarrow$  A transition temperature ( $T_m$ ), is influenced considerably by the presence of metal salts.

Melting of DNA is affected by the addition of organic solvents such as alcohol. An often used explanation is that the DNA, a polyelectrolyte, exerts a considerable influence on the surrounding solvent, and ethanol disrupts the water structure and the structural stability of DNA [308]. This is manifested by a monotonic decrease in the  $T_m$  of DNA upon the addition of ethanol, up to a certain point (around 50% Figure 9.16) [309, 310].

Alternatively, however, such effects of alcohol on  $T_m$  can be readily explained from the balance of conflicting interactions of alcohol with DNA; favorable interaction with the bases and the unfavorable interaction with the charges. The interaction between alcohol and the native DNA (possessing high charge density) is highly unfavorable: this explanation was used above for alcohol-induced DNA aggregation (i.e., aggregation



**Figure 9.16** Effect of ethanol concentration on the melting temperature ( $T_m$ ) of Na-DNA. (Data adapted from [310].)

can reduce such unfavorable interactions). An alternative way to reduce the unfavorable solvent interaction is to unfold the DNA, since unfolding and expansion result in reduction of charge density. Furthermore, unfolding will expose the bases that would favorably interact with alcohol. The combined effects are to stabilize the unfolded DNA and to decrease the  $T_m$ . At higher ethanol concentrations, DNA aggregates (as observed by the increase in  $T_m$  at approximately 65% ethanol) and eventually precipitates, due to the enhanced mutual repulsion at high alcohol concentration.

Another important factor influencing the melting properties of DNA in aqueous solutions is the nature of the counterion and its concentration. The main role of the cation is to neutralize the electrostatic repulsion among negatively charged phosphate groups in DNA. Stabilization of DNA to thermal denaturation has been reported to decrease in the order:  $Mg^{2+} > Li^+ > Cs^+, Na^+, K^+$  [311]. It appears that in an aqueous solution, there is a direct relationship between the Stokes radius of the hydrated counterion and the  $T_m$  of DNA [312]. In other words, counterions with high degrees of hydration stabilize DNA better than the less hydrated ones. Eagland [308] suggested that the structural stabilization occurred by partial dehydration of DNA.

As mentioned above, as the ethanol concentration in the cosolvent system exceeds 50%, aggregation and precipitation of DNA occurs. This induction of DNA aggregation by ethanol, as well as the stability of aggregated DNA, is greatly influenced by the nature and concentration of the counterion present in the system. The ions have been demonstrated to stabilize aggregated DNA in the following order:  $Mg^{2+} < Li^+ < K^+, Cs^+, Na^+$ . The trend suggests that the thermal stability of DNA, in an ethanol cosolvent system, varies inversely with the size of the solvated counterion, which is the opposite of what was observed in aqueous solutions. At even higher ethanol concentrations, the trend was reversed. The ability to form a thermally stable

DNA structure at high ethanol concentrations decreased in the order:  $Mg^{2+} > Li^+ > Na^+ > K^+ > Cs^+$  [303, 313]. Thus, it is important to not only be mindful of the conditions that allow for efficient purification of DNA or plasmids, but the effect of the cosolvent system (i.e., ethanol concentration, cation type and concentration) on the structural stability of the macromolecules being isolated.

## 9.5

### Mechanism

The first account on record to explain the effects of cosolvents on the solubility of proteins, to our knowledge, is that of “attraction pressure,” proposed by Traube in 1910 [68]. The “attraction pressure” concept clearly explained the pioneering observation of salting-out effects of Hofmeister [69]. However, it is now evident that there are many exceptions to this concept, and the mechanism of the effects of cosolvent on protein solubility and stability is still under extensive investigation. Explanations of the observed effects may be arbitrarily divided into two categories; the physical mechanism of cosolvent–macromolecule interactions and thermodynamic interactions. First, the physical mechanism will be discussed below.

All of the molecular interactions between solvent (water), macromolecule, and cosolvent are determined by four fundamental forces that play a major role at high concentrations of cosolvents, as summarized in Figure 9.17. The four forces include electrostatic interactions, van der Waals interactions, hydrogen bonding, and the excluded volume effect. It would be difficult to describe the effects of cosolvents based on these forces alone, and thus a more direct observation will be used. Figure 9.17 illustrates that the first three fundamental forces lead to the hydration of macromolecules and cosolvents, as well as the affinity of cosolvents for macromolecules (i.e., cosolvent binding), which in turn can be used to explain the cosolvent effects. These hydration and excluded volume mechanisms determine how the cosolvents interact with biological molecules, which ultimately determine the effects of cosolvents on the properties of macromolecules. The thermodynamic interactions between cosolvents and macromolecules reflect many different types of interactions present on the surface of macromolecules, and hence can be derived from the

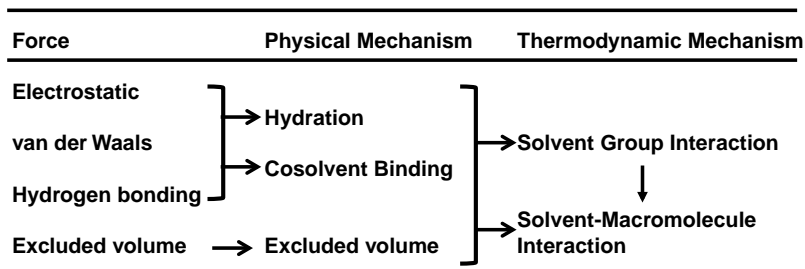


Figure 9.17 Forces and mechanism of cosolvent or water interaction with macromolecules.

analysis of cosolvent interactions with specific groups or low-molecular-weight model solutes with simple chemical structures. These physical and thermodynamic mechanisms are described below in detail.

### 9.5.1

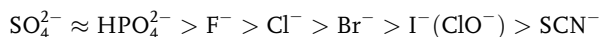
#### Physical Mechanism

##### 9.5.1.1 Hydration

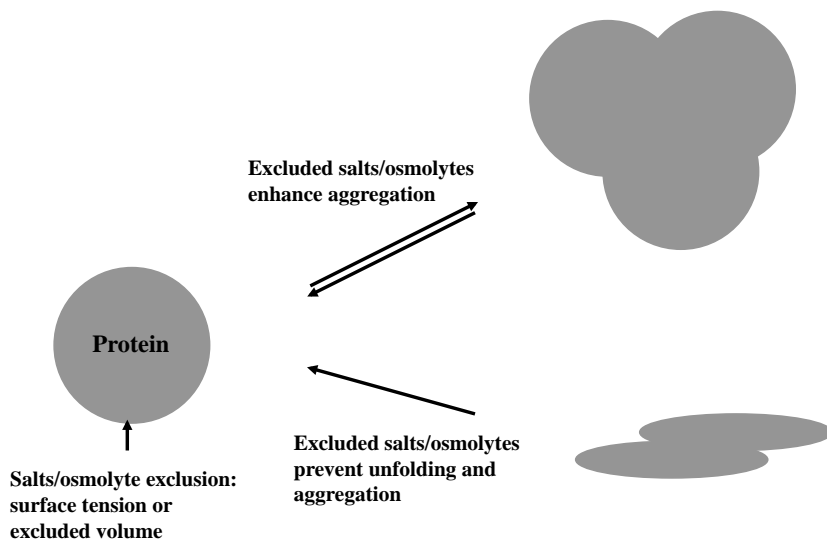
Traube [68] demonstrated that the effects of salts on the aqueous solubility of gasses have an identical order to their effects on protein solubility, as discovered previously by Hofmeister [69]. He then correlated the order of the salts to their effects on the surface tension of water. Those salts that raised the surface tension more effectively resulted in lower solubility of proteins and gasses. Sinanoglu and Abdunur [70, 71] then expanded the surface tension theory to the cavity theory to explain the stability of DNA in the presence of solvent additives. They discovered that DNA double helix was more stable in the presence of cosolvents that raised the surface tension of water. Conversely, the cosolvents that decreased the surface tension of water destabilized the double helix. The cavity theory was then applied to explain protein–protein interactions (i.e., protein stability) and protein–ligand interactions (i.e., HIC), by Melander and Horvath [72].

The surface tension theory, or cavity theory, is based on the increased surface free energy of solute molecules by the addition of salts (i.e., increase in the air-water interfacial free energy). Assuming that the protein surface represents the air-water interface, the same salts should increase the interfacial energy at the protein surface. As the interfacial free energy is proportional to the solvent accessible surface area (Figure 9.18), self-association of proteins should decrease the surface area per protein, and thus be enhanced by salts that raise the surface tension (i.e., lowering protein solubility). Protein binding to the HIC column should also reduce the surface area, and hence the same salt should enhance binding. As proposed by Traube, the surface tension increase by salts occurs through their hydration potential.

**9.5.1.1.1 Ion Hydration** Ion hydration can be determined from the elution position of salts on Sephadex G-10 gel-filtration chromatography [314–316]. Sephadex G-10 is composed of dextran that is highly cross-linked with epichlorohydrin, which produces a small pore size and hydrophobic region for the gel beads. The elution for anions occurs in the order of:



for an identical counterion (e.g.,  $\text{Na}^+$ ). Those ions eluting earlier than  $\text{Cl}^-$  (sulfate, phosphate, and fluoride) are salting-out salts and are structure stabilizing (i.e., they are water-structure makers) (Figure 9.1). However, their elution positions are earlier than those expected from their molecular weight (i.e., they are excluded from the pores more so than are expected from their size). This is ascribed to the water molecules that are tightly bound to those ions [74, 316]; more specifically, the ions possess strong affinity for water molecules and, as a result, they migrate with the



**Figure 9.18** Schematic illustration of the effect of preferentially excluded salts or osmolytes on protein aggregation and unfolding.

bound water molecules through the Sephadex G-10 column. On the other hand, those ions eluting after  $\text{Cl}^-$  are classified as salting-in salts and are water-structure breakers. They also do not elute from the column according to their molecular weight – they elute later than the expected size. Furthermore, their elution positions are dependent on the concentration of the loading salt solutions and the column temperature, indicating their binding to the column. This suggests that these ions are not only less hydrated, but also possess higher affinity for the surface of Sephadex G-10 than for water.

Ion hydration can also be measured from the viscosity of salt solutions as a function of salt concentration. The Jones–Dole viscosity  $B$  coefficient is derived from fitting the viscosity data to:

$$(\eta/\eta_0) = 1 + A_c^{0.5} + B_c$$

The  $B$  coefficient reflects ion hydration: a positive value for strongly hydrated ion and a negative value for weakly hydrated ion. Table 9.7 shows the  $B$  coefficients for several cations (left column) and anions (right column). The sign and order for anions resemble their order in the elution from a Sephadex G-10 gel filtration column. Those ions with multivalency have, in general, a higher  $B$  coefficient, consistent with their strong water-binding capability. Among the cations,  $\text{Na}^+$  and  $\text{K}^+$  salts set a boundary between the salts that possess strong and weak hydration respectively, and  $\text{Cl}^-$  sets a similar boundary for anions. As will be described below, these hydration potentials of ions are related to their interactions with proteins and macromolecules, which in turn are correlated with their effects on the properties of the macromolecules.



**Table 9.7** Jones–Dole viscosity *B* coefficient.

Cations	<i>B</i>	Anions	<i>B</i>
Mg <sup>2+</sup>	0.385	PO <sub>4</sub> <sup>3-</sup>	0.590
Ca <sup>2+</sup>	0.285	CH <sub>3</sub> CO <sub>2</sub> <sup>-</sup>	0.250
Ba <sup>2+</sup>	0.22	SO <sub>4</sub> <sup>2-</sup>	0.208
Li <sup>+</sup>	0.150	F <sup>-</sup>	0.10
Na <sup>+</sup>	0.086	HCO <sub>2</sub> <sup>-</sup>	0.052
K <sup>+</sup>	-0.007	Cl <sup>-</sup>	-0.007
NH <sub>4</sub> <sup>+</sup>	-0.007	Br <sup>-</sup>	-0.032
Rb <sup>+</sup>	-0.030	NO <sub>3</sub> <sup>-</sup>	-0.046
Cs <sup>+</sup>	-0.045	ClO <sub>4</sub> <sup>-</sup>	-0.061
		I <sup>-</sup>	-0.068
		SCN <sup>-</sup>	-0.103

Data adapted from [427, 428].

**9.5.1.1.2 Osmolyte Hydration** Just as for ions, highly polar osmolytes are also hydrated. Many osmolytes increase the surface tension of water, although less effectively than the salting-out salts [68, 317–319]. For example, one of the most common osmolytes, trimethylamine *N*-oxide (TMAO), decreases the surface tension of water slightly. Similar to hydrated salts, hydrated osmolytes are preferentially excluded from the protein surface as described below in detail, which results in increased stability and decreased solubility of proteins. However, many osmolytes have a large hydrodynamic radius, which also causes preferential exclusion due to the excluded volume effects. Thus, for osmolytes, the effects could arise from both exclusion factors; certainly for TMAO, its excluded volume effect should be significant to compensate for its ability to lower the surface tension.

**9.5.1.1.3 Protein Hydration** As with salt ions, proteins and other macromolecules are also hydrated. The amount of hydration of proteins and DNAs can be measured by the use of differential scanning calorimetry (DSC) or NMR in the frozen state. There is a layer of water that does not freeze even below  $-30^{\circ}\text{C}$ . Such water molecules can be detected and analyzed by NMR or DSC under frozen conditions. Freezing makes the frozen (and hence unbound) water molecules undetectable by NMR and there will be no phase transition of bound water during a DSC scan. An average of 0.3 g/g protein of bound water is typically observed [85, 86]. It is not a coincidence that a similar level of hydration is obtained by hydrating protein powders. Upon the addition of around 0.2–0.3 g water/g protein, proteins acquire certain properties that can be observed in solution (e.g., enzyme activity) [320]. Different rotational correlation time of water can be obtained by proton NMR, by using the value for bulk water [321]. This amount of water can cover the entire protein surface as a single molecule layer. What is the property of such bound water?

Similar to ion hydration, water molecules near the protein surface can be divided into three layers, A, B, and C: although not described in the above section on ion hydration, ions are also hydrated with water molecules in a similar manner to the

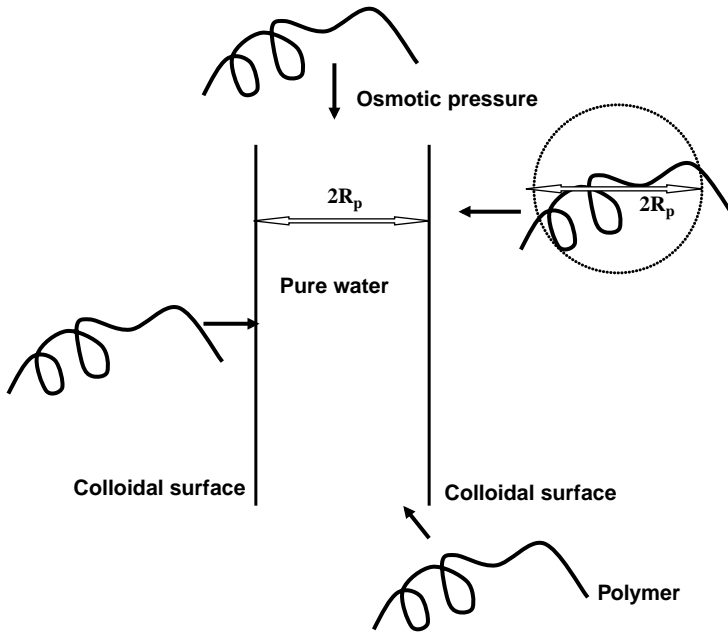
protein surface. Layer C is identical to the bulk water having a rotational correlation time of  $10^{-12}$  s (i.e., moving as bulk water). Layer A is made of tightly bound water of 0.3 g/g with a correlation time of  $10^{-8}$  to  $10^{-7}$  s that does not freeze even at  $-190$  °C. This layer constitutes, on average, a single water molecular layer at the protein surface. Proton NMR of hydrated protein powder shows a slow correlation time at a water level of around 0.25 g/g [321]. However, the protein surface is not as homogeneous as the surface of ions. Tightly bound water molecules mostly occupy the charged and polar surfaces. Layer B is the transition layer composed of several molecular layers of water and share properties from both layers A and C. Consequently, its correlation time is between those of the two adjacent layers.

DNA also binds around 0.18 g water/g DNA, based on the measurement of unfrozen water [322], which constitutes layer A. Water molecules in layer B bind to the minor grooves of DNA double helix, amounting to approximately 0.23 g/g and show a phase transition at around  $-70$  °C. This layer (B) plays a major role in the stability of the double helix.

The interactions described above are related to hydration around the charged or polar surface. The surface of proteins and other macromolecules are higher in complexity than ions and small cosolvents. These macromolecules have nonpolar surfaces that are often exposed to water. Water in the vicinity of the nonpolar surface cannot form regular hydrogen bonding with other water molecules nor with the protein surface. Such hydration is called hydrophobic hydration. This hydration is the reason for low solubility of apolar groups and compounds in solution. There is a large free energy increase in such a hydration process, due to the decrease in entropy. This entropy decrease is due to the increase in the number and strength of hydrogen bonding between water molecules. This can be depicted with the apolar surface restricting the motion of neighboring water molecules and impeding the bending motion of the hydrogen bonding between water molecules, thus stiffening the overall structure and causing the observed decrease in entropy. There are no repulsive interactions between the apolar surface and water; as a matter of fact, there are often attractive interactions. However, the entropy loss due to water structuring around the apolar surface overwhelms the attractive interaction between the apolar surface and water molecules. If any cosolvents increase such water–water interaction in the vicinity of the apolar surface, their incorporation would enhance the hydrophobic interaction. This ordering of water molecules near an apolar entity is called hydrophobic hydration. The entropy decrease accompanying the introduction of an apolar surface into water is proportional to the solvent-exposed surface area. Thus, the larger the area, the greater the entropy decrease and thus the stronger the hydrophobic interaction. This is referred to as the hydrophobic effect.

#### 9.5.1.2 Excluded Volume

Excluded volume effects are expressed in several different forms. For example, the association of colloidal particles induced by polymers has been explained by the depletion interaction mechanism [77, 80, 323, 324]. Colloidal particles are surrounded by an exclusion layer, due to the inability of the polymer to approach the colloidal surface within the distance of the polymer radius,  $R_p$ . When two colloidal

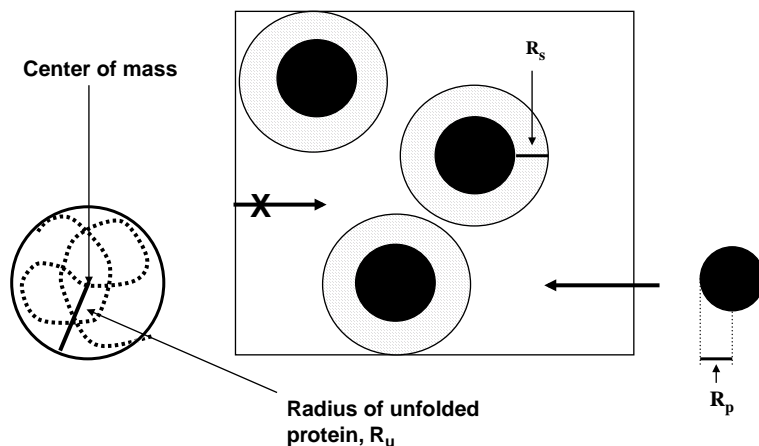


**Figure 9.19** Schematic illustration of depletion effect.  $R_p$  corresponds to the radius of unfolded protein.

particles approach each other within a distance of  $2R_p$ , the polymer cannot enter the space between the two surfaces, as illustrated in Figure 9.19, and thus are excluded from the space; the space becomes free of polymers. Therefore, a force, equivalent to the osmotic pressure of the solution (due to the exclusion of the polymer), acts on the colloidal particles and its magnitude increases with polymer concentration. This is called the “depletion effect” and forces the two colloidal particles to associate. Such a depletion effect, thus osmotic pressure, always operates between colloidal particles in the presence of excluded polymers.

The concept of macromolecular crowding describes the same effect. The interaction of native protein with the inert crowding polymers has been described by the excluded volume effect [325–327]. Such a crowding mechanism was used to explain the effects of a polymer on the hydrodynamic size of PEG11700; the addition of Ficoll gradually induced the compaction of PEG, which reduced the excluded volume effect of Ficoll. Protein stability, due to crowding, has also been explained based on the difference in excluded volume for polymers between the native and unfolded structures, although the calculation of excluded volume effect for the unfolded structure is not straightforward. Nevertheless, such a calculation resulted in good correlation with experimental data for protein stability in the presence of polymers [15].

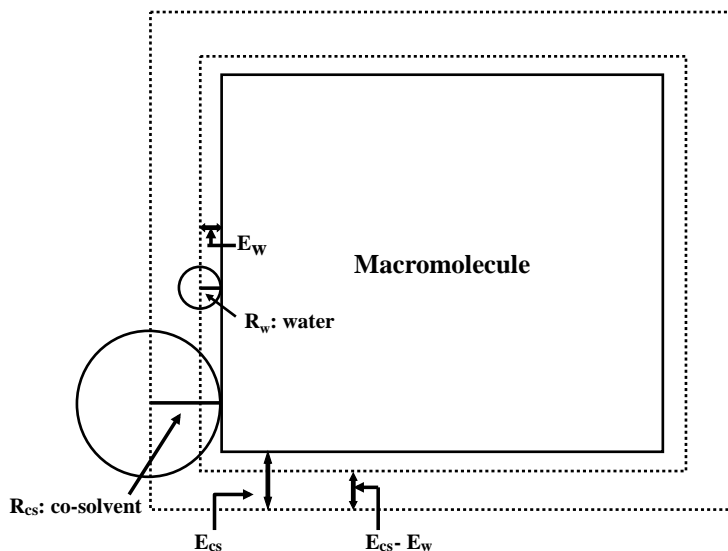
Molecular crowding effects should also occur at high concentrations of macromolecules. In other words, protein molecules should be more stable at higher protein



**Figure 9.20** Schematic illustration of molecular crowding effect. Black circles correspond to protein molecules.  $R_p$  corresponds to the radius of folded protein,  $R_s$  the hydration radius of folded protein, and  $R_u$  the radius of unfolded protein.

concentration, which suggests that the shelf-life of proteins can be enhanced at higher protein concentrations. The illustration shown in Figure 9.20 demonstrates a protein molecule occupying a finite space, which is represented by a black circle. If another solute, which has no volume, is dissolved in water, then the entire space unoccupied by the protein is available. Introduction of another protein molecule (also shown by black circle), with the radius of  $R_p$ , is much more restricted than the small molecule, since its center of mass cannot approach the surface of the protein within the radius of the newly introduced protein ( $R_s$ ), as depicted by the hatched area. More specifically, the interaction of protein molecules themselves is unfavorable due to their excluded volume. As shown in Figure 9.20, it would be much more difficult to place the unfolded protein that has a greater hydrodynamic radius ( $R_u$ ) into this system. This forces the protein to maintain the native compact structure in the crowded solution of high protein concentration, thus improving structural stability.

There are no molecules that have zero volume, which implies that all molecules are excluded from the protein surface. In fact, water molecules can exert an excluded volume effect that is not due to the osmotic pressure effect of polymer cosolvents (as this effect arises due to the exclusion of polymer from the vicinity of colloidal particles). As illustrated in Figure 9.21, water is also excluded from the surface due to its radius ( $R_w$ ), in which the translational diffusion of water is restricted [328–330]. The exclusion layer of water,  $E_w$ , is also depicted in Figure 9.21. The authors proposed that the entropically unfavorable exclusion of water molecules is the driving force for protein folding, as the side-chain interactions in folded structure reduce the solvent-accessible surface area and hence also water exclusion. Of course, this exclusion mechanism alone does not lead to protein folding. It should be noted that water is a poor solvent for proteins. If water exclusion is the sole reason for protein folding, folding should be enhanced in, for example, ethanol relative to water, as ethanol has a



**Figure 9.21** Schematic illustration of the excluded volume effect. The macromolecule (i.e., protein) is represented as the square in the center of the illustration.  $R$  is the radius of cosolvent (cs) or water (w).  $E$  is the excluded volume of cosolvent (cs) or water (w).

larger excluded volume. However, ethanol solvates the nonpolar surface of proteins, which in turn, results in protein unfolding. Furthermore, protein folding occurs through specific side-chain and peptide interactions.

The excluded volume effect of low-molecular-weight cosolvents has been clearly described by Schellman [81]. According to his analysis, interaction of cosolvents with proteins can be divided into the effects arising from binding and excluded volume. Although the excluded volume effect of low-molecular-weight cosolvent is much smaller than that of proteins and polymers, Schellman's analysis clearly demonstrates that it cannot be ignored. More specifically, unfolding effects of urea and GdnHCl, or the stabilizing effect of TMAO, cannot be simply explained from their binding affinity for proteins. In fact, the excluded volume effect partially offsets the denaturing effect of urea and GdnHCl, and reverses the same effect of TMAO: as described above, TMAO slightly decreases the surface tension of water, which explains its weak affinity for the hydrophobic air/water interface. In other words, the excluded volume effect overwhelms the favorable binding (destabilizing) interaction of TMAO with the protein, leading to an overall stabilizing effect. It should be noted that the same excluded volume effect should occur between small molecules. For example, the observed interaction between two small molecules reflects both the affinity (binding) and the excluded volume, although the latter effect is much smaller for small molecules.

The excluded volume effect of low-molecular-weight cosolvents is due to its size,  $R_{cs}$ , as depicted in Figure 9.21. However, water is also excluded from the surface.

Thus, the net effect of excluded volume for cosolvents is due to the difference between the radius of cosolvent and water (i.e.,  $R_{cs} - R_w$ ), as depicted in Figure 9.21.

## 9.5.2

### Thermodynamic Interaction

We have described above the physical mechanisms that determine the interactions between cosolvents and macromolecules, except for cosolvent binding. Cosolvents not only have affinity for water, but also possess affinity for macromolecules. In certain cases, they may demonstrate negative affinity for macromolecules (repulsion). Such bindings (or lack thereof) are difficult to measure, as they occur at high cosolvent concentrations. In other words, these weak interactions cannot be measured by isolating macromolecules bound by cosolvents to determine the amount of bound cosolvents. Isolation would lead to a dissociation of weakly bound cosolvents. Only thermodynamic equilibrium techniques can determine such binding affinity accurately. The interaction of cosolvents with model compounds can be measured from the effects of cosolvents on the solubility of model compounds, as described below. Dialysis equilibrium experiments can measure cosolvent binding for proteins, but not for small model compounds, due to the lack of appropriate dialysis membranes for low-molecular-weight compounds.

Such analysis reflects all the interactions present between the model compounds and cosolvents – cosolvent hydration (for salt ion hydration, as described above), excluded volume, and binding. If the model compounds can adequately describe the macromolecular surface, the solubility measurements should be able to predict the interaction of cosolvents with macromolecules. This assumes that the cosolvent interactions with the model compounds are additive, and that the size difference between the model compounds and macromolecules makes no contribution to the cosolvent interactions. This is unlikely, as is evident in the excluded volume effect. The observation of cosolvent exclusion from the small model compounds does not lead to the prediction of its exclusion from large protein molecules, or DNA and viruses. Direct interaction measurements of cosolvents with macromolecules can determine the overall interaction with the entire protein surface.

#### 9.5.2.1 Group Interaction: Model Compound Solubility

Both ion hydration (surface tension) and excluded volume effects can be readily observed from their effects on water (i.e., the properties of aqueous solutions). On the other hand, affinity of cosolvents for macromolecular surfaces depends on the property of the macromolecules themselves. As macromolecules are composed of smaller units (e.g., amino acids), the affinity may be inferred by studying the interactions of these smaller units (groups) with various cosolvents. What will be observed from such measurements are the sum of surface tension, or excluded volume effects, and affinity, as described by Schellman [81]. Such group interactions have been analyzed by measuring the solubility of small molecules in aqueous solutions containing cosolvents. Data of this type have been compiled by Cohn [331] and a small fraction of his published data is shown below.

**Table 9.8** Transfer free energy from water to ethanol.

Amino acid or side-chain	Transfer free energy (cal/mol)
Glycine	4630
	side-chain
Valine	-1690
Leucine	-2420
Phenylalanine	-2650
Tyrosine	-2870

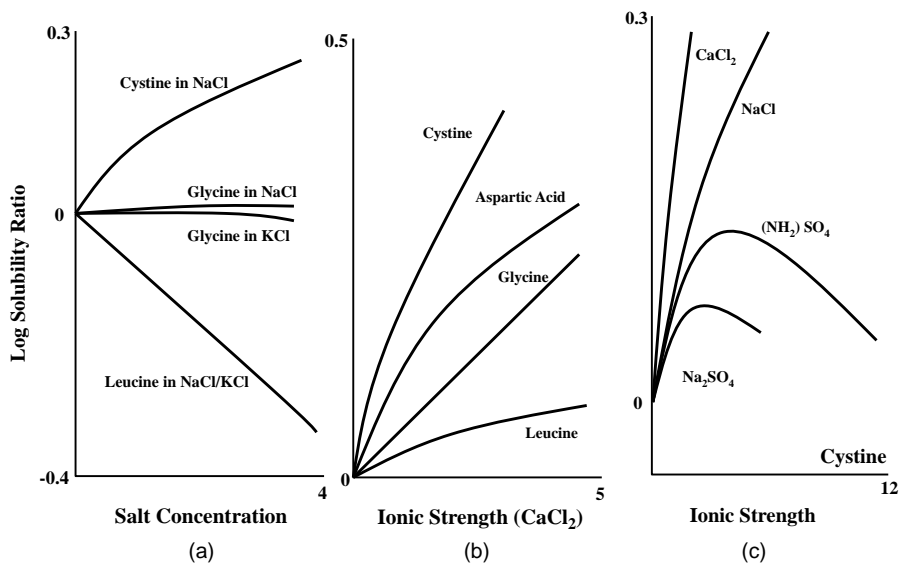
Data adapted from [332].

Not only do solubility data provide the cosolvent interactions with the small model compounds, they also demonstrate the property of groups constituting the macromolecules. Tanford [332] used 100% ethanol as an indicator of hydrophobicity of amino acid side-chains – the higher the solubility, the more hydrophobic are the side-chains. In natural environments, the hydrophobic amino acids tend to move into the hydrophobic core of folded protein structure. Some of the data are summarized in Table 9.8. Here, the solubility data are expressed as a thermodynamic interaction between the various groups and ethanol. The transfer free energy,  $\Delta G$ , is determined from:

$$\Delta G = -RT \ln S_{cs}/S_w$$

where  $S_w$  and  $S_{cs}$  are the solubility of the model compound in water and cosolvent solution, respectively,  $R$  is the gas constant, and  $T$  is the temperature in Kelvin. When the cosolvent increases the solubility of the model compound, the transfer free energy becomes negative, and the interaction between the cosolvent and the model compound becomes thermodynamically favorable. The transfer of glycine from water to ethanol is extremely unfavorable, as shown by the large positive value, indicating the repulsive interaction of ethanol and glycine. Conversely, the side-chains of valine, leucine, phenylalanine, and tyrosine interact favorably with ethanol, suggesting that these amino acid side-chains are hydrophobic in nature and thus form the hydrophobic core of the folded protein structure. Thus, solubility measurements can provide information on the property of cosolvents based on the knowledge of the chemical nature of the model compounds or the properties of the model compounds, if the solvent property is already known. A summary is given below for several systems, for which the solubility measurements have provided important insights into the interactions between cosolvents and protein groups or model compounds.

**9.5.2.1.1 Salts** Based on their hydration potential, salts are expected to exert their effects on the solubility of model compounds through the surface tension effect. If that is the case, then the solubility of amino acids should be independent of the side-chain structure and be all equivalent. Figure 9.22(a) demonstrates that this is not the case. (i) It is evident that NaCl and KCl are similar. (ii) The solubility of glycine is marginally affected by these salts, even at a high concentration (4 M). (iii) NaCl/KCl



**Figure 9.22** Plot of solubility of various amino acids in salt solution. (Data adapted from [331, 353].)

greatly decreases the solubility of leucine, indicating that these salts interact unfavorably with the side-chain of leucine, while their interactions with glycine, which has no side-chain, are close to zero. This suggests that the salts interact unfavorably with leucine, a hydrophobic amino acid. The increased solubility of cystine suggests the existence of a favorable interaction between the salt and this amino acid, although the reasons are unclear; it may be due to the structural differences between glycine and cystine, leading to a different dipole moment between these two amino acids. Except for this unexplainable data, Figure 9.22(a) clearly demonstrates the unfavorable interaction of NaCl and KCl with the hydrophobic leucine side-chain. Thus, these salts enhance hydrophobic interactions. Figure 9.22(b) indicates that CaCl<sub>2</sub> increases the solubility of glycine, different from the observed effects of NaCl/KCl, suggesting that this salt stabilizes the dipole moment of glycine. The solubility increase is less for leucine, suggesting the existence of unfavorable interactions between this salt and the leucine side-chain. This in turn implies that CaCl<sub>2</sub> interacts unfavorably with hydrophobic groups. Nevertheless, CaCl<sub>2</sub> increases the solubility of leucine, which is in contrast to the effects of NaCl/KCl (Figure 9.22a), indicating a strong salting-in effect of CaCl<sub>2</sub>. On the other hand, the interaction of CaCl<sub>2</sub> with the aspartic side-chain is favorable. Most importantly, Figure 9.22(c) compares four different salts; two salting-out salts, Na<sub>2</sub>SO<sub>4</sub> and (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, decrease the solubility of cystine after the initial salting-in effects, while NaCl and CaCl<sub>2</sub> increase its solubility. While the order clearly shows their effects on solubility, the interpretation may depend on the chemical nature of cystine. If this amino acid is assumed to be nonpolar, then the first two salting-out salts enhance the

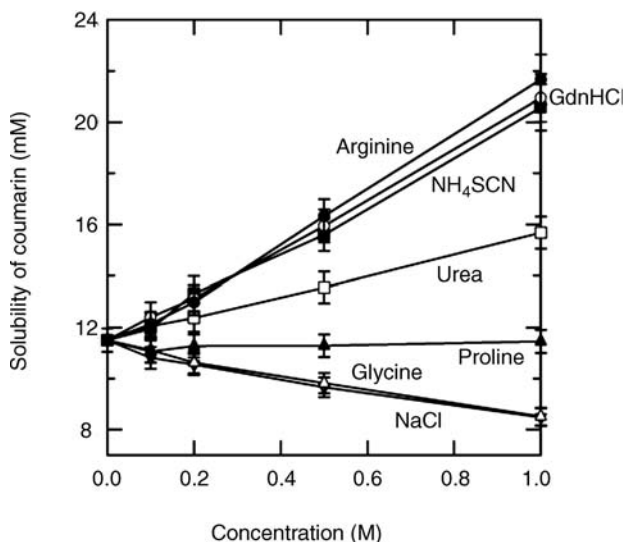


hydrophobic interaction, while  $\text{CaCl}_2$  reduces this interaction. A general dipole stabilizing effect in this case may be represented by a monotonic trend in the solubilizing effect of  $\text{NaCl}$ . The difference between  $\text{NaCl}$  and  $\text{Na}_2\text{SO}_4$  should be due to the anionic species. The effects of anionic species on the group solubility are demonstrated later for magnesium salts.

**9.5.2.1.2 Osmolytes** Osmolytes are a group of compounds that are used by halophilic organisms. They increase the osmotic pressure of cytoplasm against external osmotic pressure [333] and protect cellular macromolecules from osmotic stresses as well as other stresses, including desiccation [334–337]. They also enhance protein stability [338]. The protein stabilizing effects of osmolytes along with their relatively inert nature, in interfering with the biological function of macromolecules and cells, are the main reasons for their use in enhancing the shelf-life of purified macromolecules. Specifically, sucrose and trehalose are used most predominantly in the stabilization of pharmaceutical proteins for both liquid and lyophilized formulations. There have been extensive investigations on how these osmolytes stabilize proteins. Solubility measurements, and hence group interaction analysis, demonstrate that the interactions of osmolytes with nonpolar amino acid side-chains and hydrophobic compounds are variable, though small in magnitude [339–343]. However, the common feature among the many osmolytes is their unfavorable interactions with peptide bonds and it is considered to be the main driving force for the stabilization of proteins [340, 341, 343–345]. More specifically, the unfavorable interactions with the peptide bonds, and possibly with the hydrophobic groups, prevent proteins from unfolding and exposing additional hydrophobic side-chains and peptide bonds. Organic solvents also disfavor interaction with peptide bonds, although they strongly favor interactions with hydrophobic groups. This difference in affinity towards nonpolar groups with osmolytes and organic solvents distinguishes their effects on proteins: destabilization by organic solvents and stabilization by osmolytes.

Osmolytes interact unfavorably with peptide bonds, and possibly with hydrophobic groups, as described above. However, it should be noted that the interaction arises from contribution from both cosolvent binding and exclusion. Schellman [81] suggested that one of the osmolytes, TMAO, has binding affinity for the protein surface, although the binding effect is offset by the excluded volume effect. Thus, it is possible for the osmolytes to have affinity for certain groups in the protein molecule, which is small enough to be offset by the excluded volume effect.

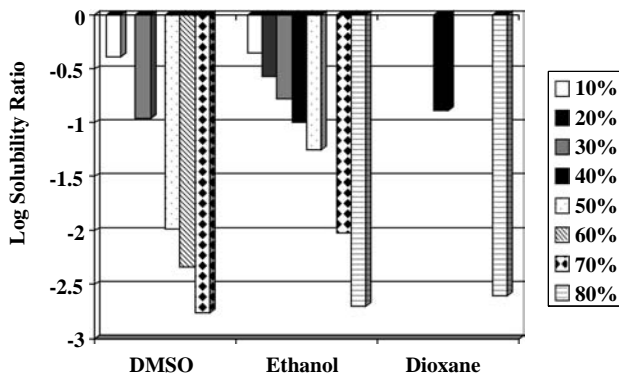
Although arginine is a highly water soluble amino acid, it is not an osmolyte that has been selected by nature [333]. As shown earlier, arginine has a unique property that is not shared by other amino acid osmolytes. For example, arginine is a strong aggregation suppressor [346–348] and shares similar behavior with a salting-in, or denaturing, cosolvent. Figure 9.23 shows the effects of arginine on the solubility of aromatic coumarin [349]. It is interesting to note that arginine increases the solubility of this model compound as effectively as  $\text{GdnHCl}$  and  $\text{NH}_4\text{SCN}$ , and even more so than urea.  $\text{NaCl}$  and glycine show a slight salting-out effect on this hydrophobic compound, while proline is ineffective. Arginine also exhibits a similar pattern of



**Figure 9.23** Plot of coumarin solubility as a function of cosolvent (amino acid) concentration. (Data reformatted from [349].)

interactions with amino acid side-chains and glycine to that observed for GdnHCl and urea [350]. If one can correlate this observation on model small molecules, then it must be concluded based on its similarity with GdnHCl that arginine is a protein denaturant. On the contrary, arginine is not a protein denaturant, probably due to its large size, as will be described later.

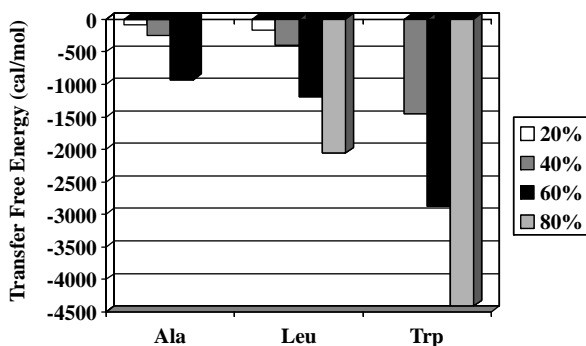
**9.5.2.1.3 Organic Solvents** Although not classified as organic solvents, GdnHCl and urea bind weakly to nonpolar and hydrophobic model compounds, which is the driving force for protein denaturation. The difference between these protein denaturants and organic solvents is their interaction with polar groups. The protein denaturants have favorable interaction with polar groups, or at least show no significant unfavorable interactions [350–352], while organic solvents demonstrate increasingly unfavorable interaction with polar glycine. Figure 9.24 shows the solubility change of glycine as a function of DMSO, ethanol, and dioxane concentration [41]. The solubility is greatly reduced by the addition of these organic solvents, suggesting the presence of unfavorable interactions of the organic solvents with dipolar glycine. The unfavorable interaction most likely occurs on proteins as well, leading to precipitation or crystallization. However, these solvents are also protein destabilizers and can often cause denaturation at high concentrations [41]. This is due to their favorable interactions with hydrophobic groups. An example of such favorable hydrophobic interaction with DMSO is shown in Figure 9.25, in which the transfer of the side-chains of alanine, leucine, and tryptophan from water to 20–80% DMSO solution is shown to be favorable. These favorable hydrophobic interactions of DMSO and other organic solvents drive the protein to unfold. As the transfer of peptide bonds



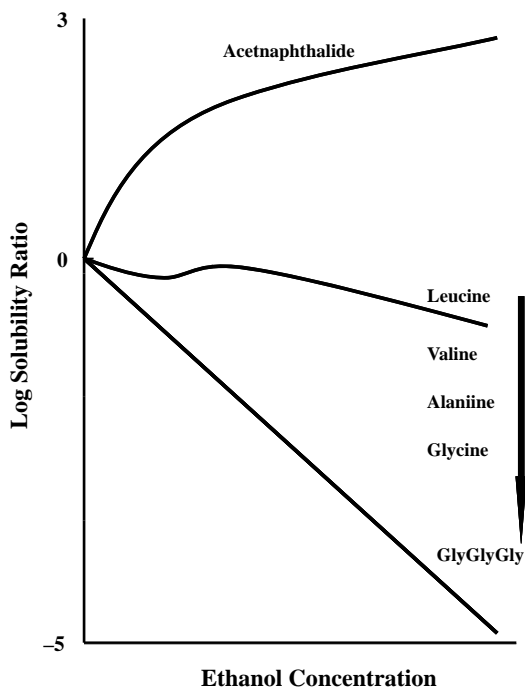
**Figure 9.24** Solubility change of glycine in different organic solvents at the indicated concentrations. Organic solvents examined include DMSO, ethanol, and dioxane. (Reformatted from [41].)

is unfavorable, these organic solvents induce the formation of  $\alpha$ -helices upon unfolding; formation of  $\alpha$ -helices reduces solvent exposure of the peptide bonds.

Cohn and Edsall [353] also compiled the effects of ethanol on the solubility of various model compounds. It should be noted that this study, as well as others from his laboratory, have led to the development of the famous Cohn fractionation of plasma proteins using cold ethanol [128]. Figure 9.26 summarizes the qualitative effects of ethanol; the solubility of glycine decreases, as observed by Nozaki and Tanford [354]. A similar conclusion can be made about the unfavorable interaction for the peptide bond (difference between triglycine and glycine) and the favorable interaction for the three hydrophobic groups (difference between alanine/valine/leucine and glycine). An aromatic compound, acetnaphthalide, shows a highly favorable interaction with ethanol, again consistent with the observation of Nozaki and Tanford that the interaction of ethanol with aromatic side-chains is highly favorable [354].



**Figure 9.25** Transfer free energy of amino acid side-chains from water to aqueous DMSO solution, ranging in concentration from 20 to 80%. Amino acids include alanine, leucine, and tryptophan. (Reformatted from [41].)



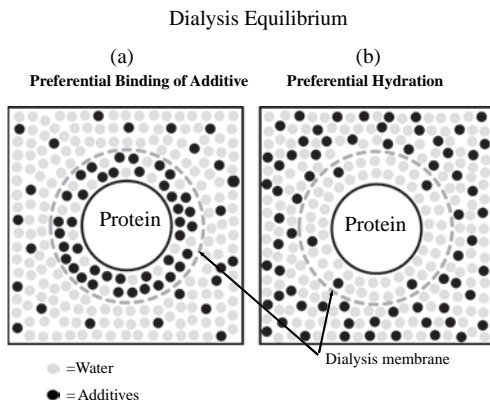
**Figure 9.26** Solubility change of amino acids and acetnaphthalide as a function of ethanol concentration. (Data adapted from [353].)

**9.5.2.1.4 PEG** Although no data are available for the effects of PEG on the solubility of model compounds, its hydrophobic property can be inferred from two observations. First, PEG decreases the surface tension of water [319]. In fact, the magnitude of surface tension depression is greater than that observed for MPD and DMSO on a weight basis [319]. The effects of PEG on the surface tension are due to the structural unit of  $\text{CH}_2\text{CH}_2\text{O}$ , with  $\text{CH}_2\text{CH}_2$  pointing towards the air phase and O pointing towards the water phase. Thus, its strong surface activity may be due to the potential orientation of PEG, as seen with surfactants. Another observation is the use of PEG as a hydrophobic resin [355, 356]. PEG can be used to bind proteins through hydrophobic interaction, although the weakness of such interactions must be enforced by salting-out salts in a practical application of PEG-conjugated resin.

### 9.5.3

#### Preferential Interaction

The combination of hydration of cosolvents and macromolecules, the excluded volume effect, and cosolvent binding cause the macromolecules to interact with the



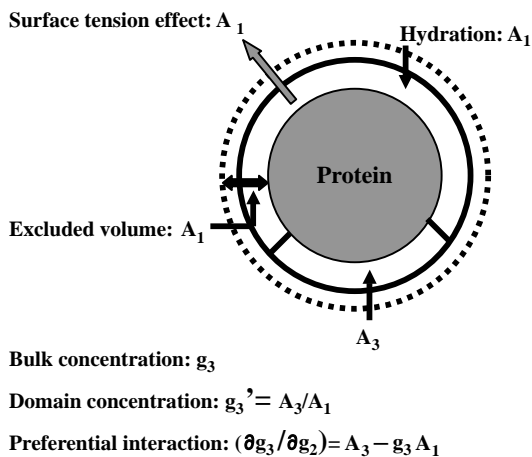
**Figure 9.27** Schematic illustration of preferential interaction: (a) preferential binding and (b) preferential hydration. (Reformatted from [372].)

solvent system. These interactions range widely in strength from weak momentary interactions to strong specific bindings, although, at high cosolvent concentrations, a large number of weak water and cosolvent interactions will potentially overwhelm the specific interactions. Such interactions are termed “preferential,” as they reflect differences in interaction between water and cosolvents. Figure 9.27 illustrates why the interaction is “preferential.” In dialysis equilibrium at high cosolvent concentration, the macromolecule (here protein) will interact with both water and cosolvent. In Figure 9.27(a), the protein molecule is surrounded by the cosolvent and hence is bound by the cosolvent. However, the observed value of binding is not the actual amount of the bound cosolvent, rather the measured value is the difference in cosolvent concentration between the interior and exterior of the dialysis bag (i.e., between the vicinity of the protein and the bulk phase). It is thus an average of every interaction that protein molecules experience, including hydration, cosolvent binding, and exclusion (e.g., due to the excluded volume effect or cosolvent hydration). When the cosolvent has little affinity for protein, the case shown in Figure 9.27(b) occurs, in which the cosolvent concentration is depleted in the vicinity of the protein molecule.

The concentration difference of a cosolvent between the vicinity of the protein and the bulk phase can be described as:

$$(\partial g_3 / \partial g_2) = A_3 - g_3 A_1 \quad (9.1)$$

where at a constant chemical potential,  $(\partial g_3 / \partial g_2)$  corresponds to the excess amount of cosolvent in the protein vicinity in grams per gram protein, and is a function of cosolvent binding,  $A_3$ , and hydration,  $A_1$ . Figure 9.28 schematically depicts the vicinity of protein surface at which the various types of cosolvent interactions can occur. There are two important points to note in this illustration. (i) The preferential interaction is also related to the bulk concentration of the cosolvent,  $g_3$ , as it is the



**Figure 9.28** Schematic illustration of various interactions of water or cosolvent with a macromolecule (protein in this case).

difference in cosolvent concentration between the protein and bulk phase. (ii) As depicted in Figure 9.28,  $A_1$  does not have to be the amount of bound water, but can represent the amount due to the excluded volume of cosolvent (shown by the dotted line) or the surface tension effect (shown by the gray arrow). As Eq. (9.1) demonstrates, in the absence of cosolvent binding, the observed preferential hydration is entirely due to hydration,  $A_1$ . Preferential hydration (i.e., excess amount of water in the protein domain) can be expressed as:

$$(\partial g_1 / \partial g_2) = -(1/g_3)(\partial g_3 / \partial g_2) \quad (9.2)$$

Thus, in the case of no cosolvent binding:

$$(\partial g_1 / g_2) = A_1 \quad (9.3)$$

and hence the observed preferential hydration corresponds to the amount of bound water, or excess water, due to the exclusion of cosolvents derived from either excluded volume, cosolvent hydration (surface tension effect), or cosolvent repulsion.

More importantly, the thermodynamic “preferential cosolvent interaction” controls the effects of cosolvents on the property of proteins (i.e., stability, solubility, and interaction). How can we relate the preferential interactions of water and cosolvents to their effects on macromolecule reactions? For a reaction described by:



where R and P both interact with water and cosolvent, preferential interaction is related to its effect on the reaction through the thermodynamic linkage relation [19]:

$$(d \ln K / d \ln a_3) = (\partial m_3 / \partial m_2)^P - (\partial m_3 / \partial m_2)^R \quad (9.4)$$

where  $(\partial m_3/\partial m_2)$  is the preferential interaction in mol/mol and hence is expressed as:

$$(\partial m_3/\partial m_2) = (M_2/M_3)(\partial g_3/\partial g_2) \quad (9.5)$$

Equation (9.4) indicates that when the preferential cosolvent binding is greater for the product, the addition of the cosolvent and hence the increase in its activity,  $a_3$ , should enhance the reaction. For example, increased preferential cosolvent binding for the unfolded structure should drive the unfolding reaction, as for urea or GdnHCl.

The above equation shows that it is the difference in preferential interaction that determines the effects of the cosolvent, and thus the reaction itself. Numerous preferential interaction measurements have been conducted, as summarized below (see also Table 9.2). However, those measurements are generally conducted on either R or P alone. Under a given condition, no measurement can be done for both R and P, which are in constant equilibrium. The measurements will be an average of both states. In fact, most of the measurements were conducted on only one state under the given solvent condition. This is clearly depicted in Figure 9.29. Preferential interaction measurements for protein stabilizers can be readily conducted on the native protein structure (namely R, marked XX), as they maintain the native protein structure. This implies that the preferential interaction property of the unfolded state (i.e., product, P) must be inferred from the surface property of the unfolded state and its interaction with the cosolvent (as marked UD), which requires the knowledge of cosolvent interactions with model group compounds. As the unfolded structure

	<b>Reactant (R)</b> Native protein	<b>Product (P)</b> Unfolded protein
<b>Stabilizer</b>	XX	UD
<b>Denaturant</b>	X (at low concentration)	X (at high concentration)
<b>Destabilizer</b>	XX	UD

	<b>Reactant (R)</b> Monomeric protein	<b>Reactant (R)</b> Aggregated protein
<b>Stabilizer</b>	XX	UD
<b>Denaturant</b>	X (at low concentration)	UD
<b>Destabilizer</b>	XX	UD
<b>Precipitant</b>	UD	X

**Figure 9.29** State of protein for which preferential interaction measurements are normally conducted. XX corresponds to possible measurements, while UD corresponds to measurements that are unlikely to be conducted.

would expose the internal hydrophobic amino acids, knowledge of cosolvent binding to nonpolar compounds would give an insight into how the various cosolvents may interact with the unfolded structure. The measurement of cosolvent interaction in the presence of stabilizers may be conducted above the melting temperature of the macromolecules. At such a condition, however, the preferential interaction measurement of the native state becomes impossible. For denaturants, the measurement can be conducted on the native state at low concentrations. At increasingly high concentrations, the measurement will have an increased contribution from the unfolded state. For destabilizers that cannot unfold macromolecules under the normal experimental condition, the measurements can only be conducted on the native state (Figure 9.29).

For an aggregation reaction (Figure 9.29, lower panel), preferential interaction is the difference between cosolvent interaction with the monomer and the aggregated state(s). For stabilizers, denaturants, and destabilizers, the interaction measurements are most likely conducted on the monomer state, as they could not enhance aggregation strongly. For a precipitant, the measurement may be conducted on the aggregated state if one can make a stable aggregated macromolecule. For example, it may be possible to conduct equilibrium measurements of protein crystals that are cross-linked. This has been performed for cross-linked crystals of ribonuclease in equilibrium with a strong protein precipitant, MPD [134]. The experiment was conducted by the addition of aqueous MPD solution of known composition to the dried cross-linked ribonuclease crystal, and then the MPD concentration was measured after equilibration. Concentration change reflects the interaction of MPD with the protein crystal; if the crystal absorbs the solution of the same composition, then there should be no change in the MPD concentration. Pittz and Bello [134] observed preferential absorption of water by the crystal.

Another approach would be to study the cause of preferential interaction. Such understanding can be made through a knowledge of hydration, excluded volume, and surface properties of macromolecules. In other words, the various binding (or exclusion) mechanisms described above can be used to explain preferential interaction, which in turn can be extrapolated to explain the possible interactions with a state that does not populate under the experimental conditions (marked as UD in Figure 9.29).

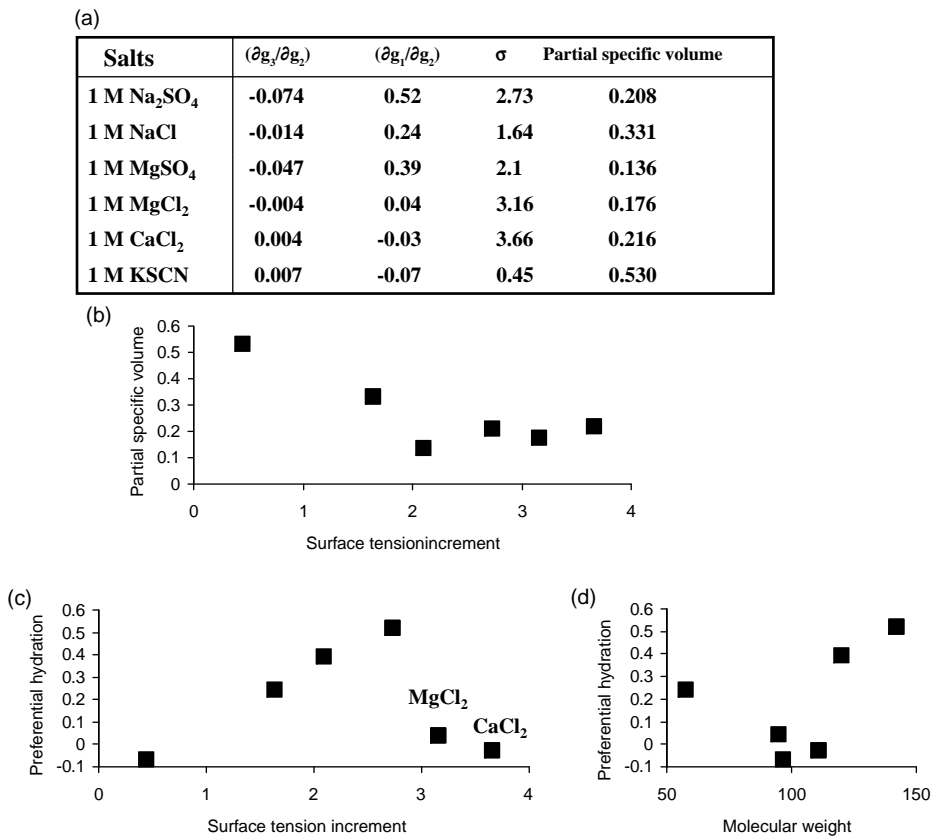
Preferential interaction parameters can be determined experimentally by equilibration techniques (e.g., equilibrium dialysis, the isopiestic method, gel filtration, ultrafiltration, and sedimentation velocity). Dialysis equilibrium is simple, in theory, but technically, it requires a highly sensitive and reliable measurement of cosolvent concentration. This is due to the high concentration of cosolvents, which gives high background noise over which the concentration difference between the bulk and protein solutions must be measured. In cases in which the density of the cosolvent differs from that of water (e.g., 0.7 versus 1 g/ml), a density meter can be used [357, 358], while a differential refractometer [358–360] can be used when the refractive index of the cosolvent differs from that of water.

A unique example of preferential interaction measurements using sedimentation velocity has been conducted for viruses. The sedimentation velocity of TMV particles



was determined as a function BSA and sucrose concentration to change the density of the solution [77]. As the density of the solute increased, the sedimentation velocity of the virus decreased linearly. From the linear dependence, the density at which the virus particles no longer sediment was determined to be 1.13 g/ml in BSA solution and 1.27 g/ml in sucrose solution. A similar observation was made for other viruses [361], with a value of 1.104 for influenza virus in BSA solution. These values should be equal to the density of viruses. However, the partial specific volume of the TMV was calculated to be 0.73 ml/g, which corresponds to a density of 1.37 g/ml. This calculated density of the virus is much heavier than the value determined by sedimentation velocity analysis in BSA and sucrose solution, which is an indication of hydration. The hydration of virus was calculated from the difference in density to be 1.39 g/g in BSA solution and 0.27 g/g in sucrose solution. Sharp *et al.* [361] used a partial specific volume of 0.823 ml/g for influenza virus; the density, 1.21 g/ml, is also heavier than the density, 1.104 g/ml, determined by velocity experiment in BSA solution. This difference results in a hydration value of 0.878 g/g. It is interesting that the observed hydration value in sucrose solution is in the same range as that observed for proteins, suggesting that the hydration mechanism of virus is similar to the hydration mechanism for proteins. Hydration is much greater in BSA solution, implying that the hydration value does not accurately reflect the amount of bound water, but rather the exclusion of BSA from the virus surface, similar to that observed for proteins (Figures 9.20 and 9.21). As these experiments are thermodynamic measurements, sedimentation velocity analysis cannot distinguish bound water from the excluded volume effect of BSA, which leads to BSA-free surface at the vicinity of the virus due to steric exclusion. Equilibrium dialysis measurements showed that TMV is also hydrated in glycerol solution: 0.16 g/g for untreated TMV and 0.205 g/g for disaggregated TMV [362, 363]. The observed smaller hydration value of the TMV in glycerol solution may be due to the smaller size of glycerol in comparison to sucrose, and BSA/glycerol has a smaller excluded volume. It is interesting to point out that these three cosolvents (glycerol, sucrose, and BSA) are unlikely to have affinity for viruses, but rather are excluded from their vicinity due to steric hindrance: if they bind to viruses, then there should be an additional hydration to compensate for the increased density of bound cosolvents.

**9.5.3.1.1 Salts** There is a clear correlation between the preferential interaction and Hofmeister series for salts. Figure 9.30 includes a table for preferential interaction measurements of several salts with BSA [364]. Salt exclusion (negative binding) of BSA was observed for  $\text{Na}_2\text{SO}_4$  and  $\text{MgSO}_4$ , followed by small exclusion for  $\text{NaCl}$ , and by minute interaction for  $\text{MgCl}_2$ ,  $\text{CaCl}_2$ , and  $\text{KSCN}$ . These observations are quantified by the preferential hydration term; there is strong preferential hydration for the first two salts, medium hydration for  $\text{NaCl}$ , and negligible to negative hydration for the last three. These interactions reasonably correlate with their order in decreasing protein solubility (Figure 9.1);  $\text{Na}_2\text{SO}_4$  and  $\text{MgSO}_4$  are salting-out salts,  $\text{NaCl}$  is neutral, and the last three are salting-in salts. More specifically, those salts that exhibit large preferential hydration exert salting-out effects and those of negligible hydration exert salting-in effects (summarized in Table 9.2). Why is such correlation observed?



**Figure 9.30** Various correlations between observed preferential interactions of cosolvent salts and their parameters. (a) Surface tension and partial specific volume of various salts obtained in preferential interaction measurements in the presence of BSA. (b) Plot demonstrating the correlation between surface

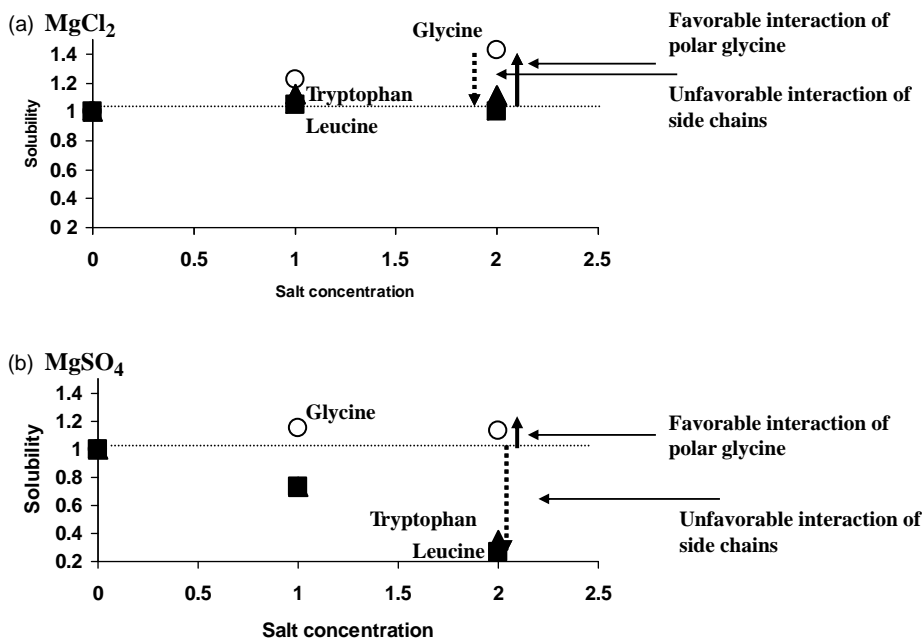
tension increment to particle specific volume of salts shown in (a). (c) Plot illustrating the correlation of surface tension increment to preferential hydration of various salts. (d) Plot demonstrating the correlation between molecular weight of salts to preferential hydration. (Reformatted from [364].)

Before getting into this discussion, it might be helpful to understand the cause of the difference in preferential hydration between the salts.

As depicted in Figure 9.28, preferential interaction is the sum of cosolvent binding or exclusion and hydration. As described earlier, the increased surface tension is due to the hydration potential of ions. The tabulated material in Figure 9.30(a) includes the surface tension and partial specific volume data for these salts. As the surface tension reflects how strongly the salt ions bind water molecules, it should also be related to the partial specific volume, which is a function of the van der Waals volume [365] per gram of salt, and the degree of electrostriction (increased water density around ions).

Assuming that the volume is similar between the salts per weight basis, those salts that increase the surface tension should exert stronger electrostriction on water molecules, and hence have smaller partial specific volume. Figure 9.30(b) appears to be consistent with this hypothesis. Namely, those salts with a larger surface tension increment possess higher levels of hydration, thus have smaller partial specific volume, which can be seen as an inverse relationship between the two parameters. If the surface tension effect is responsible for the observed hydration, then those salts with higher surface tension increment should exhibit greater levels of hydration. This is clearly shown in Figure 9.30(c). Except for the divalent cation salts, a strong correlation exists between surface tension and preferential hydration; those with higher surface tension increment are more strongly excluded. This suggests that the observed exclusion of salts (negative binding) is due to the ionic hydration mechanism. Those salts that are weakly water binding can penetrate the protein hydration layer and bind to the protein surface. Figure 9.30(d) plots the preferential hydration against molecular weight: excluded volume is the product of molecular weight and the van der Waals volume of salts [365]. There is no correlation between the molecular weight and preferential hydration, suggesting that salt exclusion is not due to the excluded volume effect.

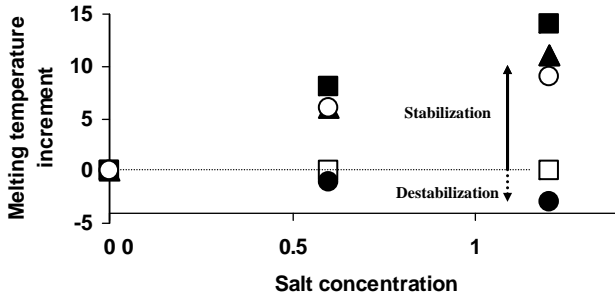
The results of two divalent cation salts reveal that they use different mechanisms to interact with the protein. Such a deviation from the surface tension mechanism is observed for other proteins as well. In essence, these salts tend to weakly bind to proteins, suggesting that their water binding properties (that cause the exclusion) are offset by their affinity for the protein surface: note that both of these salts can bind water molecules, as shown by the large surface tension increment ( $\sigma$ ) and small partial specific volume (Figure 9.30a and b). What is the cause of this affinity? Although there is no clear explanation, it has been shown that  $\text{MgCl}_2$  increases the solubility of peptides and nonpolar compounds [366, 367]. We compare  $\text{MgCl}_2$  with  $\text{MgSO}_4$  for their interaction with model compounds, as such comparison may help explain the reason why  $\text{MgCl}_2$  has some affinity for protein surfaces. In Figure 9.31, a comparison of the effects of the two salts on the solubility of glycine, tryptophan, and leucine is shown [249]. For comparison, the scales are set identical for the two panels. The solubility of glycine (open circles) is more effectively increased by  $\text{MgCl}_2$  (see longer arrow), indicating a more favorable interaction of this salt with glycine, in comparison to  $\text{MgSO}_4$ .  $\text{MgCl}_2$  interacts favorably with the dipolar compounds or groups, such as glycine. As shown by closed symbols, the solubility of tryptophan and glycine is changed very little in  $\text{MgCl}_2$ , while it is greatly reduced in  $\text{MgSO}_4$ . The difference in solubility between glycine and other amino acids is due to the side-chain. This difference between glycine and tryptophan or leucine is illustrated by the dotted arrows. The downward arrow indicates an unfavorable interaction of these side-chains with  $\text{MgCl}_2$  and  $\text{MgSO}_4$ , suggesting that these two divalent cation salts interact unfavorably with the aromatic and hydrophobic side-chains. What is important, however, is the difference between the two salts; the unfavorable interaction is much weaker for  $\text{MgCl}_2$  (compare the length of two dotted arrows). Although the reason is unclear, the difference in their observed effects on glycine and the two side-chains should be reflected on their interaction with the protein, and



**Figure 9.31** Effects of MgCl<sub>2</sub> (a) and MgSO<sub>4</sub> (b) on the solubility of glycine, tryptophan, and leucine. (Data reformatted from [249].)

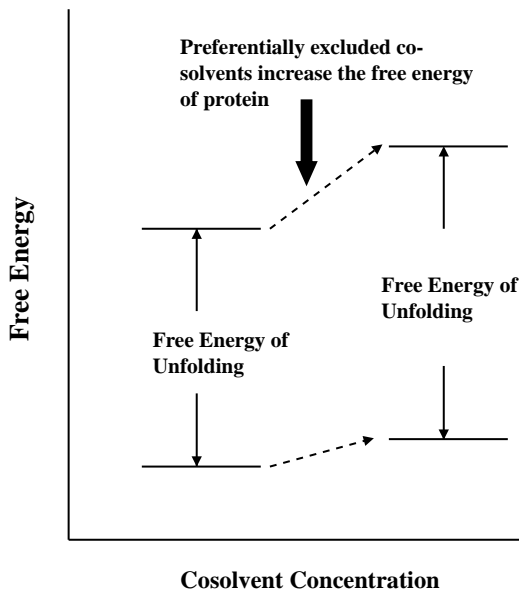
ultimately on the protein reaction. At this point, Cohn's observation should be recalled (Figure 9.22), in which CaCl<sub>2</sub> behaved differently from the other salts in altering the solubility of cystine (strong salting-in), which is consistent with the observed preferential hydration of BSA.

It may also be worth comparing the effects of these two salts on the stability of proteins, in relation to the discussion specific for these two salts. Figure 9.32 illustrates the effects of MgCl<sub>2</sub> and MgSO<sub>4</sub> on the stability of three proteins as a function of concentration [136]. It is evident from Figure 9.32 that while MgSO<sub>4</sub> has a stabilizing effect on the proteins, MgCl<sub>2</sub> is either slightly destabilizing or ineffective: melting temperatures increase by 10–15 °C at 1.2 M MgSO<sub>4</sub>. The effects of these salts on protein stability can be explained by the difference in interaction between the native and thermally unfolded states under identical conditions. As described in Figure 9.29, preferential interaction is known only for proteins in their native state; MgSO<sub>4</sub> is strongly excluded while MgCl<sub>2</sub> is unaffected. How do these salts interact with the unfolded structure? As shown in Figure 9.30, the exclusion of MgSO<sub>4</sub> is due to its hydration potential, and thus may be larger for the unfolded state. This situation is schematically illustrated in Figure 9.33. Preferential cosolvent exclusion is related to the increase in the free energy of macromolecules. Thus, the free energy of the native protein increases by the addition of MgSO<sub>4</sub>. It is inferred from the interaction



**Figure 9.32** Effects of  $\text{MgCl}_2$  and  $\text{MgSO}_4$  on the melting temperature of proteins, lysozyme, ribonuclease, and BSA: (■) lysozyme in  $\text{MgSO}_4$ , (▲) ribonuclease in  $\text{MgSO}_4$ , (○) BSA in  $\text{MgSO}_4$ , (□) lysozyme or ribonuclease in  $\text{MgCl}_2$ , and (●) BSA in  $\text{MgCl}_2$ . (Data reformatted from [136].)

mechanism that the free energy of the unfolded protein increases to a greater extent than that of the native protein in the presence of  $\text{MgSO}_4$ , which leads to a larger unfolding free energy. The nearly neutral interaction of  $\text{MgCl}_2$  with the native protein may suggest that its interaction with the unfolded structure is also neutral. Group interaction studies suggest a favorable interaction with polar groups and slightly unfavorable interaction with the nonpolar groups. The exposure of internal side-chains and peptides in the unfolded structure to the bulk solution may cause



**Figure 9.33** Free energy diagram of protein interacting with cosolvents.

Various cation-anion combination			
Cation	Anion		
	Cl <sup>-</sup>	CH <sub>3</sub> COO <sup>-</sup>	SO <sub>4</sub> <sup>-</sup>
Na <sup>+</sup>	++	+++	+++++
Mg <sup>2+</sup> , Ca <sup>2+</sup> , Ba <sup>2+</sup>	+/-	++	++++
Gdn <sup>+</sup>	--	+	++

**Figure 9.34** Preferential hydration of different cation/anion salt combinations in a concentration range of 0.5–1 M. Each symbol (+ or –) corresponds to about 0.1 g/g hydration.

cancellation of the interaction of MgCl<sub>2</sub> with the unfolded structure, leading to a neutral interaction, as observed in the preferential interaction measurements of the native protein.

In an earlier section, MgCl<sub>2</sub> was described as an effective elution cosolvent for affinity chromatography. If this salt is also neutral for both the associated and dissociated states of proteins, then it should have no effect on protein binding. It is likely that such affinity interactions between the protein and affinity resin is group (site) specific, suggesting that local environments are involved in binding. If a greater contribution of polar interaction is involved, MgCl<sub>2</sub> may have dissociating effects on specific protein binding.

A more comprehensive summary of preferential interaction data for salts is shown in Figure 9.34, which shows the qualitative range of preferential hydration of proteins in the presence of 0.5–1 M salts in different combinations. One symbol (either + or –) corresponds to about 0.1 g/g hydration. When sodium is used as a cation, all of the salt forms demonstrate preferential hydration, the degree of which increases in the order of Cl < CH<sub>3</sub>COO < SO<sub>4</sub>. Preferential hydration is consistent with the surface tension effect along with their salting-out properties. Divalent cation salts demonstrate a lower level of hydration compared to the same salts with the sodium counterion, indicating that the affinity of divalent cations reduces the overall exclusion of the salts. This is further enhanced with guanidinium salts; with GdnHCl, the overall interaction is negative hydration, thus positive binding [357].

Strong stabilization of KGF by electrolytes has been described previously in the formulation section, an effect that cannot be explained simply from the preferential exclusion mechanism. Furthermore, the data shown for certain globular proteins may not apply to any other proteins. It has been shown that β-lactoglobulin has a strong dipole moment and thereby a tendency to bind NaCl [368]. KGF is a highly basic protein, thus it may deviate in salt binding from the other proteins used for preferential interaction measurements. This suggests that actual measurements may be required to understand the salt effects on the stability of KGF. It can be speculated that the high charge density of the native KGF is stabilized by salt binding and the

binding capacity is reduced for the unfolded structure due to the expansion of the protein, and hence decreased charge density. In other words, salts bind preferentially to the native state of KGF, stabilizing the native structure.

Another interesting observation is the effects of sodium glutamate and Na<sub>2</sub>PIPES at high concentrations, resulting in enhanced tubulin polymerization. This can be readily explained from the observed preferential hydration of model proteins and tubulin in these solvent systems [258]. Although the mechanism of preferential hydration or exclusion is not clear in these cosolvent systems, preferential hydration clearly correlates with their salting-out effects (enhancing protein–protein interactions, see Table 9.2).

**9.5.3.1.2 Osmolytes** Understanding the stabilization mechanism is fundamental to understanding the reason for nature's selection of osmolytes to control the cellular osmotic pressure. The use of these cosolvents is also essential for stabilizing unstable proteins and/or enhancing protein folding. As described earlier, trehalose and sucrose may be among the most widely used cosolvents for pharmaceutical formulations, both in liquid and dried forms. Typically, glycine and mannitol are used as bulking and/or crystallizing cosolvents for freeze-dried formulations. In any case, they are commonly used in research applications that require purified proteins to maintain their functional activities. Numerous preferential interaction measurements have been performed for polyols, sugars, amino acids, and amino acid derivatives [318, 369–374]. In summary, they are all excluded from the native protein structure to a varying degree, depending on the osmolytes and proteins used. While the mechanism of osmolytes' preferential hydration is primarily due to their affinity for water, the excluded volume effect should also contribute, as described earlier. These preferential hydration effects were translated to their effects on enhanced protein association, stability, and folding.

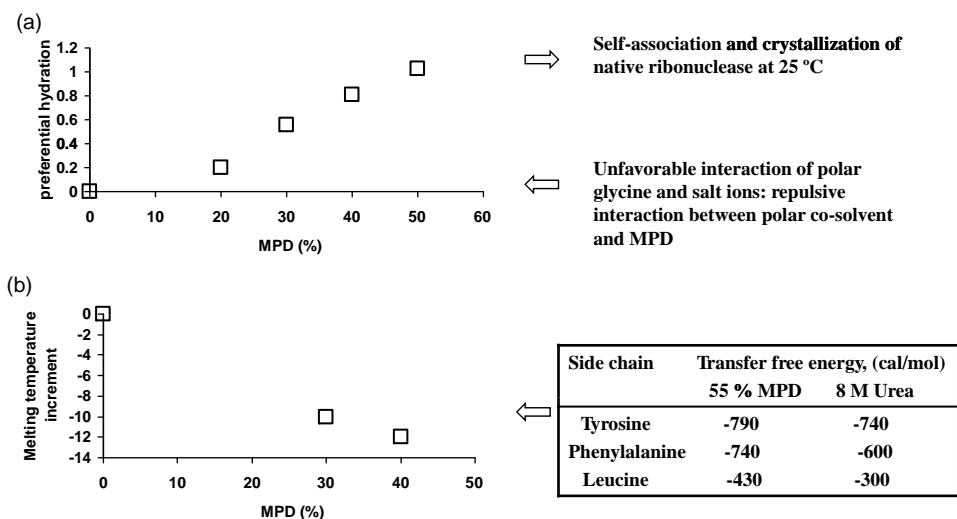
While osmolytes are useful in enhancing protein stability, they are not effective in suppressing aggregation. Rather, they can increase aggregation as schematically depicted in Figure 9.18. More specifically, unfavorable interactions are reduced upon protein aggregation and surface area reduction. For example, trehalose has been reported to suppress the aggregation of polyglutamine proteins, which are responsible for neuronal diseases [375, 376]. This observation, however, is inconsistent with the concept described above. Such aggregation suppression could possibly be an indirect consequence of protein stabilization, as the native protein should be more soluble than the unfolded protein. In this regard, arginine has been shown to be much more effective than the structure-stabilizing osmolytes. As described earlier, arginine has shown favorable interactions with most of amino acid side-chains and peptide bonds, similarly to the results for protein denaturants (i.e., urea and GdnHCl). Nevertheless, we know that arginine is not a protein denaturant. In practice, this affords an advantage to arginine over denaturants in suppressing protein aggregation. Furthermore, arginine is a natural metabolite. More importantly, the difference between arginine and denaturants, in terms of their effects on small molecules and large proteins, indicates that the effects of cosolvent may not always be deduced from their effects on the solubility of model compounds. In other

words, interaction with small molecules may not be linearly translated to their effects on large molecules. In fact, arginine has been shown to have binding affinity for the native protein surface [319]. Trout *et al.* [377–379] introduced a concept of a neutral crowder to explain the aggregation suppressive effect of arginine, which takes into consideration the weak binding and large excluded volume of arginine.

**9.5.3.1.3 Organic Solvents** Organic solvents have been used for denaturation of proteins as well as their precipitation. It has been shown above that the solubility measurements of polar and ionic compounds with organic solvents are highly unfavorable, which can cause phase separation and precipitation (Figure 9.24). In addition, organic solvents increase the solubility of nonpolar compounds, indicating that they interact favorably with the hydrophobic side-chains, and hence with the exposed surface of the unfolded protein, leading to protein denaturation (Figure 9.25). Preferential interaction measurements enforce this analysis. Strongly denaturing 2-chloroethanol demonstrated preferential binding even at concentrations of 5–20%, which was sufficient to unfold the proteins [359, 360, 380]. The unfolded structure contained more  $\alpha$ -helix content compared to the native protein, which cannot be readily explained from the preferential binding of this organic solvent. Rather, this may be due to the unfavorable interaction of peptide bonds with organic solvents, shown for other organic solvents [134, 354]. More specifically, such unfavorable interactions may cause the peptides to be buried within the internal structure, forming hydrogen bonds with other peptide sequences, and leading to the formation of  $\alpha$ -helices.

How is protein precipitation enhanced from preferential interaction? Although it is not clear whether 2-chloroethanol causes protein precipitation, it does phase-separate at 50% concentration by the addition of strong salting-out salts, such as  $\text{Na}_2\text{SO}_4$  [39]. Thus, it is possible for the highly charged proteins to precipitate upon the addition of this organic solvent. The most drastic precipitating effect was observed with MPD [39]; ribonuclease was preferentially hydrated in the presence of MPD, as shown in Figure 9.35(a). Excess water increased to 1 g/g, indicating that the protein is surrounded by, on average, 1 g of water per g protein. This is due to the highly unfavorable interaction of MPD with polar groups and ions, as summarized in Figure 9.35. Such unfavorable interactions (i.e., repulsive interactions) of MPD led to precipitation and crystallization of the protein. In fact, MPD has been used as a strong crystallizing agent [134, 135]; with the occurrence of this mutual repulsion, the crystalline protein is in the native structure, as demonstrated by the X-ray analysis of the crystal [134, 135]. Although MPD is repelled from the surface of native proteins, it does destabilize the protein [136]. If the increased surface area, caused by unfolding, enhances repulsion, then MPD should stabilize the protein according to the thermodynamic linkage function (Figure 9.18). This, however, is inconsistent with the observation. Figure 9.35(b) shows a decrease in melting temperature with increasing MPD concentration. This destabilizing effect suggests that MPD is excluded less from, or binds to, the unfolded structure. The transfer free energy associated with buffer exchange from water to 55% MPD solution is negative for tyrosine, phenylalanine, and leucine side-chains, implicating the presence of a





**Figure 9.35** Preferential hydration (a) and melting temperature (b) of protein (ribonuclease) in MPD solution. (Data adapted from [39, 134].)

favorable interaction between MPD and these hydrophobic groups. Thus, MPD may still be excluded from the unfolded structure due to charge repulsion. Possible binding to the unfolded state should offset the exclusion effect and the overall repulsive interaction should decrease upon unfolding, leading to destabilization of protein structure by MPD. Acetonitrile behaves similarly; through strong exclusion and hydrophobic binding [381].

Milder organic solvents possess properties that are between those for the two solvents described above. For example, ethanol and DMSO are preferentially excluded at low concentrations, demonstrating salting-out effects and being preferentially bound, leading to protein unfolding at higher concentrations [41]. These effects are interpreted in a similar fashion to those observed for MPD (i.e., unfavorable repulsion of organic solvents from polar compounds and favorable binding to hydrophobic groups).

**9.5.3.1.4 PEG** Polymers are excluded from the protein surface. Preferential exclusion of PEG has been observed for several proteins with PEG100 (average molecular weight of 100) to PEG6000 [179]. This exclusion was ascribed to the excluded volume of PEG. Assuming that the preferential hydration is entirely due to the excluded volume effect (i.e.,  $(\partial g_1 / \partial g_2) = A_1$ ), it would be possible to calculate the hydrodynamic radius of the excluded polymer. This was done for PEG400, PEG600, and PEG1000. The radius calculated from preferential hydration is in qualitative agreement with the theoretical hydrodynamic radius obtained from the bond length and the number of PEG (Table 9.9). Considering the number of assumptions made in the calculations (e.g., same randomness of different PEG), agreement appears to be

**Table 9.9** Calculation of hydrodynamic radius of PEG.

PEG molecular weight	Radius from preferential hydration from theory (Å) <sup>a)</sup>	
400	2.8	4.5
600	4.6	6.0
1000	7.4	7.5

Data adapted from [79].

- a) Calculated from  $(2\sigma l^2/6)^{1/2}$ , for which  $\sigma$  is the number of bonds and  $l$  is the bond length (assuming an identical bond length) [429].

reasonable, suggesting that the observed preferential hydration is, in fact, ascribed to the excluded volume effect.

Lee and Lee [382, 383] reported that the addition of PEG1000, PEG4000, and PEG6000 enhanced the binding constant of tubulin to growing microtubules. From the concentration dependence of the binding constant, they obtained an increase in PEG binding in mol/mol, upon polymerization, to be 1 for PEG1000, 0.7 for PEG4000, and 0.5 for PEG6000. This does not imply that there are more PEG binding sites on the microtubules, but rather that PEG is less preferentially excluded from the microtubules than the monomeric tubulin, as expected from the different surface area per protein monomer.

However, PEG is not an inert polymer for the protein surface. While PEG appears to have no affinity for the native protein surface (as the observed interaction with the native state is due entirely to exclusion, i.e., no binding contribution), it has affinity for the unfolded protein surface. Furthermore, PEG demonstrates affinity for hydrophobic groups, as described earlier. Thus, the preferential interaction of PEG for the unfolded protein has both contributions from excluded volume and hydrophobic binding. The hydrophobic binding contribution resulted in protein destabilization, as is observed with decrease in melting temperature [79]. In this regard, PEG acts much like organic solvents. Both cosolvents cause precipitation of native proteins due to repulsive interactions, although via a different mechanism (excluded volume effect for PEG and charge repulsion for organic solvents), and destabilize proteins via hydrophobic binding.

## 9.6

### Protein–Solvent Interactions in Frozen and Freeze-Dried Systems

#### 9.6.1

##### Frozen Systems

Many purified biological macromolecules are inherently unstable and thus are stored at low temperatures to slow their degradation. However, low-temperature storage is

not without complication. Proteins can undergo low-temperature denaturation and even further damage upon freezing. One of the key variables to control during freezing is the cooling rate. The ice crystal size, hence the surface area, depends highly on the cooling rate; the higher the ice crystal surface area, the greater the likelihood of surface denaturation. Furthermore, with the progression of water crystallization, the remaining unfrozen solution containing the protein, buffer salts, as well as any cosolvents, increases in concentration. This may result in protein aggregation [384–388], toxic effects from the high excipient concentration [389], as well as in buffer crystallization, leading to pH shift [390, 391]. To minimize these deleterious effects, stabilizers are included in the system prior to freezing.

Preferential interaction has been described previously as a mechanism to differentiate protein stabilizers from destabilizers in aqueous systems. The same mechanism can be used to classify the excipients in frozen systems (i.e., stabilizers are excluded from the protein surface, whereas destabilizers are bound to the surface). For example, sucrose and glucose (at 1 M), both of which are preferentially excluded from the surface of proteins in aqueous systems, were demonstrated to be effective in stabilizing LDH following freeze-thawing (Table 9.10) [256]. Furthermore, sucrose was demonstrated to enhance the stability of LDH following freeze-thaw in a sodium phosphate buffer, which in the absence of the sugar, resulted in enzyme instability caused by a large shift in pH [256]. Several other compounds that are preferentially excluded from the protein surface, including polyols (i.e., glycerol, sorbitol, etc.), amino acids (i.e., proline, sodium glutamate, etc.), methylamines (i.e., TMAO, sarcosine, etc.), and inorganic salts (sodium acetate,  $\text{MgSO}_4$ ,  $\text{Na}_2\text{SO}_4$ , etc.), were also shown to stabilize LDH following freeze-thaw (Table 9.10) [256].

Just as stabilizers in aqueous systems were demonstrated to be stabilizing in the frozen state, destabilizers in solution are also destabilizing in the frozen state.

**Table 9.10** Comparison of LDH activity recovered following freeze-thaw in the presence of various cosolutes at either 0.5 or 1.0 M concentration (in the absence of any co solutes, 20% of the initial activity was recovered following freeze-thawing).

Cosolutes	Cosolute concentration	
	0.5 M Recovered activity (%)	1 M Recovered activity (%)
Sucrose	65	82
Glucose	50	60
Glycerol	65	70
Sorbitol	70	75
Proline	65	70
TMAO	65	65
Sodium acetate	65	65
NaCl	0	15

Data adapted from [256].

GdnHCl and urea, both of which bind preferentially to the protein surface in solution [357, 392], were shown to destabilize LDH following freeze-thaw; 200 mM urea reduced the activity of LDH to below 27% of the initial activity, while no measurable activity was observed in 80 mM GdnHCl [256]. It is worth noting that a much lower concentration of these two compounds was required for destabilizing LDH (below 1 M) compared to the concentration required for stabilization by the preferentially excluded compounds (generally 1 M or above). Destabilization caused by binding suggests the presence of strong interactions between the compound and the protein. Stabilization, on the other hand, is expected to occur through both nonspecific and weak interactions, thus requiring a much higher concentration of the compound for the effects to be observable. As water is still present, the mechanism of cosolvent effects that operate in solution is still valid in the frozen state.

There are several exceptions to the observed parallel behavior of compounds in the aqueous and frozen systems. Although several other inorganic salts demonstrated a stabilizing effect on LDH following freeze-thaw, NaCl destabilized the enzyme and only marginally increased the activity upon increasing its concentration; at 0.5 M or lower concentration, LDH was completely inactivated, while increasing the NaCl concentration to 1 M resulted in only 15% of the initial activity being recovered (Table 9.10) [256]. Both PEG and MPD stabilize proteins following freeze-thaw; however, they have been shown to destabilize proteins in solution, although they are preferentially excluded. This apparently contradictory effect has been explained by the temperature dependence of hydrophobic interactions between proteins and these compounds [393]. Furthermore, ethylene glycol has been shown to demonstrate protein-specific stabilization; both  $\beta$ -lactoglobulin [380] and lysozyme [394] have been demonstrated to be destabilized with increasing concentration of ethylene glycol, while collagen was stabilized [370]. Thus, there appears to be more protein-specific stabilization mechanisms in the frozen state than in solution.

Similarly to proteins, viruses are sensitive to freezing conditions and their stability can be influenced by the addition of appropriate cosolvents. In solution, BDMV has been reported to release its RNA upon increasing the pH to alkaline conditions. Upon the incorporation of polyamines and cations, this release was prevented [395, 396], which was postulated to occur through the stabilization of the virus shell, the protein capsid. Upon freezing and thawing, the formation of empty capsids were observed (i.e., RNA released); however, this was prevented through the addition of cosolvents and a change in ionic strength [257]. BDMV subjected to freeze-thaw in 5 mM Tris-HCl buffer resulted in the formation of 100% empty capsids, while at 50 mM buffer concentration, the fraction of empty capsids was reduced to 56%. The percentage of empty capsids was further reduced to 20% in 100 mM Tris-HCl [257]. The addition of chloride salts, even at a low concentration of 40 mM, was shown to have an effect on the stabilization of BDMV, in the order of  $\text{Mg}^{2+} > \text{Na}^+ > \text{Ca}^{2+} \sim \text{Li}^+ \sim \text{Cs}^+ > \text{K}^+$ , with  $\text{MgCl}_2$  demonstrating 15% empty capsids, while KCl resulted in 85% empty capsids. As noted earlier, due to the freezing of water in the buffer system, the seemingly low salt concentration is actually rather high in the frozen system. Thus, to preserve the integrity of nonenveloped viruses in the frozen system, it appears that the maintenance of the protein capsid, or more specifically the

protein–protein interaction is essential. Those cosolvents that stabilize the protein–protein interaction (as well as protein structure) to numerous stresses associated with freezing (i.e., low-temperature denaturation, ice crystal formation, pH shift, etc.) would most likely confer structural stability to viruses. For an enveloped virus, such as HSV-2, freezing without appropriate protectants resulted in fusion of viral particles [397]. Thus, similarly to proteins, virus stability in the frozen system is dependent on the ability of cosolvents to confer structural stability and to reduce the degree of physical and chemical degradation (i.e., aggregation). Furthermore, the cosolvents are expected to stabilize viruses in the same manner as they stabilize proteins (i.e., preferential interaction).

### 9.6.2

#### Freeze-Dried System

Typically, degradation rates are decreased by lowering the storage temperature. Thus, the stability of purified biological macromolecules can be enhanced by storing them under refrigerated or frozen conditions. However, certain biological macromolecules may still demonstrate instability under these storage conditions. These samples may be freeze-dried in efforts to enhance their stability, by restricting their molecular mobility and reducing the rate of physical and chemical degradation.

Freeze-drying involves freezing of the protein solution, followed by the removal of crystallized water by sublimation. As described above, the interaction between cosolvents and proteins occurs in a similar fashion to that observed in aqueous solutions during the freezing process, but is completely different in the dried state from that in the solution state. The water removed by sublimation (primary drying) is generally the “bulk” water, and the water molecules comprising the hydration shell of the protein can be removed by additional drying conducted at an elevated temperature (secondary drying). Furthermore, the removal of water, particularly comprising the hydration shell of the protein, can cause additional damage [398–400]. This removal of water invalidates the use of preferential hydration or the exclusion mechanism in dried systems.

In an aqueous solution, water plays a critical role in hydrating the protein surface, whereas in its absence, proteins may aggregate or form nonnative intramolecular interactions to minimize their surface area. Stabilizers such as sugars have been proposed to cover the surface of proteins and membranes in the dried state *in lieu* of water molecules [401–404]. However, the protection conferred by the stabilizers can be neutralized if a sufficient amount of moisture is not removed, as water acts as a plasticizer (i.e., lowering the  $T_g$ ), conferring mobility to the labile biological molecule, potentially enhancing the rate of degradation [405–407].

In general, glass-forming stabilizers, such as saccharides and polymers, are included in the formulation buffered with noncrystallizing salts. Saccharides have been studied extensively to understand their efficacy in conferring stability to dehydrated biomolecules [403, 408–410]. Focus has been placed primarily on sucrose and trehalose, both of which have been found to accumulate in high concentrations in desiccation-tolerant organisms, such as tardigrades, cysts of *Artemia salina* (brine

**Table 9.11** Effect of trehalose concentration on the amide II band of freeze-dried lysozyme.

Trehalose concentration (mg/ml)	Wave number of amide II ( $\text{cm}^{-1}$ )
Hydrated protein	1543
0	1530
25	1541
50	1542
100	1542
150	1542
200	1541
300	1536
400	1531

The amide II band of hydrated lysozyme,  $1543 \text{ cm}^{-1}$ , is included as reference. Upon drying, the amide II band of lysozyme decreases, suggesting a difference in the structure or conformation of the protein. Upon inclusion of trehalose, the amide II band shifts back to the value observed in the hydrated state, suggesting the existence of direct interaction of trehalose with lysozyme. Increasing the trehalose concentration too high (400 mg/ml), however, may result in sugar crystallization, decreasing the amount of trehalose available to bind to lysozyme. This results in amide II band decrease. Data adapted from [401].

shrimp), and resurrection plants [334, 411–414]. The stabilization mechanism of sugars is quite interesting because in solution, the sugars confer stability to proteins by *not* directly interacting with the molecule (preferential exclusion or hydration) [369, 374], while in the dried state, the sugars are directly interacting with the surface of molecules, through hydrogen bonding (water replacement) [254, 401]. This is evidenced by comparing the Fourier transform IR (FTIR) spectrum of hydrated lysozyme to that for freeze-dried lysozyme, with and without trehalose [401]. Upon freeze-drying lysozyme, the amide II band broadened and shifted from  $1543$  to  $1530 \text{ cm}^{-1}$ , while lysozyme freeze-dried in the presence of trehalose (100 mg/ml) maintained the amide II band at  $1542 \text{ cm}^{-1}$  (Table 9.11). Thus, lysozyme freeze-dried in the presence of trehalose demonstrated a similar spectrum to that observed in the hydrated state. Alternatively, the FTIR spectrum of freeze-dried trehalose in the presence of lysozyme was shown to be similar to that of hydrated trehalose. Both spectra were qualitatively different from that of freeze-dried trehalose (alone) or crystalline trehalose. These observations suggest that trehalose can replace water in the hydrogen bonding network of lysozyme upon freeze-drying and, similarly, lysozyme can effectively substitute for water to hydrogen bond to trehalose. However, the ratio of sugar to protein must be optimized. An insufficient amount of sugar can lead to ineffective replacement of hydrogen bonds, while too high a concentration can lead to crystallization during sublimation, thus resulting in a decreased amount of sugar available for hydrogen bonding (Table 9.11, [401]). Insufficient replacement of hydrogen bonds by the sugar during dehydration not only leads to change in protein conformation, but also in decreased activity, as was reported for freeze-dried phosphofructokinase (PFK) [401]. In the absence of trehalose, no PFK activity was

**Table 9.12** Effect of trehalose concentration on the activity of PFK recovered following freeze-drying.

Trehalose concentration (mg/ml)	Recovered activity (%)
0	0
30	40
75	42
110	50
150	45
225	32
300	20
400	0

No activity was recovered if PFK was freeze-dried on its own. A steady level of activity was regained in the presence of trehalose, ranging from 30 to 150 mg/ml beyond which, the recovered activity decreased monotonically. No activity was observed for PFK freeze-dried at 400 mg/ml trehalose. Data adapted from [401].

detected. Upon the inclusion of 30 mg/ml trehalose, approximately 40% of the initial activity was regained, which remained constant up until 150 mg/ml trehalose, beyond which the recovered activity decreased monotonically until no activity was observed at 400 mg/ml trehalose (Table 9.12). This observation can again be explained by the effective replacement of hydrogen bonds between water and protein by trehalose during dehydration, which allows the maintenance of protein structure, and thus activity. At high trehalose concentrations (i.e., 400 mg/ml), the sugar crystallized, leaving an insufficient amount of trehalose to interact with the protein, leading to inactivation.

Furthermore, the sugars form an amorphous glassy matrix upon dehydration, serving to restrict the mobility of the biomolecule [415–417], thereby retarding the rate of chemical degradation, and to maintain the separation between the biomolecules, which in the absence of the sugars may have aggregated [418, 419]. This property of dried sugars is essential for the long-term stability of chemically unstable proteins.

Koster *et al.* [420] have reported that steric effects also play a significant role in the ability of certain saccharides and polymers to confer stability to lipid membranes. Several saccharides and polymers, varying in their dry  $T_g$  values as well as in their molecular size, were incorporated into a model cell membrane composed of phosphatidylcholines and subsequently dried. Interestingly, excipients with high  $T_g$  values, such as dextran and polyvinylpyrrolidone, were not always effective in stabilizing the lipid membrane. Koster *et al.* concluded that the cosolvents are ineffective in conferring stability to the lipid membranes because they are sterically hindered from interacting with the phospholipid head-groups. Thus, simply choosing stabilizers based on their  $T_g$  values may lead to a suboptimal formulation for lipid membranes and potentially for enveloped viruses.

As mentioned previously, the major change leading to viral inactivation is the loss of viral envelope. In order to maintain the structural integrity of viral envelopes,

similar cosolvents to those used for proteins, are included to confer stability through a variety of mechanisms. In the case of HSV-2, lyophilization in a concentrated sugar solution (27% trehalose) resulted in significantly less damage compared to those lyophilized in a dilute sugar solution (0.25% trehalose), as assessed by viral infectivity measurement and particle morphology analysis using transmission electron microscopy [421]. The most likely explanation for this observation is the water replacement hypothesis; the sugar molecules are acting as water substitutes in the dried system through direct interaction with the protein and possibly with the lipids [422, 423]. In a dilute sugar system, sufficient degree of substitution is not feasible, leading to virus inactivation. Furthermore, in a hyperosmotic medium, the internal water of viruses will dialyze out, leading to the concentration of internal contents of virus particles, including proteins. This may lower the nucleation temperature of water within the virus particle, and thus, prevent ice crystal formation and growth. This latter mechanism points to the cosolvent effects on viruses that distinguish them from those on proteins.

The residual water content of lyophilized viruses also plays a role in determining their storage stability. Lyophilized varicella zoster virus maintaining moderate levels of residual water content demonstrated improved storage stability compared to those that were dried more efficiently [266, 424]. This can be achieved by modifying the lyophilization cycle or through the incorporation of cosolvents, such as sugars that demonstrate strong water-binding properties. All of these effects will lead to an effective stabilization of virus particles, similarly to those observed for proteins, although there are subtle differences in cosolvent effects between proteins and viruses.

## 9.7

### Conclusions

Cosolvents play critical roles in the study and development of proteins, viruses, and DNA. Cosolvents are classified as stabilizers and destabilizers, or folding enhancers and aggregation suppressors; protein stabilizers enhance protein folding and protein destabilizers suppress aggregation, with few exceptions such as arginine. Such classification is generally applicable for proteins and DNA, but not necessarily for viruses due to their complex structure. Thermodynamic interactions of cosolvents with model compounds can clearly explain the effects of cosolvents on proteins, DNA, and, to a varying degree, viruses in solution and the frozen state. Cosolvents that interact unfavorably with the native structure enhance aggregation and stability, while other cosolvents that interact favorably with macromolecules enhance dissociation and often cause instability. There are few exceptions; however, as an example, organic compounds that strongly enhance aggregation are ineffective as a stabilizer. In the dried state, the thermodynamic interactions of cosolvents cannot explain the cosolvent effects, as there is little water. Instead, the direct binding of protecting cosolvents and their physical state determine the overall effects on the stability of proteins and viruses.



## References

- 1 Minton, A.P. (1981) *Biopolymers*, **20**, 2093–2120.
- 2 Ignatova, Z. and Gierasch, L.M. (2007) *Methods in Enzymology*, **428**, 353–372.
- 3 Perham, M., Stagg, L., and Wittung-Stafshede, P. (2007) *FEBS Letters*, **581**, 5065–5069.
- 4 Zhou, H.-X. (2008) *Archives of Biochemistry and Biophysics*, **469**, 76–82.
- 5 Mukherjee, S., Waegle, M.M., Chowdhury, P., Guo, L., and Gai, F. (2009) *Journal of Molecular Biology*, **393**, 227–236.
- 6 McGuffee, S.R. and Elcock, A.H. (2010) *PLoS Computational Biology*, **6**, e1000694.
- 7 Aune, K.C. and Tanford, C. (1969) *Biochemistry*, **8**, 4586–4590.
- 8 Biltonen, R.L. and Lumry, R. (1969) *Journal of American Chemical Society*, **91**, 4256–4264.
- 9 Privalov, P.L. and Khechinashvili, N.N. (1974) *Journal of Molecular Biology*, **86**, 665–684.
- 10 Pace, C.N. (1990) *Trends in Biochemical Sciences*, **15**, 14–17.
- 11 Huang, Y. and Liu, Z. (2009) *Journal of Molecular Biology*, **393**, 1143–1159.
- 12 Uversky, V.N. (2002) *European Journal of Biochemistry*, **269**, 2–12.
- 13 Uversky, V.N. (2002) *Protein Science*, **11**, 739–756.
- 14 Fink, A.L. (2005) *Current Opinion in Structural Biology*, **15**, 35–41.
- 15 Minton, A.P. (2005) *Biophysical Journal*, **88**, 971–985.
- 16 Minton, A.P. (2005) *Journal of Pharmaceutical Sciences*, **94**, 1668–1675.
- 17 Griko, Y.V. and Remeta, D.P. (1999) *Protein Science*, **8**, 554–561.
- 18 Jaenicke, R. (2000) *Journal of Biotechnology*, **79**, 193–203.
- 19 Wyman, J. Jr. (1964) *Advances in Protein Chemistry*, **19**, 223–286.
- 20 Wyman, J.G. and Gill, S.J. (1990) *Binding and Linkage: Functional Chemistry of Biological Macromolecules*, University Science Books, Mill Valley, CA.
- 21 Tsai, P.K., Volkin, D.B., Dabora, J.M., Thompson, K.C., Bruner, M.W., Gress, J.O., Matuszewska, B., Keogan, M., Bondi, J.V., and Middaugh, C.R. (1993) *Pharmaceutical Research*, **10**, 649–659.
- 22 Kretschmar, M., Mayr, E.M., and Jaenicke, R. (1999) *Journal of Molecular Biology*, **289**, 701–705.
- 23 Weisenberg, R.C. and Deery, W.J. (1976) *Nature*, **263**, 792–793.
- 24 Weisenberg, R.C. and Timasheff, S.N. (1970) *Biochemistry*, **9**, 4110–4116.
- 25 Murphy, D.B. and Borisy, G.G. (1975) *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 2696–2700.
- 26 Sloboda, R.D., Dentler, W.L., and Rosenbaum, J.L. (1976) *Biochemistry*, **15**, 4497–4505.
- 27 Lee, J.C. and Timasheff, S.N. (1975) *Biochemistry*, **14**, 5183–5187.
- 28 Frigon, R.P. and Timasheff, S.N. (1975) *Biochemistry*, **14**, 4559–4566.
- 29 Lee, J.C., Frigon, R.P., and Timasheff, S.N. (1973) *Journal of Biological Chemistry*, **248**, 7253–7262.
- 30 Lee, J.C., Frigon, R.P., and Timasheff, S.N. (1975) *Annals of the New York Academy of Sciences*, **253**, 284–291.
- 31 Lee, J.C., Tweedy, N., and Timasheff, S.N. (1978) *Biochemistry*, **17**, 2783–2790.
- 32 Green, A.A. (1932) *Journal of Biological Chemistry*, **95**, 47–66.
- 33 Hom, G.K., Lassila, J.K., Thomas, L.M., and Mayo, S.L. (2005) *Protein Science*, **14**, 1115–1119.
- 34 Cohn, E.J. (1953) *Blood Cells and Plasma Proteins: Their State in Nature* (ed. J.L. Tullis), Academic Press, New York, p. 18.
- 35 Morgenthaler, J.J. (2000) *Vox Sanguinis*, **78** (Suppl. 2), 217–221.
- 36 Okuda, M., Uemura, Y., and Tatsumi, N. (2003) *Preparative Biochemistry and Biotechnology*, **33**, 239–252.
- 37 Burnouf, T. (2007) *Transfusion Medicine Review*, **21**, 101–117.
- 38 Hamel, E., Lin, C.M., Kenney, S., and Skehan, P. (1991) *Archives of Biochemistry and Biophysics*, **286**, 57–69.

- 39 Pittz, E.P. and Timasheff, S.N. (1978) *Biochemistry*, **17**, 615–623.
- 40 Horcajada, C., Cid, E., Guinovart, J.J., Verdaguer, N., and Ferrer, J.C. (2003) *Acta Crystallographica D*, **59**, 2322–2324.
- 41 Arakawa, T., Kita, Y., and Timasheff, S.N. (2007) *Biophysical Chemistry*, **131**, 62–70.
- 42 Hjerten, S. (1973) *Journal of Chromatography*, **87**, 325–331.
- 43 Rimerman, R.A. and Hatfield, G.W. (1973) *Science*, **182**, 1268–1270.
- 44 Comings, D.E., Miguel, A.G., and Lesser, B.H. (1979) *Biochimica et Biophysica Acta*, **563** 253–260.
- 45 Doellgast, G.J. and Fishman, W.H. (1974) *Biochemical Journal*, **141**, 103–112.
- 46 Arakawa, T. (1986) *Archives of Biochemistry and Biophysics*, **248**, 101–105.
- 47 Narhi, L.O., Kita, Y., and Arakawa, T. (1989) *Analytical Biochemistry*, **182**, 266–270.
- 48 Barth, S., Huhn, M., Matthey, B., Klimka, A., Galinski, E.A., and Engert, A. (2000) *Applied and Environmental Microbiology*, **66**, 1572–1579.
- 49 Schäffner, J., Winter, J., Rudolph, R., and Schwarz, E. (2001) *Applied and Environmental Microbiology*, **67**, 3994–4000.
- 50 Cleland, J.L. and Wang, D.I. (1990) *Bio/Technology*, **8**, 1274–1278.
- 51 Buchner, J. and Rudolph, R. (1991) *Bio/Technology*, **9**, 157–162.
- 52 Rinas, U., Risse, B., Jaenicke, R., Abel, K.J., and Zettlmeissl, G. (1990) *Biological Chemistry Hoppe-Seyler*, **371**, 49–56.
- 53 Ahn, J.H., Lee, Y.P., and Rhee, J.S. (1997) *Journal of Biotechnology*, **54**, 151–160.
- 54 Arora, D. and Khanna, N. (1996) *Journal of Biotechnology*, **52**, 127–133.
- 55 Buchner, J., Pastan, I., and Brinkmann, U. (1992) *Analytical Biochemistry*, **205**, 263–270.
- 56 Chow, M.K., Amin, A.A., Fulton, K.F., Whisstock, J.C., Buckle, A.M., and Bottomley, S.P. (2006) *Protein Expression and Purification*, **46**, 166–171.
- 57 Chow, M.K., Amin, A.A., Fulton, K.F., Fernando, T., Kamau, L., Batty, C., Louca, M., Ho, S., Whisstock, J.C., Bottomley, S.P., and Buckle, A.M. (2006) *Nucleic Acids Research*, **34**, D207–212.
- 58 Yamasaki, H., Tsujimoto, K., Koyama, A.H., Ejima, D., and Arakawa, T. (2008) *Journal of Pharmaceutical Sciences*, **97**, 3067–3073.
- 59 Katsuyama, Y., Yamasaki, H., Tsujimoto, K., Koyama, A.H., Ejima, D., and Arakawa, T. (2008) *International Journal of Pharmaceutics*, **361**, 92–98.
- 60 Utsunomiya, H., Ichinose, M., Tsujimoto, K., Katsuyama, Y., Yamasaki, H., Koyama, A.H., Ejima, D., and Arakawa, T. (2009) *International Journal of Pharmaceutics*, **366**, 99–102.
- 61 Arakawa, T., Kita, Y., and Koyama, A.H. (2009) *Biotechnology Journal*, **4**, 174–178.
- 62 Patnayak, D.P., Prasad, M., Malik, Y.S., Ramakrishnan, M.A., and Goyal, S.M. (2008) *Avian Diseases*, **52**, 199–202.
- 63 Grayson, M.L., Melvani, S., Druce, J., Barr, I.G., Ballard, S.A., Johnson, P.D., Mastorakos, T., and Birch, C. (2009) *Clinical Infectious Diseases*, **48**, 285–291.
- 64 Lombardi, M.E., Ladman, B.S., Alphin, R.L., and Benson, E.R. (2008) *Avian Diseases*, **52**, 118–123.
- 65 Alphin, R.L., Johnson, K.J., Ladman, B.S., and Benson, E.R. (2009) *Poultry Science*, **88**, 1181–1185.
- 66 Rennie, P., Bowtell, P., Hull, D., Charbonneau, D., Lambkin-Williams, R., and Oxford, J. (2007) *Respiratory Research*, **8**, 38.
- 67 Okunishi, J., Okamoto, K., Nishihara, Y., Tsujitani, K., Miura, T., Matsuse, H., Yagi, T., Wada, Y., Goto, J., Seto, M., and Ikeda, M. (2010) *Yakugaku Zasshi*, **130**, 747–754.
- 68 Traube, J. (1910) *Journal of Physical Chemistry*, **14**, 452–470.
- 69 Hofmeister, F. (1888) *Archives of Experimental Pathology and Pharmacology*, **24**, 247.
- 70 Sinanoglu, O. and Abdulnur, S. (1964) *Photochemistry and Photobiology*, **3**, 333–342.

- 71 Sinanoglu, O. and Abdunur, S. (1965) *Federation of American Societies for Experimental Biology*, **24**, 512.
- 72 Melander, W. and Horvath, C. (1977) *Archives of Biochemistry and Biophysics*, **183**, 200–215.
- 73 Collins, K.D. (2006) *Biophysical Chemistry*, **119**, 271–281.
- 74 Collins, K.D. and Washabaugh, M.W. (1985) *Quarterly Review of Biophysics*, **18**, 323–422.
- 75 Cacace, M.G., Landau, E.M., and Ramsden, J.J. (1997) *Quarterly Review of Biophysics*, **30**, 241–277.
- 76 Schachman, H.K. and Lauffer, M.A. (1949) *Journal of American Chemical Society*, **71**, 536–541.
- 77 Asakura, S. and Oosawa, F. (1954) *Journal of Chemical Physics*, **22**, 1255–1256.
- 78 Minton, A.P. (1983) *Molecular and Cellular Biochemistry*, **55**, 119–140.
- 79 Arakawa, T. and Timasheff, S.N. (1985) *Biochemistry*, **24**, 6756–6762.
- 80 Lekkerkerker, H.N.W., Poon, W.C.-K., Pusey, P.N., Stroobants, A., and Warren, P.B. (1992) *Europhysics Letters*, **20**, 559–564.
- 81 Schellman, J.A. (2003) *Biophysical Journal*, **85**, 108–125.
- 82 Harano, Y. and Kinoshita, M. (2006) *Journal of Chemical Physics*, **125**, 24910–24919.
- 83 Imai, T., Harano, Y., Kinoshita, M., Kovalenko, A., and Hirata, F. (2006) *Journal of Chemical Physics*, **125**, 24911–24917.
- 84 Bull, H.B. and Breese, K. (1968) *Archives of Biochemistry and Biophysics*, **128**, 488–496.
- 85 Kuntz, I.D. (1971) *Journal of American Chemical Society*, **93**, 516–518.
- 86 Kuntz, I.D. Jr. and Kauzmann, W. (1974) *Advances in Protein Chemistry*, **28**, 239–345.
- 87 Timasheff, S.N. and Arakawa, T. (1997) *Protein Structure: A Practical Approach* (ed. D.T.E. Creighton), IRL Press, Oxford, pp. 345–364.
- 88 Frigon, R.P. and Lee, J.C. (1972) *Archives of Biochemistry and Biophysics*, **153**, 587–589.
- 89 Erickson, H.P. (1974) *Journal of Cell Biology*, **60**, 153–167.
- 90 Erickson, H.P. (1974) *Journal of Supramolecular Structure*, **2**, 393–411.
- 91 Shelanski, M.L., Gaskin, F., and Cantor, C.R. (1973) *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 765–768.
- 92 Kasai, M., Nakano, E., and Oosawa, F. (1965) *Biochimica et Biophysica Acta*, **94**, 494–503.
- 93 Lauffer, M.A. and Stevens, C.L. (1968) *Advances in Virus Research*, **13**, 1–63.
- 94 Oosawa, F. and Kasai, M. (1962) *Journal of Molecular Biology*, **4**, 10–21.
- 95 Oosawa, F. and Higashi, S. (1967) *Progress in Theoretical Biology*, **1**, 79–164.
- 96 Oosawa, F. and Asakura, S. (1975) *Thermodynamics of the Polymerization of Actin*, Academic Press, New York.
- 97 Lee, J.C. and Timasheff, S.N. (1977) *Biochemistry*, **16**, 1754–1764.
- 98 Herzog, W. and Weber, K. (1977) *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 1860–1864.
- 99 Bulinski, J.C. and Borisy, G.G. (1979) *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 293–297.
- 100 Bulinski, J.C. and Borisy, G.G. (1980) *Journal of Biological Chemistry*, **255**, 11570–11576.
- 101 Hamel, E. and Lin, C.M. (1981) *Archives of Biochemistry and Biophysics*, **209**, 29–40.
- 102 Hamel, E. and Lin, C.M. (1981) *Proceedings of the National Academy of Sciences of the United States of America*, **78**, 3368–3372.
- 103 Hamel, E., Del Campo, A.A., Lowe, M.C., and Lin, C.M. (1981) *Journal of Biological Chemistry*, **256**, 11887–11894.
- 104 Hamel, E., Del Campo, A.A., Lowe, M.C., Waxman, P.G., and Lin, C.M. (1982) *Biochemistry*, **21**, 503–509.
- 105 Bela, N. and Jencks, W.P. (1965) *Journal of American Chemical Society*, **87**, 2480–2488.
- 106 Wakabayashi, K., Hotani, H., and Asakura, S. (1969) *Biochimica et Biophysica Acta*, **175**, 195–203.

- 107 Aune, K.C. and Timasheff, S.N. (1970) *Biochemistry*, **9**, 1481–1484.
- 108 Senczuk, A.M., Klinke, R., Arakawa, T., Vedantham, G., and Yigzaw, Y. (2009) *Biotechnology and Bioengineering*, **103**, 930–935.
- 109 Durham, A.C.H. (1972) *Journal of Molecular Biology*, **67**, 289–296.
- 110 Finch, J.T., Leberman, R., Chang, Y.-S., and Klug, A. (1966) *Nature*, **212**, 349–350.
- 111 Durham, A.C.H. and Klug, A. (1972) *Journal of Molecular Biology*, **67**, 315–332.
- 112 Zaks, A. and Klibanov, A.M. (1984) *Science*, **224**, 1249–1251.
- 113 Zaks, A. and Klibanov, A.M. (1985) *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 3192–3196.
- 114 Zaks, A. and Klibanov, A.M. (1988) *Journal of Biological Chemistry*, **263**, 3194–3201.
- 115 Volkin, D.B., Staubli, A., Langer, R., and Klibanov, A.M. (1991) *Biotechnology and Bioengineering*, **37**, 843–853.
- 116 Vermue, M.H. and Tramper, J. (1995) *Pure and Applied Chemistry*, **67**, 345–373.
- 117 Douzou, P., Sireix, R., and Travers, F. (1970) *Proceedings of the National Academy of Sciences of the United States of America*, **66**, 787–792.
- 118 Douzou, P. (1971) *Biochimie*, **53**, 17–23.
- 119 Debey, P., Balny, C., and Douzou, P. (1973) *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 2633–2636.
- 120 Greeves, M.A. and Fink, A.L. (1980) *Journal of Biological Chemistry*, **255**, 3248–3250.
- 121 Fink, A.L. (1974) *Journal of Biological Chemistry*, **249**, 5027–5032.
- 122 Nakano, T. and Fink, A.L. (1990) *Journal of Biological Chemistry*, **265**, 12356–12362.
- 123 Kuwajima, K., Yamaya, H., and Sugai, S. (1996) *Journal of Molecular Biology*, **264**, 806–822.
- 124 Chaudhuri, T.K., Arai, M., Terada, T.P., Ikura, T., and Kuwajima, K. (2000) *Biochemistry*, **39**, 15643–15651.
- 125 Schanda, P., Forge, V., and Brutscher, B. (2007) *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 11257–11262.
- 126 Nishimura, C., Dyson, H.J., and Wright, P.E. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4765–4770.
- 127 Nishimura, C., Dyson, H.J., and Wright, P.E. (2010) *Journal of Molecular Biology*, **396**, 1319–1328.
- 128 Cohn, E.J., Strong, L.E., Hughes, W.L., Mulford, D.J., Ashworth, J.N., Melin, M., and Taylor, H.L. (1946) *Journal of the American Chemical Society*, **68**, 459–475.
- 129 Kistler, P. and Nitschmann, H. (1962) *Vox Sanguinis*, **7**, 414–424.
- 130 Burnouf, T. (1991) *Bioseparation*, **1**, 383–396.
- 131 Burnouf, T. and Radosevich, M. (2001) *Journal of Biochemical and Biophysical Methods*, **49**, 575–586.
- 132 Mellanby, J. (1907) *Journal of Physiology*, **36**, 288–333.
- 133 Nydegger, U. (2008) *Pipette*, **2**, 13.
- 134 Pittz, E.P. and Bello, J. (1971) *Archives of Biochemistry and Biophysics*, **146**, 513–524.
- 135 Pittz, E.P. and Bello, J. (1973) *Archives of Biochemistry and Biophysics*, **156**, 437–447.
- 136 Arakawa, T., Bhat, R., and Timasheff, S.N. (1990) *Biochemistry*, **29**, 1924–1931.
- 137 Potts, W.M. and Vogt, V.M. (1987) *Virology*, **160**, 494–497.
- 138 Thompson, N.E., Aronson, D.B., and Burgess, R.R. (1990) *Journal of Biological Chemistry*, **265**, 7069–7077.
- 139 Berry, M.J., Davies, J., Smith, C.G., and Smith, I. (1991) *Journal of Chromatography*, **587**, 161–169.
- 140 Thompson, N.E., Hager, D.A., and Burgess, R.R. (1992) *Biochemistry*, **31**, 7003–7008.
- 141 Durkee, K.H., Roh, B.H., and Doellgast, G.J. (1993) *Protein Expression and Purification*, **4**, 405–411.
- 142 Agraz, A., Duarte, C.A., Costa, L., Perez, L., Paez, R., Pujol, V., and Fontirrochi, G. (1994) *Journal of Chromatography A*, **672**, 25–33.
- 143 Narhi, L.O., Caughey, D.J., Horan, T., Kita, Y., Chang, D., and Arakawa, T. (1997) *Analytical Biochemistry*, **253**, 236–245.

- 144 Narhi, L.O., Caughey, D.J., Horan, T.P., Kita, Y., Chang, D., and Arakawa, T. (1997) *Analytical Biochemistry*, **253**, 246–252.
- 145 Kummer, A. and Li-Chan, E.C.Y. (1998) *Journal of Immunological Methods*, **211**, 125–137.
- 146 Geren, C.R., Magee, S.C., and Ebner, K.E. (1976) *Archives of Biochemistry and Biophysics*, **172**, 149–155.
- 147 Marcus, S.L., Smith, S.W., Racevskis, J., and Sarkar, N.H. (1979) *Journal of Biological Chemistry*, **254**, 4809–4813.
- 148 Doellgast, G.J. and Kohlhaw, G.B. (1972) *Federation American Society for Experimental Biology*, **31**, 424.
- 149 Rahimi-Laridjani, I., Grimminger, H., and Lingens, F. (1973) *FEBS Letters*, **30**, 185–187.
- 150 Memoli, V.A. and Doellgast, G.J. (1975) *Biochemical and Biophysical Research Communications*, **66**, 1011–1016.
- 151 Regnier, F.E. and Gooding, K.M. (1980) *Analytical Biochemistry*, **103**, 1–25.
- 152 Chicz, R.M. and Regnier, F.E. (1990) *Methods in Enzymology*, **182**, 392–421.
- 153 Rubinstein, M. (1979) *Analytical Biochemistry*, **98**, 1–7.
- 154 Fausnaugh, J.L., Pfannkoch, E., Gupta, S., and Regnier, F.E. (1984) *Analytical Biochemistry*, **137**, 464–472.
- 155 Alpert, A.J. (1990) *Journal of Chromatography*, **499**, 177–196.
- 156 Shukla, A.A., Hubbard, B., Tressel, T., Guhan, S., and Low, D. (2007) *Journal of Chromatography B*, **848**, 28–39.
- 157 Roettger, B.F., Myers, J.A., Ladish, M.R., and Regnier, F.E. (1989) *Biotechnology Progress*, **5**, 79–88.
- 158 Gagnon, P., Mayes, T., and Danielsson, A. (1997) *Journal of Pharmaceutical and Biomedical Analysis*, **16**, 587–592.
- 159 Gagnon, P. and Grund, E. (1996) *BioPharm Journal*, **9**, 54, 56 58–64.
- 160 Ejima, D., Yumioka, R., Arakawa, T., and Tsumoto, K. (2005) *Journal of Chromatography A*, **1094**, 49–55.
- 161 Janado, M., Shimada, K., and Nishida, T. (1976) *Journal of Biochemistry*, **79**, 513–520.
- 162 Mevarech, M., Leicht, W., and Werber, M.M. (1976) *Biochemistry*, **15**, 2383–2387.
- 163 Leicht, W. and Pundak, S. (1981) *Analytical Biochemistry*, **114**, 186–192.
- 164 Mayhew, S.G. and Howell, L.G. (1971) *Analytical Biochemistry*, **41**, 466–470.
- 165 Fujita, T., Suzuki, Y., Yamauti, J., Takagahara, I., Fujii, K., Yamashita, J., and Horio, T. (1980) *Journal of Biochemistry*, **87**, 89–100.
- 166 Janado, M. and Nishida, T. (1981) *Journal of Solution Chemistry*, **10**, 489–500.
- 167 Lanyi, J.K. (1974) *Bacteriological Reviews*, **38**, 272–290.
- 168 Kessel, M. and Klink, F. (1981) *European Journal of Biochemistry*, **114**, 481–486.
- 169 Leicht, W. (1978) *European Journal of Biochemistry*, **84**, 133–139.
- 170 Wang, W. (1999) *International Journal of Pharmaceutics*, **185**, 129–188.
- 171 Jones, A.J.S. (1983) *Advanced Drug Delivery Reviews*, **10**, 29–90.
- 172 Chang, B.S. and Hershenson, S. (2002) *Pharmaceutical Biotechnology*, **13**, 1–25.
- 173 Hartmann, W.K., Saptharishi, N., Yang, X.Y., Mitra, G., and Soman, G. (2004) *Analytical Biochemistry*, **325**, 227–239.
- 174 Regnier, F.E. (1983) *Science*, **222**, 245–252.
- 175 Stulik, K., Pacakova, V., and Ticha, M. (2003) *Journal of Biochemical and Biophysical Methods*, **56**, 1–13.
- 176 Pacakova, V., Stulik, K., Hau, P.T., Jelinek, I., Vins, I., and Sykora, D. (1995) *Journal of Chromatography A*, **700**, 187–193.
- 177 Watson, E. and Kenney, W.C. (1988) *Journal of Chromatography*, **436**, 289–298.
- 178 Gagnon, P. (1996) *Purification Tools for Monoclonal Antibodies*. Validated Biosystems, Tucson, AZ.
- 179 Ricker, R.D. and Sandoval, L.A. (1996) *Journal of Chromatography A*, **743**, 43–50.
- 180 Arakawa, T., Ejima, D., Li, T., and Philo, J.S. (2010) *Journal of Pharmaceutical Sciences*, **99**, 1674–1692.
- 181 Welling, G.W., Groen, G., Slopsema, K., and Welling-Wester, S. (1985) *Journal of Chromatography*, **326**, 173–178.
- 182 Klyushnichenko, V.E. and Wulfson, A.N. (1993) *Pure and Applied Chemistry*, **65**, 2265–2272.

- 183 Tsumoto, K., Ejima, D., Nagase, K., and Arakawa, T. (2007) *Journal of Chromatography A*, **1154**, 81–86.
- 184 Stahlberg, J., Jonsson, B., and Horvath, C. (1991) *Analytical Chemistry*, **63**, 1867–1874.
- 185 Arakawa, T., Tsumoto, K., Ejima, D., Kita, Y., Yonezawa, Y., and Tokunaga, M. (2007) *Journal of Biochemical and Biophysical Methods*, **70**, 493–498.
- 186 Gagnon, P. (2009) *Current Pharmaceutical Biotechnology*, **10**, 434–439.
- 187 Gagnon, P. and Beam, K. (2009) *Current Pharmaceutical Biotechnology*, **10**, 440–446.
- 188 Sii, D. and Sadana, A. (1991) *Journal of Biotechnology*, **19**, 83–98.
- 189 Stead, C.V. (1991) *Bioseparation*, **2**, 129–136.
- 190 Mattiasson, B., Galaev, I.Y., and Garg, N. (1996) *Journal of Molecular Recognition*, **9**, 509–514.
- 191 Kumar, A., Galaev, I.Y., and Mattiasson, B. (2000) *Journal of Chromatography B*, **741**, 103–113.
- 192 Arakawa, T., Ejima, D., Tsumoto, K., Ishibashi, M., and Tokunaga, M. (2007) *Protein Expression and Purification*, **52**, 410–414.
- 193 Labrou, N. and Clonis, Y.D. (1994) *Journal of Biotechnology*, **36**, 95–119.
- 194 Inagaki, M., Watanabe, M., and Hidaka, H. (1985) *Journal of Biological Chemistry*, **260**, 2922–2925.
- 195 Valsasina, B., Kalisz, H.M., and Isacchi, A. (2004) *Expert Review of Proteomics*, **1**, 303–315.
- 196 Clonis, Y.D. (2006) *Journal of Chromatography A*, **1101**, 1–24.
- 197 Hage, D.S. (1999) *Clinical Chemistry*, **45**, 593–615.
- 198 Bras, G.L., Teschner, W., Deville-Bonne, D., and Garel, J.R. (1989) *Biochemistry*, **28**, 6836–6841.
- 199 Deville-Bonne, D., Le Bras, G., Teschner, W., and Garel, J.R. (1989) *Biochemistry*, **28**, 1917–1922.
- 200 Godl, K., Wissing, J., Kurtenbach, A., Habenberger, P., Blencke, S., Gutbrod, H., Salassidis, K., Steingerlach, M., Missio, A., Cotten, M., and Daub, H. (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15434–15439.
- 201 Caughey, D.J., Narhi, L.O., Kita, Y., Meng, S.Y., Wen, D., Chen, W., Ratzkin, B.J., Fujimoto, J., Iwahara, T., Yamamoto, T., and Arakawa, T. (1999) *Journal of Chromatography B*, **728**, 49–57.
- 202 Fahrner, R.L., Whitney, D.H., Vanderlaan, M., and Blank, G.S. (1999) *Biotechnology and Applied Biochemistry*, **30**, 121–128.
- 203 Hahn, R., Schlegel, R., and Jungbauer, A. (2003) *Journal of Chromatography B*, **790**, 35–51.
- 204 Ejima, D., Tsumoto, K., Fukada, H., Yumioka, R., Nagase, K., Arakawa, T., and Philo, J.S. (2007) *Proteins*, **66**, 954–962.
- 205 Arakawa, T., Philo, J.S., Tsumoto, K., Yumioka, R., and Ejima, D. (2004) *Protein Expression and Purification*, **36**, 244–248.
- 206 Ejima, D., Yumioka, R., Tsumoto, K., and Arakawa, T. (2005) *Analytical Biochemistry*, **345**, 250–257.
- 207 Yumioka, R., Tsumoto, K., Arakawa, T., and Ejima, D. (2010) *Protein Expression and Purification*, **70**, 218–223.
- 208 Arakawa, T., Kita, Y., Sato, H., and Ejima, D. (2009) *Protein Expression and Purification*, **63**, 158–163.
- 209 Arakawa, T., Futatsumori-Sugai, M., Tsumoto, K., Kita, Y., Sato, H., and Ejima, D. (2010) *Protein Expression and Purification*, **71**, 168–173.
- 210 Umetsu, M., Tsumoto, K., Nitta, S., Adschiri, T., Ejima, D., Arakawa, T., and Kumagai, I. (2005) *Biochemical and Biophysical Research Communications*, **328**, 189–197.
- 211 Tsumoto, K., Umetsu, M., Kumagai, I., Ejima, D., and Arakawa, T. (2003) *Biochemical and Biophysical Research Communications*, **312**, 1383–1386.
- 212 Zardeneta, G. and Horowitz, P.M. (1994) *Analytical Biochemistry*, **223**, 1–6.
- 213 Rudolph, R. and Lilie, H. (1996) *FASEB Journal*, **10**, 49–56.
- 214 Maachupalli-Reddy, J., Kelley, B.D., and De Bernardes Clark, E. (1997) *Biotechnology Progress*, **13**, 144–150.

- 215 Lilie, H., Schwarz, E., and Rudolph, R. (1998) *Current Opinion in Biotechnology*, **9**, 497–501.
- 216 Carrio, M.M. and Villaverde, A. (2002) *Journal of Biotechnology*, **96**, 3–12.
- 217 Burgess, W.H. and Maciag, T. (1989) *Annual Review of Biochemistry*, **58**, 575–602.
- 218 Dabora, J.M., Sanyal, G., and Middaugh, C.R. (1991) *Journal of Biological Chemistry*, **266**, 23637–23640.
- 219 Volkin, D.B., Tsai, P.K., Dabora, J.M., Gress, J.O., Burke, C.J., Linhardt, R.J., and Middaugh, C.R. (1993) *Archives of Biochemistry and Biophysics*, **300**, 30–41.
- 220 Chen, B.-L., Arakawa, T., Morris, C.F., Kenney, W.C., Wells, C.M., and Pitt, C.G. (1994) *Pharmaceutical Research*, **11**, 1581–1587.
- 221 Chen, B.-L., Arakawa, T., Hsu, E., Narhi, L.O., Tressel, T.J., and Chien, S.L. (1994) *Journal of Pharmaceutical Sciences*, **83**, 1657–1661.
- 222 Chen, B.-L. and Arakawa, T. (1996) *Journal of Pharmaceutical Sciences*, **85**, 419–426.
- 223 Maity, H., Karkaria, C., and Davagnino, J. (2009) *Current Pharmaceutical Biotechnology*, **10**, 609–625.
- 224 Maity, H., O'Dell, C., Srivastava, A., and Goldstein, J. (2009) *Current Pharmaceutical Biotechnology*, **10**, 761–766.
- 225 Wait, D.A. and Sobsey, M.D. (1983) *Applied and Environmental Microbiology*, **46**, 379–385.
- 226 Shields, P.A. and Farrah, S.R. (1986) *Applied and Environmental Microbiology*, **51**, 211–213.
- 227 Wallis, C. and Melnick, J.L. (1967) *Journal of Virology*, **1**, 472–477.
- 228 Downing, L.A., Bernstein, J.M., and Walter, A. (1992) *Journal of Virological Methods*, **38**, 215–228.
- 229 Adams, A. (1973) *Journal of General Virology*, **20**, 391–394.
- 230 Black, P.H., Crawford, E.M., and Crawford, L.V. (1964) *Virology*, **24**, 381–387.
- 231 Farrah, S.R., Shah, D.O., and Ingram, L.O. (1981) *Proceedings of the National Academy of Sciences of the United States of America*, **78**, 1229–1232.
- 232 Vaehri, A., von Bonsdorff, C.-H., Vesikari, T., Hovi, T., and Väänänen, P. (1969) *Journal of General Virology*, **5**, 39–46.
- 233 Gao, G., Qu, G., Burnham, M.S., Huang, J., Chirmule, N., Joshi, B., Yu, Q.-C., Marsh, J.A., Conceicao, C.M., and Wilson, J.M. (2000) *Human Gene Therapy*, **11**, 2079–2091.
- 234 Auricchio, A., O'Connor, E., Hildinger, M., and Wilson, J.M. (2001) *Molecular Therapy*, **4**, 372–374.
- 235 Njayou, M. and Quash, G. (1991) *Journal of Virological Methods*, **32**, 67–77.
- 236 Izadpanah, K. and Shepherd, R.J. (1966) *Virology*, **28**, 463–476.
- 237 Fernie, B.F. and Gerin, J.L. (1980) *Virology*, **106**, 141–144.
- 238 Sokolov, N.N., Zhukova, T.Y., Heinman, V.Y., Alexandrova, G.I., and Smorodintsev, A.A. (1971) *Archives of Virology*, **35**, 356–363.
- 239 Smrekar, F., Ciringer, M., Peterka, M., Podgornik, A., and Strancar, A. (2008) *Journal of Chromatography B*, **861**, 177–180.
- 240 Gwaltney, J.M. Jr. and Calhoun, A.M. (1970) *Applied Microbiology*, **20**, 390–392.
- 241 Totsuka, A., Ohtaki, K., and Tagaya, I. (1978) *Journal of General Virology*, **38**, 519–533.
- 242 Farrah, S.R. and Bitton, G. (1978) *Applied and Environmental Microbiology*, **36**, 982–984.
- 243 Arakawa, T., Kita, Y.A., and Narhi, L.O. (1991) *Methods of Biochemical Analysis*, **35**, 87–125.
- 244 Hopkins, S.R. (1967) *Avian Diseases*, **11**, 261–267.
- 245 Nishibe, Y., Inoue, Y.K., and Melnick, J.L. (1986) *Journal of Medical Virology*, **20**, 105–109.
- 246 Wallis, C. and Melnick, J.L. (1965) *Journal of Bacteriology*, **90**, 1632–1637.
- 247 Wallis, C., Melnick, J.L., and Rapp, F. (1965) *Virology*, **26**, 694–699.
- 248 Rapp, F., Butel, J.S., and Wallis, C. (1965) *Journal of Bacteriology*, **90**, 132–135.
- 249 Arakawa, T. and Timasheff, S.N. (1984) *Biochemistry*, **23**, 5912–5923.

- 250 Arakawa, T., Bhat, R., and Timasheff, S.N. (1990) *Biochemistry*, **29**, 1914–1923.
- 251 Ozaki, Y. and Melnick, J.L. (1963) *Journal of Immunology*, **90**, 429–437.
- 252 Wallis, C., Yang, C.S., and Melnick, J.L. (1962) *Journal of Immunology*, **89**, 41–46.
- 253 Nakamura, M. and Ueno, Y. (1964) *Proceedings of the Society for Experimental Biology and Medicine*, **117**, 700–704.
- 254 Carpenter, J.F., Arakawa, T., and Crowe, J.H. (1992) *Developments in Biological Standardization*, **74**, 225–238.
- 255 Arakawa, T., Kita, Y., and Carpenter, J.F. (1991) *Pharmaceutical Research*, **8**, 285–291.
- 256 Carpenter, J.F. and Crowe, J.H. (1988) *Cryobiology*, **25**, 244–255.
- 257 Savithri, H.S., Munshi, S.K., Suryanarayana, S., Divakar, S., and Murthy, M.R.N. (1987) *Journal of General Virology*, **68**, 1533–1542.
- 258 Arakawa, T. and Timasheff, S.N. (1984) *Journal of Biological Chemistry*, **259**, 4979–4986.
- 259 Kaushik, J.K. and Bhat, R. (1999) *Protein Science*, **8**, 222–233.
- 260 Ludwig, C. (1961) *Archives of Experimental Veterinary Medicine*, **15**, 482–483.
- 261 Ludwig, C. (1964) *Monatsh Veterinary Medicine Journal*, **19**, 44.
- 262 Fellows, O.N. (1965) *Applied Microbiology*, **13**, 496–499.
- 263 Scott, E.M. and Woodside, W. (1976) *Journal of Clinical Microbiology*, **4**, 1–5.
- 264 Bovarnick, M.R., Miller, J.C., and Snyder, J.C. (1950) *Journal of Bacteriology*, **59**, 509–522.
- 265 Calnek, B.W., Hitchner, S.B., and Adlinder, H.K. (1970) *Applied Microbiology*, **20**, 723–726.
- 266 Grose, C., Friedrichs, W.E., and Smith, K.O. (1981) *Intervirology*, **15**, 154–160.
- 267 Tannock, G.A., Hierholzer, J.C., Bryce, D.A., Chee, C.F., and Paul, J.A. (1987) *Journal of Clinical Microbiology*, **25**, 1769–1771.
- 268 Adams, M.H. (1948) *Journal of General Physiology*, **31**, 417–431.
- 269 Wolf, K., Quimby, M.C., and Carlson, C.P. (1969) *Applied Microbiology*, **17**, 623–624.
- 270 Plowright, W., Rampton, C.S., Taylor, W.P., and Herniman, K.A. (1970) *Research in Veterinary Science*, **11**, 71–81.
- 271 Mariner, J.C., House, J.A., Sollod, A.E., Stem, C., van den Ende, M., and Mebus, C.A. (1990) *Veterinary Microbiology*, **21**, 195–209.
- 272 Suzuki, M. (1970) *Journal of Hygiene*, **68**, 29–41.
- 273 Collier, L.H. (1955) *Journal of Hygiene*, **53**, 76–101.
- 274 Cho, C. and Obayashi, Y. (1956) *Bulletin of the World Health Organization*, **14**, 657–669.
- 275 Obayashi, Y. and Cho, C. (1957) *Bulletin of the World Health Organization*, **17**, 255–274.
- 276 Huang, J., Garmise, R.J., Crowder, T.M., Mar, K., Hwang, C.R., Hickey, A.J., Mikszta, J.A., and Sullivan, V.J. (2004) *Vaccine*, **23**, 794–801.
- 277 Evengard, B., Ehrnst, A., von Sydow, M., Pehrson, P.O., Lundbergh, P., and Linder, E. (1989) *AIDS*, **3**, 591–595.
- 278 Brorson, K., Krejci, S., Lee, K., Hamilton, E., Stein, K., and Xu, Y. (2003) *Biotechnology and Bioengineering*, **82**, 321–329.
- 279 Nowak, T., Gregersen, J.-P., Klockmann, U., Cummins, L.B., and Hilfenhaus, J. (1992) *Journal of Medical Virology*, **36**, 209–216.
- 280 Vermeer, A.W., Bremer, M.G., and Norde, W. (1998) *Biochimica et Biophysica Acta*, **1425**, 1–12.
- 281 Wallis, C. and Menick, J.L. (1962) *Virology*, **16**, 504–506.
- 282 Moorer, W.R. (2003) *International Journal of Dental Hygiene*, **1**, 138–142.
- 283 Croughan, W.S. and Behbehani, A.M. (1988) *Journal of Clinical Microbiology*, **26**, 213–215.
- 284 Roberts, P.L. and Lloyd, D. (2007) *Biologicals*, **35**, 343–347.
- 285 Macinga, D.R., Sattar, S.A., Jaykus, L.-A., and Arbogast, J.W. (2008) *Applied and Environmental Microbiology*, **74**, 5047–5052.



- 286 Bidawid, S., Malik, N., Adegburin, O., Sattar, S.A., and Farber, J.M. (2004) *Journal of Food Protection*, **67**, 103–109.
- 287 Doultree, J.C., Druce, J.D., Birch, C.J., Bowden, D.S., and Marshall, J.A. (1999) *Journal of Hospital Infection*, **41**, 51–57.
- 288 Springthorpe, V.S., Grenier, J.L., Lloyd-Evans, N., and Sattar, S.A. (1986) *Journal of Hygiene*, **97**, 139–161.
- 289 Sattar, S.A., Springthorpe, V.S., Karim, Y., and Loro, P. (1989) *Epidemiology and Infection*, **102**, 493–505.
- 290 Perham, M., Liao, J., and Wittung-Stafshede, P. (2006) *Biochemistry*, **45**, 7740–7749.
- 291 Ferreira, G.N.M., Cabral, J.M.S., and Prazeres, D.M.F. (2000) *Bioseparation*, **9**, 1–6.
- 292 Chandra, G., Patel, P., Kost, T.A., and Gray, J.G. (1992) *Analytical Biochemistry*, **203**, 169–172.
- 293 Wicks, I.P., Howell, M.L., Hancock, T., Kohsaka, H., Olee, T., and Carson, D.A. (1995) *Human Gene Therapy*, **6**, 317–323.
- 294 Raymond, G.J., Bryant, P.K. 3rd, Nelson, A., and Johnson, J.D. (1988) *Analytical Biochemistry*, **173**, 125–133.
- 295 Schluep, T. and Cooney, C.L. (1998) *Nucleic Acids Research*, **26**, 4524–4528.
- 296 Umemo, D., Kano, T., and Maeda, M. (1998) *Analytica Chimica Acta*, **365**, 101–108.
- 297 Diogo, M.M., Queiroz, J.A., Monteiro, G.A., and Prazeres, D.M. (1999) *Analytical Biochemistry*, **275**, 122–124.
- 298 Diogo, M.M., Queiroz, J.A., Monteiro, G.A., Martins, S.A., Ferreira, G.N., and Prazeres, D.M. (2000) *Biotechnology and Bioengineering*, **68**, 576–583.
- 299 Tseng, W.-C. and Ho, F.-L. (2003) *Journal of Chromatography B*, **791**, 263–272.
- 300 Girod, J.C., Johnson, W.C. Jr., Huntington, S.K., and Maestre, M.F. (1973) *Biochemistry*, **12**, 5092–5096.
- 301 Herbeck, R., Yu, T.-J., and Peticolas, W.L. (1976) *Biochemistry*, **15**, 2656–2660.
- 302 Lang, D. (1973) *Journal of Molecular Biology*, **78**, 247–254.
- 303 Rupprecht, A., Piskur, J., Schultz, J., Nordenskiöld, L., Song, Z., and Lahajnar, G. (1994) *Biopolymers*, **34**, 897–920.
- 304 Humphreys, G.O., Willshaw, G.A., and Anderson, E.S. (1975) *Biochimica et Biophysica Acta*, **383**, 457–463.
- 305 Lis, J.T. (1980) *Methods in Enzymology*, **65**, 347–353.
- 306 Sauer, M.-L., Kollars, B., Geraets, R., and Sutton, F. (2008) *Analytical Biochemistry*, **380**, 310–314.
- 307 Zasloff, M., Ginder, G.D., and Felsenfeld, G. (1978) *Nucleic Acids Research*, **5**, 1139–1152.
- 308 Eagland, D. (1975) *Water – A Comprehensive Treatise*, vol. 4 (ed. F. Franks), Plenum Press, New York, pp. 305–518.
- 309 Matsuoka, Y., Nomura, A., Tanaka, S., Baba, Y., and Kagemoto, A. (1990) *Thermochimica Acta*, **163**, 147–154.
- 310 Piskur, J. and Rupprecht, A. (1995) *FEBS Letters*, **375**, 174–178.
- 311 Gruenwedel, D.W., Hsu, C.-H., and Lu, D.S. (1971) *Biopolymers*, **10**, 47–68.
- 312 Dix, D.E. and Straus, D.B. (1972) *Archives of Biochemistry and Biophysics*, **152**, 299–310.
- 313 Schultz, J., Rupprecht, A., Song, Z., Piskur, J., Nordenskiöld, L., and Lahajnar, G. (1994) *Biophysical Journal*, **66**, 810–819.
- 314 Washabaugh, M.W. and Collins, K.D. (1986) *Journal of Biological Chemistry*, **261**, 12477–12485.
- 315 Collins, K.D. (1995) *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5553–5557.
- 316 Kiriukhin, M.Y. and Collins, K.D. (2002) *Biophysical Chemistry*, **99**, 155–168.
- 317 Pappenheimer, J.R., Lepie, M.P., and Wyman, J. Jr. (1936) *Journal of American Chemical Society*, **58**, 1851–1855.
- 318 Arakawa, T. and Timasheff, S.N. (1983) *Archives of Biochemistry and Biophysics*, **224**, 169–177.
- 319 Kita, Y., Arakawa, T., Lin, T.Y., and Timasheff, S.N. (1994) *Biochemistry*, **33**, 15178–15189.
- 320 Rupley, J.A. and Careri, G. (1991) *Advances in Protein Chemistry*, **41**, 37–172.

- 321 Fullerton, G.D., Ord, V.A., and Cameron, I.L. (1986) *Biochimica et Biophysica Acta*, **869**, 230–246.
- 322 Mrevlishvili, G.M. (1998) *Introduction to Molecular Physiology of Water* (eds H. Uedaira and T. Tataru), Medical Sciences International, Tokyo, p. 119.
- 323 Vrij, A. (1976) *Pure and Applied Chemistry*, **48**, 471–483.
- 324 Sacanna, S., Irvine, W.T.M., Chaikin, P.M., and Pine, D.J. (2010) *Nature*, **464**, 575–578.
- 325 Laurent, T.C. (1963) *Acta Chemica Scandinavica*, **17**, 2664–2668.
- 326 Laurent, T.C. and Ogston, A.G. (1963) *Biochemical Journal*, **89**, 249–253.
- 327 Laurent, T.C. (1963) *Biochemical Journal*, **89**, 253–257.
- 328 Harano, Y. and Kinoshita, M. (2004) *Chemical Physics Letters*, **399**, 342–348.
- 329 Harano, Y. and Kinoshita, M. (2005) *Biophysical Journal*, **89**, 2701–2710.
- 330 Kinoshita, M. (2009) *International Journal of Molecular Sciences*, **10**, 1064–1080.
- 331 Cohn, E.J. (1943) *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions* (eds E.J. Cohn and J.T. Edsall), Reinhold, New York, pp. 236–275.
- 332 Tanford, C. (1962) *Journal of American Chemical Society*, **84**, 4240–4247.
- 333 Yancey, P.H., Clark, M.E., Hand, S.C., Bowlus, R.D., and Somero, G.N. (1982) *Science*, **217**, 1214–1222.
- 334 Crowe, J.H., Carpenter, J.F., and Crowe, L.M. (1998) *Annual Review of Physiology*, **60**, 73–103.
- 335 Crowe, J.H., Carpenter, J.F., Crowe, L.M., and Anchordoguy, T.J. (1990) *Cryobiology*, **27**, 219–231.
- 336 Crowe, J.H., Hoekstra, F.A., and Crowe, L.M. (1992) *Annual Review of Physiology*, **54**, 579–599.
- 337 Bolen, D.W. (2001) *Methods in Molecular Biology*, vol. **168** (ed. K.P. Murphy), Humana, Totowa, NJ, pp. 17–36.
- 338 Arakawa, T. and Timasheff, S.N. (1985) *Biophysical Journal*, **47**, 411–414.
- 339 Lakshmi, T.S. and Nandi, P.K. (1976) *Journal of Physical Chemistry*, **80**, 249–252.
- 340 Gekko, K. (1981) *Journal of Biochemistry*, **90**, 1633–1641.
- 341 Gekko, K. and Koga, S. (1984) *Biochimica et Biophysica Acta*, **786**, 151–160.
- 342 Uedaira, H. (1977) *Bulletin of the Chemical Society of Japan*, **50**, 1298–1304.
- 343 Qu, Y., Bolen, C.L., and Bolen, D.W. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 9268–9273.
- 344 Bolen, D.W. and Baskakov, I.V. (2001) *Journal of Molecular Biology*, **310**, 955–963.
- 345 Auton, M. and Bolen, D.W. (2004) *Biochemistry*, **43**, 1329–1342.
- 346 Arakawa, T., Kita, Y., Ejima, D., Tsumoto, K., and Fukada, H. (2006) *Protein and Peptide Letters*, **13**, 921–927.
- 347 Arakawa, T. and Tsumoto, K. (2003) *Biochemical and Biophysical Research Communications*, **304**, 148–152.
- 348 Shiraki, K., Kudou, M., Fujiwara, S., Imanaka, T., and Takagi, M. (2002) *Journal of Biochemistry*, **132**, 591–595.
- 349 Arakawa, T., Kita, Y., and Koyama, A.H. (2008) *International Journal of Pharmaceutics*, **355**, 220–223.
- 350 Arakawa, T., Ejima, D., Tsumoto, K., Obeyama, N., Tanaka, Y., Kita, Y., and Timasheff, S.N. (2007) *Biophysical Chemistry*, **127**, 1–8.
- 351 Nozaki, Y. and Tanford, C. (1963) *Journal of Biological Chemistry*, **238**, 4074–4081.
- 352 Nozaki, Y. and Tanford, C. (1970) *Journal of Biological Chemistry*, **245**, 1648–1652.
- 353 Cohn, E.J. and Edsall, J.T. (1943) *Proteins, Amino Acids, and Peptides as Ions and Dipolar Ions* (eds E.J. Cohn and J.T. Edsall), Reinhold, New York, pp. 196–216.
- 354 Nozaki, Y. and Tanford, C. (1971) *Journal of Biological Chemistry*, **246**, 2211–2217.
- 355 Chang, J.-P. and Yuan, Y. (1987) *Biomedical Chromatography*, **2**, 20–23.
- 356 Li, J.-J., Venkataramana, M., Wang, A.-Q., Sanyal, S., Janson, J.-C., and Su, Z.-G. (2005) *Protein Expression and Purification*, **40**, 327–335.
- 357 Lee, J.C. and Timasheff, S.N. (1974) *Biochemistry*, **13**, 257–265.
- 358 Lee, J.C., Gekko, K., and Timasheff, S.N. (1979) *Methods in Enzymology*, **61**, 26–49.

- 359 Inoue, H., and Timasheff, S.N. (1968) *Journal of the American Chemical Society*, **90**, 1890–1898.
- 360 Inoue, H. and Timasheff, S.N. (1972) *Biopolymers*, **11**, 737–743.
- 361 Sharp, D.G., Taylor, A.R., McLean, I.W. Jr., Beard, D., and Beard, J.W. (1944) *Science*, **100**, 151–153.
- 362 Jaenicke, R. and Lauffer, M.A. (1969) *Biochemistry*, **8**, 3077–3082.
- 363 Jaenicke, R. and Lauffer, M.A. (1969) *Biochemistry*, **8**, 3083–3092.
- 364 Arakawa, T. and Timasheff, S.N. (1982) *Biochemistry*, **21**, 6545–6552.
- 365 Zhao, Y.H., Abraham, M.H., and Zissimos, A.M. (2003) *Journal of Organic Chemistry*, **68**, 7368–7373.
- 366 Robinson, D.R. and Jencks, W.P. (1965) *Journal of American Chemical Society*, **87**, 2470–2479.
- 367 Robinson, D.R. and Jencks, W.P. (1965) *Journal of American Chemical Society*, **87**, 2462–2470.
- 368 Arakawa, T. and Timasheff, S.N. (1987) *Biochemistry*, **26**, 5147–5153.
- 369 Lee, J.C. and Timasheff, S.N. (1981) *Journal of Biological Chemistry*, **256**, 7193–7201.
- 370 Gekko, K. and Koga, S. (1983) *Journal of Biochemistry*, **94**, 199–205.
- 371 Gekko, K. and Morikawa, T. (1981) *Journal of Biochemistry*, **90**, 39–50.
- 372 Na, G.C. and Timasheff, S.N. (1981) *Journal of Molecular Biology*, **151**, 165–178.
- 373 Gekko, K. and Timasheff, S.N. (1981) *Biochemistry*, **20**, 4667–4676.
- 374 Arakawa, T. and Timasheff, S.N. (1982) *Biochemistry*, **21**, 6536–6544.
- 375 Tanaka, M., Machida, Y., Niu, S., Ikeda, T., Jana, N.R., Doi, H., Kurosawa, M., Nekooki, M., and Nukina, N. (2004) *Nature Medicine*, **10**, 148–154.
- 376 Tanaka, M., Machida, Y., and Nukina, N. (2005) *Journal of Molecular Medicine*, **83**, 343–352.
- 377 Baynes, B.M. and Trout, B.L. (2004) *Biophysical Journal*, **87**, 1631–1639.
- 378 Baynes, B.M., Wang, D.I., and Trout, B.L. (2005) *Biochemistry*, **44**, 4919–4925.
- 379 Shukla, D., Shinde, C., and Trout, B.L. (2009) *Journal of Physical Chemistry B*, **113**, 12546–12554.
- 380 Timasheff, S.N. and Inoue, H. (1968) *Biochemistry*, **7**, 2501–2513.
- 381 Gekko, K., Ohmae, E., Kameyama, K., and Takagi, T. (1998) *Biochimica et Biophysica Acta*, **1387**, 195–205.
- 382 Lee, J.C. and Lee, L.L. (1979) *Biochemistry*, **18**, 5518–5526.
- 383 Lee, J.C. and Lee, L.L. (1981) *Journal of Biological Chemistry*, **256**, 625–631.
- 384 Heller, M.C., Carpenter, J.F., and Randolph, T.W. (1999) *Biotechnology and Bioengineering*, **63**, 166–174.
- 385 Eckhardt, B.M., Oeswein, J.Q., and Bewley, T.A. (1991) *Pharmaceutical Research*, **8**, 1360–1364.
- 386 Sarciaux, J.-M., Mansour, S., Hageman, M.J., and Nail, S.L. (1999) *Journal of Pharmaceutical Sciences*, **88**, 1354–1361.
- 387 Kreilgaard, L., Jones, L.S., Randolph, T.W., Frokjaer, S., Flink, J.M., Manning, M.C., and Carpenter, J.F. (1996) *Journal of Pharmaceutical Sciences*, **87**, 1593–1603.
- 388 Chang, B.S., Kendrick, B.S., and Carpenter, J.F. (1996) *Journal of Pharmaceutical Sciences*, **85**, 1325–1330.
- 389 Lovelock, J.E. (1954) *Biochemical Journal*, **56**, 265–270.
- 390 Gomez, G., Pikal, M.J., and Rodriguez-Hornedo, N. (2001) *Pharmaceutical Research*, **18**, 90–97.
- 391 Pikal-Cleland, K.A., Rodriguez-Hornedo, N., Amidon, G.L., and Carpenter, J.F. (2000) *Archives of Biochemistry and Biophysics*, **384**, 398–406.
- 392 Timasheff, S.N. and Xie, G. (2003) *Biophysical Chemistry*, **105**, 421–448.
- 393 Arakawa, T., Carpenter, J.F., Kita, Y.A., and Crowe, J.H. (1990) *Cryobiology*, **27**, 401–415.
- 394 Gekko, K. (1982) *Journal of Biochemistry*, **91**, 1197–1204.
- 395 Virudachalam, R., Sitaraman, K., Heuss, K.L., Markley, J.L., and Argos, P. (1983) *Virology*, **130**, 351–359.
- 396 Virudachalam, R., Sitaramant, K., Heusst, K.L., Argost, P., and Markley, J.L. (1983) *Virology*, **130**, 360–371.

- 397 Hansen, R.K., Zhai, S., Skepper, J.N., Johnston, M.D., Alpar, H.O., and Slater, N.K. (2005) *Biotechnology Progress*, **21**, 911–917.
- 398 Carpenter, J.F., Pikal, M.J., Chang, B.S., and Randolph, T.W. (1997) *Pharmaceutical Research*, **14**, 969–975.
- 399 Prestrelski, S.J., Tedeschi, N., Arakawa, T., and Carpenter, J.F. (1993) *Biophysical Journal*, **65**, 661–671.
- 400 Tang, X.C. and Pikal, M.J. (2005) *Pharmaceutical Research*, **22**, 1167–1175.
- 401 Carpenter, J.F. and Crowe, J.H. (1989) *Biochemistry*, **28**, 3916–3922.
- 402 Allison, S.D., Chang, B., Randolph, T.W., and Carpenter, J.F. (1999) *Archives of Biochemistry and Biophysics*, **365**, 289–298.
- 403 Crowe, J.H., Crowe, L.M., and Chapman, D. (1984) *Science*, **223**, 701–703.
- 404 Crowe, J.H., Whittam, M.A., Chapman, D., and Crowe, L.M. (1984) *Biochimica et Biophysica Acta*, **769**, 151–159.
- 405 Roy, M.L., Pikal, M.J., Rickard, E.C., and Maloney, A.M. (1992) *Developments in Biological Standardization*, **74**, 323–339.
- 406 Pikal, M.J., Dellerman, K., and Roy, M.L. (1992) *Developments in Biological Standardization*, **74**, 21–37.
- 407 Hill, J.J., Shalaev, E.Y., and Zografi, G. (2005) *Journal of Pharmaceutical Sciences*, **94**, 1636–1667.
- 408 Cleland, J.L., Lam, X., Kendrick, B., Yang, J., Yang, T.H., Overcashier, D., Brooks, D., Hsu, C., and Carpenter, J.F. (2001) *Journal of Pharmaceutical Sciences*, **90**, 310–321.
- 409 Crowe, L.M. and Crowe, J.H. (1992) *Developments in Biological Standardization*, **74**, 285–294.
- 410 Chang, B.S., Beauvais, R.M., Dong, A., and Carpenter, J.F. (1996) *Archives of Biochemistry and Biophysics*, **331**, 249–258.
- 411 Crowe, L.M. and Crowe, J.H. (1992) *Advances in Space Research*, **12**, 239–247.
- 412 Hengherr, S., Heyer, A.G., Kohler, H.R., and Schill, R.O. (2008) *FEBS Journal*, **275**, 281–288.
- 413 Clegg, J.S. (1965) *Comparative Biochemistry and Physiology*, **14**, 135–143.
- 414 Winkler, A. (2002) *Phytochemistry*, **60**, 437–440.
- 415 Sun, W.Q. and Davidson, P. (1998) *Biochimica et Biophysica Acta*, **1425**, 235–244.
- 416 Yoshioka, S., Aso, Y., and Kojima, S. (1997) *Pharmaceutical Research*, **14**, 736–741.
- 417 Chang, L.L., Shepherd, D., Sun, J., Ouellette, D., Grant, K.L., Tang, X.C., and Pikal, M.J. (2005) *Journal of Pharmaceutical Sciences*, **94**, 1427–1444.
- 418 Duddu, S.P., Zhang, G., and Dal Monte, P.R. (1997) *Pharmaceutical Research*, **14**, 596–600.
- 419 Yoshioka, S., Miyazaki, T., Aso, Y., and Kawanishi, T. (2007) *Pharmaceutical Research*, **24**, 1660–1667.
- 420 Koster, K.L., Lei, Y.P., Anderson, M., Martin, S., and Bryant, G. (2000) *Biophysical Journal*, **78**, 1932–1946.
- 421 Zhai, S., Hansen, R.K., Taylor, R., Skepper, J.N., Sanches, R., and Slater, N.K. (2004) *Biotechnology Progress*, **20**, 1113–1120.
- 422 Costantino, H.R., Firouzabadian, L., Hogeland, K., Wu, C., Beganski, C., Carrasquillo, K.G., Cordova, M., Griebenow, K., Zale, S.E., and Tracy, M.A. (2000) *Pharmaceutical Research*, **17**, 1374–1383.
- 423 Crowe, J.H., Crowe, L.M., Carpenter, J.F., Rudolph, A.S., Wistrom, C.A., Spargo, B.J., and Anchordoguy, T.J. (1988) *Biochimica et Biophysica Acta*, **947**, 367–384.
- 424 Bennett, P.S., Maigetter, R.Z., Olson, M.G., Provost, P.J., Scattergood, E.M., and Schofield, T.L. (1992) *Developments in Biological Standardization*, **74**, 215–221.
- 425 Fukumoto, F. (2008) *Journal of General Plant Pathology*, **74**, 164–170.
- 426 Obayashi, Y., Ota, S., and Arai, S. (1961) *Journal of Hygiene*, **59**, 77–91.
- 427 Robinson, J.B. Jr., Strottmann, J.M., and Stellwagen, E. (1981) *Proceedings of the National Academy of Sciences of the United States of America*, **78**, 2287–2291.
- 428 Krestov, G.A. (1991) *Thermodynamics of Solvation: Solution and Dissolution, Ions and Solvents, Structure and Energetics*, Horwood, New York.
- 429 Tanford, C. (1961) *Physical Chemistry of Macromolecules*, Wiley, New York.

## 10

### Role of Cysteine

Lalla A. Ba, Torsten Burkholz, Thomas Schneider, and Claus Jacob

#### 10.1

##### Sulfur: A Redox Chameleon with Many Faces

The adult human, with an average weight of 70 kg, contains around 175 g of sulfur [1]. This amount places sulfur firmly in the “Top 10” of chemical elements present in the human body – it actually occupies position seven after oxygen, carbon, hydrogen, nitrogen, calcium, and phosphorus (when ranked according to mass). While its abundance in the body is similar to that of sodium (105 g) and potassium (140 g), its *in vivo* chemistry differs significantly from that associated with such “inorganic” elements. Sulfur is able to appear as inorganic sulfur (e.g. as hydrogen sulfide or as sulfide ions) or as sulfide ions; yet, it is also part of a multitude of “organic” sulfur chemotypes, some of which are listed in Table 10.1.

As “inorganic” sulfur, the element is found in metal/sulfur clusters, which contain sulfide ( $S^{2-}$ ) anions [1]. Furthermore, hydrogen sulfide ( $H_2S/HS^-$ ) has recently gained prominence as the “third gaseous transmitter” besides nitric oxide ( $^*NO$ ) and carbon monoxide (CO). Indeed, several human enzymes, such as cystathionine  $\beta$ -synthase and cystathionine  $\gamma$ -lyase, are able to generate hydrogen sulfide inside the human body [2]. The latter appears to be present in blood in micromolar concentrations and specific biochemical targets for this transmitter have now been identified. (It should be noted that we refer here to endogenous hydrogen sulfide, which is produced by the human body on purpose. In contrast, exogenous hydrogen sulfide and hydrogen sulfide formed by bacteria in the human gut are a completely different matter. It appears that this kind of hydrogen sulfide does not enter the human bloodstream, but is mostly released from the human body.) This area of “inorganic” sulfur biochemistry is currently attracting considerable attention, with several recent reviews covering aspects of the latest developments [3–5].

Nonetheless, most sulfur in the human body is found as part of small “organic” molecules. Here, the amino acids cysteine and methionine stand out, although other sulfur-containing small molecules, such as lipoic acid, are equally important and should also be considered. Unlike lipoic acid and related molecules, however, cysteine and methionine form part of numerous peptides, proteins, and enzymes,

**Table 10.1** Brief overview of reactive sulfur species commonly encountered in biological systems.

Reactive species	Formula	Sulfur oxidation state	Occurrence
Hydrogen sulfide	H <sub>2</sub> S/HS <sup>-</sup>	-2	generated enzymatically in liver and brain
Thiol	RSH	-2	GSH and proteins
Disulfide radical anion	RSSR <sup>*-</sup>	-1.5 (overall)	intermediate of RS <sup>*</sup> radical reaction
Disulfide radical cation	(R <sub>2</sub> S:·SR <sub>2</sub> ) <sup>*+</sup>	-1.5 (overall)	intermediate of methionine oxidation
Thiyl radical	RS <sup>*</sup>	-1	one-electron oxidation of RSH (e.g., in enzymes)
Perthiol	RSSH	-1	formed by reduction of polysulfides
Perthiyl radical	RSS <sup>*</sup>	-1, 0	intermediate in perthiol reactions
Disulfide	RSSR	-1	GSSG and numerous proteins
Trisulfide	RSSSR	-1, 0, -1	GSSSG, diallyl polysulfides and leinamycin products
Tetrasulfide	RSSSSR	-1, 0, 0, -1	diallyltetrasulfide in garlic
Pentasulfide	RSSSSSR	-1, 0, 0, 0, -1	varacin
Hexasulfide	RSSSSSR	-1, 0, 0, 0, 0, -1	unclear, 1,2,3,4,5,6-hexathiepane in shiitake mushrooms
Sulfenic acid	RSOH	0	Prdx and NADH oxidase
Sulfinic acid	RS(O)OH	+2	Prdx and proteins oxidized during oxidative stress
Sulfonic acid	RS(O) <sub>2</sub> OH	+4	taurin and overoxidized proteins
Sulfate	SO <sub>4</sub> <sup>2-</sup>	+6	glucosinolates and heparin sulfate
Thiosulfinate	RS(O)SR	+1, -1	allicin and Prdx
Thiosulfonate	RS(O) <sub>2</sub> SR	+3, -1	various natural products
Thiocyanate	RSCN	-2	products of myrosinase

The various species are ordered according to increasing formal oxidation states (assuming an oxidation state of R = +1). Important biological occurrences are provided for each chemotype. Note that this list is neither complete nor is it exhaustive with regard to the natural occurrences of these modifications *in vivo*.

where these amino acids exhibit unique properties, which may differ considerably from those of “free” cysteine and methionine [6]. Importantly, the various properties associated with these (and other) amino acids *inside* proteins are governed and “fine-tuned” by the microenvironment provided by the protein itself. This influence of the protein microenvironment is illustrated, for instance, by individual pK<sub>s</sub> and E<sup>0</sup> values of cysteine residues, which may differ considerably from those of “free” cysteine (this matter will be discussed in more detail in Section 10.2).

Cysteine and methionine also stand out among other amino acids due to the extraordinary chemical and biochemical flexibility of the sulfur chemotype [4]. The valence electron configuration of sulfur is 2s<sup>2</sup>2p<sup>4</sup>, which implies around 10

possible oxidation states, ranging from  $-2$  (in  $\text{H}_2\text{S}$  and thiols) to  $+6$  (in  $\text{SO}_4^{2-}$ ), and including various intermediate and fractional oxidation states (Table 10.1). Amazingly, most of these oxidation states are reasonably stable under physiological conditions and many of them can therefore be found in biological systems [7]. The variety of sulfur redox states is mirrored by a high flexibility of sulfur to undergo redox transformations (Figure 10.1): sulfur is able to participate in one- and two-electron transfers, radical reactions, nucleophilic substitutions, hydride transfers, and oxygen atom transfers. The ability to change between oxidation states under physiological conditions rather easily turns sulfur into a true “redox chameleon.” At the same time, sulfur in different oxidation states, including thiols, polysulfanes, sulfenic, and sulfinic acids, is able to bind to metal ions, which further complicates the intracellular sulfur chemistry, and provides the basis for complicated redox and metal-binding control networks.

The diversity of the sulfur chemotype and the “fine-tuning” of chemical properties by the protein microenvironment together provide the foundation for an extraordinary chemistry and biochemistry of cysteine and methionine inside the living cell. Cysteine-based biochemical processes, for instance, form the basis for “simple” redox transformations, catalysis, metal binding, protein structure (via disulfides *and* metal/sulfur complexes), spontaneous chemical protection of sensitive residues, redox sensing and response, antioxidant defense, and extensive cellular signaling networks, to name just a few.

Some of these “functionalities” associated with cysteine are summarized in Table 10.2, which albeit incomplete, reflects the diversity of cysteine (bio-)chemistry rather nicely. In essence, there are examples of enzymes containing catalytic, active-site cysteine residues in all major classes of enzymes, including oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), and ligases (EC 6). The impact of cysteine is most pronounced, however, in oxidoreductases and hydrolases. Cysteine serves as an electron donor/acceptor and participates in redox catalysis in human enzymes such as the peroxiredoxins (Prdxs) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). It is also found in bacterial enzymes such as NADH oxidase and NADH peroxidase, in each case with distinct oxidation states [8]. Equally important, cysteine is present in several key hydrolytic enzymes, where it acts as nucleophile and catalyzes the cleavage of esters and amides. In this particular role, it is found in various caspases, which are involved in apoptosis, and in cdc25 enzymes, a family of phosphatases that are key guardians of the (human) cell cycle [9]. Caspases and cdc25 enzymes control cell survival and proliferation, and, not surprisingly, also provide major targets for anticancer drug development.

In a slightly different role, cysteine provides two distinct types of structural features in proteins and enzymes – disulfide bridges and zinc/sulfur complexes – and hence ensures three-dimensional stability of these macromolecules.

In the following sections, we cover some of the chemistry and associated biochemical/biological implications of cysteine. This discussion is guided by the desire to showcase the extent of the cysteine biochemistry known to date. For this purpose, various examples from human and mammalian biochemistry have been chosen to serve as highlights. Some of these discoveries are still quite speculative and





**Table 10.2** Cysteine is present at the active site of numerous proteins and enzymes.

Class	Protein	Active site
Oxidoreductase (EC 1)	glutathione disulfide reductase	Cys58 and Cys63 (human enzyme)
Transferase (EC 2)	glutathione-S-transferase	Cys47 and Cys101 (human GSTP1)
Hydrolase (EC 3)	caspase	Cys163 (in yeast caspase-3)
Lyase (EC 4)	isocitrate lyase	Cys195 ( <i>E. coli</i> enzyme)
Isomerase (EC 5)	protein disulfide isomerase	Cys38 and Cys35 (rat enzyme)
Ligase (EC 6)	ubiquitin ligase	Cys26 (poxviral enzyme)

There are examples of active-site cysteine enzymes in every major class of enzymes. Although the individual catalytic cycles associated with these enzymes obviously differ considerably, there are also common features, such as a high nucleophilicity of the cysteine thiol(ate) groups involved (see also Figure 10.2).

## 10.2

### Three Faces of Thiols: Nucleophilicity, Redox Activity, and Metal Binding

To begin with, it is essential to deal with the extensive list of sulfur chemotypes, associated oxidation states, redox processes, and biological functions (Table 10.1 and Figure 10.1). In order to do so, one must simplify to some extent. The most commonly encountered sulfur oxidation states inside the human body are thiols and disulfides, which are provided by cysteine, and a couple of dialkylsulfide derivatives based on methionine (e.g., sulfide, sulfoxide, sulfone). We will deal with the thiols first, which themselves in general show three distinct “faces” in biology – as nucleophiles, as redox agents, and as ligands for certain metal ions.

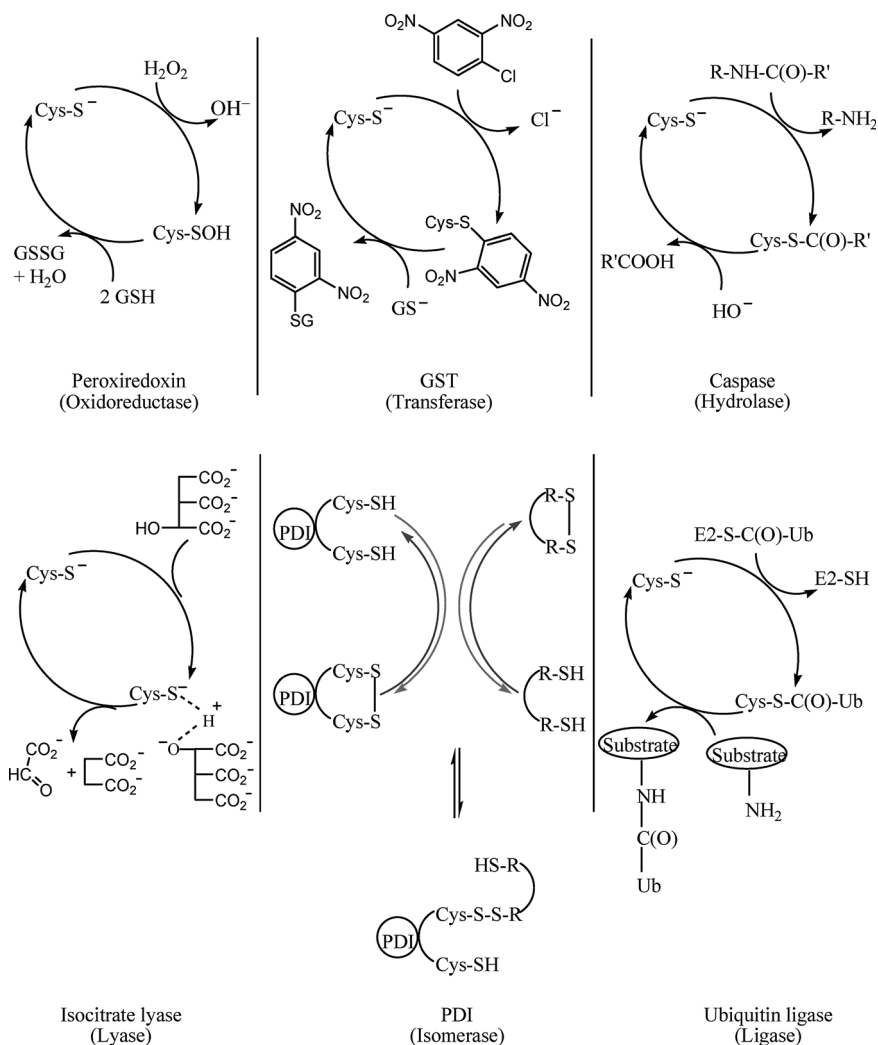
Cysteine in proteins and enzymes may be present as a thiol (RSH) or thiolate ( $RS^-$ ), depending on the pH of the cytosol or cellular compartment, on the one hand, and the  $pK_a$  value of the cysteine in question, on the other hand. In some ways, the nucleophilicity of cysteine is then a reflection of its  $pK_a$  value. The “free” form of cysteine has a  $pK_a$  of about 8.5, and under physiological conditions both the protonated and deprotonated forms of cysteine play a role. In proteins, the situation is more complicated, of course, with significant variations in  $pK_a$  values of cysteine residues, largely depending on the protein microenvironment in which these residues are found. Such a microenvironment may, for instance, assist deprotonation of the cysteine residue (i.e., endow cysteine with a lower  $pK_a$  value). Indeed, cysteine residues with a  $pK_a$  value as low as 5 have been postulated to exist at the active site of some enzymes, which is a dramatic deviation from the value of 8.5 for “free” cysteine. For instance, Cys63-SH in yeast glutathione disulfide reductase has a  $pK_a$  value of 4.8 [10]. A similar effect of the microenvironment has been observed regarding the thiol oxidation potential  $E^0$  of certain cysteine residues (Table 10.3).

The important role of cysteine as a nucleophile is perhaps best illustrated by human thiol-S-transferase enzymes (Figure 10.2). Glutathione-S-transferases (GSTs) are highly important human liver enzymes, which facilitate the removal of toxic, lipophilic molecules from the human body by conjugating such molecules to the

**Table 10.3** Brief summary of thiol : disulfide oxidoreductases.

Protein	Motif in active site	Redox potential (mV versus NHE)	Biological function
Glutathione		–240	redox buffer, reduces oxidative stress, substrate of many enzymes
Thioredoxin (12 kDa)	-Cys–Gly–Pro–Cys-	–270	protein reduction
Glutaredoxin (9 kDa)	-Cys–Pro–Tyr–Cys-	–233 (Grx-1); –198 (Grx-3)	catalyzes reduction of ribonucleotide reductase by GSH
Tryparedoxin (16 kDa)	-Cys–Pro–Pro–Cys-	–249	utilizes trypanothione in place of GSH
Protein disulfide isomerase (57 kDa)	-Cys–Gly–His–Cys-	–190	catalyzes reduction and reformation of disulfide bonds
DsbA (21 kDa)	-Cys–Pro–His–Cys-	–125	catalyzes formation/rearrangement of protein disulfide bonds from thiols/misformed disulfide bonds

Although these proteins and enzymes all contain two active-site cysteine residues, the individual motifs at the active site differ. This leads to distinct – and distinctively different – microenvironments and associated thiol/disulfide redox potentials. Interestingly, the resulting range of potentials, from a rather reducing –270 mV in the case of Trx to a more oxidizing –125 mV for DsbA, can be employed by the human cell for different purposes. Trx reduces (accidentally) oxidized proteins, while DsbA forms and rearranges disulfide bonds. Note that not all redox active thiol/disulfide proteins and enzymes are shown, and that the redox potentials provided for such proteins are often rough estimates only.



**Figure 10.2** Comparison of the cysteine-based catalytic cycles in different classes of enzymes. Prdx represents an oxidoreductase, GST a transferase, caspase a hydrolase, isocitrate lyase a lyase, PDI an isomerase, and ubiquitin ligase a ligase enzyme. While the individual substrates and subsequently formed products differ considerably, the six enzymes share quite a few common features from a chemical perspective.

The catalytically active thiol(ate) acts as a good base or nucleophile, which attacks an electrophilic center at the substrate. The subsequent substitution results in various protein-bound, thiolate-based intermediates, such as sulfenic acids, disulfides, thioethers (RSR') and thioesters (RSC(O)R'), which are usually cleaved by a second nucleophilic substitution.

small, cysteine-containing tripeptide glutathione ( $\gamma$ -glutamyl-cysteinyl-glycine, GSH) [11]. The resulting *S*-conjugated, glutathiolated molecules are fairly hydrophilic and can be excreted from the body. This detoxification systems works rather well for a wide range of lipophilic aromatic compounds, including particularly

unpleasant agents, such as halogenated benzene derivatives (the substrate to measure GST activity *in vitro* is actually 1-chloro-2,4-dinitrobenzene).

The nucleophilic character of the thiol (or thiolate) is also at the center of cysteine's ability to catalyze hydrolysis reactions (e.g., in caspases). Hydrolytic enzymes generally require a nucleophile in order to attack and subsequently cleave the target (ester or amide) bond. Here, nature provides various alternatives, including hydroxyl (ate) groups via a catalytic serine, or water/hydroxide ( $\text{HO}^-$ ) ions bound to a catalytic zinc ion ( $\text{Zn}^{2+}$  serves as Lewis acid). The thiol(ate) of cysteine is a special case. While it may also perform this hydrolytic task, it is redox sensitive at the same time and catalytic activity may be lost upon oxidation or thiol modification (e.g., by  $\bullet\text{NO}$ ). A similar catalytic mechanism is found in cysteine-based phosphatases, such as the cdc25 enzymes, which control the (human) cell cycle. These phosphatases employ cysteine as nucleophile, and generate a phospho-cysteine intermediate as part of their catalytic cycle [9, 12].

Nucleophilicity of the thiol(ate) of cysteine is also at the heart of thiol/disulfide exchange reactions. The latter are special substitution reactions, whereby nucleophilic attack of a thiol(ate) at the sulfur–sulfur bond of a disulfide results in a redox reaction. Although this kind of “hidden” redox reaction does not even involve the “flow” of (free) electrons, it is probably by far the most common and frequently encountered type of redox process *in vivo*. As Figure 10.2 illustrates, this thiol/disulfide exchange chemistry is found in protein disulfide isomerase (PDI) and also plays a role in the catalytic cycle of Prdx enzymes.

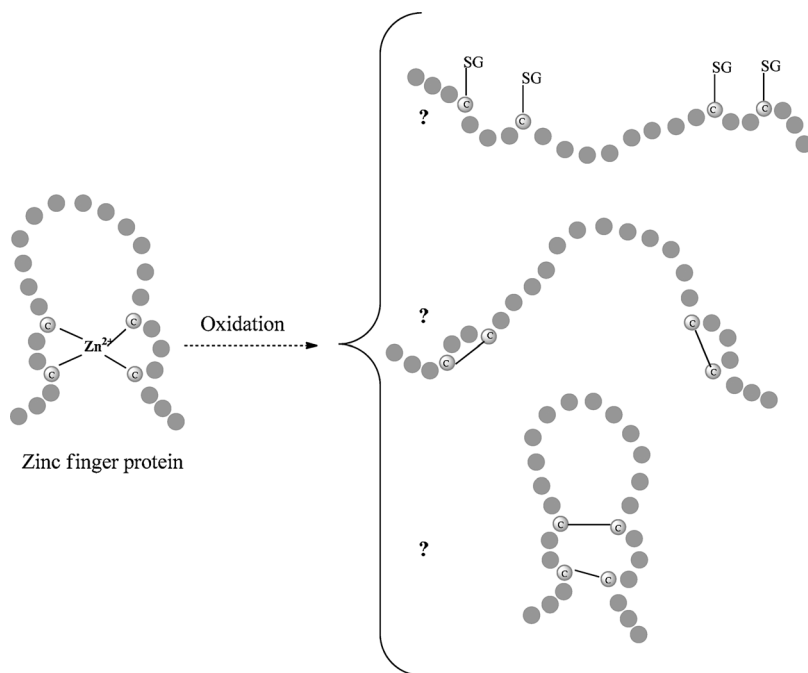
The biochemical importance of this kind of intracellular redox chemistry is highlighted further by the glutathione/glutathione disulfide redox pair. GSH is present in most human cells in cytosolic concentrations of around 10 mM. (Without any intention to belittle the overarching importance of metal ion-based redox chemistry in the human body, none of the iron-, copper-, or manganese-based redox systems occurs at such high concentrations in the human cell. Neither are such metalloproteins able to change between 10 oxidation states or employ a host of different redox mechanisms. Such metal ions – on their own – also do not possess the kind of features that allow sulfur to take on an *active* role (e.g., to “respond” to redox changes).) Apart from GSH being a substrate for GST and many enzymes, the GSH/GSSG pair forms a main constituent of the cellular redox buffering system. The standard redox potential of glutathione  $E^0$  (GSH/GSSG) is  $-240$  mV versus normal hydrogen electrode (NHE), and the relative concentrations of GSH and GSSG vary from 100 : 1 in most cells to 1 : 1 in cells or cellular organelles with a more oxidizing redox environment [13–15]. In this respect, the glutathione redox pair is an “active” redox buffer – it not only maintains a reducing, negative intracellular redox potential, but also reacts with cysteine residues in proteins and enzymes. The resulting S-glutathiolation and deglutathiolation processes, which will be discussed later in Section 10.4 in more detail, may not only protect and deprotect redox sensitive cysteine residues, they may also regulate protein function and enzyme activity and participate in (redox) signal transduction. Furthermore, GSH may “play” with its sulfur chemistry. Under severe conditions of oxidative stress, for instance, the thiol (formal oxidation state of  $-2$ ) may be oxidized not just to disulfide (oxidation state

of  $-1$ ), but also to sulfonic acid (formal oxidation state of  $+4$ ), which implies the formal donation of a total of six electrons per cysteine.

The thiol/disulfide pair is not limited to glutathione, of course. It also forms an integral part of many cysteine-based (redox) proteins and enzymes, and serves as a “handle” to control protein function and enzyme activity. While the underlying thiol/disulfide chemistry is now fairly well understood, the full extent of biochemical and biological implications is still only emerging.

Apart from nucleophilicity and redox activity, the thiol group is also able to coordinate to metal ions, such as iron, copper, zinc, and cadmium [1, 16]. Generally, thiol(ate)s are “soft” ligands and the metal ions they are able to bind to are often of a soft or intermediate nature (hence ions such as  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Ca}^{2+}$  and  $\text{Mg}^{2+}$  are usually *not* affected by the presence of thiol ligands). This metal/ligand chemistry, together with disulfide bonds, electrostatic interactions, and hydrogen bonds, forms the basis for cysteine’s role in protein structure. While electrostatic interactions, for instance of the thiolate group ( $\text{RS}^-$ ), play some role at the active site of cysteine enzymes, their impact on protein structure in general is rather limited. Similarly, hydrogen bonds involving thiols are also not particularly common. In contrast, disulfide bridges and metal/sulfur complexes play an overwhelming role in protein structure.

We will briefly consider the metal/sulfur complexes, whose integrity is frequently governed by sulfur redox chemistry. Compared to electrostatic interactions or hydrogen bonds, the coordinative bond found in metal/sulfur complexes is strong. Such complexes are found in many proteins and enzymes, where they mostly fulfill a role as a structural or catalytic entity, but on occasion are also involved in the maintenance of (essential) metal ion homeostasis or assist in the detoxification of toxic metal ions (such as cadmium). Iron/sulfur clusters, for instance, are widespread in biology [1]. They can be found in lower organisms, such as the ferredoxin proteins in *Clostridium pasteurianum* and *Azotobacter vinelandii* [17, 18]. Iron/sulfur proteins mostly function as redox proteins, although some of them, such as aconitase, may also – maybe additionally – serve as iron sensors [19]. (It is not unusual for one protein or enzyme to fulfill two distinct biological functions. Apart from aconitase, we will later encounter the Prdxs, which appear to “switch” between catalytic and chaperone activity.) Zinc/sulfur proteins are of similar importance. The metallothionein (MT) proteins, for instance, are omnipresent in mammalian kidney and liver, where they form an integral part of zinc trafficking and possibly cadmium and copper detoxification [20]. These MT-1 and MT-2 proteins are around 60 amino acids in size, and their 20 cysteine residues form two distinct zinc/sulfur clusters (i.e., a  $\text{Zn}_4\text{Cys}_{11}$   $\alpha$ -cluster and a  $\text{Zn}_3\text{Cys}_9$   $\beta$ -cluster). A related cluster structure exists in MT-3, which occurs in human brain and strongly binds  $\text{Cu}^+$  ions in copper/sulfur clusters. (It should be mentioned that the precise biological function or functions of the MT proteins is still not fully understood. Apart from an involvement in cadmium detoxification, zinc homeostasis, as metal ion chaperone, or DNA-binding protein, the MT proteins may also act as powerful antioxidants, with each MT molecule being oxidized from the thiol to sulfonic acid state formally able to donate  $20 \times 6 = 120$  electrons!) Other prominent zinc/sulfur clusters are found in the zinc finger



**Figure 10.3** Schematic representation of a representative CCCC zinc finger (e.g., as found in human steroid hormone receptors). The structural integrity of the finger is provided by the tetrahedrally coordinated  $Zn^{2+}$  ion at the finger's base. Removal of the  $Zn^{2+}$  ion (e.g., with the help of a (strong) metal chelator) or oxidation of the cysteine residues results in the disintegration of the  $Zn^{2+}$  complex and loss of structural integrity. Structural features of the

oxidized form or forms of such fingers are often not well characterized. While an open, random structure may be assumed (see top), other alternatives are possible, depending on the oxidizing agent used and the disulfides ultimately formed. It is worth mentioning that intramolecular disulfides may generate a structure similar to the original finger, while intermolecular disulfides may well result in a complex, polymeric product (not shown).

proteins [21, 22]. These proteins feature loop-like finger structures, which are held together at their base by tetrahedrally coordinated  $Zn^{2+}$  ions bound to cysteine (C) and histidine (H) residues (Figure 10.3). Various forms of these fingers exist, with either CCHH, CCCH, or CCCC coordination of  $Zn^{2+}$ . This structural finger motif is often found in transcription factors, where the “finger” is used to insert into the major groove of target DNA.

From a chemist's perspective, the zinc finger is a rather interesting structure. While the cysteine (and histidine) residues are firmly held in their respective positions by the metal ion, the finger is also sensitive to a range of external influences. Chelators, for instance, are able to remove the metal ion, which results in opening and disintegration of the finger [23, 24]. This effect is exploited in the area of drug development, whereby the chelator would remove the  $Zn^{2+}$  ion, disintegrate the finger, and hence prevent its activity (e.g., in promoting cell division).  $Zn^{2+}$  in these

fingers may also “exchange” for other metal ions, such as  $\text{Cd}^{2+}$ , which may affect the stability or function of the finger [25]. Most importantly, however, the zinc finger is also sensitive towards oxidation. Although thiols bound to metal ions are generally less prone to oxidation when compared to “free” thiols, reasonably strong oxidants or electrophiles are able to modify the thiol ligands, destroy the cluster, and release the metal ion. Oxidation of cysteine ligands in metal complexes is therefore seen as a highly promising avenue in drug development [26, 27].

Interestingly, oxidation of a zinc/sulfur complex may result in disulfide bond formation at the site originally characterized by the metal/cysteine complex (Figure 10.3). This “exchange” of one sulfur-based structural element for another may have significant biochemical implications, which are still hardly explored. To date, oxidation of a structural zinc/sulfur site is considered as being detrimental to the structural integrity and function of the protein affected. Nonetheless, it is possible that the disulfide formed as part of this process may retain some of the structural features of the original metal/sulfur complex, albeit probably in a modified form. This issue of “exchanging” a metal/cysteine complex for disulfide(s) requires further investigation.

Before turning our attention to disulfide bonds, we should briefly mention that in addition to thiols, sulfenic ( $\text{RSOH}$ ) and sulfinic acids ( $\text{RSO}_2\text{H}$ ) are also able to bind metal ions. The bacterial nitrile hydratase, which occurs in *Rhodococcus* sp., for instance, contains a nonheme  $\text{Fe}^{3+}$  or noncorrinoid  $\text{Co}^{3+}$  ion firmly coordinated to an oxygen atom of a cysteine sulfinic acid [28–30].

### 10.3

#### Towards a Dynamic Picture of Disulfide Bonds

While thiols are often at the center of nucleophilic attacks, redox reactions, and catalysis, disulfide bonds provide an enormous potential for protein architecture, which in turn can be used for regulatory and signaling purposes. Disulfides may be formed within a single protein as intramolecular disulfide bonds. They may also occur between two separate protein molecules, linking them together as either homodimers (a dimer consisting of two identical subunits) or heterodimers (a dimer consisting of two different subunits). Insulin is a fine example of a molecule where two short peptide chains (called the  $\alpha$ -chain and  $\beta$ -chain) are held together by two well-defined disulfide bonds [31]. In theory, disulfides may be used to connect not only two but three or more proteins, forming oligo- or polymeric structures. To the best of our knowledge, such extensive disulfide-bridged polymers have not (yet) been identified in the human cell. Nonetheless, there has been speculation that oxidized MTs – at least *in vitro* – may form random, polymeric structures held together by a multitude of intra- and intermolecular bonds (Jacob, C., Maret, W., and Vallee, B.L., unpublished results). (Since the MT molecule contains 20 cysteines out of 60 amino acids, its oxidation (e.g., by hydrogen peroxide) will result in a complicated mixture of products. At high MT concentrations, formation of polymeric oxidized MT is a realistic possibility. Alternatively, *S*-glutathiolation of MT or the formation of

“overoxidized” MT containing sulfenic, sulfinic, or sulfonic acid residues is also possible.)

The full extent and impact of disulfide bond formation in proteins and enzymes is only just emerging. Historically, there were a few apparent “paradoxes” concerning disulfide bond formation, which had to be understood first. We will briefly mention some of them, since they provide valuable insights into the biological chemistry of the disulfide bond and in turn help us to avoid some basic misunderstandings. We will primarily focus on “chemical” aspects, since Chapter 11 of this volume provides an extensive discussion of disulfide bonds in proteins and enzymes.

First of all, there has been reasonable doubt that disulfide bonds would be stable inside the cell (i.e., under reducing conditions and in the presence of a large excess of GSH). While disulfides may well survive in the oxidizing conditions outside the cell, they may surely be reduced readily by GSH via a thiol/disulfide exchange? This is not the case – the redox potential of the cysteine-based thiol/disulfide redox pair in proteins differs considerably, depending on the microenvironment provided by the protein (Table 10.3). (Traditionally, biochemists tend to make a direct correlation between the  $pK_a$  value of a thiol and its oxidation potential. In this view, thiols with a low  $pK_a$  are very acidic, hence easily deprotonated and more reactive (as nucleophile), and therefore exhibit a more negative oxidation potential. While the comparison of  $pK_a$  and  $E^0$  has considerable merit, one should not consider this as a direct causal relationship between  $pK_a$ , on the one hand, and  $E^0$ , on the other. *De facto*, a low  $pK_a$  and a negative  $E^0$  have common causes – a particular electronic state of the thiol(ate) that is associated with good electron (pair) donor properties.) While thioredoxin (Trx) proteins, for instance, exhibit a rather reducing potential of around  $-270$  mV, other proteins, such as PDIs, are considerably more oxidizing, with a redox potential of around  $-190$  mV [4, 32, 33]. Table 10.3 provides a brief and highly selective overview of some of the thiol/disulfide redox potentials of proteins found inside the human cell. While these potentials are all associated with thiol groups of cysteine, there are dramatic differences between them (145 mV and more). Certain proteins, which contain highly reducing cysteine residues, may therefore be able to form disulfide bonds even in the presence of high concentrations of GSH. In contrast, disulfides formed by other cysteine residues, which are less reducing, are prone to reduction by GSH as long as the disulfide is accessible. Other factors may also contribute to disulfide stability. Certain disulfides in proteins may be stable under reducing conditions since they are sterically inaccessible. Furthermore, each cellular compartment exhibits its own, specific redox environment. While the cytosol may be dominated by GSH, certain organelles, such as the endoplasmic reticulum, provide conditions appropriate for disulfide formation.

Interestingly, the different individual redox potentials of cysteine residues, which are fine-tuned by the respective protein microenvironments, also answer another puzzling question: why are disulfides not formed “randomly,” but only between specific cysteine residues? Clearly, since not all cysteine residues are equal, disulfide bonds will not be formed just between *any* two cysteine residues, but in a more or less orderly fashion between residues which, among other criteria, are amenable to oxidation. In fact, we know now that disulfide formation is a tightly controlled



process. It involves a host of proteins endowed with the sole task of forming, guarding and correcting disulfides, among them thiol : disulfide oxidoreductases, such as the above-mentioned PDIs, Trx, and glutaredoxin [34]. Interestingly, PDIs are even able to repair proteins by rearranging incorrectly formed disulfide bonds.

The third, and maybe most intriguing question addresses the “dynamic” nature of disulfide bonds. Disulfides in proteins have traditionally been considered as resilient and “static” features. Recent studies have shown, however, that rather than presenting a fixed element of protein structure, many – but not all – intra- and intermolecular disulfides appear to be transformed regularly. Similar to initial disulfide bond formation, these transformation processes do not occur randomly; they are tightly controlled in order to fulfill a range of regulatory and signaling tasks.

This dynamic picture of disulfide formation is perhaps illustrated best when considering oxidative stress – a biochemical condition characterized by elevated levels of reactive oxygen species (ROS) and an impaired antioxidant defense [35, 36]. Under those conditions, *S*-glutathiolation (i.e., the formation of mixed protein–glutathione disulfides (PrSSG)) is observed in many proteins and enzymes. While this process modifies cysteine residues, and hence may adversely impact on protein function and enzyme activity, it also has some benefits. On the one hand, the modified protein may relay the redox signal to other proteins and ultimately trigger an antioxidant response. On the other hand, *S*-thiolation provides a simple, yet effective protection of the cysteine thiol from “overoxidation” [37]. *S*-Thiolation is reversible and the “protection group” may be removed once the cell has stabilized and ROS and GSH levels have returned to normal. Amazingly, “deprotection” (i.e., deglutathiolation) may occur almost “automatically” in the presence of normal levels of GSH. Such aspects of *S*-thiolation will be discussed in more detail in Section 10.4.

In essence, disulfide formation, transformation, and cleavage are processes nowadays considered as part of a complicated, dynamic intracellular steady state, which can tolerate various disturbances, such as oxidative stress, and is able to respond to such disturbances by triggering extensive control, regulatory, and feedback mechanisms [38]. Within this context, disulfides may regulate protein function and enzyme activity via several avenues. In some proteins, intramolecular disulfide bonds form essential structural elements. Their reduction would lead to an unfolding or misfolding of the protein, and hence loss of function. Similarly, intermolecular disulfide bonds between (subunits) of proteins may be essential to lock together individual peptide chains. Loss of such disulfides would also result in impaired function and activity. Indeed, there are various examples of enzymes that *gain* activity upon oxidation and disulfide formation. Yap1, for instance, is an enzyme found in *Saccharomyces cerevisiae*. It is activated by formation of a structurally important disulfide bond between Cys303 and Cys598 [39, 40]. Oxidation of a cysteine residue (by H<sub>2</sub>O<sub>2</sub>) also plays a decisive role in the activation of nuclear factor NFκB: it triggers dissociation of the IκB–NFκB complex, the release of NFκB, and subsequent translocation to the nucleus, where NFκB controls the transcription of a host of genes [37, 41].

In contrast, many enzymes contain active-site cysteine residues and modification of these catalytically active thiol(ate)s (e.g., by *S*-thiolation) may result in significant loss of activity. There are numerous examples of enzymes “inactivated” by disulfide bond formation, including GAPDH and H<sup>+</sup>-ATPase [42, 43]. Importantly, the underlying thiol/disulfide chemistry is mostly reversible and reactivation is generally possible in the presence of reducing agents, such as GSH or the (non-natural) dithiothreitol, which is commonly used to protect protein preparations from oxidation. The circumstance that activity may be lost *and* restored by *S*-thiolation provides an almost ideal basis for redox feedback, control, and signaling processes. In contrast, some proteins, such as the MTs, appear to be oxidized *irreversibly*, possibly due to the formation of polymers which contain inaccessible disulfides difficult to reduce (see above).

#### 10.4

#### Chemical Protection and Regulation via *S*-Thiolation

Apart from the formation of intramolecular disulfide bonds, which may influence the activity of an enzyme by altering its structure, *S*-thiolation by small molecules, such as glutathione, plays an important role in protein (bio-)chemistry [38]. In this respect, *S*-thiolation may serve various roles, from a “simple” protection of thiols against “overoxidation” to a regulatory element, which senses the intracellular redox state and triggers an antioxidant response. There is also growing evidence that disulfide bond formation plays a major role in intracellular signaling. In any case, the disulfide “chemistry” at the center of *S*-thiolation is strikingly simple from a chemical point of view, yet highly efficient in a biological context [44].

Let us consider the various avenues of *S*-thiolation first. Figure 10.4 summarizes *S*-thiolation pathways and their biochemical implications in the context of intracellular oxidative stress, as exemplified by *S*-glutathiolation. First of all, *S*-glutathiolation of a protein cysteine residue (PrSH) may occur by thiol/disulfide exchange with GSSG resulting in a mixed disulfide (PrSSG) and GSH. GSSG is normally present in cells in submillimolar concentrations, yet this concentration may increase due to oxidative stress. *S*-Glutathiolation of proteins by GSSG is therefore the direct consequence of a disturbed intracellular redox homeostasis, whereby the GSH:GSSG ratio shifts towards more oxidizing values. Biochemically, changes in protein function or enzyme activity due to *S*-glutathiolation may be seen as an immediate response towards oxidative stress. They may, for instance, protect the most vulnerable cysteine thiols from overoxidation to sulfenic (RSOH), sulfinic (RS(O)OH), or sulfonic acids (RS(O)<sub>2</sub>OH). Interestingly, the thiols most sensitive towards ROS are actually also the ones to react with GSSG first, which makes this type of “chemical protection” of protein thiols by *S*-glutathiolation particularly effective.

Importantly, *S*-glutathiolation and subsequent changes in protein function or enzyme activity may also trigger extensive cellular signaling pathways. The latter may ultimately activate an antioxidant defense or direct the cell towards apoptosis.

Once a normal cellular redox environment with sufficient amounts of GSH has been restored, GSH may spontaneously cleave the disulfide bond, remove the glutathione moiety, and restore the protein to its normal structure and function. Although the full extent of intracellular protein S-thiolation (e.g., under normal conditions and during oxidative stress) is only just emerging, and many aspects of this simple protection and response mechanism are still poorly understood, it highlights the importance and dynamic character of intracellular disulfide bond formation.

Apart from S-glutathiolation via GSSG, sulfur chemistry in proteins provides several additional avenues leading to protein mixed disulfides, some of which may, however, be less common. Figure 10.4 illustrates some of these pathways, using a particular color coding for oxidative damage, protection, and restoration (rescue) of the initial thiol or disulfide oxidation state. If a thiol is attacked by ROS, such as  $\text{H}_2\text{O}_2$ , a highly reactive sulfenic acid ( $\text{PrSOH}$ ) is initially formed. The latter may react via several avenues. It may “dimerize” to form a thiosulfinate ( $\text{PrS(O)SPr}$ ) or become oxidized to form a sulfinic acid ( $\text{PrS(O)OH}$ ). Most likely, however, is the reaction with another thiol, such as  $\text{PrSH}$  or GSH, to form water and a mixed disulfide,  $\text{PrSSPr}$  or  $\text{PrSSG}$ , respectively. This disulfide formation reaction is extraordinarily fast under physiological conditions. As a result, the cysteine that has initially been attacked by an oxidant has ultimately reacted to gain a (reversible) protection from further oxidation in the form of a disulfide.

A related chemical protection mechanism based on a protein sulfenic acid has been observed in protein tyrosine phosphatase B, where the undesired oxidation of the active-site cysteine residue results in the formation of a sulfenic acid. The latter subsequently reacts with a nitrogen atom of a backbone amide to form a rather unusual cyclic sulfenyl amide ( $\text{RSN(H)R}'$ ) [45, 46]. The sulfenyl amide provides a temporary protection from further oxidation. It also reacts with thiols to form (the more stable) disulfide – a reaction that cleaves the ring and restores the backbone to its original form. The disulfide may then be reduced to the thiol.

A similar chemical protection mechanism is observed for thiols oxidized to thiyl radicals ( $\text{PrS}^\bullet$ ). The latter may react rapidly with  $\text{PrSH}$  or GSH to form disulfide radical anions,  $\text{PrSSPr}^{\bullet-}$  or  $\text{PrSSG}^{\bullet-}$ , respectively. These radicals are extraordinarily reactive; they act as reducing agents, donate one electron, and form disulfides (i.e.,  $\text{PrSSPr}$  or  $\text{PrSSG}$ ). In this case, initial oxidation of the thiol has also resulted in (reversible) thiol protection. In addition, one electron has been donated to fight off oxidants. (Unfortunately, the chemistry surrounding disulfide radical anions is more complicated. It has been postulated that the one electron donated is actually mostly accepted by dioxygen, which is reduced to the superoxide radical anion ( $\text{O}_2^{\bullet-}$ ), itself a ROS. Rather than fighting off oxidants, the electron donated from the disulfide radical anion may therefore actually *increase* oxidative stress. This issue is contentious, and the processes will ultimately depend on which oxidants are present in the cell and at which concentrations. Only then is it possible to decide which of them acquires the electron in question.)



The chemistry surrounding oxidative stress and ROS frequently raises the question if the disulfide bond is indeed an efficient protecting group against overoxidation. At first sight, the answer is “no.” Disulfides are also sensitive towards oxidation by certain ROS, such as  $\text{H}_2\text{O}_2$  and hydroxyl radicals ( $\text{HO}^\bullet$ ), resulting in thiosulfonates ( $\text{RS(O)SR}$ ) and thiosulfonates ( $\text{RS(O)}_2\text{SR}$ ). The biochemical importance of this chemistry “beyond” the disulfide state, however, has long been ignored. Nonetheless, it provides several interesting insights, some of which will also be discussed in the following sections. If a disulfide is (accidentally) oxidized to a thiosulfinate, the latter reacts readily with thiols to form a disulfide and sulfenic acid, which in turn reacts with another thiol to form a second disulfide [7]. In total, a thiosulfinate, such as  $\text{PrS(O)SPr}$ , reacts with two equivalents of thiol, such as GSH, to form two S-glutathiolated thiols  $\text{PrSSG}$  and  $\text{H}_2\text{O}$ . In effect, this returns the disulfide to its original (oxidation) state (albeit in the form of two mixed disulfides, which may regroup to  $\text{PrSSPr}$  and  $\text{GSSG}$ ) and fends off the oxidative damage caused by ROS. A similar sequence of reactions may occur if the thiosulfonate ( $\text{PrS(O)}_2\text{SPr}$ ) is formed, although the issue of reducing sulfinic acids  $\text{PrS(O)OH}$  in proteins under physiological conditions is still a matter of ongoing investigations.

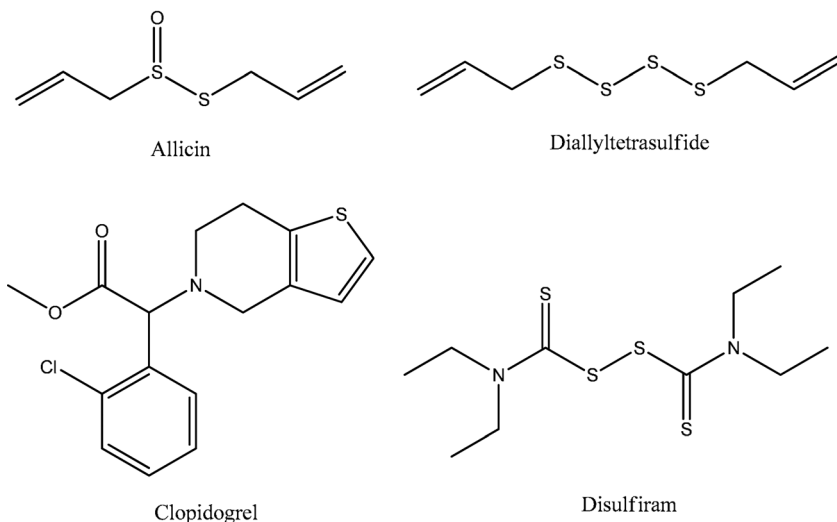
Importantly, the disulfide attacked by ROS in both cases (i.e., as thiosulfinate and thiosulfonate) returns to its original state by a sequel of rapid – and spontaneous – chemical transformations, which in essence provide some protection even in the presence of highly oxidizing ROS able to “overoxidize” disulfides. Nonetheless, in the presence of overwhelming amounts of ROS, disulfide formation may only stem the tide of oxidants for a certain while and ultimately the cell will die via apoptotic or necrotic pathways, both of which have been associated with oxidative stress on numerous occasions.

It should be mentioned that S-thiolation is not limited to events associated with oxidative stress. Various natural products and drug molecules also react with thiols to form disulfides. Allicin, for instance, is a highly reactive thiosulfinate ( $\text{RS(O)SR}$ ) found in garlic (Figure 10.5) [47, 48]. It reacts readily and indiscriminately with cysteine residues in proteins and enzymes, such as  $\alpha$ -tubulin and certain caspases. Together with diallyldisulfide and diallylpolysulfides [49], which also S-thiolate cysteine residues in proteins, allicin has been considered as a selective anticancer and antimicrobial agent. The underlying S-thiolation processes, which unfortunately are still little understood, appear to trigger (redox-controlled) signaling cascades that ultimately lead to (selective) cell death.

Many issues surrounding the dynamic nature of S-thiolation are still not fully understood. While the underlying sulfur chemistry is slowly emerging, the extent of such reactions *in vivo* and the processes they may subsequently trigger at the level of proteins and whole cells are still vastly unknown.

---

← redox systems, that initial oxidation of a thiol or disulfide initiates the chemistry that in the end protects these groups against overoxidation and allows restoration of the thiol or disulfide state after oxidative stress is removed.



**Figure 10.5** Chemical structures of some of the thiol-specific reagents mentioned in this chapter. While the thiosulfinate allicin reacts readily with most thiols to form (mixed) disulfides, the reactivity of the allyltri- and tetrasulfide is still under investigation. While these polysulfides may react with thiols, they may also form the entry point to an extensive perthiol chemistry, which may result in the formation of  $O_2^{\cdot-}$  and  $H_2O_2$ . The latter may attack thiols and ultimately also cause disulfide formation. Disulfiram, a dithiourea, has been

used to treat alcoholism in the past. This compound and its metabolic products react with cysteine residues, including the active-site cysteine in acetaldehyde dehydrogenase. Clopidogrel, on the other hand, is not itself reactive towards thiols. It represents a prodrug, which is modified by cytochrome P450 enzymes. The resulting thiol reacts with the platelet ADP receptor  $P2Y_{12}$  to form a disulfide bond and to inhibit platelet aggregation. It is used in several arterial and vascular diseases associated with undesired blood clotting.

## 10.5

### “Dormant” Catalytic Sites

A recent study by Cha *et al.* has revealed another, rather intriguing aspect of cysteine (bio-)chemistry [50, 51]. When incubating human serum albumin (HSA) with hydrogen or lipid peroxides in the presence of a reducing agent (such as Trx/Trx reductase/NADPH) and some other “factors” (e.g., palmitoyl-CoA), the authors observed a catalytic conversion of the peroxide and thiol-based reducing agent (Trx) to water and disulfide. Albeit the catalytic activity was comparably weak, it resembled the one of cysteine- and selenocysteine-based peroxidases (i.e., Prdx and glutathione peroxidase (GPx), respectively). After ruling out an involvement of the one “free” cysteine in HSA (i.e., Cys34), the authors assigned this catalytic activity to one of the disulfide bonds in the protein (Cys392–Cys438).

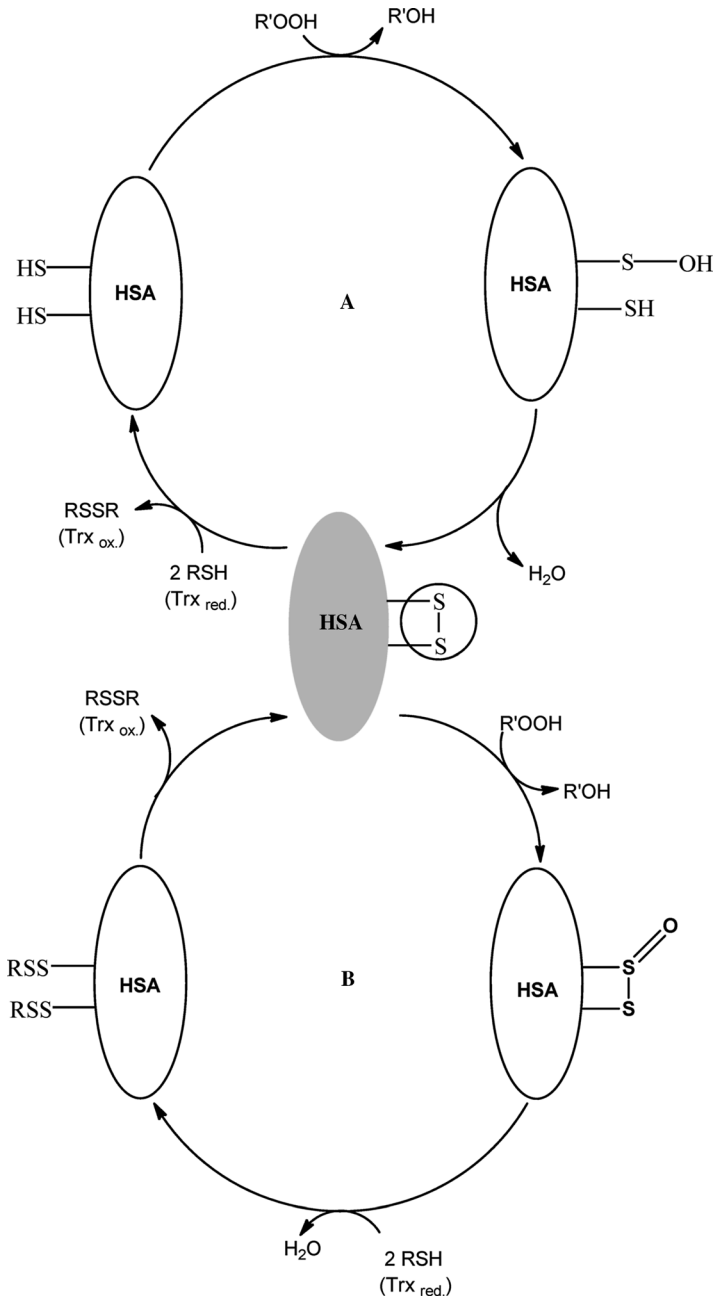
One may consider certain aspects of these studies as preliminary and in many parts also as speculative, especially with regard to a particular relevance *in vivo* (see below).

Nonetheless, these findings reveal that certain disulfide-bearing proteins may become “activated,” whereby some of their structural disulfides suddenly take on a role as catalytic sites. The possible chemistry behind such an “activated” disulfide bond is illustrated in Figure 10.6, although it must be pointed out that the two mechanisms shown are still hypothetical. Projected on the whole cell, this may point towards a dormant peroxidase-like activity in some proteins, which are otherwise wholly unrelated to redox catalysis [35, 52]. Once activated, such a dormant activity admittedly may be less effective when compared to the “usual” peroxidases, yet may be considerably more widespread and abundant – and hence of similar importance. It is possible, albeit highly speculative at this time, that such a second – or maybe even final – line of antioxidant peroxidase defense exists in the human cell based on “dormant” disulfides, which can be activated under severe conditions of oxidative stress.

## 10.6 Peroxiredoxin/Sulfiredoxin Catalysis and Control Pathway

While peroxidase-like catalysis in HSA is still a matter of debate, enzymes such as Prdx and GPx are well-known to catalyze the reduction of peroxides in the presence of (sacrificial) thiols. Nonetheless, many aspects of this antioxidant catalytic activity are still not fully understood. In fact, the last decade has unraveled a series of quite surprising aspects associated with Prdx, in general, and their catalytic cysteine residues, in particular. Figure 10.7 provides a schematic overview of the catalytic cycle of human Prdx enzymes [53]. Although forms with either one (1-Cys) or two (2-Cys) active-site cysteine residues exist, the general “chemistry” of the catalytic cycle(s) remains the same. Firstly, an active-site thiol(ate) (RSH) is oxidized by peroxide to form a sulfenic acid (RSOH) and water. This reaction proceeds via a nucleophilic attack of the thiol at  $\text{H}_2\text{O}_2$ . In a subsequent reaction, and again via nucleophilic attack, a second thiol, either in the form of a second active-site cysteine (in the case of 2-Cys Prdx) or as an external thiol (in the case of 1-Cys Prdx), replaces  $\text{HO}^-$  in RSOH to form a disulfide and water. Finally, an external, sacrificial thiol enters the fray to restore the enzyme thiol via thiol/disulfide exchange (i.e., nucleophilic substitution) and to form a dispensable disulfide. The latter may be re-reduced, for instance with the help of a reductase consuming NADPH. GSSG, for example, is reduced by glutathione disulfide reductase, and oxidized Trx is reduced by Trx reductase. These enzymes consume NADPH and hence connect the redox state of the cell with its energy household.

While the well-known antioxidant peroxidase GPx exhibits a similar catalytic cycle – based on a selenol rather than thiol – there are notable differences between GPx, on the one hand, and the Prdx enzymes, on the other. These differences in sulfur and selenium chemistry are not just coincidental, but endow Prdx with a wide range of additional features that are, to the best of today’s knowledge, absent in GPx. As Figure 10.7 illustrates, the catalytic cycle of Prdx passes through a sulfenic acid. As mentioned before, the latter is extraordinarily sensitive towards oxidation to a sulfinic



**Figure 10.6** Dormant catalytic sites in disulfide-containing proteins and enzymes. Although speculative at this time, there is initial evidence for a more widespread peroxidase-like activity based on disulfide-containing proteins such as HSA. Under certain conditions, a disulfide may be reduced to two thiols, which may enter a “classical” Prdx-like catalytic cycle

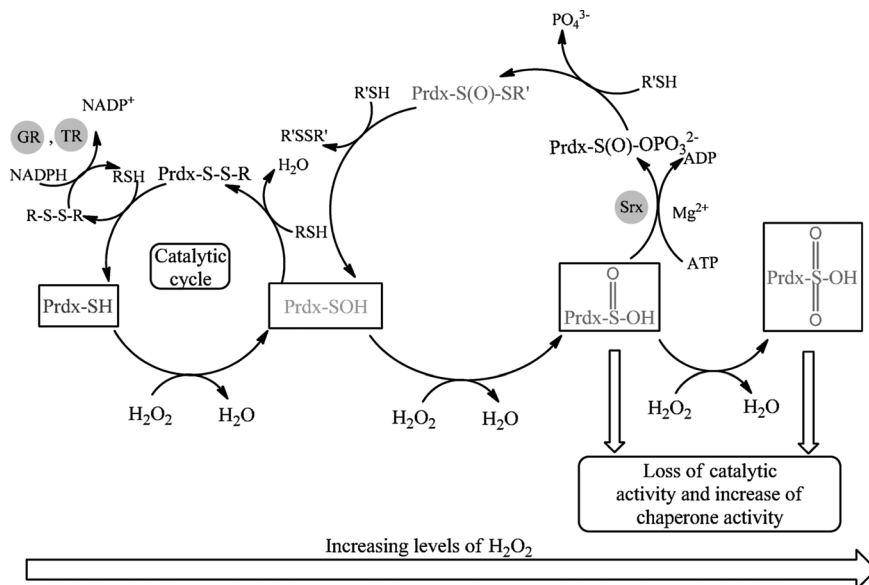


acid. It is therefore not surprising that laboratory preparations of Prdx often contain a sulfinic acid modification where a thiol, sulfenic acid, or disulfide may have been expected. Indeed, the first crystal structure of human Prdx, solved by Jennifer Littlechild and colleagues at Exeter, showed an unusual, decameric toroidal arrangement of Prdx subunits in the sulfinic acid state [55]. While this kind of “overoxidation” of the active-site cysteine – and the unusual shape – may have been seen as artifacts at the time, it has soon become apparent that there is more to the Prdx story than initially thought. We will highlight just some of the more striking aspects of Prdx sulfur chemistry here, while referring the reader to recent specialist reviews for more details.

First of all, the matter of sulfenic or sulfinic acid oxidation state in Prdx is not a question of chance. It is rather a reflection of  $H_2O_2$  concentrations, and hence a direct response to levels of  $H_2O_2$ , ROS and oxidative stress [10, 56]. In essence, Prdxs function as antioxidant peroxidases at elevated, yet tolerable concentrations of  $H_2O_2$ . If  $H_2O_2$  levels rise above a critical threshold, however, the sulfinic acid state will be formed, Prdx “switches off” its antioxidant activity and opens the “floodgate” for  $H_2O_2$  to overwhelm the cell and induce apoptosis. In this picture, the sulfenic acid switch in Prdx serves as a pivotal sensor of oxidative stress, which decides between mounting an antioxidant defense or capitulation before  $H_2O_2$ .

While this initial model of a sulfenic acid switch in Prdx as gatekeeper of an all-important  $H_2O_2$  floodgate explains many aspects of Prdx (bio-)chemistry, it has recently been modified and expanded considerably. We will consider here just two of the key elements, namely the switch between antioxidant catalytic activity and chaperone activity, and the – additional – control of this switch by the sulfiredoxin (Srx) proteins. First of all, it appears that the sulfenic acid switch serves not to just “switch off” antioxidant catalysis designed to remove  $H_2O_2$ . It rather seems that the sulfenic-to-sulfinic acid transition switches between two distinct activities of Prdx. While the catalytic, antioxidant activity is indeed switched off, another activity of (oxidized) Prdx as molecular chaperone is switched on [57, 58]. To date, little is known about this chaperone activity of Prdx. It appears that the Prdx chaperone is able to protect (certain) proteins from damage, for instance from oxidation or thermal degradation. The chaperone activity of Prdx proteins has been studied using heat-sensitive enzymes such as malate dehydrogenase [58]. It is possible that there is a connection between chaperone activity and the decameric structure of (oxidized) Prdx, and that the Prdx chaperone fulfils an important biochemical task, which is only required once the cell is overwhelmed by  $H_2O_2$ . One may speculate, for instance, that the Prdx chaperone protects proteins against high concentrations of  $H_2O_2$  or other ROS, or that it acts as a removal system of irreversibly damaged proteins. One should mention that recent studies have identified phosphorylation

← (cycle A). In contrast, oxidation of the disulfide may result in a highly reactive thiosulfinate, which may also participate in a catalytic process removing  $H_2O_2$  and consuming RSH (cycle B, which is less common than cycle A). It is possible that such processes occur in proteins which are normally redox-inactive yet react in the presence of particularly high concentrations of oxidants.



**Figure 10.7** Redox control network centered around the Prdx enzymes. Under normal conditions (i.e., in the presence of modest levels of oxidants) Prdx catalyzes the removal of H<sub>2</sub>O<sub>2</sub> in the presence of thiols. The catalytic cycle involves an active-site cysteine that passes through the thiol(ate), sulfenic acid, and disulfide oxidation states. In the presence of high levels of H<sub>2</sub>O<sub>2</sub>, the sulfenic acid may become “overoxidized” to a sulfinic acid. This process abolishes the catalytic activity of Prdx, changes the protein structure (formation of a decamer), and switches on a distinct chaperone function. A similar switch from peroxidase to chaperone function is observed when the active-site

cysteine residue is oxidized to sulfonic acid, probably under conditions of extreme oxidative stress. While sulfonic acid formation appears to be irreversible, however, the active-site sulfenic acid may be reduced back to the sulfenic acid and thiol in the presence of Srx. This protein, its associated activity, and its mode of action were only discovered in 2003 [54]. The reduction of sulfinic to sulfenic acid includes several highly unusual sulfur chemotypes, such as a transient sulfenic acid phosphoryl ester and a thiosulfinate. Srx not only restores catalytic activity in Prdx, it also connects intracellular H<sub>2</sub>O<sub>2</sub> levels with both GSH and Trx, resulting in an extensive redox sensing, response, and signaling network.

as an additional control of Prdx catalytic and chaperone activity, which further complicates this matter [59, 60].

A second twist of the initial picture of the sulfenic acid switch has emerged in the form of the Srx proteins, which are able to reduce the cysteine sulfinic acid in Prdx to sulfenic acid and hence restore the enzyme’s original antioxidant peroxidase activity. The Srx proteins were first described in yeast by Michel Toledano and colleagues in 2003 [54]. Similar proteins, including members of the sestrin protein family, have now been identified in humans [61]. The discovery of these proteins must be seen as a major turning point in biological cysteine chemistry and biochemistry. For instance, the reduction of a sulfinic acid to sulfenic acid has previously been considered as highly difficult – and even impossible – to achieve under physiological conditions. Indeed, the chemistry employed by Srx to reduce the sulfinic acid must be described

as extraordinary – a true masterpiece of biological chemistry (Figure 10.7). Rather than employing a more classical electron or hydride transfer mechanism to yield a direct reduction, the sulfinic acid is modified first to form a sulfinic acid phosphoryl ester ( $\text{RS(O)OPO}_3^{2-}$ ). The latter can subsequently be attacked by a thiol, which replaces the phosphate ( $\text{PO}_4^{3-}$ ) and forms a thiosulfinate ( $\text{RS(O)SR'}$ ). The thiosulfinate is highly reactive. It is reduced rapidly by thiols such as GSH to form the desired sulfenic acid and a disulfide.

One should note that the “undesired” oxygen atom of the sulfinic acid (when compared to the sulfenic acid), which has long been considered as irreversibly stuck to the sulfur atom, is actually removed rather gently as part of a phosphate leaving group. Furthermore, one must also point out that the thiosulfinate formed as an integral part of the reductive mechanism is probably the first example of this sulfur chemotype in *human* biology. The “normal” presence of a thiosulfinate in Prdx has had a major impact on the scientific community – it has led to the acceptance of this and similar sulfur chemotypes in human biochemistry.

The discovery of sulfinic acid reduction in Prdx has many biochemical implications. It may be seen as part of a rather extensive feedback loop that controls Prdx activity with the assistance of Srx and in response to varying levels of  $\text{H}_2\text{O}_2$  – and possibly dependent on intracellular thiols and ATP, which are also involved in sulfinic acid reduction. Ultimately, a rather complicated network of activities and controls centered around the various Prdx enzymes is emerging, incorporating various oxidation state-dependent activities, redox control by  $\text{H}_2\text{O}_2$ , thiols, cofactors, and proteins [37]. While this network seems to stretch well beyond the initial floodgate model, most of the key players are still little understood or not even identified. It is also still not known, for instance, if the sulfinic acid in Prdx is a “special case,” or if sulfinic acid reduction is more common *in vivo*. Interestingly, recent attempts to identify such sulfur chemotypes in proteins and enzymes with the help of specifically designed chemical probes in combination with modern proteomic techniques have pointed towards the widespread presence of such sulfur oxidation states *in vivo*. We will discuss some of these sulfur modifications in Section 10.7.

Furthermore, a recent report by Sue Goo Rhee and colleagues discusses the possibility of a yet higher oxidation state of Prdx, where the cysteine sulfinic acid is oxidized to sulfonic acid (Figure 10.7) [57]. It appears that this oxidative modification occurs in the presence of exceptionally high concentrations of  $\text{H}_2\text{O}_2$  and may require specific, still unidentified cellular components, since this oxidation is difficult to perform *in vitro*. Like the sulfinic acid modification, the sulfonic acid form of the Prdx prefers an oligomeric, toroidal structure, exhibits little catalytic activity, yet acts as a chaperone.

It should be mentioned that the selenium redox chemistry at the center of GPx does not appear to endow this protein with such “additional,” Prdx-like functions. Although the selenenic acid ( $\text{RSeOH}$ ), which forms part of the GPx catalytic cycle, may in theory be “overoxidized” to seleninic acid ( $\text{RSe(O)OH}$ ), it seems that seleninic acid is reduced rather easily to selenenic acid, selenylsulfide ( $\text{RSeSR'}$ ), or selenol ( $\text{RSeH}$ ). There are indications that this reduction occurs spontaneously in the presence of GSH [62]. This is, of course, in sharp contrast to the sulfinic acid and

its respective reduction, which does not occur spontaneously, but may proceed with the help of Srx, and hence can be employed for rather eloquent sensing and response purposes – while seleninic acid apparently cannot.

## 10.7

### Higher Sulfur Oxidation States: From the Shadows to the Heart of Biological Sulfur Chemistry

At the turn of the millennium, cysteine chemistry may have been considered by some as a well-understood matter of thiols and disulfides with a few exotic exceptions, such as thyl radicals, at the periphery. This notion has changed dramatically during the last decade, which has provided fertile ground for biological sulfur research. Key events shaping this process have been among others: the detailed studies of Prdx (beginning around 2000; e.g., with the publication of the first X-ray crystal structure of a Prdx enzyme) [55]; the discovery of Srx (in 2003) [54]; the modern models describing regulatory aspects of dynamic disulfide formation and *S*-glutathiolation [36]; studies confirming the link between  $\text{NO}$  and *S*-nitrosylation [63]; the in-depth studies of hydrogen sulfide as third gaseous transmitter [2]; the identification of extensive cysteine modifications in proteins and enzymes with the help of specific probes and modern proteomic techniques [64]; the idea of a widespread peroxidase activity based on “activated” disulfide (e.g., in HSA) [50]; and the renewed interest in sulfur-containing natural products and drugs for applications in medicine and as “green” antimicrobial agents or pesticides [47].

The more or less unified concept of the various reactive sulfur species, which has been emerging since 2001, reflects these developments [7]. One of its key assumptions is the presence of various sulfur oxidation states which coexist under physiological conditions and are linked to each other in a controlled – and biologically important, beneficial manner. Unfortunately, it is often difficult, if not impossible, to unambiguously identify some of these chemotypes. While some of them, such as thiosulfonates and thiosulfonates, exhibit an almost identical chemical reactivity, others, such as sulfenic acids, are still difficult to “trap” due to their low reactivity [64–69]. Our knowledge regarding most of these less common sulfur species is therefore still limited. Fortunately, many sulfur chemotypes, such as sulfenic and sulfonic acids and thiosulfonates, all of which were in the shadows a decade ago, are now reasonably well understood. We will therefore briefly consider some of the remaining, *in vivo* less commonly observed and often little understood sulfur oxidation states.

Sulfenic acids have been mentioned on various occasions. One aspect of sulfenic acid chemistry still needs to be covered – the reduction of sulfenic acids to thiols. In Prdx, this reduction is achieved by nucleophilic attack of two thiols and via an intermediate disulfide. Importantly, the bacterial enzymes NADH oxidase and NADH peroxidase, which occur in *Streptococcus faecalis*, also contain a sulfenic acid, which is formed by oxidation of a thiol. In these enzymes, reduction to thiol proceeds in one step – by hydride ( $\text{H}^-$ ) transfer from NADH and via FAD [8, 70, 71]. If a similar hydride transfer-based reduction of a sulfenic or maybe even a sulfonic acid exist in

**Table 10.4** Examples of sulfur-based radicals associated with biological systems.

Radical species	Chemical formula	Radical center (formal oxidation state)	Formation pathway
Thiyl radical	$RS^\bullet$	S (-1)	RSH oxidation $R^\bullet + RSH$
Sulfide radical cation	$RSR^\bullet+$	S (-1)	RSR oxidation
Disulfide radical anion	$RSSR^{\bullet-}$	S (-1.5)	$RS^\bullet + RSH$
Disulfide radical cation	$R_2S \cdot :SR_2^{\bullet+}$	S (-1.5)	$RSR^\bullet+ + R_2S$
Perthiyl radical	$RSS^\bullet$	S (0)	RSSSR cleavage RSSH oxidation
Sulfinyl radical	$RSO^\bullet$	S (+1)	$RSH + ROO^\bullet$
Sulfonyl radical	$RS(O)_2^\bullet$	S (+3)	$RSO^\bullet$ isomerization
Sulfur trioxide radical anion	$SO_3^{\bullet-}$	S (+5)	$SO_3^{2-}$ oxidation
Sulfur pentoxide radical anion	$SO_5^{\bullet-}$	O (0)	$SO_3^{\bullet-} + O_2$
Thiyl peroxy radical	$RSOO^\bullet$	O (0)	$RS^\bullet + O_2$
Sulfonyl peroxy radical	$RS(O)_2OO^\bullet$	O (0)	$RS(O)_2^\bullet + O_2$

Chemical structures and main formation pathways are provided. Please note that the radical center in some of these radicals is not the sulfur, but the oxygen atom. In some instances, a precise assignment of the radical center is not possible, with electron density distributed between sulfur and oxygen atoms.

biology is still unclear. Indeed, sulfonic acids ( $RS(O)_2OH$ ) are still considered as irreversibly oxidized forms of cysteine – although the discovery of  $Srx$  may caution us against any absolute statement within this context.

Similarly, sulfur-based radicals are still only partially understood. While recent progress in electron paramagnetic resonance spectroscopy has enabled the detection and rather thorough characterization of thiyl radicals ( $RS^\bullet$ ) in proteins and enzymes, other sulfur-based radicals are still mostly a matter of speculation. Table 10.4 provides a brief glimpse at some of the sulfur-based radicals, merely to showcase their chemistry and to underline the possibility of their formation under physiological conditions [72, 73]. Thiyl radicals are part of catalytic cycles found in human enzymes, such as ribonucleotide reductase [74]. A range of other radicals, such as the thiyl peroxy radical, are formed from the thiyl radical by reaction with dioxygen. Similarly, the disulfide radical anion is also formed from the thiyl radical, in this case by reaction with a thiol (see above). The disulfide radical cation ( $R_2S \cdot :SR_2^{\bullet+}$ ), in contrast, results when a sulfide radical reacts with a sulfide. The disulfide radical cation stands out because of its two-center-three electron bond, and, like most sulfur-based radicals, provides plenty of scope for future investigations at the interface between (radical) chemistry, analytical biochemistry, protein chemistry, and cell biology.

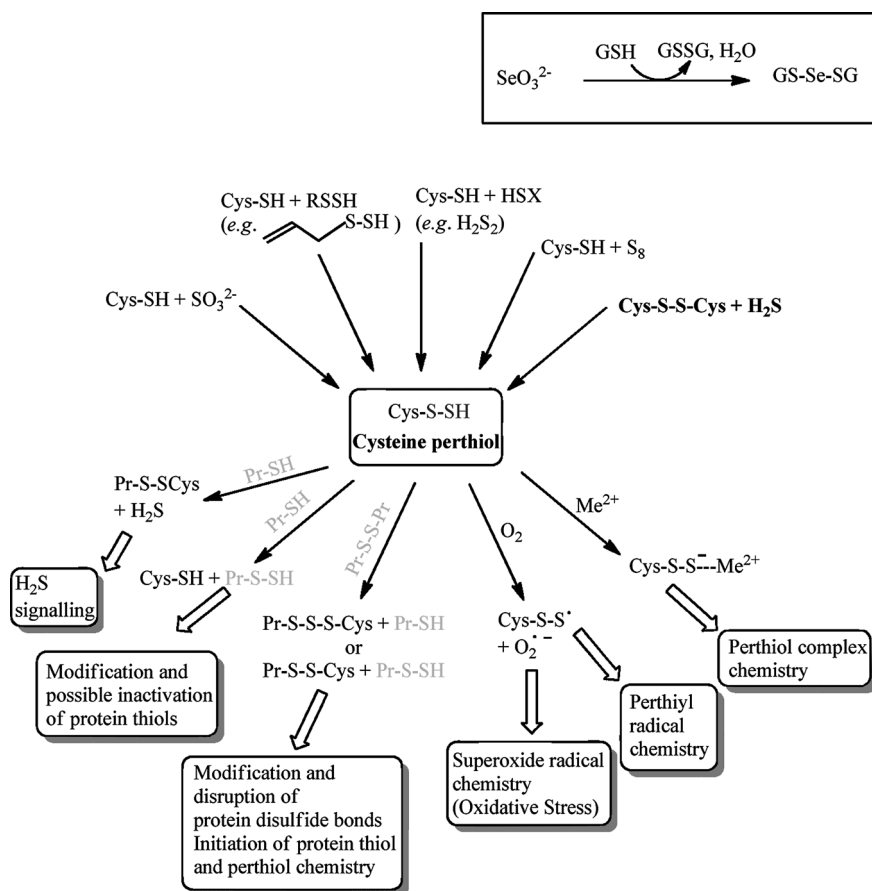
The discovery of a “normal” thiosulfinate in Prdx has provided a renewed interest in disulfide-S-oxides. Indeed, thiosulfinates are neither particularly novel nor exceptional: this chemotype is found frequently in the plant kingdom (e.g., in the compound allicin). Allicin formation in garlic is a tightly controlled process that begins with the sulfoxide alliin ((+)-S-(2-propenyl)-L-cysteine sulfoxide), involves the C-S-lyase enzyme alliinase, and proceeds via allyl sulfenic acid. Alliin itself is a cysteine derivative, which is formed from S-allyl cysteine by oxidation. The chemistry

surrounding alliin is therefore, to some extent, also a cysteine chemistry, involving the thiol, alkylsulfide (RSR'), and sulfoxide (RS(O)R') chemotype [47, 75]. The over-oxidized form of alliin containing a cysteine-based sulfone (RS(O)<sub>2</sub>R') does not appear to play a role under normal conditions. (Sulfoxides and sulfones are, however, rather important in the biochemistry of methionine. "Accidental" oxidation of this amino acid during oxidative stress is not particularly rare, and the human cell contains special methionine sulfoxide reductase enzymes that are responsible for the reduction of such sulfoxides in proteins and hence the restoration of protein function [76]. In contrast, it appears that methionine sulfone is a "dead end" and cannot be reduced under physiological conditions – neither spontaneously nor enzymatically.)

Compared to the thiosulfinate, the related thiosulfonate (RS(O)<sub>2</sub>SR') is less commonly observed. It is present in some natural plant products, such as *Scorodocarpus borneensis* [77]. Some studies have also considered thiosulfinate and thiosulfonate formation under severe conditions of oxidative stress [78]. As far as we know, the chemistry of thiosulfonates resembles that of thiosulfates: they also react readily with thiols, albeit the sulfinic acid generated in this reaction is less reactive when compared to the sulfenic acid produced from the thiosulfinate. Overall, the role of thiosulfonates in biology is still a matter of discussion and future investigation.

In contrast, the polysulfides (RS<sub>x</sub>R,  $x \geq 3$ ) have recently emerged from the shadows to become a focus of biochemical and medicinal research [5]. This is partially due to a series of studies, published between 2001 and 2007, which indicate that diallylpolysulfides are cytotoxic and able to kill bacteria, fungi, nematodes, and – in particular – cancer cells with comparable ease, while leaving healthy human cells largely unaffected [49, 79–81]. The underlying chemistry, especially inside the cell, is only just emerging and in future may explain some of the remaining open issues (e.g., as related to selectivity). There are indications that cysteine may form similar trisulfides with a sulfur atom "linking" two cysteines (Figure 10.8). Indeed, the existence of a selenotrisulfide GSSeSG as an integral part of human selenium metabolism has been known for a while [82]. Whether CysSSSCys-like trisulfides and related perthiols (RSSR; occasionally also referred to as persulfides) are formed inside the human cell (e.g., by reaction of thiols or disulfides with sulfite (SO<sub>3</sub><sup>2-</sup>), perthiols (RSSH), sulfides, hydrogen sulfide, or sulfur (S<sub>8</sub>)) is still a question that remains to be answered. It is also entirely unclear which role such trisulfides may play in proteins – if any. One may speculate that such structures, if formed in proteins, may have dramatic effects on the proteins affected.

Figure 10.8 provides some hypothetical reactions that may possibly be associated with these trisulfides. On the one hand, trisulfides are reactive and prone to reduction. Once reduced, however, the perthiol formed (RSSH) is itself highly reactive. It may act as reducing agent (e.g., reducing dioxygen to O<sub>2</sub><sup>•-</sup>). Alternatively, the perthiol could be reduced further to thiol and hydrogen sulfide. It may also facilitate a "sulfur transfer" to other thiols, as has been shown for the perthiol in the human liver enzyme rhodanese that catalyzes the detoxification of cyanide by sulfur transfer and is the only human enzyme known to date to contain a perthiol with a particular function [83, 84]. Trisulfides, and in particular perthiols (whose pK<sub>a</sub> is



**Figure 10.8** Overview of a potential perthiol (RSSH) chemistry centered around cysteine. To date, the liver enzyme rhodanese is the only human enzyme associated with an active-site perthiol. Nonetheless, there are indications that perthiol chemistry may be more widespread in biology than previously thought. Cysteine perthiols may be formed from thiols by the reaction with sulfite ( $\text{SO}_3^{2-}$ ), in analogy with the formation of GSSeSG from selenite ( $\text{SeO}_3^{2-}$ , see insert). More likely, however, is a formation by “sulfur transfer” from another perthiol (e.g., allyl perthiol, which is formed from several natural products) or in the presence of an (electrophilic) sulfide derivative (e.g.,  $\text{H}_2\text{S}_2$ ). Interestingly, certain disulfides may be reduced by hydrogen sulfide to form a perthiol. Hydrogen sulfide may also attack

thiols to form a perthiol, yet this reaction would require the presence of an oxidant. In any case, a cysteine perthiol, once formed, is rather reactive. It may interact with metal ions, cleave disulfides with subsequent formation of other reactive perthiols, or enter a rather complicated perthiyl radical chemistry, which may form superoxide radicals ( $\text{O}_2^{\cdot -}$ ) and catalytically enhance oxidative stress. It should be emphasized that the formation as well as reaction pathways of such cysteine perthiols have a major impact on the proteins affected: active-site cysteine residues may be modified, protein functions may be changed, metal centers may be blocked, disulfides may be cleaved (disruption of protein structure), and an unusual and potentially damaging perthiol redox chemistry may be initiated.

considerably lower when compared to that of the corresponding thiols), also appear to bind to metal ions, although this matter has been hardly explored yet, possibly due to the high reactivity of perthiols *in vitro* and *in vivo*. In any case, the formation of a reactive trisulfide in a protein may trigger structural changes (once the sulfur–sulfur bond is broken) and endow the protein with an unusual, yet quite aggressive perthiol chemistry.

## 10.8

### Cysteine as a Target for Oxidants, Metal Ions, and Drug Molecules

It is hardly surprising that cysteine and the biological chemistry associated with it are frequently considered as targets for exogenous substances. Natural products as part of our daily nutrition, but also environmental pollutants, antimicrobial agents, and certain drug molecules appear to interact with cysteine residues in peptides, proteins, and enzymes. The diversity of cysteine-reactive molecules, and the amount of data available, only allow us to mention a few selected highlights, which are shown in Figure 10.5.

One should point out from the beginning that most agents able to attack the thiol group of cysteine will probably react with GSH. Such substances will be “sequestered” by the glutathione system, and subsequently pose little harm to proteins and enzymes. The latter only become the target of cysteine-specific reagents once GSH has been “depleted”, or if they contain particularly reactive cysteine residues (e.g., residues with a rather negative oxidation potential). Alternatively, certain enzymes are known to bind and actually activate cysteine-specific reagents. In this case, the enzyme is the preferred target of the activated reagent, since it is encountered first. The enzyme subsequently commits “suicide.”

Compounds that are particularly reactive towards cysteine are electrophiles and oxidizing agents. Disulfides, for instance, are fairly reactive reagents that can be employed to (reversibly) modify cysteine residues. Disulfides are fairly specific and do not react with most other reducing agents found in the cell. A similar reactivity is found for the family of thiuram compounds, which has gained a certain prominence in the treatment of alcoholism (e.g., disulfiram) [85]. Indeed, the biological activity of this compound is closely related to the modification of cysteine residues in liver enzymes such as acetaldehyde dehydrogenase, which are subsequently inhibited by this agent.

In contrast, the antiplatelet agent and P2Y<sub>12</sub> receptor antagonist clopidogrel, which inhibits the formation of blood clots in arterial and vascular disorders, is a sulfur-based prodrug activated by cytochrome P450 enzymes [86–88]. Upon oxidation, the molecule forms a highly reactive thiol, which ultimately binds covalently to the platelet ADP receptor P2Y<sub>12</sub> (located at the surface of the thrombocyte) via a disulfide. This modification in essence inhibits platelet aggregation and cross-linking, which in turn prevents the formation of blood clots.

Disulfide-S-oxides may be considered as “activated” disulfides. Indeed, thiosulfonates and thiosulfonates are highly reactive and readily, yet also selectively modify



thiol groups in proteins and enzymes. This reactivity is perhaps illustrated best in the case of the garlic ingredient allicin. The widespread antimicrobial (and even anticancer) activity associated with this compound appears to be due to (reversible) modification of key cellular proteins and enzymes [47, 89, 90].

Polysulfides, on the other hand, also seem to react with thiols, although the extent, rates, and specificity of these reactions are still controversial. Diallyltrisulfide and diallyltetrasulfide (from garlic) have on occasion been considered as cysteine-modifying agents [49]. It is possible, however, that the perthiols formed from these polysulfides inside cells generate  $O_2^{\bullet-}$  radicals, which would ultimately leave a similar footprint of oxidative cysteine modification.

Other chemical species able to modify cysteine residues are peroxides, redox active metal ions as well as selenium and tellurium compounds, such as ebselen. The organochalcogen compounds are mentioned since they combine an extraordinary reactivity with high selectivity for the thiol group. The importance of selenium and tellurium reflects a certain exclusiveness that exists within the chalcogen group of the periodic system [4]. This special relationship between the elements oxygen, sulfur, selenium, and tellurium not only plays a role in the enzyme GPx, but also provides the rationale for drug design and development [91–93].

In the pharmaceutical area, cysteine provides an excellent target. Cysteine modification may be used to “switch” certain proteins either on or off. Zinc finger proteins, for instance, may be switched off by cysteine oxidation, which in turn would prevent cell proliferation. This strategy is of particular interest in cancer therapy [91]. Similarly, zinc fingers present in the coat of the HIV virus could be destroyed by oxidation, hence impairing the virus [26, 94]. In contrast, certain processes may also be switched on by oxidation of cysteine. It is known, for instance, that oxidants are able to induce apoptosis. Controlled oxidation of key signaling enzymes may therefore provide a strategy to kill cancer cells. Indeed, various so-called sensor/effector agents have been developed which mimic the catalytic cycle of GPx, consume a wide variety of thiols and appear to push certain cells over a critical oxidative stress “threshold” [91].

In the past, other sulfur oxidation states (apart from thiol) have hardly been considered as drug targets. This is rather unfortunate, since these less common oxidation states provide unique and selective targets for novel drugs. Disulfides, for instance, may be targeted in order to disrupt protein structure. In this context, reductive agents may be used, such as highly reducing, disulfide-specific thiols. Such thiols would not be reduced and sequestered by GSH, and hence may be even more resilient when compared to thiol-specific *oxidants*. Such agents may also be useful to interfere with thiosulfonates and thiosulfonates, which are more reactive than disulfides. Similarly, sulfenic acids, which are less common *in vivo* when compared to thiols and disulfides, may provide highly specific targets for drugs. We have already discussed the overarching importance of the “sulfenic acid switch” in Prdx in Section 10.6. Agents targeting this particular sulfenic acid may be used to selectively switch off certain Prdxs and hence alter the redox state of these cells. Ultimately, this kind of sulfenic acid chemistry may be used to induce apoptosis in cancer cells by irreversibly modifying the relevant Prdx enzymes.

## 10.9

### Conclusions and Outlook

The previous sections have been an attempt to describe some of the basic aspects of biological cysteine chemistry and to highlight biochemical events associated with this chemistry. Overall, the amino acid cysteine is a rather complicated matter. Compared to its chalcogen analogs serine and selenocysteine, it is either more abundant (compared to selenocysteine) or considerably more chemically diverse (compared to serine). The numerous oxidation states and chemotypes of cysteine, and the various distinct physical and chemical properties associated with each of them, still hide many secrets and potential surprises. In any case, they provide numerous opportunities for future research at the interface of chemistry, biochemistry, biology, medicine, and pharmacology.

We will conclude this chapter by highlighting some of the areas that at this point in time appear to be particularly intriguing. First of all, there is the open question regarding the sulfur oxidation states of proteins and enzymes inside the cell. Are most of the relevant cysteine residues in the thiol or disulfide state – as thought previously – or are other oxidation states, such as sulfenic and sulfinic acids, present? If yes, how widespread are such “unusual” modifications and is there a dynamic equilibrium between them?

This question brings us to the still vastly incomplete, dynamic picture of a (continuous) change of the sulfur oxidation states of cysteine residues in proteins and enzymes which together form the cellular “thiolstat” [95]. This dynamic model is supported by studies that have observed widespread *S*-glutathiolation and deglutathiolation of cysteine residues in proteins. The full extent of these processes, the involvement of oxidation states other than the thiol and disulfide state, and the biochemical impact of such dynamic modification processes are still unknown.

Related to this emerging dynamic picture is the still unsettled issue of redox-controlled “protein activation,” as has been postulated to occur in the case of HSA. The idea of proteins switching their activity due to disulfide bond formation or reduction is rather novel. Surely, there are examples of so-called “moonlighting” proteins with more than one function in the literature, yet the extent of redox switching appears to exceed by far just a handful of proteins. In fact, any protein containing a disulfide bond could, in theory, be “switched” to a cysteine enzyme by reduction of the disulfide, which yields two thiols (i.e., the basic feature for Prdx-like peroxidase catalysis).

The whole area of cysteine-based switches, which goes well beyond the disulfide or sulfenic acid switch, is still hardly explored to date. For instance, how widespread are sulfenic acids (i.e. the so-called cellular “sulfenome” [96]), is there a sulfinic/sulfonic acid switch, is sulfonic acid formation irreversible? These are questions that have recently been posed by new discoveries, such as a sulfonic acid in Prdx, yet have barely been addressed so far.

Ultimately, the answers to such questions are not just of interest in the context of basic research. They will allow us to better understand many physiological processes, including the development of diseases. In the future, an insight into these

fundamental biochemical processes may provide us with a firm and rational foundation for innovative drug design and development.

### Acknowledgments

The authors would like to thank the University of Saarland, the Ministry of Economics and Science of Saarland and the Deutsche Forschungsgemeinschaft (grant JA 1741/2-1) for financial support. We also acknowledge the support of the European Union in form of the RedCat Initial Training Network (grant 215009) and the Corena Interreg IVa Programme.

### References

- 1 Kaim, W. and Schwederski, B. (1994) *Bioinorganic Chemistry: Inorganic Elements in the Chemistry of Life*, John Wiley & Sons, Ltd, Chichester.
- 2 Jacob, C., Anwar, A., and Burkholz, T. (2008) *Planta Medica*, **74**, 1580.
- 3 Giles, N.M., Watts, A.B., Giles, G.I., Fry, F.H., Littlechild, J.A., and Jacob, C. (2003) *Chemistry and Biology*, **10**, 677.
- 4 Jacob, C., Giles, G.I., Giles, N.M., and Sies, H. (2003) *Angewandte Chemie International Edition*, **42**, 4742.
- 5 Münchberg, U., Anwar, A., Mecklenburg, S., and Jacob, C. (2007) *Organic and Biomolecular Chemistry*, **5**, 1505.
- 6 Giles, N.M., Giles, G.I., and Jacob, C. (2003) *Biochemical and Biophysical Research Communications*, **300**, 1.
- 7 Giles, G.I., Tasker, K.M., and Jacob, C. (2001) *Free Radical Biology and Medicine*, **31**, 1279.
- 8 Miller, H., Mande, S.S., Parsonage, D., Sarfaty, S.H., Hol, W.G., and Claiborne, A. (1995) *Biochemistry*, **34**, 5180.
- 9 Rudolph, J. (2005) *Antioxidants and Redox Signaling*, **7**, 761.
- 10 Poole, L.B., Karplus, P.A., and Claiborne, A. (2004) *Annual Review of Pharmacology and Toxicology*, **44**, 325.
- 11 Salinas, A.E. and Wong, M.G. (1999) *Current Medicinal Chemistry*, **6**, 279.
- 12 Kristjansdottir, K. and Rudolph, J. (2004) *Chemistry and Biology*, **11**, 1043.
- 13 Jocelyn, P.C. (1967) *European Journal of Biochemistry*, **2**, 327.
- 14 Akerboom, T.P.M., Bilzer, M., and Sies, H. (1982) *Journal of Biological Chemistry*, **257**, 4248.
- 15 Fahey, R.C. (1977) *Advances in Experimental Medicine and Biology*, **86**, 1.
- 16 Jacob, C., Maret, W., and Vallee, B.L. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3489.
- 17 Busse, S.C., La Mar, G.N., and Howard, J.B. (1991) *Journal of Biological Chemistry*, **266**, 23714.
- 18 Stephens, P.J., Morgan, T.V., Devlin, F., Penner-Hahn, J.E., Hodgson, K.O., Scott, R.A., Stout, C.D., and Burgess, B.K. (1985) *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 5661.
- 19 Holm, R.H. (1992) *Advances in Inorganic Chemistry*, **38**, 1.
- 20 Vašák, M. and Hasler, D.W. (2000) *Current Opinion in Chemical Biology*, **4**, 177.
- 21 Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003) *Nucleic Acids Research*, **31**, 532.
- 22 Gamsjaeger, R., Liew, C.K., Loughlin, F.E., Crossley, M., and Mackay, J.P. (2007) *Trends in Biochemical Sciences*, **32**, 63.
- 23 Nyborg, J.K. and Peersen, O.B. (2004) *Biochemical Journal*, **381**, e3.
- 24 Matt, T., Martinez-Yamout, M.A., Dyson, H.J., and Wright, P.E. (2004) *Biochemical Journal*, **381**, 685.
- 25 Hartwig, A., Asmuss, M., Blessing, H., Hoffmann, S., Jahnke, G., Khandelwal, S.,

- Pelzer, A., and Burkler, A. (2002) *Food and Chemical Toxicology*, **40**, 1179.
- 26 Rice, W.G., Supko, J.G., Malspeis, L., Buckheit, R.W. Jr., Clanton, D., Bu, M., Graham, L., Schaeffer, C.A., Turpin, J.A., Domagala, J., Gogliotti, R., Bader, J.P., Halliday, S.M., Coren, L., Sowder, R.C. 2nd, Arthur, L.O., and Henderson, L.E. (1995) *Science*, **270** 1194.
- 27 Lee, N., Gorelick, R.J., and Musier-Forsyth, K. (2003) *Nucleic Acids Research*, **31**, 4847.
- 28 Tsujimura, M., Dohmae, N., Odaka, M., Chijimatsu, M., Takio, K., Yohda, M., Hoshino, M., Nagashima, S., and Endo, I. (1997) *Journal of Biological Chemistry*, **272**, 29454.
- 29 Bonnet, D., Stevens, J.M., de Sousa, R.A., Sari, M.A., Mansuy, D. and Artaud, I. (2001) *Journal of Biochemistry*, **130**, 227.
- 30 Murakami, T., Nojiri, M., Nakayama, H., Odaka, M., Yohda, M., Dohmae, N., Takio, K., Nagamune, T., and Endo, I. (2000) *Protein Science*, **9**, 1024.
- 31 Sparrow, L.G., McKern, N.M., Gorman, J.J., Strike, P.M., Robinson, C.P., Bentley, J.D., and Ward, C.W. (1997) *Journal of Biological Chemistry*, **272**, 29460.
- 32 Nishinaka, Y., Masutani, H., Nakamura, H., and Yodoi, J. (2001) *Redox Report*, **6**, 289.
- 33 Lundstrom, J. and Holmgren, A. (1993) *Biochemistry*, **32**, 6649.
- 34 Ginalski, K., Kinch, L., Rychlewski, L., and Grishin, N.V. (2004) *Trends in Biochemical Sciences*, **29**, 339.
- 35 Jacob, C. and Winyard, P.G. (eds) (2009) *Redox Signaling and Regulation in Biology and Medicine*, 1st edn, Wiley-VCH Verlag GmbH, Weinheim.
- 36 Ghezzi, P., Romines, B., Fratelli, M., Eberini, I., Gianazza, E., Casagrande, S., Laragione, T., Mengozzi, M., and Herzenberg, L.A. (2002) *Molecular Immunology*, **38**, 773.
- 37 Winyard, P.G., Moody, C.J., and Jacob, C. (2005) *Trends in Biochemical Sciences*, **30**, 453.
- 38 Ghezzi, P. and Di Simplicio, P. (2009) *Redox Signaling and Regulation in Biology and Medicine* (eds C. Jacob and P.G. Winyard), Wiley-VCH Verlag GmbH, Weinheim, p. 123.
- 39 Delaunay, A., Pflieger, D., Barrault, M.B., Vinh, J., and Toledano, M.B. (2002) *Cell*, **111**, 471.
- 40 Barford, D. (2004) *Current Opinion in Structural Biology*, **14**, 679.
- 41 Janssen-Heininger, Y.M., Poynter, M.E., and Baeuerle, P.A. (2000) *Free Radical Biology and Medicine*, **28**, 1317.
- 42 Batthyany, C., Schopfer, F.J., Baker, P.R., Duran, R., Baker, L.M., Huang, Y., Cervenansky, C., Branchaud, B.P., and Freeman, B.A. (2006) *Journal of Biological Chemistry*, **281**, 20450.
- 43 Feng, Y. and Forgac, M. (1994) *Journal of Biological Chemistry*, **269**, 13224.
- 44 Jacob, C., Doering, M., and Burkholz, T. (2009) *Redox Signaling and Regulation in Biology and Medicine* (eds C. Jacob and P.G. Winyard), Wiley-VCH Verlag GmbH, Weinheim, p. 63.
- 45 Salmeen, A., Andersen, J.N., Myers, M.P., Meng, T.C., Hinks, J.A., Tonks, N.K., and Barford, D. (2003) *Nature*, **423**, 769.
- 46 van Montfort, R.L., Congreve, M., Tisi, D., Carr, R., and Jhoti, H. (2003) *Nature*, **423**, 773.
- 47 Jacob, C. (2006) *Natural Product Reports*, **23**, 851.
- 48 Jacob, C. and Anwar, A. (2008) *Physiologia Plantarum*, **133**, 469.
- 49 Hosono, T., Fukao, T., Ogihara, J., Ito, Y., Shiba, H., Seki, T., and Ariga, T. (2005) *Journal of Biological Chemistry*, **280**, 41487.
- 50 Cha, M.K. and Kim, I.H. (2006) *Archives of Biochemistry and Biophysics*, **445**, 19.
- 51 Lee, H. and Kim, I.H. (2001) *Free Radical Biology and Medicine*, **30**, 327.
- 52 Jacob, C., Knight, I., and Winyard, P.G. (2006) *Biological Chemistry*, **387**, 1385.
- 53 Szabo, K.E., Line, K., Eggelton, P., Littlechild, J.A., and Winyard, P.G. (2009) *Redox Signaling and Regulation in Biology and Medicine* (eds C. Jacob and P.G. Winyard), Wiley-VCH Verlag GmbH, Weinheim, p. 143.
- 54 Biteau, B., Labarre, J., and Toledano, M.B. (2003) *Nature*, **425**, 980.
- 55 Schröder, E., Littlechild, J.A., Lebedev, A.A., Errington, N., Vagin, A.A., and Isupov, M.N. (2000) *Structure*, **8**, 605.
- 56 Jacob, C., Holme, A.L., and Fry, F.H. (2004) *Organic and Biomolecular Chemistry*, **2**, 1953.

- 57 Lim, J.C., Choi, H.I., Park, Y.S., Nam, H.W., Woo, H.A., Kwon, K.S., Kim, Y.S., Rhee, S.G., Kim, K., and Chae, H.Z. (2008) *Journal of Biological Chemistry*, **283**, 28873.
- 58 Jang, H.H., Kim, S.Y., Park, S.K., Jeon, H.S., Lee, Y.M., Jung, J.H., Lee, S.Y., Chae, H.B., Jung, Y.J., Lee, K.O., Lim, C.O., Chung, W.S., Bahk, J.D., Yun, D.J., and Cho, M.J. (2006) *FEBS Letters*, **580** 351.
- 59 Wood, Z.A., Schroder, E., Harris, J.R., and Poole, L.B. (2003) *Trends in Biochemical Sciences*, **28**, 32.
- 60 Chang, T.-S., Jeong, W., Choi, S.Y., Yu, S., Kang, S.W., and Rhee, S.G. (2002) *Journal of Biological Chemistry*, **277**, 25370.
- 61 Budanov, A.V., Sablina, A.A., Feinstein, E., Koonin, E.V., and Chumakov, P.M. (2004) *Science*, **304**, 596.
- 62 Jacob, C., Maret, W., and Vallee, B.L. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 1910.
- 63 Torta, F., Usuelli, V., Malgaroli, A., and Bachi, A. (2008) *Proteomics*, **8**, 4484.
- 64 Saurin, A.T., Neubert, H., Brennan, J.P., and Eaton, P. (2004) *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 17982.
- 65 Eaton, P., Byers, H.L., Leeds, N., Ward, M.A., and Shattock, M.J. (2002) *Journal of Biological Chemistry*, **277**, 9806.
- 66 Brennan, J.P., Wait, R., Begum, S., Bell, J.R., Dunn, M.J., and Eaton, P. (2004) *Journal of Biological Chemistry*, **279**, 41352.
- 67 Charles, R.L., Schröder, E., May, G., Free, P., Gaffney, P.R., Wait, R., Begum, S., Heads, R.J., and Eaton, P. (2007) *Molecular and Cellular Proteomics*, **6**, 1473.
- 68 Eaton, P., Jones, M.E., McGregor, E., Dunn, M.J., Leeds, N., Byers, H.L., Leung, K.-Y., Ward, M.A., Pratt, J.R., and Shattock, M.J. (2003) *Journal of the American Society of Nephrology*, **14**, S290.
- 69 Eaton, P. (2006) *Free Radical Biology and Medicine*, **40**, 1889.
- 70 Claiborne, A., Miller, H., Parsonage, D., and Ross, R.P. (1993) *FASEB Journal*, **7**, 1483.
- 71 Parsonage, D., Miller, H., Ross, R.P., and Claiborne, A. (1993) *Journal of Biological Chemistry*, **268**, 3161.
- 72 Bonifacic, M. and Asmus, K.D. (1984) *International Journal of Radiation Biology and Related Studies in Physics, Chemistry, and Medicine*, **46**, 35.
- 73 Abedinzadeh, Z. (2001) *Canadian Journal of Physiology and Pharmacology*, **79**, 166.
- 74 Kolberg, M., Strand, K.R., Graff, P., and Andersson, K.K. (2004) *Biochimica et Biophysica Acta*, **1699**, 1.
- 75 Miron, T., Shin, I., Feigenblat, G., Weiner, L., Mirelman, D., Wilchek, M., and Rabinkov, A. (2002) *Analytical Biochemistry*, **307**, 76.
- 76 Lowther, W.T., Brot, N., Weissbach, H., and Matthews, B.W. (2000) *Biochemistry*, **39**, 13307.
- 77 Lim, H., Kubota, K., Kobayashi, A., and Sugawara, F. (1998) *Phytochemistry*, **48**, 787.
- 78 Huang, K.-P. and Huang, F.L. (2002) *Biochemical Pharmacology*, **64**, 1049.
- 79 Antosiewicz, J., Herman-Antosiewicz, A., Marynowski, S.W., and Singh, S.V. (2006) *Cancer Research*, **66**, 5379.
- 80 Jakubíková, J. and Sedlák, J. (2006) *Neoplasma*, **53**, 191.
- 81 Ha, M.W., Ma, R., Shun, L.P., Gong, Y.H., and Yuan, Y. (2005) *World Journal of Gastroenterology*, **11**, 5433.
- 82 Ganther, H.E. (1999) *Carcinogenesis*, **20**, 1657.
- 83 Iciek, M. and Wlodek, L. (2001) *Polish Journal of Pharmacology*, **53**, 215.
- 84 Westley, J. (1973) *Advances in Enzymology and Related Areas of Molecular Biology*, **39**, 327.
- 85 Petersen, E.N. (1992) *Acta Psychiatrica Scandinavica*, **369**, 7.
- 86 Savi, P., Zachayus, J.-L., Delesque-Touchard, N., Labouret, C., Herve, C., Uzabiaga, M.-F., Pereillo, J.M., Culouscou, J.-M., Bono, F., Ferrara, P., and Herbert, J.-M. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11069.
- 87 Pereillo, J.-M., Maftouh, M., Andrieu, A., Uzabiaga, M.-F., Fedeli, O., Savi, P., Pascal, M., Herbert, J.-M., Maffrand, J.P., and Picard, C. (2002) *Drug Metabolism and Disposition*, **30**, 1288.
- 88 Savi, P., Pereillo, J.M., Uzabiaga, M.F., Combalbert, J., Picard, C., Maffrand, J.P.,

- Pascal, M., and Herbert, J.M. (2000) *Thrombosis and Haemostasis*, **84**, 891.
- 89 Shadkchan, Y., Shemesh, E., Mirelman, D., Miron, T., Rabinkov, A., Wilchek, M., and Oshero, N. (2004) *Journal of Antimicrobial Chemotherapy*, **53**, 832.
- 90 Fry, F.H., Okarter, N., Baynton-Smith, C., Kershaw, M.J., Talbot, N.J., and Jacob, C. (2005) *Journal of Agricultural and Food Chemistry*, **53**, 574.
- 91 Fry, F.H. and Jacob, C. (2006) *Current Pharmaceutical Design*, **12**, 4479.
- 92 Mecklenburg, S., Collins, C.A., Döring, M., Burkholz, T., Abbas, M., Fry, F.H., Pourzand, C., and Jacob, C. (2008) *Phosphorus Sulfur and Silicon and the Related Elements*, **183**, 863.
- 93 Shabaan, S., Ba, L.A., Abbas, M., Burkholz, T., Denkert, A., Gohr, A., Wessjohann, L.A., Sasse, F., Weber, W., and Jacob, C. (2009) *Chemical Communications*, 4702.
- 94 Rice, W.G., Turpin, J.A., Schaeffer, C.A., Graham, L., Clanton, D., Buckheit, R.W. Jr., Zaharevitz, D., Summers, M.F., Wallqvist, A., and Covell, D.G. (1996) *Journal of Medicinal Chemistry*, **39**, 3606.
- 95 Jacob C., Jamier V., Ba L. A. (2010) *Current Opinion in Chemical Biology*, **15**, 149.
- 96 Leonard S.E., Reddie K. G., Carroll K. S. (2009) *ACS Chemical Biology*, **4**, 783.

## 11

# Role of Disulfide Bonds in Peptide and Protein Conformation

*Keith K. Khoo and Raymond S. Norton*

### 11.1

#### Introduction

The native structures of proteins and peptides are stabilized by a number of interactions that dictate directly or indirectly the folding, conformation, and flexibility of the molecule. Most of these interactions, such as hydrogen bonds and hydrophobic interactions, are noncovalent and relatively weak. Covalent interactions, on the other hand, are generally stronger, and are thought to exert a greater influence on stability and conformation, particularly in peptides, which tend to have fewer and weaker hydrophobic interactions because of their smaller size. The most common example of a covalent interaction is the disulfide bond, formed between the sulfur atoms of two cysteine residues [1]. Disulfide bonds are typically present in extracellular proteins and peptides, such as growth factors, hormones, enzymes, and toxins, and have also been found in several thermostable intracellular proteins of archaeal microbes [2]. They can play a role in covalently linking subunits in protein complexes (e.g., the heavy and light chains of antibodies or the two peptide chains of insulin) [3]. Structurally, the strong covalent links formed by disulfide bonds are thought to confer additional conformational stability on proteins such as keratin that have a high disulfide content [4]. Disulfides also play a catalytic role in enzymes such as thioredoxin, which acts as a cellular redox sensor via the oxidation status of its thiol groups [5–7]. Many disulfides in proteins appear to have no direct functional role, rather their main purpose is to maintain the conformation of the protein. In certain cases, conformational changes associated with the reduction and oxidation of these bonds may allow a protein to switch between different functions [8, 9]. The relationship between disulfide bonds and the conformation of peptides and proteins is thus an intriguing one. This chapter provides a broad overview of the roles of disulfide bonds in peptides, focusing on their roles in the folding and stability of proteins, as well as how different disulfide bonding patterns and connectivity give rise to different protein topologies and conformations, and ultimately a diverse range of protein functions. An understanding of these roles also has important applications in the field of protein engineering and drug design.

## 11.2

### Probing the Role of Disulfide Bonds

Several strategies have been applied to probe the roles of disulfide bonds. Typically, the disulfide bond under study is removed by replacing the selected cysteine pair with alanine, aminobutyric acid, or serine residues, and the effects on folding, stability, conformation, and activity of the peptide are monitored [10, 11]. The effects of removing a disulfide bond upon protein folding are often studied in refolding experiments using folding buffers, the most common being a mixture of oxidized and reduced glutathione, which allows the disulfides to exchange reversibly between their oxidized and reduced forms [12, 13].

Protein stability can also be measured in reductive unfolding experiments, where the protein is subjected to a strong reducing agent such as dithiothreitol (DTT) and both its unfolding kinetics and conformation are monitored as it unfolds through the progressive loss of disulfide bonds [13]. Nuclear magnetic resonance (NMR), X-ray crystallography, circular dichroism, or fluorescence spectroscopy can be used to assess the consequences of removing individual disulfide bonds upon protein structure and function. NMR spectroscopy can provide additional information on the conformational flexibility and stability of the protein [14]. Other studies have focused more on the evolution of disulfide bonds in stabilizing known protein folds, focusing on the prevalence, distribution, and position of these bonds in small disulfide-rich domains, as well as factors leading to different disulfide connectivities in various protein folds, such as conserved cysteine frameworks, amino acid sequences, and intercysteine loop size and composition [15, 16].

## 11.3

### Contribution of Disulfide Bonds to Protein Stability

It is generally accepted that disulfide bonds enhance the thermodynamic stability of proteins, making them less susceptible to denaturation and degradation in the extracellular environment, and more resistant to extremes of temperature and pH. Stability here refers to the stability of the native fold to temperature or other denaturants rather than dynamic stability, which will be discussed below. Precisely how these disulfide bonds confer stability has been the subject of some debate. One explanation is that disulfide bonds reduce the conformational freedom of the protein in its unfolded state, reducing the entropy of this state and thus destabilizing it relative to the folded state [1, 17]. Another posits that disulfide bonds destabilize the unfolded state of the protein by sterically preventing an effective hydrogen bonding network from forming [18]. While such theories focus on destabilization of the unfolded state, it has also been argued that disulfides can destabilize the folded state of a protein by reducing its conformational freedom, taking into account that the folded state is not entirely static [19, 20]. Disulfide bridges in flexible regions decrease the native state entropy more so than bridges between rigid regions. These entropic effects are potentially countered by similar but opposite enthalpic effects: bridges



between rigid regions may cause torsional strain and local repacking, whereas those between flexible regions are more readily accommodated with little loss of favorable enthalpic interactions. Thus, native state entropic and enthalpic effects must be considered as well as unfolded state entropic effects. As noted by Betz [19], crystallographic data on novel disulfides introduced into subtilisin BPN' [21, 22] indicate that disulfides are more readily accommodated by flexible regions rather than rigid ones, suggesting that enthalpic native state effects may dominate negative entropic considerations.

Loop sizes formed by native disulfides are usually large and their disulfide connectivities often result in disulfide bonds being formed between cysteines that are not sequential in the peptide sequence. These factors potentially lead to inefficient formation of the native disulfides, as seen in the cone shell peptide  $\omega$ -MVIIA [23], as they are difficult to form based on random probability. Yet such linkages are often observed, especially in small peptides, as they contribute to the relative stability of the native fold by greatly reducing the conformational entropy of the unfolded peptide chain. Unfolding of a protein is usually measured by exposing the native protein to strong reducing conditions, such as DTT, and monitoring the loss of disulfide bonds and native structure. As expected, the most solvent exposed and unprotected disulfide is usually the most susceptible to reduction and the resulting partially reduced structure largely retains native structure. Further reduction of buried disulfides usually results in a loss of native conformation. In a statistical study of proteins in the Protein Data Bank (PDB [24]), the stability of proteins was correlated with chain length and the presence of disulfide bonds [25]; this study found that the number of disulfide bonds per residue was negatively correlated with the length of the protein chain in smaller proteins (less than 200 residues), but no correlation was observed in larger proteins. A positive correlation was found between the number of disulfide bonds and the favorable free energy that stabilizes the native state relative to its unfolded state, thus affirming the stabilizing effect disulfide bonds confer on proteins in general, and especially on short proteins with fewer hydrophobic interactions. Indeed, it is not uncommon for 20–50 residue peptides such as toxins to have three or more disulfide bonds stabilizing their structure. This stabilizing effect of disulfide bonds has been used to advantage in engineering protein scaffolds, as discussed below.

## 11.4

### Role of Disulfide Bonds in Protein Folding

It is well established that the information required for the folding of a protein is encrypted within its primary amino acid sequence [26–28]. The formation of disulfide bonds is often an integral component of the folding pathway of a protein. Folding yields and efficiency are often determined by the ability of a protein to form the native disulfide bonds, but what exactly determines the formation of these bonds is not always obvious, and is complicated by the presence of other interactions and by whether these interactions promote the preferential formation of native disulfides.

The mechanism of protein folding has been studied extensively, with some of the best-studied models being small proteins with multiple disulfide bonds, such as bovine pancreatic trypsin inhibitor (BPTI) and bovine pancreatic ribonuclease A (RNase A) [12]. The isolation of stable disulfide intermediates in the folding pathway of these polypeptides has led to a more comprehensive understanding of the role of disulfides in driving protein folding, as well as their overall importance relative to other interactions [29, 30]. Several theories have been proposed regarding the role played by disulfide bonds in protein folding. One generally accepted model is a stepwise model in which a protein folds to its native structure via stable disulfide intermediates that are partially folded and form a native-like subdomain [31]. This was observed for BPTI, in which a stable intermediate containing a single disulfide bond between residues Cys30 and Cys51 was identified as a stable intermediate in the folding pathway, and was shown to adopt a partially native conformation comprising the structurally crucial  $\alpha$ -helical and  $\beta$ -sheet regions of BPTI, but with the N-terminal region remaining unfolded [32]. Subsequently, the partially folded intermediate serves as a template to direct formation of the remaining disulfides or causes steric inhibition to prevent the formation of non-native disulfides, allowing the protein to fold into its native conformation [31, 33].

In the 25-residue peptide  $\omega$ -MVIIA, by contrast, no singly or doubly disulfide-bonded intermediates formed preferentially during the initial phase of folding, pointing towards the lack of conformational specificity in driving native disulfide bond formation [23, 34]. However, once two native disulfide bonds had formed, formation of the final disulfide was favored, suggesting that other native interactions (hydrophobic, electrostatic, etc.) became more prominently involved once the polypeptide chain was significantly restrained. Thus, the disulfides play an important role in driving this polypeptide into a more compact conformation, allowing subsequent interactions to complete the final stages of folding.

In contrast to the case of small peptides such as  $\omega$ -MVIIA that lack a hydrophobic core, the exact role played by disulfide bonds in the folding of larger peptides and proteins such as BPTI and RNase A is complicated by contributions from a larger number of noncovalent native interactions. Indeed, such interactions are solely responsible for driving folding of proteins lacking disulfide bonds. An important question that arises is whether hydrophobic interactions or disulfide formation drive the folding of disulfide-rich proteins. Several protein folding models suggest that conformational sampling of the native structure, either locally or globally, allows the formation of disulfide bonds which then stabilize the native conformation [35–38]. Such a mechanism implies that disulfide bonds do not direct folding, but rather stabilize the native conformation. An alternative model proposes that disulfide formation drives protein folding [39]. Local changes induced by the formation of a single disulfide bond can direct the global folding of a protein, which often occurs cooperatively. Disulfide formation can bring about a direct global effect by bringing together residues that are far apart in the sequence and promoting the formation of a folding nucleus [40]. As proposed by Wedemeyer *et al.* for RNase A [13], many non-native disulfide intermediates form randomly during the initial phase of folding and these non-native species go through several rearrangements until a specific set of

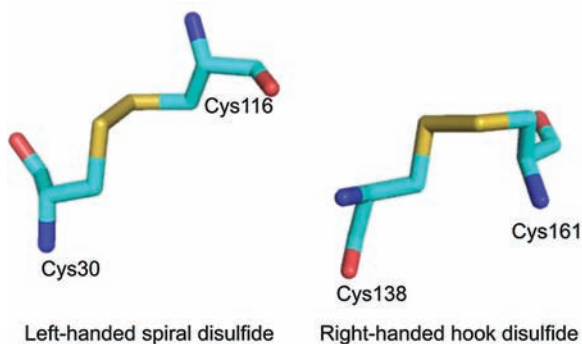
native disulfide bonds is formed that are able to direct folding. Once this occurs, the resulting structure is folded such that the disulfides are protected from rearrangement by being buried within the structure; subsequently, the remaining disulfides are formed preferentially.

Less well understood is the role of disulfide bonds in protein folding *in vivo*. Several factors potentially affect the formation of disulfides and protein folding in the cell, as reviewed recently by Bulaj and Walewska [41]. The presence of folding catalysts and molecular chaperones can have a strong influence on the folding of a protein into its native state. In particular, the enzyme protein disulfide isomerase is responsible for oxidative folding in the endoplasmic reticulum of the cell, playing an important role in the formation of natively paired disulfide bonds [42, 43]; this enzyme is present in organisms such as the cone snails that elaborate disulfide-rich peptides [44]. Other factors such as precursor sequences may also be involved in oxidative folding in the cell. Precursor sequences had little effect on the folding of  $\omega$ -conotoxin MVIIA and potato carboxypeptidase inhibitor [45, 46], but studies on other propeptides demonstrated a direct involvement of the precursor sequence in folding. The peptide hormone guanylin is an interesting example with relevance to disulfide-coupled folding as its prosequence contains a disulfide bond in addition to the two native disulfide bonds of the mature hormone [47]. Lauber's group demonstrated that the disulfide bond in the prosequence was necessary for stabilizing the tertiary structure of the prohormone, which in turn favors the formation of the native disulfide bonds, as opposed to the cysteine residues of the prosequence being involved in the formation of non-native disulfide intermediates during the oxidative folding of proguanylin [47]. In bacterial cells, the enzyme DsbA catalyzes the oxidative folding of proteins by initiating disulfide bond formation. The mechanism of this enzyme was described recently by Kadokura *et al.* [48], who found that the oxidative folding process was even affected by whether the protein was translocated cotranslationally or post-translationally.

## 11.5

### Role of Individual Disulfide Bonds in Protein Structure

The structure of the disulfide linkage itself plays an important role in protein structure as a consequence of the covalent geometry of the sulfur-sulfur bond. The covalently linked sulfur atoms are 2.0 Å apart and the C-S-S-C dihedral angle is preferably in the 90–100° range [1]. As summarized by Richardson [1], the disulfide bond adopts preferred conformations based on its cysteine side-chain  $\chi$  angles, with most disulfides having either a left-handed spiral conformation ( $\approx\chi_1 = -60^\circ$ ,  $\chi_2 = -90^\circ$ ,  $\chi_3 = -90^\circ$ ,  $\chi_2' = -90^\circ$ ,  $\chi_1' = -60^\circ$ ) or a right-handed hook conformation ( $\approx\chi_1 = -60^\circ$ ,  $\chi_2 = +120^\circ$ ,  $\chi_3 = +90^\circ$ ,  $\chi_2' = -50^\circ$ ,  $\chi_1' = -60^\circ$ ) (Figure 11.1). The C $^\alpha$  distances between the two connecting cysteines are also constrained to less than 6.5 Å in most cases as a result of the rigid torsion and preferred conformations of the disulfide bond [1]. Thus, disulfide bonds have a significant influence on protein structure in their immediate vicinity.



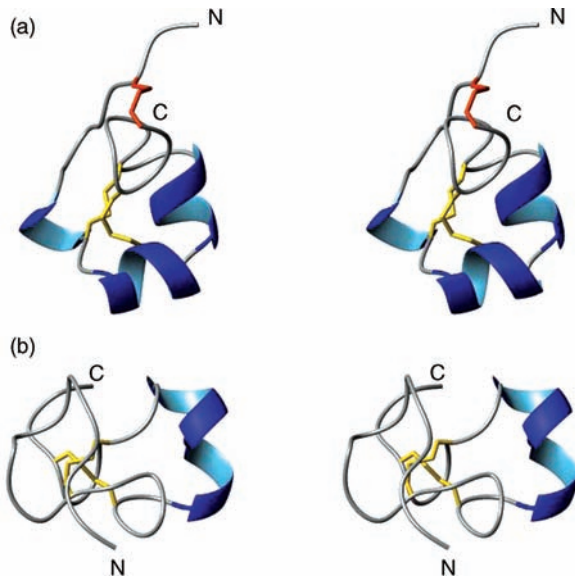
**Figure 11.1** Two major conformations of disulfide bonds: left-handed spiral formed by the Cys30–Cys116 disulfide of hen egg white lysozyme (PDB ID: 1LZ1) and right-handed

hook formed by the Cys138–Cys161 disulfide of carboxypeptidase A (PDB ID: 5CPA). Disulfide bonds are colored yellow. (Figure prepared using PyMOL; [www.pymol.org](http://www.pymol.org)).

Disulfide bonds are generally well conserved in protein families with similar folds, yet several mutational studies in proteins containing multiple disulfide bonds have shown that individual disulfides are not always necessary for maintaining the native fold. Removing a specific disulfide bond may have minimal effects on the structure and function of the protein. The crystal structure of BPTI lacking the Cys14–Cys38 disulfide bond, for example, was almost identical to the native structure [49]. The Cys13–Cys33 disulfide-deficient analog of the scorpion toxin charybdotoxin (ChTx) was also reported to adopt a native conformation [50]. In many cases, deletion of a disulfide bond was accompanied by local changes in structure but the global fold remained intact. Removing the Cys6–Cys20 disulfide bond in epidermal growth factor (EGF) [10] or the Cys1–Cys9 disulfide bond of the  $\mu$ -conotoxin  $\mu$ -KIIIA [51] resulted in local conformational changes in the N-terminal region but preservation of the overall fold. Disulfide deletion can sometimes alter local structural features of a protein. In the scorpion toxin  $\alpha$ -KTx6, removing the fourth disulfide bond caused a local change in its two-stranded  $\beta$ -sheet from a twisted to a nontwisted conformation [52], while, in insulin-like growth factor I, removing the Cys47–Cys52 disulfide bond resulted in the local unfolding of a proximal  $\alpha$ -helix [53].

Other studies, by contrast, have shown that deleting a specific disulfide bond results in significant changes in structure. For example, NMR analysis of disulfide-deficient analogs of ShK toxin indicated a drastic loss of native structure for each of the three disulfide-deficient analogs, with the analog missing the Cys3–Cys35 disulfide bond adopting a well-defined but altered structure [54] (Figure 11.2). Similarly, a loss of native structure was reported upon elimination of the Cys3–Cys13 disulfide bond of  $\alpha$ -conotoxin GI [55] and the Cys44–Cys59 disulfide bond of enterotoxin B [56].

In small proteins and peptides with multiple disulfide bonds, the most solvent-exposed disulfide can usually be removed without a drastic loss in structure, but disulfides that are buried tend to be more crucial in maintaining the native fold of the



**Figure 11.2** Stereoviews of (a) ShK (PDB ID: 1ROO) and (b) [Abu3,35]ShK<sub>12-28,17-32</sub> (PDB ID: 1C2U). Cys12–Cys28 and Cys17–Cys32 disulfide bonds are shown in gold, and the Cys3–Cys35 disulfide bond of ShK is shown in orange. The two molecules were superimposed over backbone heavy atoms of residues 14–24 and are shown in the same orientation.

structure [54, 57]. However, in most cases the removal of two or more disulfides results in a complete loss of native structure.

Less well studied are the conformational changes associated with the reduction of disulfides in extracellular proteins upon entry into the reducing environment of the cell [58]. It is thought that disulfides in these proteins act as redox switches rather than as stabilizers of native structure [59]. These disulfides are thought to adopt a strained conformation, storing potential energy that is released upon cleavage of this bond following cell entry [58]. The release of energy results in a conformational change in the protein which may be necessary for its function. In other redox-sensitive proteins such as Hsp33 in *Escherichia coli*, cysteine residues are oxidized to form disulfide bonds in response to oxidative stress. In this manner, the functions of such proteins are regulated by the conformational change associated with formation of these bonds [8].

## 11.6 Disulfide Bonds in Protein Dynamics

The conformation of a globular protein is intimately linked to its biological activity and it is expected that changes in conformation will have functional consequences. However, disulfide deletion studies have shown that this does not always hold true. Removal of a disulfide bond may affect the biological activity, but not the overall fold

of a protein. For example, deletion of the 6–20 disulfide bridge in EGF resulted in a loss of activity despite retention of a native-like fold [10]. The conotoxin  $\mu$ -KIIIA provides an intermediate example, with activity being retained on the voltage-gated sodium channels Na<sub>v</sub>1.2 and Na<sub>v</sub>1.4 following removal of the first disulfide bond, and the structure also remaining native-like [51], but the selectivity profile for the other channel subtypes being different and the binding kinetics to the Na<sub>v</sub>1.2 and Na<sub>v</sub>1.4 channels faster. In contrast, there are several examples of disulfide-deficient analogs retaining their activity despite the loss of native structure. Deletion of the Cys3–Cys11 disulfide bond in endothelin-1, for example, resulted in a switch from a “cyclic” to “linear” conformation, but with retention of biological activity [60]. Similarly the Cys3–Cys35 disulfide-deficient analog of ShK retained its potassium channel blocking activity even though its structure was significantly different from that of the native toxin [54]. Such examples reflect the fact that flexibility and conformational stability also play important roles in activity, and that disulfide bonds contribute significantly to these physical attributes of a protein. Conformational stability here refers to dynamic stability of the protein as opposed to thermodynamic stability discussed above.

NMR is a powerful method for determining the stability of peptide or protein structures [14, 61]. In addition, NMR relaxation parameters can be measured to quantify conformational flexibility [62–65]. In barnase, a small globular protein lacking native disulfide bonds, two disulfides were introduced, leading to a local decrease in flexibility of the protein [66]. Similarly, a surface disulfide bond engineered into the cleft region of cytochrome *b*<sub>5</sub> dampened the dynamics of this enzyme [67]. NMR relaxation measurements carried out on oxidized and reduced forms of the *E. coli* enzymes thioredoxin [68] and glutaredoxin-1 [69] indicated that in both cases the disulfide bonds restricted the internal motions in these proteins. In BPTI, removal of the Cys14–Cys38 disulfide bond had only a minor effect on the backbone dynamics in wild-type BPTI, but a significant effect when removed in the Y35G mutant [70]. This observation highlights the cooperation between disulfide bonding and hydrophobic interactions in the dynamic stability of a structured polypeptide such as BPTI.

Intrinsically unstructured proteins (IUPs), a class of proteins that generally lack disulfide bonds, are unstructured and highly dynamic [71]. The 221-residue merozoite surface protein 2 (MSP2) is an IUP with a single disulfide bond near the C-terminus. NMR relaxation studies measuring the backbone dynamics of MSP2 indicated local motional restriction in the region around this disulfide bond, in contrast to the largely unstructured bulk of the protein [72]. Beyond this local restriction, however, the disulfide bond was predicted to contribute no further to the local structure of the C-terminal region.

In  $\mu$ -KIIIA, molecular dynamics simulations were used to characterize conformational flexibility resulting from the removal of individual disulfide bonds [51]. The molecular dynamics simulations indicated that native  $\mu$ -KIIIA explored less conformational space and was less flexible compared with its disulfide-deficient analogs, suggesting that the disulfide bonds impart conformational stability to the toxin. In relation to binding activity, increased conformational flexibility of the Cys1–Cys9 and

Cys2–Cys15 disulfide-deficient analogs may enable them to sample a larger volume of conformational space and facilitate eventual binding to the channel, explaining their retention of binding affinity for sodium channels. Indeed, the increased flexibility of the analog lacking the Cys1–Cys9 disulfide may explain its increased binding kinetics to the Na<sub>v</sub>1.2 and Na<sub>v</sub>1.4 channel subtypes by enhancing its ability to associate with and dissociate from its target channel [51]. A similar outcome was reported in conkunitzin-S1, where the addition of a disulfide bond to the native toxin was thought to make the structure more rigid, decreasing the binding kinetics of the toxin to the Shaker potassium channel target without affecting the overall blocking activity against this channel [73].

## 11.7

### Disulfide Bonding Patterns and Protein Topology

Beyond the mere presence or absence of disulfide bonds, the location of such bonds within a protein plays an important role in protein topology. Several studies have analyzed the relationship between disulfide bonding patterns and protein structure, typically drawing on the conservation and evolution of disulfide bonding patterns within a family of homologous proteins or proteins with similar folds to reveal these relationships [15, 16]. Other studies have focused on comparing proteins that may not necessarily be homologous in amino acid sequence, but have structures with similar disulfide topology [74, 75]. Such large-scale comparative studies of both protein sequences and structures provide a better understanding of the conservation of disulfide bonding patterns and the factors that determine these patterns. A recent study classified structurally about 3000 small disulfide-rich protein domains into different fold groups and explored the disulfide bonding patterns within each group [15]. There was generally a high level of conservation in disulfide bonding patterns within each fold group, but also several naturally occurring variations, giving insight into the roles of disulfide bonds in maintaining these folds.

#### 11.7.1

##### Conservation and Evolution of Disulfide Bonding Patterns

In a recent analysis examining the conservation of disulfide bonds within homologous protein domains, only 54% were found to be conserved, indicating a poor relationship between sequence identity and disulfide bond conservation [16]. Conserved disulfide bonds are presumed to play important structural or functional roles. It is not uncommon to observe conserved disulfide bonding patterns among non-homologous proteins with similar folds. Conversely, proteins with high sequence identity do not always share a conserved disulfide bonding pattern. For example, disintegrins are a group of homologous proteins from snake venoms that display a high level of sequence similarity, and yet adopt different disulfide bonding patterns and topologies. In general, similar bonding patterns result in similar topologies, although different topologies can sometimes be observed. Obtustatin, for example,

adopts a more compact fold compared to the extended conformations of echistatin and other disintegrins, despite having a similar disulfide bonding pattern [76, 77]. On the other hand, the disintegrin salmosin displays a disulfide bonding pattern and protein fold that differ from other disintegrins [78]. The disintegrin family has been divided into five different groups based on the length of the amino acid sequence and the number of disulfide bonds present. It was hypothesized based on the conserved cysteine residues in each subgroup that a selective loss of disulfide bonds occurred during evolution, resulting in structural diversity within the disintegrin family [79].

### 11.7.2

#### Conservation of Disulfide Bonds

Cysteine residues linked in disulfide bonds are often conserved or mutated in pairs [17]. One common variation observed within a fold group is that some members have one more or one less disulfide bond compared to other family members. Konkunitzin-S1, which belongs to the Kunitz domain family of proteins, possesses only two of the three disulfide bonds normally conserved in other members of this family (BPTI, dendrotoxins) [73], yet its structure is essentially identical. The converse is observed in the inhibitory cystine knot (ICK) family [80], where the spider toxins robustoxin,  $\omega$ -agatoxin IVB, and  $\mu$ -agatoxin I all have an additional disulfide bond not seen in most other members of this structural family [81]. The additional disulfide bond in robustoxin [82] links a loop between the first two  $\beta$ -sheets of the ICK fold to the C-terminal end, while in  $\omega$ -agatoxin IVB and  $\mu$ -agatoxin I the additional bond links a  $\beta$ -hairpin between the two  $\beta$ -sheets of the fold (Figure 11.3).

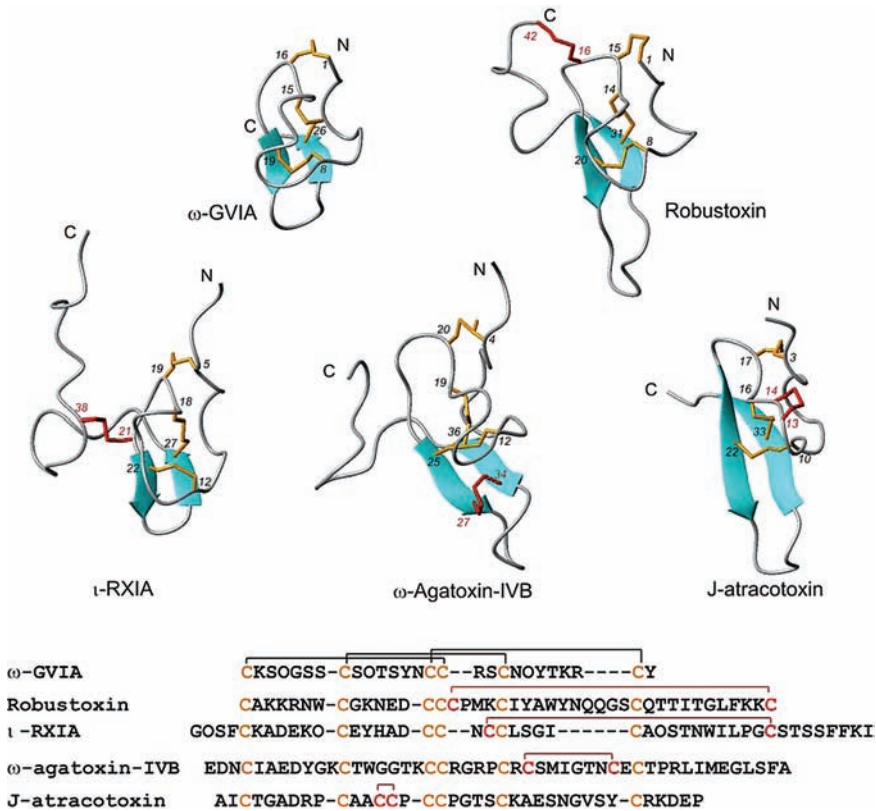
In some families, only one half of a cysteine pair making up a disulfide bond may be conserved. Among the thrombospondin repeat (TSR) domains, two out of the three disulfide bonds are conserved in most of its members [84], but the third disulfide bond varies in that one of the cysteine residues contributing to the disulfide bond is located in a different position in the amino acid sequence. Structurally, however, this “migrated” cysteine residue lies in a similar spatial region. Based on the unique disulfide bonding pattern observed across this family, Tan *et al.* classified the TSRs into two groups and functional differences have been demonstrated between these two groups despite their similar folds.

### 11.7.3

#### Cysteine Framework and Disulfide Connectivity

The cysteine framework (arrangement of cysteine residues in the amino acid sequence) is often conserved within similar fold groups, even though other amino acids may vary considerably. Indeed, the high level of conservation of cysteine frameworks has provided a basis for classifying certain groups of proteins such as the conotoxins [85, 86]. Within conotoxin superfamilies that have similar cysteine frameworks, additional diversity is often generated through the different bonding patterns of the cysteine residues, leading to several different disulfide scaffolds for a particular cysteine arrangement [86]. For instance, the  $\alpha$ -conotoxins and the





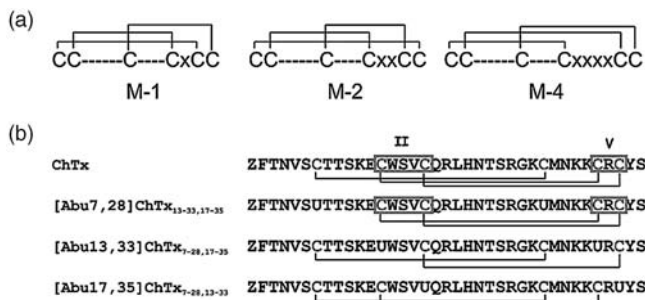
**Figure 11.3** Structures and sequences of peptides with the ICK fold. Structures were aligned based on the disulfide bonding pattern of GVIA as in the sequence alignment shown.

Additional disulfide bonds in the peptides with four disulfide bonds are highlighted in red.  $\beta$ -sheets were determined based on the ribbon macro of MOLMOL [83].

$\chi/\lambda$ -conotoxins have different disulfide connectivities even though they share the same (-CC-C-C-) cysteine framework, resulting in different conformations and biological targets [87]. Such a difference could be due to the presence of proline residues in the sequence, as discussed below.

The  $\iota$ -conotoxin RXIA and the spider toxin J-atracotoxin share the same cysteine framework (-C-C-CC-CC-C-C-), but have different disulfide connectivities [88] (Figure 11.3). On the other hand, different cysteine frameworks can sometimes support the same pattern of disulfide connectivities. Robustoxin [82] and versutoxin [89], despite having a different cysteine framework from  $\iota$ -RXIA, share the same disulfide bonding pattern and are structurally more similar to  $\iota$ -RXIA than is J-atracotoxin, even though the latter shares a similar arrangement of cysteine residues with  $\iota$ -RXIA [88].

The number and type of residues between cysteine residues can be an important factor in determining disulfide connectivity and overall fold. For example, the



**Figure 11.4** (a) Disulfide connectivity patterns of the M-superfamily branch of conotoxins (M-1, M-2 and M-4). X in the third inter-cysteine loop represents any residue. (b) Amino acid sequences of the scorpion toxin ChTx and its three disulfide-deficient analogs in which each of the three cysteine pairs was individually

replaced with Abu (aminobutyric acid).

Z represents pyroglutamate and U represents aminobutyric acid in the sequences above.

Conserved inter-cysteine loop spacings in the second and last inter-cysteine loops of ChTx and the first disulfide-deficient analog are highlighted in the box.

M-superfamily of conotoxins, which has the cysteine framework (-CC-C-C-CC-) [85], has been subdivided into five groups (M-1 to M-5) based on the number of residues between the fourth and fifth cysteines [90, 91]. The structures in the M-1 (C1-C5/C2-C4/C3-C6), M-2 (C1-C6/C2-C4/C3-C5) and M-4/5 (C1-C4/C2-C5/C3-C6) branches have distinctive disulfide connectivity patterns (as indicated in parentheses) (Figure 11.4a), suggesting that the number of residues in the last cysteine loop might determine the disulfide connectivity in this family [92, 93]. Unsurprisingly, tertiary structures differ among these groups. The M-1 conotoxin mr3e possesses a double-turn motif while the M-2 conotoxin mr3a adopts a triple-turn motif and folds into a more globular structure compared to mr3e [92].

Similarly, it has been proposed that inter-cysteine spacings could be responsible for the selective formation of specific disulfide bonds in the scorpion toxin ChTx [50]. Separation of the cysteines in the second and last inter-cysteine loops of ChTx by three and one residues, respectively, (C...CxxxC...C...CxC) was proposed to direct preferential formation of native disulfide pairings by disfavoring certain cysteine pairings (Figure 11.4b). Cysteine mutational studies suggest that the highly conserved inter-cysteine spacing of the second and last inter-cysteine loops is not essential for maintaining the  $\alpha/\beta$  scorpion fold of ChTx, as the native fold was achieved in a disulfide-deficient analog lacking the conserved inter-cysteine spacings. Rather, the inter-cysteine spacings seem to contribute to the folding efficiency of the toxin into its native fold, evident from the higher yield of structures with native disulfide connectivity obtained in the disulfide-deficient analog ([Abu7,28]ChTx<sub>13-33,17-35</sub>) having the conserved spacings CxxxC and CxC [50].

The nature of the residues in each inter-cysteine loop, as well as the number, plays a role in directing disulfide connectivities. In particular, the presence of proline, which is known to favor a turn in conventional folding patterns, seems to affect the final disulfide bonding pattern of some proteins. This was demonstrated in the

$\alpha$ -conotoxin ImI, in which substitution of the conserved proline in the first inter-cysteine loop resulted in a non-native “ribbon” disulfide bonding pattern and conformation, as opposed to the native globular fold [87]. Similarly, mutation of two proline residues in maurotoxin resulted in a rearrangement of disulfide bond pairings accompanied by slight changes in conformation [94]. Zhang and Snyder showed that substituting amino acids other than proline had negligible effects upon the disulfide pattern and conformation [95].

#### 11.7.4

##### **Non-Native Disulfide Connectivities**

It is evident that disulfide connectivity plays an important role in the conformation of a peptide or protein. This has been highlighted by studies examining the conformations of peptides containing non-native disulfide connectivity patterns. Several studies have focused on the  $\alpha$ -conotoxins, which have two disulfide bonds with a C1–C3, C2–C4 connectivity pattern and adopt a globular fold [96]. In the conotoxin  $\alpha$ -GI, however, two additional non-native disulfide bond isomers can be formed, the “ribbon” isomer (C1–C4, C2–C3 disulfide connectivity) and the “beads” isomer (C1–C2, C3–C4), each adopting a different fold from that of the native [96].

The role of disulfide connectivity in protein conformational flexibility is less clear cut. In general, the ribbon and bead isomers are expected to display greater flexibility, as the structures of these isomers are not as tightly folded as the globular isomer [96]. This was demonstrated for the disulfide isomers of  $\alpha$ -conotoxins GI and AuIB, in which solution structures of the ribbon and bead isomers were less well-defined compared with the native globular isomer [97]. Yet the opposite was reported for  $\alpha$ -conotoxin BuIA, wherein the native globular isomer adopted multiple conformations in solution but the ribbon isomer had a well-defined conformation [98]. Intriguingly, the conformationally flexible ribbon isomer of  $\alpha$ -AuIB was more potent than native [97], but the well-defined ribbon isomer of  $\alpha$ -BuIA was inactive [98].

The disulfide connectivity pattern for the somatomedin B domain of human vitronectin has also been studied in great detail. This 35-residue domain is unique in that its four disulfide bonds are closely arranged in the center of the domain, replacing the typical hydrophobic core found in larger proteins [99]. However, the close proximity of these four disulfides within the core of this domain has made it difficult to determine their exact connectivities, leading to different connectivities being proposed by different groups [100–103]. Intriguingly, the different disulfide bonding patterns reported in the literature were compatible with the same fold of the somatomedin B domain. Kamikubo *et al.* explored this further by computing conformational energies of alternate disulfide-bonded forms during structure calculations [101]; the different disulfide bonding patterns proved to have comparable energies and were all compatible with the NMR-derived structure. The alternative disulfide connectivities were subsequently disproved experimentally by studying the functionality and stability of synthetic somatomedin B peptides with alternate disulfide bonding patterns [99]. Only one of the three previously proposed disulfide

connectivity patterns produced an active and stable structure and was taken to be the native disulfide bonding pattern.

## 11.8

### Applications

Studies of the role of disulfide bonds in protein stability and folding have led to a better understanding of the relationships between protein structure and disulfide bonding patterns, and the experience gathered in these studies has been useful for predicting protein structure and folds by computational means [74, 75]. Several approaches have been described to predict disulfide connectivity and protein structures from protein sequences with increased accuracy [104–108].

Disulfide bonds are also important in peptide and protein engineering. The loss of disulfide restraints must be taken into account when attempting to minimize a peptide structure in which these disulfides contribute to its overall stability. In minimized analogs of ShK, for example, lactam bridges were utilized to compensate for the removal of the disulfides [109]. In disulfide bridge-based PEGylation [110], a solvent-accessible disulfide is reduced to release two free cysteine thiols across which the poly(ethylene glycol) (PEG) group can then be attached to the protein using thiol-selective chemistry. In this strategy, designed to make the protein therapeutically more bioavailable, it is important that the protein structure be maintained following reduction of the disulfide bond. In protein engineering, disulfide bonds may be added to commercially important proteins or therapeutics to increase the thermodynamic and proteolytic stability of the protein [111]. For example, Matsumura *et al.* demonstrated that introducing disulfides at strategic locations of T4 phage lysozyme could significantly stabilize the protein against thermal unfolding [112, 113]. Disulfide bonds can be added as cyclic constraints to conformationally constrain linear peptides that would otherwise be flexible in solution. Such a modification, by limiting conformational freedom, represents a valuable strategy for enhancing receptor selectivity and increasing binding affinity. This approach has been used to optimize linear peptides targeting the opioid receptors [114]. Introduction of a disulfide restraint to an opioid receptor peptide increased its selectivity for the  $\mu$ -opioid receptor, presumably by locking the peptide into a conformation active for the  $\mu$  receptor but not the  $\delta$  and  $\kappa$  subtypes. A similar observation was reported for  $\beta$ -melanocyte-stimulating hormone-derived melanocortin-4 receptor (MC4R) peptide agonists, in which disulfide cyclization improved selectivity and potency for the MC4R over the MC1, 3 and 5 receptor subtypes [115]. Some disulfide-stabilized frameworks such as the highly stable ICK motif, which can accommodate variable intercysteine amino acid sequences while maintaining its overall fold, have potential applications as protein scaffolds for drug design or engineering novel polypeptides [81, 116].

Although disulfides can be incorporated to increase protein stability, the disulfides themselves are susceptible to reduction in certain extracellular environments such as the blood [117]. To overcome this problem, disulfides can be replaced with disulfide

mimetics such as dicarba or diselenium bridges, which are less susceptible to degradation. One potential drawback of such strategies is the difficulty of reproducing the covalent geometry of the disulfide bond. Compared with the sulfur–sulfur bond in disulfide bridges, selenium–selenium bond lengths in diselenium bridges are slightly longer (around 2.02 Å) [117], while the carbon–carbon bond lengths in dicarba bridges are shorter (around 1.34 Å) [118]. In separate studies, diselenium and dicarba bonds were incorporated in the  $\alpha$ -conotoxin ImI [117, 118]. For both substitutions, the conformation of the analog was close to that of the native structures, with slight local differences in the vicinity of the substituted residues arising from the different covalent geometries. Nevertheless, biological activity against the nicotinic acetylcholine receptor was retained for both modified peptides, supporting their future use in the design of stable scaffolds for drugs. Furthermore, disulfide substitution with diselenium bridges could also be an effective strategy for enhancing folding yield and kinetics, as recently demonstrated in  $\omega$ -GVIA, in which diselenium-substituted analogs of  $\omega$ -GVIA displayed a higher propensity to adopt the native fold and folded with a half-time approximately 5-fold shorter than native  $\omega$ -GVIA [119].

In addition to stabilizing structural folds, disulfides may also be used as an effective means of covalently linking two or more subunits together. This approach was employed to attach cell-penetrating peptides (CPPs) to cargo proteins, facilitating the entry of these proteins into the cell for in-cell NMR studies [120]. The reduction-sensitive disulfide link was subsequently cleaved in the reducing environment of the cytoplasm, allowing the cargo protein to detach from the CPP and distribute uniformly around the cell [120].

Yet another useful application of disulfide bonds is the technique of “disulfide trapping” [121, 122]. This method takes advantage of formation of a disulfide bond between a pair of cysteine residues substituted in strategic locations in a protein with known structure in order to detect long-range backbone motions when the two cysteine residues collide with one another to form the disulfide bond. Patterns and rates of disulfide formation can be analyzed to obtain information about the trajectories and frequencies of backbone and domain motions [122].

## 11.9

### Conclusions

Disulfide bonds play important roles in the conformation and stability of proteins and peptides. Studies of their roles through both experimental means and analysis of evolutionary patterns have led to a better understanding of how disulfide bonds contribute to various biological processes in nature such as folding and dynamic stability, and their importance in maintaining the native conformation of a protein. It is evident that the roles and importance of individual disulfide bonds vary across different protein folds, with some being more important for the proper function of a protein. The location of a disulfide within a structure may also influence the role it plays in stabilizing or folding the protein.

Gaps remain in our understanding of how disulfides contribute to protein folding and stability, leading to different theories being proposed. In particular, it is difficult to differentiate completely the relative contributions of disulfide bonds and other noncovalent interactions. It should also be noted that most studies are conducted *in vitro* rather than *in vivo* and may not accurately represent how disulfides stabilize a protein or assist its folding in the cell. Nonetheless, an understanding of how disulfide bonds contribute to the folding, stability, and function of a protein *in vitro* has important implications for the field of protein engineering and therapeutic design, particularly in achieving higher yields of product or designing a stable scaffold for therapeutic use. Several effective strategies have already been developed to either compensate for the loss of structure-stabilizing disulfides in protein minimization or incorporate disulfides or disulfide mimetics to stabilize protein scaffolds. With future advances in techniques and more in-depth studies, the roles of disulfide bonds will be understood even more clearly.

### Acknowledgements

We thank Charles Galea, Zhihe Kuang, Jeff Babon, and Chris MacRaid for valuable comments on this chapter. R.S.N. acknowledges fellowship support from the Australian National Health and Medical Research Council.

### References

- Richardson, J.S. (1981) The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, **34**, 167–339.
- Mallick, P., Boutz, D.R., Eisenberg, D., and Yeates, T.O. (2002) Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9679–9684.
- Chang, S.-G., Choi, K.-D., Jang, S.-H., and Shin, H.-C. (2003) Role of disulfide bonds in the structure and activity of human insulin. *Molecules and Cells*, **16**, 323–330.
- Parbhu, A.N., Bryson, W.G., and Lal, R. (1999) Disulfide bonds in the outer layer of keratin fibers confer higher mechanical rigidity: correlative nano-indentation and elasticity measurement with an AFM. *Biochemistry*, **38**, 11755–11761.
- Berndt, C., Lillig, C.H., and Holmgren, A. (2008) Thioredoxins and glutaredoxins as facilitators of protein folding. *Biochimica et Biophysica Acta*, **1783**, 641–650.
- Choi, H.-J., Kim, S.-J., Mukhopadhyay, P., Cho, S., Woo, J.-R., Storz, G., and Ryu, S.-E. (2001) Structural basis of the redox switch in the OxyR transcription factor. *Cell*, **105**, 103–113.
- Holmgren, A. (1995) Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. *Structure*, **3**, 239–243.
- Barford, D. (2004) The role of cysteine residues as redox-sensitive regulatory switches. *Current Opinion in Structural Biology*, **14**, 679–686.
- Hogg, P.J. (2003) Disulfide bonds as switches for protein function. *Trends in Biochemical Sciences*, **28**, 210–214.

- 10 Barnham, K.J., Torres, A.M., Alewood, D., Alewood, P.F., Domagala, T., Nice, E.C., and Norton, R.S. (1998) Role of the 6–20 disulfide bridge in the structure and activity of epidermal growth factor. *Protein Science*, **7**, 1738–1749.
- 11 Flinn, J.P., Pallaghy, P.K., Lew, M.J., Murphy, R., Angus, J.A., and Norton, R.S. (1999) Role of disulfide bridges in the folding, structure and biological activity of  $\omega$ -conotoxin GVIA. *Biochimica et Biophysica Acta*, **1434**, 177–190.
- 12 Mamathambika, B.S. and Bardwell, J.C. (2008) Disulfide-linked protein folding pathways. *Annual Review of Cell and Developmental Biology*, **24**, 211–235.
- 13 Wedemeyer, W.J., Welker, E., Narayan, M., and Scheraga, H.A. (2000) Disulfide bonds and protein folding. *Biochemistry*, **39**, 7032.
- 14 Kay, L.E. (2005) NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance*, **173**, 193–207.
- 15 Cheek, S., Krishna, S.S., and Grishin, N.V. (2006) Structural classification of small, disulfide-rich protein domains. *Journal of Molecular Biology*, **359**, 215–237.
- 16 Thangudu, R.R., Manoharan, M., Srinivasan, N., Cadet, F., Sowdhamini, R., and Offmann, B. (2008) Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families. *BMC Structural Biology*, **8**, 55.
- 17 Thornton, J.M. (1981) Disulphide bridges in globular proteins. *Journal of Molecular Biology*, **151**, 261–287.
- 18 Doig, A.J. and Williams, D.H. (1991) Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability. *Journal of Molecular Biology*, **217**, 389–398.
- 19 Betz, S.F. (1993) Disulfide bonds and the stability of globular proteins. *Protein Science*, **2**, 1551–1558.
- 20 Tidor, B. and Karplus, M. (1993) The contribution of cross-links to protein stability: a normal mode analysis of the configurational entropy of the native state. *Proteins*, **15**, 71–79.
- 21 Katz, B.A. and Kossiakoff, A. (1986) The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. *Journal of Biological Chemistry*, **261**, 15480–15485.
- 22 Katz, B. and Kossiakoff, A.A. (1990) Crystal structure of subtilisin BPN' variants containing disulfide bonds and cavities: concerted structural rearrangements induced by mutagenesis. *Proteins*, **7**, 343–357.
- 23 Price-Carter, M., Hull, M.S., and Goldenberg, D.P. (1998) Roles of individual disulfide bonds in the stability and folding of an  $\omega$ -conotoxin. *Biochemistry*, **37**, 9851–9861.
- 24 Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., and Zardecki, C. (2002) The Protein Data Bank. *Acta Crystallographica D*, **58**, 899–907.
- 25 Bastolla, U. and Demetrius, L. (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Engineering, Design and Selection*, **18**, 405–415.
- 26 Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
- 27 Anfinsen, C.B., Haber, E., Sela, M., and White, F.H.Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, **47**, 1309–1314.
- 28 Creighton, T.E. (1990) Protein folding. *Biochemical Journal*, **270**, 1–16.
- 29 Creighton, T.E. (1988) Toward a better understanding of protein folding pathways. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 5082–5086.
- 30 Creighton, T.E. (1997) Protein folding coupled to disulfide bond formation. *Biological Chemistry*, **378**, 731–744.
- 31 Staley, J.P. and Kim, P.S. (1994) Formation of a native-like subdomain in a

- partially folded intermediate of bovine pancreatic trypsin inhibitor. *Protein Science*, **3**, 1822–1832.
- 32 van Mierlo, C.P.M., Kemmink, J., Neuhaus, D., Darby, N.J., and Creighton, T.E. (1994) <sup>1</sup>H-NMR analysis of the partly-folded non-native two-disulphide intermediates (30–51,5–14) and (30–51,5–38) in the folding pathway of bovine pancreatic trypsin inhibitor. *Journal of Molecular Biology*, **235**, 1044–1061.
  - 33 Creighton, T.E., Darby, N.J., and Kemmink, J. (1996) The roles of partly folded intermediates in protein folding. *FASEB Journal*, **10**, 110–118.
  - 34 Price-Carter, M., Bulaj, G., and Goldenberg, D.P. (2002) Initial disulfide formation steps in the folding of an  $\omega$ -conotoxin. *Biochemistry*, **41**, 3507–3519.
  - 35 Creighton, T.E. (1977) Kinetics of refolding of reduced ribonuclease. *Journal of Molecular Biology*, **113**, 329–341.
  - 36 Creighton, T.E. (1979) Intermediates in the refolding of reduced ribonuclease A. *Journal of Molecular Biology*, **129**, 411–431.
  - 37 Creighton, T.E. (1986) Disulfide bonds as probes of protein folding pathways. *Methods in Enzymology*, **131**, 83–106.
  - 38 Frech, C. and Schmid, F.X. (1995) Influence of protein conformation on disulfide bond formation in the oxidative folding of ribonuclease T<sub>1</sub>. *Journal of Molecular Biology*, **251**, 135–149.
  - 39 Welker, E., Wedemeyer, W.J., Narayan, M., and Scheraga, H.A. (2001) Coupling of conformational folding and disulfide-bond reactions in oxidative folding of proteins. *Biochemistry*, **40**, 9059–9064.
  - 40 Wedemeyer, W.J., Welker, E., Narayan, M., and Scheraga, H.A. (2000) Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.
  - 41 Bulaj, G. and Walewska, A. (2009) Oxidative folding of single-stranded disulfide-rich peptides, in *Oxidative Folding of Peptides and Proteins* (eds J. Buchner and L. Moroder), Royal Society of Chemistry, London, pp. 274–296.
  - 42 Winter, J., Klappa, P., Freedman, R.B., Lilie, H., and Rudolph, R. (2002) Catalytic activity and chaperone function of human protein-disulfide isomerase are required for the efficient refolding of proinsulin. *Journal of Biological Chemistry*, **277**, 310–317.
  - 43 Gruber, C.W., Cemazar, M., Clark, R.J., Horibe, T., Renda, R.F., Anderson, M.A., and Craik, D.J. (2007) A novel plant protein-disulfide isomerase involved in the oxidative folding of cystine knot defense proteins. *Journal of Biological Chemistry*, **282**, 20435–20446.
  - 44 Bulaj, G., Buczek, O., Goodsell, I., Jimenez, E.C., Kranski, J., Nielsen, J.S., Garrett, J.E., and Olivera, B.M. (2003) Efficient oxidative folding of conotoxins and the radiation of venomous cone snails. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (Suppl. 2), 14562–14568.
  - 45 Bronsoms, S., Villanueva, J., Canals, F., Querol, E., and Aviles, F.X. (2003) Analysis of the effect of potato carboxypeptidase inhibitor pro-sequence on the folding of the mature protein. *European Journal of Biochemistry*, **270**, 3641–3650.
  - 46 Buczek, O., Olivera, B.M., and Bulaj, G. (2004) Propeptide does not act as an intramolecular chaperone but facilitates protein disulfide isomerase-assisted folding of a conotoxin precursor. *Biochemistry*, **43**, 1093–1101.
  - 47 Lauber, T., Schulz, A., Rösch, P., and Marx, U.C. (2004) Role of disulfide bonds for the structure and folding of proguanylin. *Biochemistry*, **43**, 10050–10057.
  - 48 Kadokura, H. and Beckwith, J. (2009) Detecting folding intermediates of a protein as it passes through the bacterial translocation channel. *Cell*, **138**, 1164–1173.
  - 49 Zakharova, E., Horvath, M.P., and Goldenberg, D.P. (2008) Functional and structural roles of the Cys14–Cys38 disulfide of bovine pancreatic trypsin inhibitor. *Journal of Molecular Biology*, **382**, 998–1013.
  - 50 Drakopoulou, E., Vizzavona, J., Neyton, J., Anriot, V., Bouet, F., Virelizier, H.,



- Menez, A., and Vita, C. (1998) Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation. *Biochemistry*, **37**, 1292–1301.
- 51 Khoo, K.K., Feng, Z.-P., Smith, B.J., Zhang, M.-M., Yoshikami, D., Olivera, B.M., Bulaj, G., and Norton, R.S. (2009) Structure of the analgesic  $\mu$ -conotoxin KIIIA and effects on the structure and function of disulfide deletion. *Biochemistry*, **48**, 1210–1219.
- 52 Carrega, L., Mosbah, A., Ferrat, G., Beeton, C., Andreotti, N., Mansuelle, P., Darbon, H., De Waard, M., and Sabatier, J.-M. (2005) The impact of the fourth disulfide bridge in scorpion toxins of the  $\alpha$ -KTx6 subfamily. *Proteins*, **61**, 1010–1023.
- 53 Hua, Q.-X., Narhi, L., Jia, W., Arakawa, T., Rosenfeld, R., Hawkins, N., Miller, J.A., and Weiss, M.A. (1996) Native and non-native structure in a protein-folding intermediate: spectroscopic studies of partially reduced IGF-I and an engineered alanine model. *Journal of Molecular Biology*, **259**, 297–313.
- 54 Pennington, M.W., Lanigan, M.D., Kalman, K., Mahnir, V.M., Rauer, H., McVaugh, C.T., Behm, D., Donaldson, D., Chandy, K.G., Kem, W.R., and Norton, R.S. (1999) Role of disulfide bonds in the structure and potassium channel blocking activity of ShK toxin. *Biochemistry*, **38**, 14549–14558.
- 55 Mok, K.H. and Han, K.-H. (1999) NMR solution conformation of an antitoxic analog of  $\alpha$ -conotoxin GI: identification of a common nicotinic acetylcholine receptor  $\alpha 1$ -subunit binding surface for small ligands and  $\alpha$ -conotoxins. *Biochemistry*, **38**, 11895–11904.
- 56 Arriaga, Y.L., Harville, B.A., and Dreyfus, L.A. (1995) Contribution of individual disulfide bonds to biological action of *Escherichia coli* heat-stable enterotoxin B. *Infection and Immunity*, **63**, 4715–4720.
- 57 Takeda, K., Ogawa, K., Ohara, M., Hamada, S., and Moriyama, Y. (1995) Conformational changes of  $\alpha$ -lactalbumin induced by the stepwise reduction of its disulfide bridges: the effect of the disulfide bridges on the structural stability of the protein in sodium dodecyl sulfate solution. *Journal of Protein Chemistry*, **14**, 679–684.
- 58 Wouters, M.A., Lau, K.K., and Hogg, P.J. (2004) Cross-strand disulphides in cell entry proteins: poised to act. *Bioessays*, **26**, 73–79.
- 59 Wouters, M.A., George, R.A., and Haworth, N.L. (2007) “Forbidden” disulfides: their role as redox switches. *Current Protein and Peptide Science*, **8**, 484–495.
- 60 Hewage, C.M., Jiang, L., Parkinson, J.A., Ramage, R., and Sadler, I.H. (1999) Solution structure of a novel ET<sub>B</sub> receptor selective agonist ET1–21 [Cys(Acm)<sup>1,15</sup>, Aib<sup>3,11</sup>, Leu<sup>7</sup>] by nuclear magnetic resonance spectroscopy and molecular modelling. *Journal of Peptide Research*, **53**, 223–233.
- 61 Kempf, J.G. and Loria, J.P. (2002) Protein dynamics from solution NMR: theory and applications. *Cell Biochemistry and Biophysics*, **37**, 187–211.
- 62 Jarymowycz, V.A. and Stone, M.J. (2006) Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chemical Reviews*, **106**, 1624–1671.
- 63 Tolman, J.R. and Ruan, K. (2006) NMR residual dipolar couplings as probes of biomolecular dynamics. *Chemical Reviews*, **106**, 1720–1736.
- 64 Yao, S., Zhang, M.-M., Yoshikami, D., Azam, L., Olivera, B.M., Bulaj, G., and Norton, R.S. (2008) Structure, dynamics, and selectivity of the sodium channel blocker  $\mu$ -conotoxin SIIIA. *Biochemistry*, **47**, 10940–10949.
- 65 Yao, S., Smith, D.K., Hinds, M.G., Zhang, J.G., Nicola, N.A., and Norton, R.S. (2000) Backbone dynamics measurements on leukemia inhibitory factor, a rigid four-helical bundle cytokine. *Protein Science*, **9**, 671–682.
- 66 Clarke, J., Hounslow, A.M., Bond, C.J., Fersht, A.R., and Daggett, V. (2000) The effects of disulfide bonds on the

- denatured state of barnase. *Protein Science*, **9**, 2394–2404.
- 67 Storch, E.M., Grinstead, J.S., Campbell, A.P., Daggett, V., and Atkins, W.M. (1999) Engineering out motion: a surface disulfide bond alters the mobility of tryptophan 22 in cytochrome *b<sub>5</sub>* as probed by time-resolved fluorescence and <sup>1</sup>H-NMR experiments. *Biochemistry*, **38**, 5065–5075.
- 68 Stone, M.J., Chandrasekhar, K., Holmgren, A., Wright, P.E., and Dyson, H.J. (1993) Comparison of backbone and tryptophan side-chain dynamics of reduced and oxidized *Escherichia coli* thioredoxin using nitrogen-15 NMR relaxation measurements. *Biochemistry*, **32**, 426–435.
- 69 Kelley, J.J. 3rd, Caputo, T.M., Eaton, S.F., Laue, T.M., and Bushweller, J.H. (1997) Comparison of backbone dynamics of reduced and oxidized *Escherichia coli* glutaredoxin-1 using <sup>15</sup>N NMR relaxation measurements. *Biochemistry*, **36**, 5029–5044.
- 70 Beeser, S.A., Oas, T.G., and Goldenberg, D.P. (1998) Determinants of backbone dynamics in native BPTI: cooperative influence of the 14–38 disulfide and the Tyr35 side-chain. *Journal of Molecular Biology*, **284**, 1581–1596.
- 71 Tompa, P. (2002) Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, **27**, 527–533.
- 72 Zhang, X., Perugini, M.A., Yao, S., Adda, C.G., Murphy, V.J., Low, A., Anders, R.F., and Norton, R.S. (2008) Solution conformation, backbone dynamics and lipid interactions of the intrinsically unstructured malaria surface protein MSP2. *Journal of Molecular Biology*, **379**, 105–121.
- 73 Bayrhuber, M., Vijayan, V., Ferber, M., Graf, R., Korukottu, J., Imperial, J., Garrett, J.E., Olivera, B.M., Terlau, H., Zweckstetter, M., and Becker, S. (2005) Conkunitzin-S1 is the first member of a new Kunitz-type neurotoxin family. Structural and functional characterization. *Journal of Biological Chemistry*, **280**, 23766–23770.
- 74 Chuang, C.-C., Chen, C.-Y., Yang, J.-M., Lyu, P.-C., and Hwang, J.-K. (2003) Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **53**, 1–5.
- 75 Mas, J.M., Aloy, P., Marti-Renom, M.A., Oliva, B., Blanco-Aparicio, C., Molina, M.A., de Llorens, R., Querol, E., and Aviles, F.X. (1998) Protein similarities beyond disulphide bridge topology. *Journal of Molecular Biology*, **284**, 541–548.
- 76 Atkinson, R.A., Saudek, V., and Pelton, J.T. (1994) Echistatin: the refined structure of a disintegrin in solution by <sup>1</sup>H NMR and restrained molecular dynamics. *International Journal of Peptide and Protein Research*, **43**, 563–572.
- 77 Paz Moreno-Murciano, M., Monleon, D., Marcinkiewicz, C., Calvete, J.J., and Celda, B. (2003) NMR solution structure of the non-RGD disintegrin obtustatin. *Journal of Molecular Biology*, **329**, 135–145.
- 78 Shin, J., Hong, S.Y., Chung, K., Kang, I., Jang, Y., Kim, D.-s., and Lee, W. (2003) Solution structure of a novel disintegrin, salmosin, from *Agkistrondon halys* venom. *Biochemistry*, **42**, 14408–14415.
- 79 Calvete, J.J. (2005) Structure–function correlations of snake venom disintegrins. *Current Pharmaceutical Design*, **11**, 829–835.
- 80 Pallaghy, P.K., Nielsen, K.J., Craik, D.J., and Norton, R.S. (1994) A common structural motif incorporating a cystine knot and a triple-stranded beta-sheet in toxic and inhibitory polypeptides. *Protein Science*, **3**, 1833–1839.
- 81 Norton, R.S. and Pallaghy, P.K. (1998) The cystine knot structure of ion channel toxins and related polypeptides. *Toxicol*, **36**, 1573–1583.
- 82 Pallaghy, P.K., Alewood, D., Alewood, P.F., and Norton, R.S. (1997) Solution structure of robustoxin, the lethal neurotoxin from the funnel-web spider *Atrax robustus*. *FEBS Letters*, **419**, 191–196.
- 83 Koradi, R., Billeter, M., and Wüthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular

- structures. *Journal of Molecular Graphics*, **14**, 51–55.
- 84 Tan, K., Duquette, M., Liu, J.-h., Dong, Y., Zhang, R., Joachimiak, A., Lawler, J., and Wang, J.h. (2002) Crystal structure of the TSP-1 type 1 repeats: a novel layered fold and its biological implication. *Journal of Cell Biology*, **159**, 373–382.
- 85 Norton, R.S. and Olivera, B.M. (2006) Conotoxins down under. *Toxicon*, **48**, 780–798.
- 86 Bulaj, G. and Olivera, B.M. (2008) Folding of conotoxins: formation of the native disulfide bridges during chemical synthesis and biosynthesis of *Conus* peptides. *Antioxidants and Redox Signaling*, **10**, 141–156.
- 87 Kang, T.S., Radic, Z., Talley, T.T., Jois, S.D.S., Taylor, P., and Kini, R.M. (2007) Protein folding determinants: structural features determining alternative disulfide pairing in  $\alpha$ - and  $\gamma/\lambda$ -conotoxins. *Biochemistry*, **46**, 3338–3355.
- 88 Buczek, O., Wei, D., Babon, J.J., Yang, X., Fiedler, B., Chen, P., Yoshikami, D., Olivera, B.M., Bulaj, G., and Norton, R.S. (2007) Structure and sodium channel activity of an excitatory I<sub>1</sub>-superfamily conotoxin. *Biochemistry*, **46**, 9929–9940.
- 89 Fletcher, J.I., Chapman, B.E., Mackay, J.P., Howden, M.E.H., and King, G.F. (1997) The structure of versutoxin ( $\delta$ -atracotoxin-Hv1) provides insights into the binding of site 3 neurotoxins to the voltage-gated sodium channel. *Structure*, **5**, 1525–1535.
- 90 Corpuz, G.P., Jacobsen, R.B., Jimenez, E.C., Watkins, M., Walker, C., Colledge, C., Garrett, J.E., McDougal, O., Li, W., Gray, W.R., Hillyard, D.R., Rivier, J., McIntosh, J.M., Cruz, L.J., and Olivera, B.M. (2005) Definition of the M-conotoxin superfamily: characterization of novel peptides from molluscivorous *Conus* venoms. *Biochemistry*, **44**, 8176–8186.
- 91 Jacob, R.B. and McDougal, O.M. (2010) The M-superfamily of conotoxins: a review. *Cellular and Molecular Life Sciences*, **67**, 17–27.
- 92 Du, W.-H., Han, Y.-H., Huang, F.-j., Li, J., Chi, C.-W., and Fang, W.-H. (2007) Solution structure of an M-1 conotoxin with a novel disulfide linkage. *FEBS Journal*, **274**, 2596–2602.
- 93 Han, Y.-H., Wang, Q., Jiang, H., Liu, L., Xiao, C., Yuan, D.-D., Shao, X.-X., Dai, Q.-Y., Cheng, J.-S., and Chi, C.-W. (2006) Characterization of novel M-superfamily conotoxins with new disulfide linkage. *FEBS Journal*, **273**, 4972–4982.
- 94 Carlier, E., Fajloun, Z., Mansuelle, P., Fathallah, M., Mosbah, A., Oughideni, R., Sandoz, G., Di Luccio, E., Geib, S., Regaya, I., Brocard, J., Rochat, H., Darbon, H., Devaux, C., Sabatier, J.M., and de Waard, M. (2001) Disulfide bridge reorganization induced by proline mutations in maurotoxin. *FEBS Letters*, **489**, 202–207.
- 95 Zhang, R.M. and Snyder, G.H. (1991) Factors governing selective formation of specific disulfides in synthetic variants of  $\alpha$ -conotoxin. *Biochemistry*, **30**, 11343–11348.
- 96 Gehrmann, J., Alewood, P.F., and Craik, D.J. (1998) Structure determination of the three disulfide bond isomers of  $\alpha$ -conotoxin GI: a model for the role of disulfide bonds in structural stability. *Journal of Molecular Biology*, **278**, 401–415.
- 97 Dutton, J.L., Bansal, P.S., Hogg, R.C., Adams, D.J., Alewood, P.F., and Craik, D.J. (2002) A new level of conotoxin diversity, a non-native disulfide bond connectivity in  $\alpha$ -conotoxin AuIB reduces structural definition but increases biological activity. *Journal of Biological Chemistry*, **277**, 48849–48857.
- 98 Jin, A.-H., Brandstatter, H., Nevin, S.T., Tan, C.C., Clark, R.J., Adams, D.J., Alewood, P.F., Craik, D.J., and Daly, N.L. (2007) Structure of  $\alpha$ -conotoxin BuIA: influences of disulfide connectivity on structural dynamics. *BMC Structural Biology*, **7**, 28.
- 99 Kjaergaard, M., Gårdsvoll Jørgensen, H., Hirschberg, D., Nielbo, S., Mayasundari, A., Peterson, C.B., Jansson, A., Jørgensen, T.J.D., Poulsen, F.M., and Ploug, M. (2007) Solution structure of recombinant somatomedin B domain from vitronectin produced in *Pichia pastoris*. *Protein Science*, **16**, 1934–1945.
- 100 Horn, N.A., Hurst, G.B., Mayasundari, A., Whittemore, N.A., Serpersu, E.H.,

- and Peterson, C.B. (2004) Assignment of the four disulfides in the N-terminal somatomedin B domain of native vitronectin isolated from human plasma. *Journal of Biological Chemistry*, **279**, 35867–35878.
- 101 Kamikubo, Y., De Guzman, R., Kroon, G., Curriden, S., Neels, J.G., Churchill, M.J., Dawson, P., Oldziej, S., Jagielska, A., Scheraga, H.A., Loskutoff, D.J., and Dyson, H.J. (2004) Disulfide bonding arrangements in active forms of the somatomedin B domain of human vitronectin. *Biochemistry*, **43**, 6519–6534.
- 102 Mayasundari, A., Whittemore, N.A., Serpersu, E.H., and Peterson, C.B. (2004) The solution structure of the N-terminal domain of human vitronectin: proximal sites that regulate fibrinolysis and cell migration. *Journal of Biological Chemistry*, **279**, 29359–29366.
- 103 Zhou, A., Huntington, J.A., Pannu, N.S., Carrell, R.W., and Read, R.J. (2003) How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration. *Nature Structural Biology*, **10**, 541–544.
- 104 Vincent, M., Passerini, A., Labbe, M., and Frasconi, P. (2008) A simplified approach to disulfide connectivity prediction from protein sequences. *BMC Bioinformatics*, **9**, 20.
- 105 Tsai, C.-H., Chan, C.-H., Chen, B.-J., Kao, C.-Y., Liu, H.-L., and Hsu, J.-P. (2007) Bioinformatics approaches for disulfide connectivity prediction. *Current Protein and Peptide Science*, **8**, 243–260.
- 106 Song, J., Yuan, Z., Tan, H., Huber, T., and Burrage, K. (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, **23**, 3147–3154.
- 107 Chen, B.-J., Tsai, C.-H., Chan, C.-h., and Kao, C.-Y. (2006) Disulfide connectivity prediction with 70% accuracy using two-level models. *Proteins*, **64**, 246–252.
- 108 Zhao, E., Liu, H.-L., Tsai, C.-H., Tsai, H.-K., Chan, C.-h., and Kao, C.-Y. (2005) Cysteine separations profiles on protein sequences infer disulfide connectivity. *Bioinformatics*, **21**, 1415–1420.
- 109 Lanigan, M.D., Pennington, M.W., Lefievre, Y., Rauer, H., and Norton, R.S. (2001) Designed peptide analogs of the potassium channel blocker ShK toxin. *Biochemistry*, **40**, 15528–15537.
- 110 Brocchini, S., Godwin, A., Balan, S., Choi, J.-w., Zloh, M., and Shaunak, S. (2008) Disulfide bridge based PEGylation of proteins. *Advanced Drug Delivery Reviews*, **60**, 3–12.
- 111 Bulaj, G. (2005) Formation of disulfide bonds in proteins and peptides. *Biotechnology Advances*, **23**, 87–92.
- 112 Matsumura, M., Becktel, W.J., Levitt, M., and Matthews, B.W. (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 6562–6566.
- 113 Wetzel, R., Perry, L.J., Baase, W.A., and Becktel, W.J. (1988) Disulfide bonds and thermal stability in T4 lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 401–405.
- 114 Janecka, A. and Kruszynski, R. (2005) Conformationally restricted peptides as tools in opioid receptor studies. *Current Medicinal Chemistry*, **12**, 471–481.
- 115 Yan, L.Z., Hsiung, H.M., Heiman, M.L., Gadski, R.A., Emmerson, P.J., Hertel, J., Flora, D., Edwards, P., Smiley, D., Zhang, L., Husain, S., Kahl, S.D., DiMarchi, R.D., and Mayer, J.P. (2007) Structure–activity relationships of  $\beta$ -MSH derived melanocortin-4 receptor peptide agonists. *Current Topics in Medicinal Chemistry*, **7**, 1052–1067.
- 116 Craik, D.J., Daly, N.L., and Waine, C. (2001) The cystine knot motif in toxins and implications for drug design. *Toxicon*, **39**, 43–60.
- 117 Armishaw, C.J., Daly, N.L., Nevin, S.T., Adams, D.J., Craik, D.J., and Alewood, P.F. (2006)  $\alpha$ -selenoconotoxins, a new class of potent  $\alpha_7$  neuronal nicotinic receptor antagonists. *Journal of Biological Chemistry*, **281**, 14136–14143.
- 118 MacRaid, C.A., Illesinghe, J., van Lierop, B.J., Townsend, A.L., Chebib, M., Livett, B.G., Robinson, A.J., and Norton, R.S.

- (2009) Structure and activity of (2,8)-dicarba-(3,12)-cystino  $\alpha$ -ImI, an  $\alpha$ -conotoxin containing a nonreducible cystine analog. *Journal of Medicinal Chemistry*, **52**, 755–762.
- 119 Gowd, K.H., Yarotsky, V., Elmslie, K.S., Skalicky, J.J., Olivera, B.M., and Bulaj, G. (2010) Site-specific effects of diselenide bridges on the oxidative folding of  $\omega$ -selenoconotoxin GVIA. *Biochemistry*, **49**, 2741–2752.
- 120 Inomata, K., Ohno, A., Tochio, H., Isogai, S., Tenno, T., Nakase, I., Takeuchi, T., Futaki, S., Ito, Y., Hiroaki, H., and Shirakawa, M. (2009) High-resolution multi-dimensional NMR spectroscopy of proteins in human cells. *Nature*, **458**, 106–109.
- 121 Gloor, S.L. and Falke, J.J. (2009) Thermal domain motions of CheA kinase in solution: disulfide trapping reveals the motional constraints leading to trans-autophosphorylation. *Biochemistry*, **48**, 3631–3644.
- 122 Bass, R.B., Butler, S.L., Chervitz, S.A., Gloor, S.L., and Falke, J.J. (2007) Use of site-directed cysteine and disulfide chemistry to probe protein structure and dynamics: applications to soluble and transmembrane receptors of bacterial chemotaxis. *Methods in Enzymology*, **423**, 25–51.



## 12

# Quantitative Mass Spectrometry-Based Proteomics

*Shao-En Ong*

### 12.1

#### Introduction

Genomics is the study of the genetic material of an organism, with special interest in understanding how genes are regulated to effect biological outcomes. This includes analyses of the primary DNA sequence or the expression of mRNA that is subsequently translated to proteins – the primary effector molecules of the cell. Proteomics, the study of the protein complement in a given cell or tissue, has lagged behind genomics for several reasons.

- i) While the sensitivity of proteomic analyses is good and constantly improving, there is no equivalent of the polymerase chain reaction (PCR) in proteomics to amplify the amount of protein in a sample.
- ii) Proteins may exist in distinct splice isoforms and are often decorated with a variety of post-translational modifications (PTMs) that complicate proteomic analyses, particularly when even subtle changes in PTMs occurring in a subpopulation of the protein present may be biologically significant.
- iii) Proteins are structurally and functionally diverse (e.g., they may take on distinct roles in different cellular compartments, exhibit a large dynamic range of expression, and often operate as members of larger multiprotein complexes). Proteins are therefore more heterogeneous in sample preparation and functionally important subproteomes often need to be enriched out of a whole cell for analysis.

Despite these challenges, it is clear that our need to study and understand protein function is critically important in understanding and interpreting genomic data. Mass spectrometry (MS)-based proteomics is now the method of choice for the unbiased analysis of proteins in biological samples and has effectively replaced Edman degradation for protein identification. Accurate mass measures of intact peptides and their fragments generated by gas-phase fragmentation in the MS allow rapid and sensitive identification of peptides ([1–3] and also see Chapter 1). While robust genomic analysis platforms like the widely used Affymetrix GeneChip®

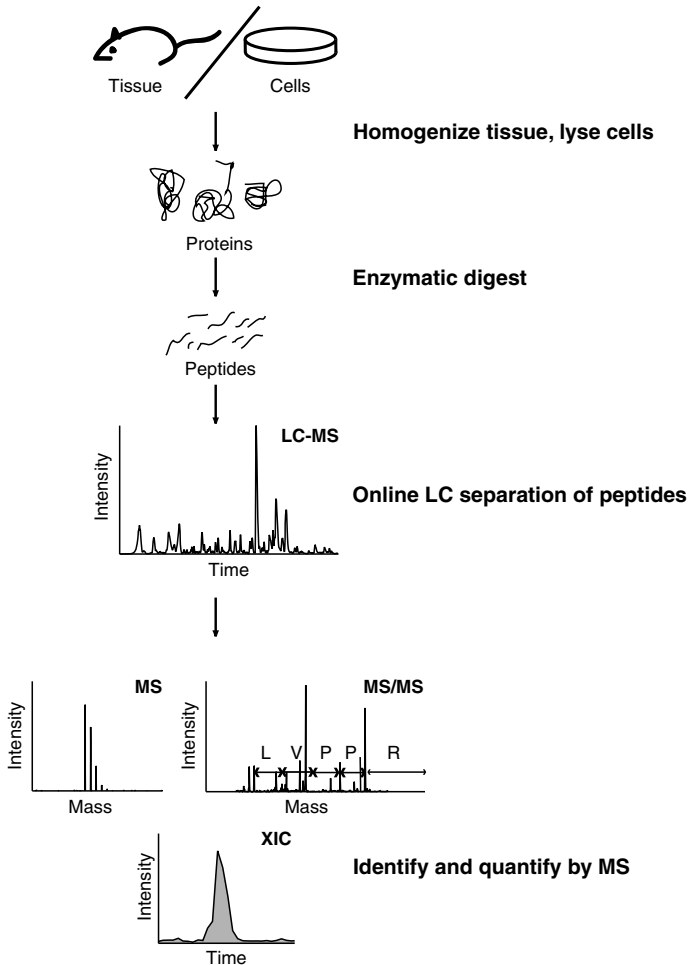
provide broad standardization of reagents, tools, protocols, and data analysis for the analysis of mRNA transcript abundance, proteomic analyses range from specific studies of single proteins to large-scale analyses of complex mixtures, often relying on specialized techniques and workflows to access specific proteomes such as post-translationally modified proteins and peptides. As a result, MS-based proteomics methods have been hard to standardize; MS experiments with different objectives, focused on either discovery or accurate quantification of known analytes, are often best performed with specific MS instrumentation. The field of proteomics is rapidly evolving, driven by novel approaches and new instrumentation. This chapter focuses on the importance of quantitative MS-based proteomics approaches that have revolutionized our study of biological systems, providing sensitive and unbiased quantitative analyses in complex systems such as temporal changes in protein phosphorylation in growth factor signaling or applications such as the specific identification of protein–protein interactions. There is further discussion of their application to the analysis of small molecule–protein interactions as well as its impact on the study of biological systems.

## 12.2

### Quantification in Biological MS

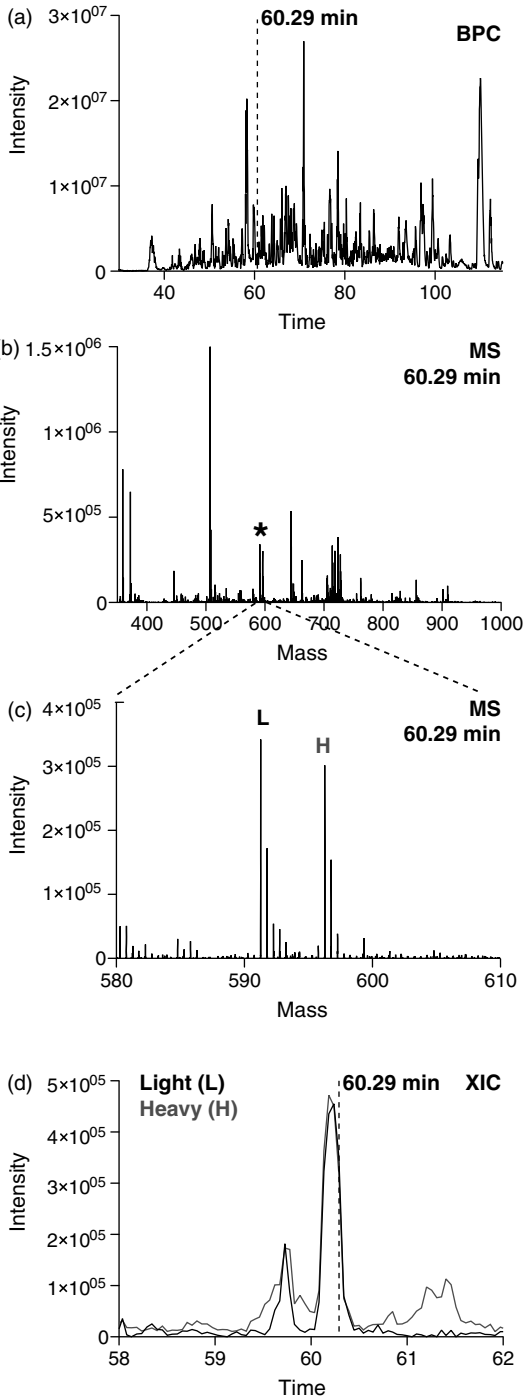
Most, if not all, experimental designs rely on the comparison of a perturbed system or state to that of a reference. The comparison of two biological samples (e.g., drug-treated versus normal cells) is the most direct approach for identifying molecular differences within these states. In a typical “shotgun” or “discovery” MS proteomics experiment (Figure 12.1), proteins are digested with proteolytic enzymes like trypsin that cleave after arginine and lysine residues to generate peptides. The shotgun proteomics approach is so named because the identification of proteins in samples is inferred by the observation of its digested peptide fragments, akin to the assembly of larger genome sequences from fragments in shotgun DNA sequencing [4, 5]. Soft ionization methods like electrospray are used to convert peptides in the liquid phase of liquid chromatography (LC) into the gas phase. Peptides elute at specific times in reversed-phase LC according to their hydrophobicity and retention on C18 matrices, allowing the MS instrument to sample complex peptide mixtures. Charged peptide ions detected by electron multipliers or array detectors in MS instruments generate signals proportional to the abundance of the peptide in the ion beam. Peptides of unique chemical composition have different mass and carry different numbers of charge-bearing amino acids allowing them to be resolved separately and identified by their corresponding mass-to-charge ( $m/z$ ) ratios. The MS instrument detects these peptides and selects them sequentially for a second tandem MS/MS experiment that isolates and fragments the peptide to yield sequence information. As modern MS instruments have dramatically improved sensitivity, duty cycle, mass accuracy and resolution, thousands of peptides resolved in LC time and  $m/z$  space can be sampled for peptide identification and quantification with nanoflow LC-MS analyses.





**Figure 12.1** Typical analytical workflow in shotgun proteomics. From a sample source of tissue or cultured cells, proteins are extracted in a buffer compatible with downstream analyses. If samples are complex, an intervening protein fractionation step such as molecular weight fractionation on a sodium dodecyl sulfate–polyacrylamide gel electrophoresis gel is often used (not shown). Proteins are proteolytically digested to peptides using enzymes with high cleavage site specificity like trypsin. Sample fractionation at the peptide level with an orthogonal approach such as strong

cation-exchange chromatography is also possible used here (not shown). Peptides are separated by their hydrophobicity on an online reversed-phase liquid chromatographic column coupled to a mass spectrometer. Intact peptide masses are detected and subsequent tandem MS/MS experiments are used to fragment peptides to generate sequence-specific information. Extracted peptide signals (extraction ion chromatogram (XIC)) are integrated and peak areas (gray) used to quantify peptides and proteins.



Quantifying proteins in complex mixtures depends on the identification of peptides and their subsequent assignment to proteins through database searching and protein grouping algorithms [6]. Peptides with identical sequences derived from protein isoforms or even unrelated sequences are often present in a sample, especially in complex protein mixtures. If the sample is not fractionated at the level of intact protein, peptide abundance information from a single peptide species may be summed across all sources of the peptide. Protein identification and quantification with unique peptides is therefore preferable for increased specificity [7], and several software packages incorporate elements to address the issue of protein inference in identification and quantification. Signal intensities or other proxies of protein abundance of identified peptides from a given protein are extracted from MS data and then summed to obtain a value for the protein (Figure 12.2).

The overall process of MS analysis is not inherently quantitative. A propagation of quantitative errors can arise from the multiple processing steps required for MS analyses. For example, losses occurring during preparation of the protein sample in a microcentrifuge tube, losses through sample introduction and the ionization process, losses of ion packets in transit from the source region to the detector, and saturation of the detector, all affect the signal intensity recorded by the MS instrument. As such, MS-based quantification is necessarily based on the relative comparison of ion signal intensities for a specific analyte. This requires the comparison between two samples analyzed in the same way, often referred to as “label-free” quantification, or within a single analysis through the incorporation of stable isotopes. An overview of different quantitative proteomics methods indicating the different sample processing stages where quantitative comparisons are made is shown in Figure 12.3.

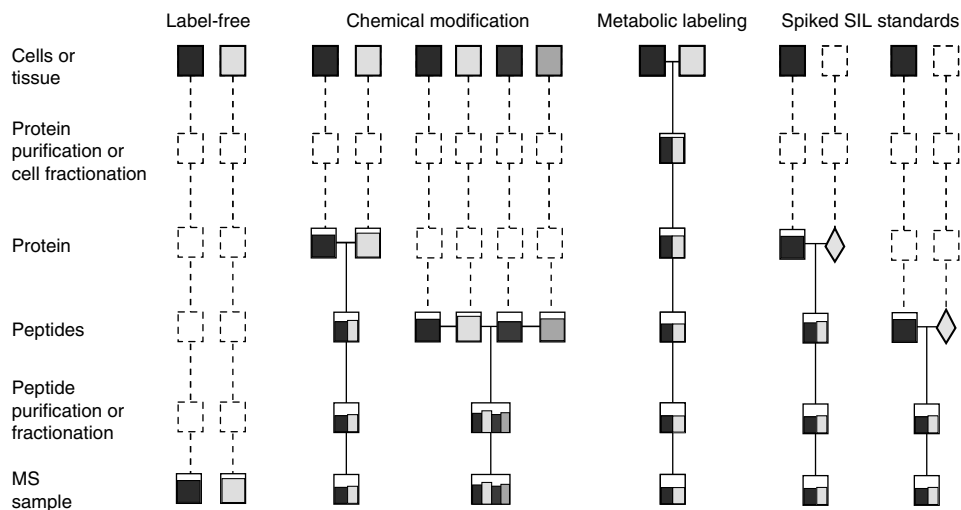
### 12.2.1

#### Label-Free Approaches in Quantitative MS Proteomics

There are several label-free approaches described in quantitative proteomics and these fall in two broad categories. The first set of approaches rely on the comparison of LC-MS chromatographic signal intensities of peptides across multiple samples,

**Figure 12.2** Quantifying peptides in nano-scale LC-MS analyses. Tens to hundreds of thousands of peptides may be present in a sample from cellular preparations. If sample complexity is excessive, fractionation at the protein or peptide level may be necessary. (a) Peptides from a single gel slice of a gel-based LC-MS (GeLC-MS) experiment are separated in an hour-long LC-MS run. The dotted line indicates the location of the MS scan at 60.29 min described in the following text. (b) In a single MS scan, several peptide species are detectable by the MS instrument based on their

charge state. Tryptic peptides in acidic LC solvent often carry a positive charge at the N-terminus of the peptide, and in the basic amino acid residues of arginine, lysine, and histidine. (c) The zoomed view of  $m/z$  580–610 is a single MS scan showing a doubly charged ( $z = 2$ ) light and heavy peptide pair separated by 5.0  $m/z$ . The mass delta is therefore 10 Da. (d) The XIC shows integrated peak signals for the light and heavy peptides over the duration of the LC run. The dotted line indicates the single MS scan, which is around 1 s of the total peak width of approximately 22 s.



**Figure 12.3** Comparing experimental workflows for quantitative proteomics. Label-free and SIL methods differ in their experimental workflows, particularly at the points where samples can be combined and analyzed together. Colored boxes and their levels indicate different samples and the respective amounts compared. Dotted lines and boxes denote stages where relative differences in quantitative yields or losses are not captured in the experimental workflow. A horizontal line connecting two boxes indicates a stage where samples are combined (SIL experiments only).

Samples that can be combined at an earlier stage in the workflow have less intersample variability. A label-free approach does not use stable isotope labels and thus quantitative comparisons occur only after each sample is completely analyzed. Chemical modification or metabolic labeling introduce SIL labels in the entire sample and allow global proteome quantitation. Spiked heavy exogenous SIL standards (yellow diamond) are designed to target only a subset of the proteins in the sample.

not unlike how one would compare UV absorbance traces in LC separation except LC-MS separates thousands of analytes in single runs in both the mass and LC retention timescale [8]. The second category is spectral counting; this relies on the assumption that sampling and identification rate of peptides by the MS instrument is proportional to the abundance of the protein in the sample [4, 9, 10]. As both categories do not use stable isotopes, there are no additional labeling steps in experimental workflows. Since no mass difference is added through stable isotope labeling, quantitative comparisons are performed between two or more complete proteomic analyses. The challenge, therefore, is to execute multistep analyses precisely in the same way without introducing bias or quantitative errors that cannot be accounted for. Furthermore, even with the greatest care, experimental variables such as sensitivity and mass calibration of the MS instrument, LC solvents and column separation material can be expected to change slightly over the course of routine analyses over a period of days or even months. As a result, peptide features in both the mass and retention time scales stretch and skew, requiring normalization across samples by spiking-in exogenous species or some internal marker [11]. Algorithms

that map peptide features, or “landmarks,” between MS data files are critically important to improve quantitative comparisons of LC chromatographic peaks in label-free quantification [12–15].

Spectral counting, which uses the numbers of peptide sequencing events as proxies for protein abundance, is a simple and straightforward approach to protein quantification. However, some estimates of protein abundance, including the number of peptide MS spectra per protein, are highly dependent on the duty cycle of the MS instrument, the size of proteins, and the complexity of the sample peptide mixture; it is therefore very important to keep MS acquisition parameters identical and to compare samples of similar complexity. Spectral counting approaches work best for larger and more abundant proteins where many peptides are sampled and compared; it is not feasible to compare proteins identified by one or two peptides using this approach [16]. Nevertheless, label-free quantification without the aid of chromatographic peak intensities is very easy and convenient to do postacquisition and is certainly useful in providing some level of quantitative comparison between lists of proteins identified in certain proteomic analyses. However, it is widely recognized that spectral counting approaches are less accurate and precise in comparison to quantitative methods using extracted LC-MS peak intensities and stable isotope labeling (SIL) approaches.

### 12.2.2

#### **SIL in Quantitative Proteomics**

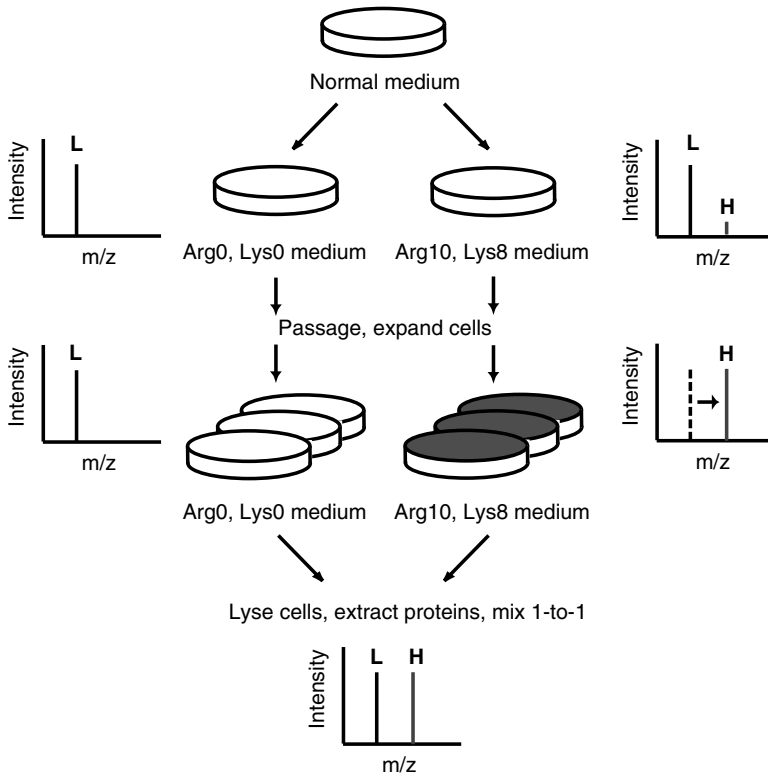
Stable, nonradioactive, isotopes, like  $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{18}\text{O}$ , and  $^2\text{H}$ , have been applied in small-molecule quantification in clinical chemistry for many decades. Through SIL, compounds are generated with identical chemical compositions, which but are isotopically distinct. These isotopologs are distinct in mass by the number of neutrons, but otherwise display very similar retention characteristics in LC.  $^{13}\text{C}$ -,  $^{15}\text{N}$ -, and  $^{18}\text{O}$ -labeled peptides have minimal effect on retention times in LC, but deuterium ( $^2\text{H}$ )-labeled peptides exhibit a more noticeable shift in LC retention times, preceding the normal stable isotope-labeled peptides by seconds or even minutes over the course of an hour-long LC run. Quantification of deuterated samples may be less accurate than  $^{13}\text{C}$ - or  $^{15}\text{N}$ -labeled samples as electrospray ionization conditions during the elution of each isotopolog pair may differ, but this effect is likely to be small in comparison to other sources of quantitative variation. In MS, the mass envelope (isotopic cluster) of a peptide species depends on its chemical composition and size. Tryptic peptides in the human proteome have an average length of about 11 amino acid residues, and most peptides observed in MS have masses between 600 and 2500 Da. For most peptides of this size, the monoisotopic peak (the mass peak containing the dominant isotope within the population) has the highest intensity and the isotopic distributions are small such that each isotopic envelope contains three major peaks (Figures 12.1 and 12.2c). Accordingly, at least four SIL-enriched atoms (e.g., six  $^{13}\text{C}$  in  $^{13}\text{C}_6$ -lysine) should be incorporated in the SIL peptide to separate it from its counterpart with a normal isotope composition in

the mass scale (Figure 12.2c). The mass separation of the “light” and “heavy” peptide isotopic clusters allows proper integration of signal intensities from each peptide form – a requirement for good quantitative accuracy.

There are two primary approaches in using stable isotope labels in quantitative proteomics; (i) global quantification by introducing SIL throughout whole proteome samples and quantifying relative peptide and protein abundance between isotopologs, and (ii) quantification with spiked internal standards by introducing known amounts of exogenous SIL peptides and comparing their abundance to levels of endogenous peptides. For the former approach, the main difference is the stage at which labels are introduced globally in the proteome sample – metabolic incorporation before harvest of protein and postharvest chemical derivatization of proteins or peptides (Figure 12.3).

There are obvious advantages and disadvantages in either approach to globally label protein and peptide mixtures. Metabolic labeling is very simple, robust, and cost-effective, but is only possible with living cells. Chemical labeling can be performed with any protein sample, but the derivatization step to incorporate stable isotope labels occurs only at a late stage after proteins have been digested to generate peptides. As quantitative proteomics workflows improve and becoming increasingly robust, the choice of labeling method will depend more on the sample, the type of experimental comparison required, and the costs associated with instrument analysis time and reagents (Figure 12.3) (for recent reviews, see [17, 18]).

SIL with amino acids in cell culture (stable isotope labeling with amino acids in cell culture SILAC) is the most widely used metabolic labeling approach for quantitative proteomics [19–21]. Cells are grown in a medium containing stable isotope-containing amino acids like  $^{13}\text{C}_6$ -arginine and  $^{13}\text{C}_6$   $^{15}\text{N}_2$ -lysine, incorporating isotopic labels in newly synthesized protein as cells divide and proteins turn over through normal growth (Figure 12.4). The incorporation of stable isotopes in amino acids is convenient, especially compared to  $^{15}\text{N}$  labeling, because the protease used to generate peptides can be coupled with the SILAC amino acid used. For example, arginine and lysine labeling in combination with trypsin digests constrains the number of stable isotopes incorporated and also provides confirmation of the peptide sequence [22, 23]. SILAC labeling generates two cell populations with proteins that are essentially identical except in mass, and can therefore be mixed even as live cells and processed together in every downstream biochemical step. It has become increasingly popular because the metabolic labeling process is robust, simple to perform, essentially complete in the entire proteome, and inexpensive compared to chemical labeling. The method was originally described for mammalian cell culture [19], but has been since been applied in yeast [20, 24], bacteria [25, 26], and even higher organisms like the chicken [27], mouse [28], and newt [29]. In addition to the more conventional case versus control SILAC experiment, the SILAC method has also been used to label cellular proteomes for the generation of whole-proteome SIL standards that can be combined with unlabeled protein or peptide samples for quantification [30, 31]. Furthermore, a variant of the SILAC approach was developed to study PTMs in protein. As many amino acids differ in mass by a methyl group, proteomic analyses of methylated peptides has a higher chance of false-positive

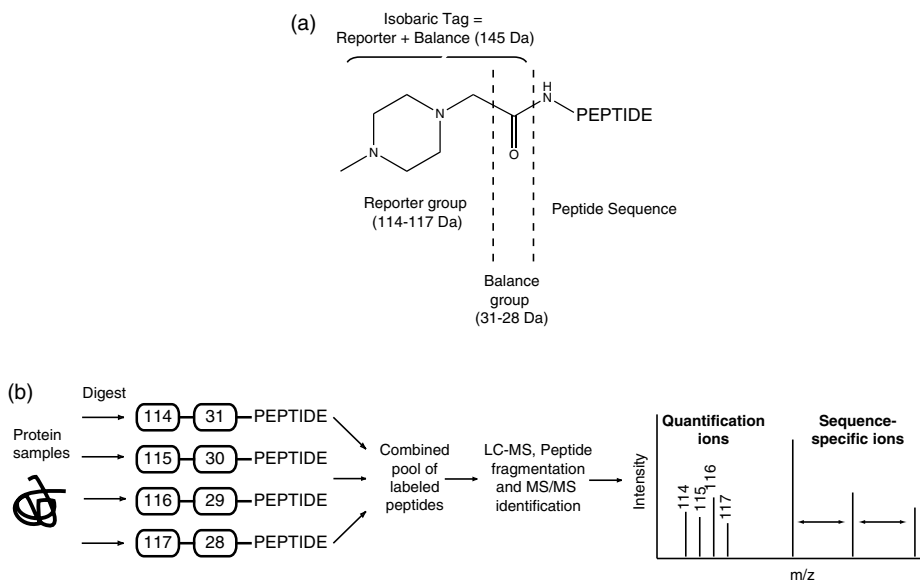


**Figure 12.4** SIL by amino acids in cell culture. Cells growing in normal culture medium are passaged into SILAC light and heavy culture media [19, 23]. As cells grow and divide in the heavy medium, they incorporate the heavy amino acids  $^{13}\text{C}_6\ ^{15}\text{N}_4$ -arginine (Arg10) and  $^{13}\text{C}_6\ ^{15}\text{N}_2$ -lysine (Lys8) throughout their proteome. Cells are passaged and can be expanded to generate the desired number of

dishes. After five cell doublings, cells should incorporate in excess of 96% of the heavy amino acid by dilution of the parental light forms along. In practice this is much higher based on protein turnover and passaging of cells. The light peptides are exchanged to heavy peptide forms in the heavy state and when a 1:1 mixture of protein is combined, each peptide is observed in a light and heavy pair.

identification [32]. Labeling methyl groups with heavy methyl SILAC facilitates the identification and quantification of protein methylation [32, 33].

Postprotein harvest incorporation of stable isotope labels can be performed with chemical derivatization of amino acid functional groups like thiols (ICAT (isotope-coded affinity tags)) [34], carboxyl groups [35], and amines (iTRAQ (isobaric tag for relative and absolute quantification) [36], TMT (tandem mass tags) [37], and dimethyl labeling [38, 39]) or through proteolytic  $^{18}\text{O}$ -labeling of the C-termini of peptides [40]. There are several popular chemical labeling reagents and chemistries used in quantitative proteomics, many available as kits from commercial vendors. Beyond the simple introduction of stable isotope labels for quantitation, chemical labeling allows new functionalities to be added to labeled peptides, such as a biotin moiety for

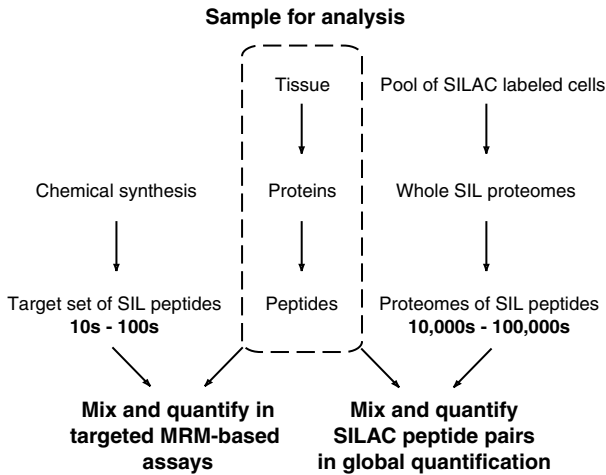


**Figure 12.5** Chemical labeling with iTRAQ. (a) The iTRAQ reagent is an example of a chemical derivatization reagent that targets primary amines, like the N-termini and  $\epsilon$ -amino groups of lysines in peptides. It is designed to label and incorporate mass tags that can distinguish peptides from multiple samples in a single MS analysis. In contrast to other quantitative methods like SILAC or ICAT, which quantify peptides at the intact precursor level, iTRAQ is designed as isobaric tags that contain two parts: reporter groups (114–117 Da) that fragment in the MS/MS spectrum for quantification and balance groups (31–28 Da) to bring each

reporter–balance tag to a combined mass of 145 Da. (b) Abundances of proteins from four states can be compared at the same time. Extracted proteins are digested separately and labeled with the four iTRAQ reagents. After peptide labeling is completed, the four samples can be combined and analyzed together. As labeled peptides from multiple samples do not resolve in the MS scan and do not increase sample complexity, this makes the general approach very conducive for multiplexing, and the latest generation of iTRAQ reagents allows up to eight samples to be labeled and combined for simultaneous quantification.

specific enrichment on avidin matrices [34]. This ability to design chemical labeling tags allows for unique approaches in quantitative proteomics. For instance, with metabolic labeling like SILAC or chemical tags like ICAT, quantification occurs in the full-scan MS mode comparing the signal intensities of intact peptides (Figure 12.1). In contrast, chemical labels like iTRAQ [36] or TMT [37] are designed to be equal in mass. When these reagents are used to label peptides from different samples, peptides do not separate by mass and therefore do not increase the number of detectable peptide forms in the MS scan. Upon fragmentation of the peptide in MS/MS experiments, however, each label generates a fragment ion reporter with a distinct mass allowing quantification between samples (Figure 12.5). This labeling strategy then allows a larger number of samples, up to eight with newer versions of iTRAQ, to be tagged and combined in a single MS analysis. Many opportunities to innovate in chemical labeling strategies still remain, particularly in developing





**Figure 12.6** Quantification using exogenous stable isotope-labeled peptide standards. The tissue sample to be analyzed is common to both forks in the workflow and is marked in the dotted box. Tissue samples are first processed to extract proteins and digested with trypsin to generate complex mixtures of peptides. In a targeted MRM-based assay (left branch) [52, 82], known amounts of chemically synthesized SIL peptides matching peptides from target proteins are introduced to the sample and serve as relative internal standards

in peptide quantification. In an alternate workflow, pools of SILAC labeled cells are combined; extracted proteins are digested with the same enzyme (trypsin) to generate a whole-proteome SIL peptide standard containing 10 000 to 100 000s of peptides [30]. This SIL proteome standard can be adjusted to match the cellular characteristics of the sample to be quantified. A large stock of a suitable proteome standard could be a common internal reference spiked into hundreds of experiments.

reagents that combine quantification and other attractive features to target specific enzyme classes [41]. Reagents that bind *in vivo* protein complexes, and then allow specific capture and enrichment of these complexes *in vitro*, would also have tremendous potential in proteomics.

For several decades, small molecules containing stable isotopes have been used in the pharmaceutical industry as internal standards to quantify levels of drugs in complex samples like human plasma [42–44]. Desiderio and Kai extended this to peptide quantification in the early 1980s [45]. In the 2000s, nanoflow-LC-MS with lower flow rates (200 nl/min) and faster and more sensitive MS instruments [46, 47] based on triple-quadrupole MS [48, 49] provided a robust, sensitive, and specific platform for peptide quantification in complex mixtures [50, 51]. This multiple-reaction monitoring (MRM) approach is becoming increasingly popular for targeted analyses of specific proteins in the biomarker verification field [52, 53] as well as to monitor levels of kinases and phosphatases in cell signaling [54]. Depending on the experimental design, just one or up to a few hundred SIL peptides can be spiked into samples as internal standards for quantification (Figure 12.6, left branch). Quantitative MRM-MS assays are typically specifically designed to increase the number of distinct peptide species monitored in a single MS run without compromising

specificity and sensitivity of the assay. It is now possible to monitor the abundances of a few hundred peptides with between two to five peptide precursor-fragment transitions per peptide (MRMs) in an hour-long LC-MS run.

In an alternative approach, proteins from cells heavily labeled with SILAC were digested to generate a whole-proteome SIL internal standard and spiked in to quantify peptides from unlabeled tissue samples [30, 31, 55]. Instead of a targeted analysis of specific peptides, the whole-proteome SIL internal standard uses precursor MS ion quantification of any identified peptide using discovery-mode LC-MS proteomic analyses typical of other global analyses. As several hundreds of thousands of peptides would be present in the whole-proteome SIL standard, spiking the heavy labeled peptides into the unlabeled sample will double, at a minimum, the number of peptide forms present. The whole-proteome SIL standard would therefore be less useful in MRM-based analyses because MS signals from coeluting peptides interfering with the target analyte are detrimental to quantitative accuracy. Instead, the high accuracy and mass resolution of MS and MS/MS spectra in the high-performance MS instruments commonly used in shotgun proteomics today allows the global quantification of thousands of proteins from tissue samples using the reference proteome SIL standard (Figure 12.6, right branch) [30].

The first chapter of MS-based proteomics in the 1990s was marked by the identification of proteins separated to near-homogeneity on gels; improving methodologies and instrumentation have since driven the development of new experimental workflows, and proteomics experiments are increasingly designed in a quantitative format [7]. Accurate quantification of peptides and proteins by mass spectrometry is readily accessible with relatively low additional cost or effort by the researcher. The returns on invested effort are many-fold:

- i) Quantitative data allow comparative experiments at the level of single proteins, organelles, cells, tissues, and whole organisms.
- ii) Whole-proteome labeling allows quantitation of all identified peptides, including PTMs such as phosphorylated peptides [56].
- iii) Multiplexing of protein populations facilitates expression profiling data such as timecourse measurements of protein abundances for thousands of proteins [56–58].
- iv) Quantitative labeling of populations of proteins can serve as analytical reagents in comparing experimental workflows (e.g., to measure the relative yield of peptides from a protein digestion protocol [59]).
- v) Quantitative measures of protein abundance can serve as markers to distinguish proteins of interest from background proteins [57, 60–63].

### 12.3

#### Identifying Proteins Interacting with Small Molecules with Quantitative Proteomics

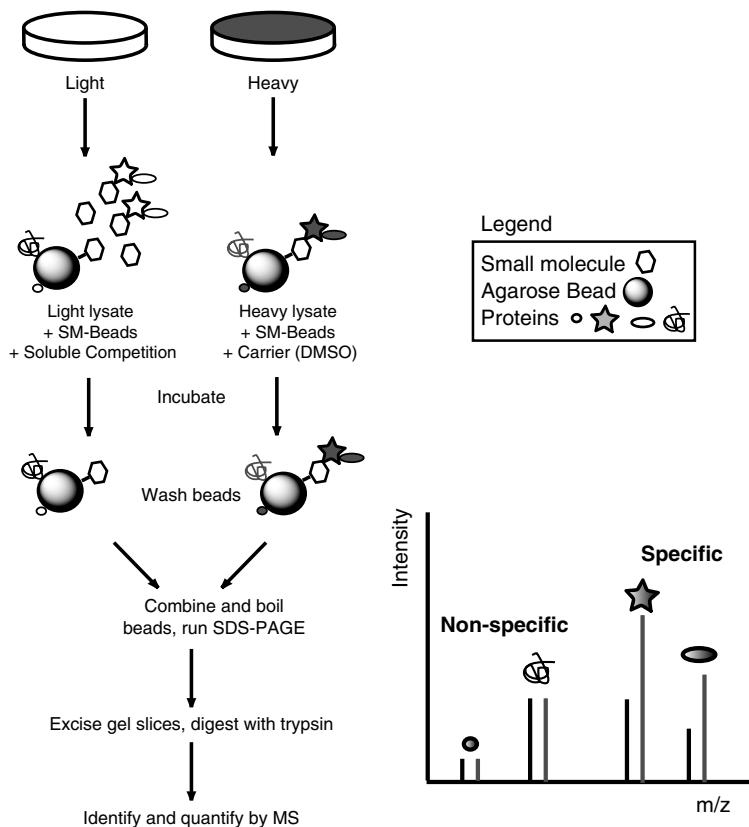
The biochemical enrichment of proteins with affinity baits has been applied to study protein–protein interactions for decades. It is commonly used to detect and validate protein interactions in a coimmunoprecipitation experiment using antibodies and Western blotting as readout. With proteomics, the affinity pull-down experiment

provides an opportunity to identify novel protein interactions with baits in an unbiased manner, making this experiment an incredibly powerful method to functionally interrogate molecular baits. More recently, quantitative proteomics has been applied to this experimental paradigm to distinguish proteins bound specifically to the molecular bait versus nonspecific protein interactions with linker regions or the solid support itself [60, 61, 63, 64]. There are many ways to immobilize molecular baits like protein, nucleic acids, and small molecules to solid-phase matrices for use in affinity enrichment, including biotinylation, direct covalent coupling, and antibodies. In order to be a useful tool, however, immobilized baits would need to retain the relevant binding sites for interacting proteins. This is difficult to ascertain and even cross-validation with known protein interactions will not guarantee that *in vitro* protein interactions with the affinity matrix will fully recapitulate the behavior of the molecule *in vivo*.

Nevertheless, the affinity capture of proteins interacting with solid phase matrices is, in many cases, the most straightforward and established approach to identifying mechanisms of action of small molecules in target deconvolution [65, 66] (and reviewed in [67, 68]). As gene expression- and image-based phenotype screens to identify novel bioactive molecules increase in scope and popularity [69], the numbers of small molecules eliciting interesting biological phenotypes but with unknown mechanisms of action will steadily increase. The unbiased proteomic approach is therefore extremely important as its application to identify proteins interacting with small-molecule affinity baits may reveal novel interactors and inform downstream experiments [60, 62, 70, 71].

Enrichment of proteins by small-molecule matrices can be compared using quantitative proteomics, discriminating interacting proteins from ones that bind nonspecifically to beads. The primary issue at hand is that current LC-MS analyses are very sensitive and most pull-down experiments would yield many false-positive, nonspecifically bound proteins. By using the quantitative ratios to identify bona fide protein–small molecule interactions among the nonspecific interactions, the need for *ad hoc* optimization of experimental conditions in each pull-down experiment is diminished. Importantly, the usual tradeoff between sensitivity and specificity of the affinity enrichment experiment is mitigated by the use of the quantitative information.

As discussed in Section 12.2, SILAC is one of several quantitative proteomics approaches applied to chemical proteomics [62, 72]. SILAC is a convenient approach particularly where cell-based phenotypic screens were used in the discovery of the small molecules. In a SILAC labeling experiment using affinity enrichment with a kinase inhibitor (Figure 12.7), equal amounts of light and heavy SILAC-labeled cell lysates are presented to the kinase inhibitor affinity matrices, but an excess of soluble kinase inhibitor is coincubated with the light lysate and beads. The carrier solvent, dimethylsulfoxide (DMSO), is added to the pull-down containing heavy cell lysate and beads as a control. As kinases bind to the soluble form of the kinase inhibitor present in excess, the relative amounts of kinase bound to beads in the light sample will be lower than in the heavy sample. When the beads from both light and heavy pull-downs are combined and proteins analyzed by LC-MS, target kinases will have differential heavy/light ratios and proteins that do not interact with the small

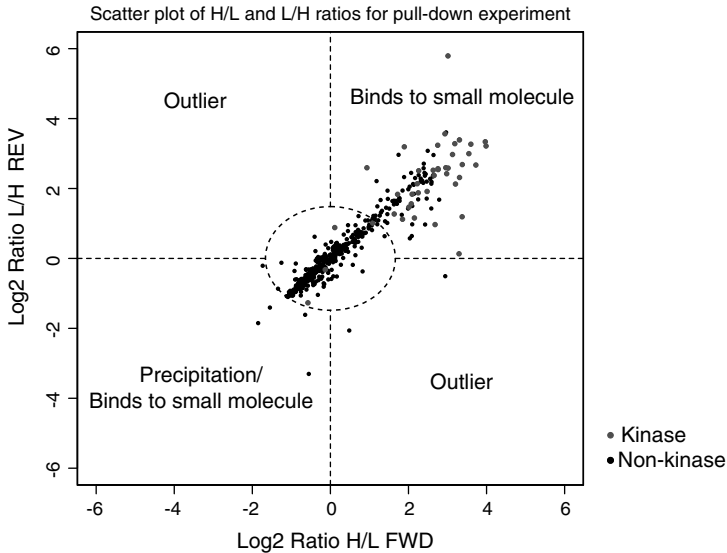


**Figure 12.7** Identifying proteins interacting with an immobilized kinase inhibitor. Equal amounts of light and heavy SILAC labeled cell lysates will be incubated with the kinase inhibitor affinity matrices, but an excess of soluble kinase inhibitor is first introduced to the light lysate. The carrier solvent, DMSO, is added to the pull-down containing heavy cell lysate as a control. Small-molecule affinity beads are introduced to both pull-downs in equal amounts to bind target protein. As kinases also bind to the soluble form of the kinase inhibitor present

in excess, the relative amounts of kinase bound to beads in the light sample will be lower than in the heavy sample. The beads from both light and heavy pull-downs are combined and eluted proteins analyzed by LC-MS; target kinases have differential heavy/light ratios and proteins that do not interact with the small molecule will have ratios close to 1 : 1. This allows proteins interacting with the small molecule to be easily distinguished from proteins bound nonspecifically to the solid-phase matrix.

molecule will have ratios close to 1 : 1. Quantitative ratios allow estimates of protein ratio significance to be calculated with statistical tools – a very useful approach for prioritizing protein hits for downstream validation. Replicate experiments are commonly performed where combinations of SIL states and experiments are swapped, thus requiring ratios of true-positive hits to invert in replicate label-swap experiments and this provides yet another strong discriminating filter (Figure 12.8).

The application of quantitative proteomics to the important area of target identification of small-molecule bioactives is just one example of how modern



**Figure 12.8** Ratios and experimental design provide increased confidence in target selection. Quantitative proteomics experiments comparing two states such as pull-down or differential expression experiments are most useful when changes in protein abundances occur in just a subset of all quantified proteins. Each data point is a quantified protein with SILAC ratios from two experiments as the x- and y- coordinate, respectively. The majority of proteins in the dataset have unchanged ratios, in the example here with a kinase inhibitor

affinity pull-down (described in Figure 12.7) the central distribution of proteins within the dotted circle has  $\log_2$  heavy/light ratios near zero. Kinases (gray dots) enriched by the kinase inhibitor affinity matrix are found in the top right of the plot indicating high heavy/light ratios in one experiment and the high inverted light/heavy ratios in the label-swap replicate. This allows clear discrimination of the targets from the null distribution of nonspecific interactions with the kinase inhibitor beads (see also [62, 64]).

quantitative MS is enhancing classical biochemical approaches. Many well-established and powerful biochemical methodologies can now be performed with MS as a readout, allowing simultaneous evaluation of thousands of proteins instead of single-protein studies with UV absorbance measurements. Through this renaissance in applying biochemical methods, researchers are applying quantitative proteomics methods to obtain a more global analysis of subcellular proteomes like the mitochondrion [73] or nuclear bodies like the nucleolus [74].

## 12.4 Conclusions

The toolbox for MS-based proteomics has been expanding rapidly over the past decade and analyses have become increasingly sophisticated. Instead of routine protein identification, quantitative proteomics is allowing the encoding of functional states into stable isotope-labeled protein populations. We can now identify thousands

of proteins and simultaneously measure the changes in abundance of these proteins in response to biological perturbation using quantitative proteomics. In the future, all proteomics studies should be designed as quantitative analyses since it requires little additional expense, particularly in comparison to the high cost of MS instrumentation. Quantitative proteomics is also becoming increasingly accessible to researchers as there are many robust chemical or metabolic labeling methods available today, and commercial and academic software [75–79] to support MS quantification analyses are much improved.

Our ability to study and understand protein function at the proteome level is the key to our success in integrating genomic and proteomic datasets. Next-generation sequencing is already driving a new revolution in gene expression studies [80, 81], and proteomics methods will need to continue to evolve and improve. The ability to combine both these domains will bring tremendous insight and power to our analyses of gene regulation, and is one of the first important steps towards developing a systems-wide understanding of biology.

## References

- 1 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
- 2 Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, **5**, 699–711.
- 3 Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.
- 4 Washburn, M.P., Wolters, D., and Yates, J.R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.
- 5 Marcotte, E.M. (2007) How do shotgun proteomics algorithms identify proteins? *Nature Biotechnology*, **25**, 755–757.
- 6 Nesvizhskii, A.I. and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Molecular and Cellular Proteomics*, **4**, 1419–1440.
- 7 Ong, S.-E. and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology*, **1**, 252–262.
- 8 Chelius, D. and Bondarenko, P.V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *Journal of Proteome Research*, **1**, 317–323.
- 9 Liu, H., Sadygov, R.G., and Yates, J.R. 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, **76**, 4193–4201.
- 10 Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, **25**, 117–124.
- 11 Callister, S.J., Barry, R.C., Adkins, J.N., Johnson, E.T., Qian, W.-j., Webb-Robertson, B.-J.M., Smith, R.D., and Lipton, M.S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of Proteome Research*, **5**, 277–286.
- 12 Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T.A., Hill, L.R., Norton, S., Kumar, P., Anderle, M., and Becker, C.H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, **75**, 4818–4826.
- 13 Wiener, M.C., Sachs, J.R., Deyanova, E.G., and Yates, N.A. (2004) Differential mass spectrometry: a label-free LC-MS method for finding significant differences in

- complex peptide and protein mixtures. *Analytical Chemistry*, **76**, 6085–6096.
- 14 Jaffe, J.D., Mani, D.R., Leptos, K.C., Church, G.M., Gillette, M.A., and Carr, S.A. (2006) PEPPer, a platform for experimental proteomic pattern recognition. *Molecular and Cellular Proteomics*, **5**, 1927–1941.
  - 15 Mueller, L.N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Müller, M. (2007) SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, **7**, 3470–3480.
  - 16 Zybailov, B., Coleman, M.K., Florens, L., and Washburn, M.P. (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Analytical Chemistry*, **77**, 6218–6224.
  - 17 Domon, B. and Aebersold, R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology*, **28**, 710–721.
  - 18 Mallick, P. and Kuster, B. (2010) Proteomics: a pragmatic perspective. *Nature Biotechnology*, **28**, 695–709.
  - 19 Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics*, **1**, 376–386.
  - 20 Jiang, H. and English, A.M. (2002) Quantitative analysis of the yeast proteome by incorporation of isotopically labeled leucine. *Journal of Proteome Research*, **1**, 345–350.
  - 21 Zhu, H., Pan, S., Gu, S., Bradbury, E.M., and Chen, X. (2002) Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Communications in Mass Spectrometry*, **16**, 2115–2123.
  - 22 Ibarrola, N., Kalume, D.E., Gronborg, M., Iwahori, A., and Pandey, A. (2003) A proteomic approach for quantitation of phosphorylation using stable isotope labeling in cell culture. *Analytical Chemistry*, **75**, 6043–6049.
  - 23 Ong, S.-E. and Mann, M. (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols*, **1**, 2650–2660.
  - 24 Gruhler, A., Olsen, J.V., Mohammed, S., Mortensen, P., Faergeman, N.J., Mann, M., and Jensen, O.N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Molecular and Cellular Proteomics*, **4**, 310–327.
  - 25 Deng, W., de Hoog, C.L., Yu, H.B., Li, Y., Croxen, M.A., Thomas, N.A., Puente, J.L., Foster, L.J., and Finlay, B.B. (2010) A comprehensive proteomic analysis of the type III secretome of *Citrobacter rodentium*. *Journal of Biological Chemistry*, **285**, 6790–6800.
  - 26 Soufi, B., Kumar, C., Gnad, F., Mann, M., Mijakovic, I., and Macek, B. (2010) Stable isotope labeling by amino acids in cell culture (SILAC) applied to quantitative proteomics of *Bacillus subtilis*. *Journal of Proteome Research*, **9**, 3638–3646.
  - 27 Hayter, J.R., Doherty, M.K., Whitehead, C., McCormack, H., Gaskell, S.J., and Beynon, R.J. (2005) The subunit structure and dynamics of the 20S proteasome in chicken skeletal muscle. *Molecular and Cellular Proteomics*, **4**, 1370–1381.
  - 28 Krüger, M., Moser, M., Ussar, S., Thievensen, I., Luber, C.A., Forner, F., Schmidt, S., Zanivan, S., Fässler, R., and Mann, M. (2008) SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell*, **134**, 353–364.
  - 29 Looso, M., Borchardt, T., Kruger, M., and Braun, T. (2010) Advanced identification of proteins in uncharacterized proteomes by pulsed *in vivo* stable isotope labeling-based mass spectrometry. *Molecular and Cellular Proteomics*, **9**, 1157–1166.
  - 30 Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods*, **7**, 383–385.
  - 31 Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T., and Oda, Y. (2005) Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards. *Nature Biotechnology*, **23**, 617–621.

- 32 Ong, S.-E., Mittler, G., and Mann, M. (2004) Identifying and quantifying *in vivo* methylation sites by heavy methyl SILAC. *Nature Methods*, **1**, 119–126.
- 33 Zee, B.M., Levin, R.S., Xu, B., LeRoy, G., Wingreen, N.S., and Garcia, B.A. (2010) *In vivo* residue-specific histone methylation dynamics. *Journal of Biological Chemistry*, **285**, 3341–3350.
- 34 Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.
- 35 Goodlett, D.R., Keller, A., Watts, J.D., Newitt, R., Yi, E.C., Purvine, S., Eng, J.K., von Haller, P., Aebersold, R., and Kolker, E. (2001) Differential stable isotope labeling of peptides for quantitation and *de novo* sequence derivation. *Rapid Communications in Mass Spectrometry*, **15**, 1214–1221.
- 36 Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlett-Jones, M., He, F., Jacobson, A., and Pappin, D.J. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics*, **3**, 1154–1169.
- 37 Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, **75**, 1895–1904.
- 38 Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A.J.R. (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nature Protocols*, **4**, 484–494.
- 39 Hsu, J.-L., Huang, S.-Y., Chow, N.-H., and Chen, S.-H. (2003) Stable-isotope dimethyl labeling for quantitative proteomics. *Analytical Chemistry*, **75**, 6843–6852.
- 40 Miyagi, M. and Rao, K.C.S. (2007) Proteolytic  $^{18}\text{O}$ -labeling strategies for quantitative proteomics. *Mass Spectrometry Reviews*, **26**, 121–136.
- 41 Cravatt, B.F., Wright, A.T., and Kozarich, J.W. (2008) Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annual Review of Biochemistry*, **77**, 383–414.
- 42 Baty, J.D., Robinson, P.R., and Wharton, J. (1976) A method for the estimation of acetanilide, paracetamol and phenacetin in plasma and urine using mass fragmentography. *Biomedical Mass Spectrometry*, **3**, 60–63.
- 43 MacCoss, M.J., Fukagawa, N.K., and Matthews, D.E. (1999) Measurement of homocysteine concentrations and stable isotope tracer enrichments in human plasma. *Analytical Chemistry*, **71**, 4527–4533.
- 44 Bakhtiar, R., Lohne, J., Ramos, L., Khemani, L., Hayes, M., and Tse, F. (2002) High-throughput quantification of the anti-leukemia drug STI571 (Gleevec) and its main metabolite (CGP 74588) in human plasma using liquid chromatography-tandem mass spectrometry. *Journal of Chromatography B*, **768**, 325–340.
- 45 Desiderio, D.M. and Kai, M. (1983) Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biomedical Mass Spectrometry*, **10**, 471–479.
- 46 Hager, J.W. and Yves Le Blanc, J.C. (2003) Product ion scanning using a Q-q-Q linear ion trap (Q TRAP) mass spectrometer. *Rapid Communications in Mass Spectrometry*, **17**, 1056–1064.
- 47 Hopfgartner, G., Varesio, E., Tschäppät, V., Grivet, C., Bourgonne, E., and Leuthold, L.A. (2004) Triple quadrupole linear ion trap mass spectrometer for the analysis of small molecules and macromolecules. *Journal of Mass Spectrometry*, **39**, 845–855.
- 48 Yost, R.A. and Enke, C.G. (1978) Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society*, **100**, 2274–2275.
- 49 Yost, R.A., Perchalski, R.J., Brotherton, H.O., Johnson, J.V., and Budd, M.B. (1984) Pharmaceutical and clinical



- analysis by tandem mass spectrometry. *Talanta*, **31**, 929–935.
- 50 Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6940–6945.
- 51 Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W., and Guild, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and <sup>13</sup>C-labeled peptide standards. *Proteomics*, **4**, 1175–1186.
- 52 Addona, T.A., Abbatiello, S.E., Schilling, B., Skates, S.J., Mani, D.R., Bunk, D.M., Spiegelman, C.H., Zimmerman, L.J., Ham, A.-J.L., Keshishian, H., Hall, S.C., Allen, S., Blackman, R.K., Borchers, C.H., Buck, C., Cardasis, H.L., Cusack, M.P., Dodder, N.G., Gibson, B.W., Held, J.M., Hiltke, T., Jackson, A., Johansen, E.B., Kinsinger, C.R., Li, J., Mesri, M., Neubert, T.A., Niles, R.K., Pulsipher, T.C., Ransohoff, D., Rodriguez, H., Rudnick, P.A., Smith, D., Tabb, D.L., Tegeler, T.J., Variyath, A.M., Vega-Montoto, L.J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J.R., Zhao, L., Anderson, N.L., Fisher, S.J., Liebler, D.C., Paulovich, A.G., Regnier, F.E., Tempst, P., and Carr, S.A. (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nature Biotechnology*, **27**, 633–641.
- 53 Keshishian, H., Addona, T., Burgess, M., Kuhn, E., and Carr, S.A. (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Molecular and Cellular Proteomics*, **6**, 2212–2229.
- 54 Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., and Aebersold, R. (2010) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nature Methods*, **7**, 43–46.
- 55 Neubert, T.A. and Tempst, P. (2010) Super-SILAC for tumors and tissues. *Nature Methods*, **7**, 361–362.
- 56 Olsen, J.V., Blagoev, B., Gnäd, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.
- 57 Blagoev, B., Ong, S.-E., Kratchmarova, I., and Mann, M. (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature Biotechnology*, **22**, 1139–1145.
- 58 Zhang, Y., Wolf-Yadlin, A., Ross, P.L., Pappin, D.J., Rush, J., Lauffenburger, D.A., and White, F.M. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Molecular and Cellular Proteomics*, **4**, 1240–1250.
- 59 Ong, S.-E., Mortensen, P., and Mann, M. (2005) Comparing in-solution and in-gel enzymatic digestion efficiencies with SILAC. 53rd American Society for Mass Spectrometry Conference, San Antonio, TX.
- 60 Oda, Y., Owa, T., Sato, T., Boucher, B., Daniels, S., Yamanaka, H., Shinohara, Y., Yokoi, A., Kuromitsu, J., and Nagasu, T. (2003) Quantitative chemical proteomics for identifying candidate drug targets. *Analytical Chemistry*, **75**, 2159–2165.
- 61 Blagoev, B., Kratchmarova, I., Ong, S.-E., Nielsen, M., Foster, L.J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnology*, **21**, 315–318.
- 62 Ong, S.-E., Schenone, M., Margolin, A.A., Li, X., Do, K., Doud, M.K., Mani, D.R., Kuai, L., Wang, X., Wood, J.L., Tolliday, N.J., Koehler, A.N., Marceau, L.A., Golub, T.R., Gould, R.J., Schreiber, S.L., and Carr, S.A. (2009) Identifying the proteins to which small-molecule probes and drugs bind in cells. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 4617–4622.
- 63 Ranish, J.A., Yi, E.C., Leslie, D.M., Purvine, S.O., Goodlett, D.R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by

- quantitative proteomics. *Nature Genetics*, **33**, 349–355.
- 64 Ong, S.-E. (2010) Unbiased identification of protein–bait interactions using biochemical enrichment and quantitative proteomics. *Cold Spring Harbor Protocols*, **2010**, pdb prot5400.
- 65 Cuatrecasas, P. (1970) Agarose derivatives for purification of protein by affinity chromatography. *Nature*, **228**, 1327–1328.
- 66 Harding, M.W., Galat, A., Uehling, D.E., and Schreiber, S.L. (1989) A receptor for the immunosuppressant FK506 is a *cis-trans* peptidyl–prolyl isomerase. *Nature*, **341**, 758–760.
- 67 Rix, U. and Superti-Furga, G. (2009) Target profiling of small molecules by chemical proteomics. *Nature Chemical Biology*, **5**, 616–624.
- 68 Terstappen, G.C., Schlüpen, C., Raggiacchi, R., and Gaviraghi, G. (2007) Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery*, **6**, 891–903.
- 69 Stockwell, B.R. (2004) Exploring biology with small organic molecules. *Nature*, **432**, 846–854.
- 70 Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C., Reader, V., Sweetman, G., Bauer, A., Bouwmeester, T., Hopf, C., Kruse, U., Neubauer, G., Ramsden, N., Rick, J., Kuster, B., and Drewes, G. (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnology*, **25**, 1035–1044.
- 71 Godl, K., Wissing, J., Kurtenbach, A., Habenberger, P., Blencke, S., Gutbrod, H., Salassidis, K., Stein-Gerlach, M., Missio, A., Cotten, M., and Daub, H. (2003) An efficient proteomics method to identify the cellular targets of protein kinase inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **100** 15434–15439.
- 72 Daub, H., Olsen, J.V., Bairlein, M., Gnad, F., Oppermann, F.S., Körner, R., Greff, Z., Keri, G., Stemmann, O., and Mann, M. (2008) Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Molecular Cell*, **31**, 438–448.
- 73 Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.-E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., Hill, D.E., Vidal, M., Evans, J.G., Thorburn, D.R., Carr, S.A., and Mootha, V.K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–123.
- 74 Andersen, J.S., Lam, Y.W., Leung, A.K.L., Ong, S.-E., Lyon, C.E., Lamond, A.I., and Mann, M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77–83.
- 75 Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, **26**, 1367–1372.
- 76 MacCoss, M.J., Wu, C.C., Liu, H., Sadygov, R., and Yates, J.R. 3rd (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Analytical Chemistry*, **75**, 6912–6921.
- 77 MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, **26**, 966–968.
- 78 Mortensen, P., Gouw, J.W., Olsen, J.V., Ong, S.-E., Rigbolt, K.T.G., Bunkenborg, J., Cox, J., Foster, L.J., Heck, A.J.R., Blagoev, B., Andersen, J.S., and Mann, M. (2010) MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *Journal of Proteome Research*, **9**, 393–403.
- 79 Muth, T., Keller, D., Puetz, S.M., Martens, L., Sickmann, A., and Boehm, A.M. (2010) jTraQX: a free, platform independent tool for isobaric tag quantitation at the protein level. *Proteomics*, **10**, 1223–1225.
- 80 Hawkins, R.D., Hon, G.C., and Ren, B. (2010) Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, **11**, 476–486.
- 81 Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680.
- 82 Rifai, N., Gillette, M.A., and Carr, S.A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnology*, **24**, 971–983.

## 13

# Two-Dimensional Gel Electrophoresis and Protein/Polypeptide Assignment

Takashi Manabe and Ya Jin

### 13.1

#### Introduction

The techniques of two-dimensional gel electrophoresis (2-DE) were developed in the late 1960s and are now widely employed for the global analyses of proteins/polypeptides in complex protein systems. During these last 40 years, the techniques were improved for better resolution, higher reproducibility, and wider application. The applicability of 2-DE techniques in protein analysis was reinforced by the advent of mass spectrometry (MS) for the assignment of protein/polypeptide spots on 2-DE gels. MS techniques became popular in the 1990s, and provided much higher sensitivity and throughput in protein/polypeptide assignment compared with chemical amino acid sequencing. Further, they enabled simultaneous assignment of multiple polypeptides in one gel spot, which implies that protein complexes on 2-DE gels would also be assigned.

In this chapter, the following three points will be covered:

- Aim of protein analysis and development of 2-DE techniques.
- Current status of 2-DE techniques.
- Development of protein assignment techniques on 2-DE gels and current status of mass spectrometric techniques.

The term “proteins” will be used for functional proteins that retain their tertiary/quaternary structures and the term “polypeptides” will be used for single polypeptide chains, irrespective of their secondary and tertiary structures.

### 13.2

#### Aim of Protein Analysis and Development of 2-DE Techniques

Since the total amino acid sequencing of insulin by Sanger *et al.* in the early 1950s [1, 2], the importance of the determination of amino acid sequence in protein analysis has been recognized. In the 1960s and 1970s, the research aims in biochemistry

laboratories were focused on the analysis of the structure (molecular mass, isoelectric point (pI), subunit structure, amino acid sequence, etc.) and function (enzyme activity, specific binding, etc.) of single proteins. For this purpose, large-scale purification of the target proteins, retaining their biological functions, was the prerequisite. The proteins in the starting materials (such as cells and organs) were subjected to crude fractionation by salt or cold organic solvent precipitation, separated by ion-exchange chromatography utilizing the differences in protein net charge and by gel-permeation chromatography utilizing the differences in protein size, and then crystallized to ensure further purification. Generally, native proteins were prepared in quantities of 100 µg to grams, especially when the full-length amino acid sequence and X-ray crystallographic analysis were the aim. The precipitation methods could separate protein samples into several fractions and each method of liquid chromatography could separate into a further 10–20 fractions; thus, the target protein could be purified about 500- to 1000-fold after several weeks of labor.

The techniques of one-dimensional polyacrylamide gel electrophoresis (1-DE), including discontinuous buffer-gel electrophoresis (disk gel electrophoresis) [3, 4], sodium dodecylsulfate (SDS) gel electrophoresis [5, 6], and gel isoelectric focusing (IEF) [7, 8], were also developed in the 1960s. The applicable protein quantity of these analytical techniques ranged from 1 to 100 µg and could not be used directly for large-scale protein preparation, but they could reproducibly separate proteins into 50–100 stained bands on polyacrylamide gels within one to several hours. The advantages of the 1-DE techniques – high resolution, high reproducibility, and short analysis time – were immediately recognized and applied for the analysis of various protein systems. In 1970, Laemmli [9] reported an improved technique of SDS gel electrophoresis introducing a discontinuous buffer system and used it for the analysis of polypeptides of bacteriophage T4. This landmark work demonstrated that all the polypeptides of T4 phage – some of which had been assigned their gene loci in the phage DNA – could be separated according to their size differences and visualized on a polyacrylamide gel, hence the processes to construct the molecular architecture of the phage could be studied using this technique. (The studies on the molecular morphology of T4 phage have been reviewed [10].) In the same year, Kaltschmidt and Wittmann [11] reported the analysis of all the component proteins in the *Escherichia coli* ribosome – 21 proteins in the 30S subunit and 34 proteins in the 50S subunit – aiming at the reconstruction of structure and function of the ribosome using a technique of 2-DE. These works were followed by the 2-DE analysis of *E. coli* polypeptides reported by O'Farrell [12]. *E. coli* polypeptides were separated into about 1100 spots under denaturing conditions and the results suggested the possibility of the analysis of total polypeptides present in a cell, opening the way to reconstruct the complex biological structures and functions of living organisms.

O'Farrell's work did not include information on the identity of the separated polypeptides, such as their gene loci, amino acid sequences, or biological properties, because at that time the information could be obtained only from purified proteins. The elaborate works by Neidhardt *et al.* on the preparation of an *E. coli* Protein Index undertaken from the late 1970s to 1983 [13] provided the assignment of 160 spots on the 2-DE gel, which also suggested the limitations of protein assignment techniques

available in this period (i.e., coelectrophoresis of purified proteins and immunochemical assignment). However, the development of DNA sequencing techniques in the 1970s dramatically accelerated the rate of sequencing and the size of the DNA sequence database grew from 10 kbp in 1977 to 1.5 Mbp in 1987, 1.1 Gbp in 1997, and 99 Gbp in 2008, and continues to grow at an exponential rate (<http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>). The MS techniques of protein assignment, which employ the information in DNA sequence/amino acid sequence/protein function databases, enabled assignment of almost all the stained spots on 2-DE gels. The 2-DE techniques can now provide not only high-resolution 2-D maps of proteins/polypeptides in complex protein systems, but also the structural and functional information of each spot on the map. The development of the techniques for protein/polypeptide assignment is summarized in Section 13.4.

About 4400 polypeptides are predicted from the 4.6-Mbp sequence of *E. coli* strain K12 (<http://genprotec.mbl.edu/overview.html>) and about 20 000–25 000 protein-coding genes are estimated from the human 3-billion-base pair sequence [14]. These numbers are larger than the polypeptide spots resolved by denaturing 2-DE, so the techniques have been improved to attain higher resolution, reproducibility, and sensitivity for the analysis of polypeptides as gene products. On the other hand, the limitations of the denaturing 2-DE specified for the analysis of polypeptides have become more and more obvious as studies on complex protein systems accumulate. Just as polypeptide amino acid sequences could not be predicted from DNA sequence data alone, the biological functions of proteins and protein complexes could not be predicted from amino acid sequence data alone. Therefore, it is important to develop methods to separate and analyze functional proteins and protein complexes, setting the aim of protein analysis at the reconstruction of the biological structures and functions in living organisms [15]. 2-DE techniques aiming at the analyses of functional proteins and protein complexes have also been developed. The current status of 2-DE techniques is summarized in the next section.

### 13.3

#### Current Status of 2-DE Techniques

As reviewed in the previous section, 2-DE techniques have been most commonly used for the separation of polypeptide chains in order to correlate genes to their translation products (i.e., polypeptides). However, the increasing interest in reconstructing the complex biological structures and functions led the development of 2-DE techniques that aim to separate proteins holding their native structures and biological functions. In this section, the 2-DE techniques are classified into the following three categories:

- Denaturing 2-DE for the separation of polypeptides.
- Nondenaturing 2-DE for the separation of biologically active proteins and protein complexes.
- Blue-native 2-DE for the detection of protein–protein interactions.

**Table 13.1** Charge number and  $pK_a$  values of dissociable groups in simple proteins.

Name of dissociable groups	Charge number when totally ionized	$pK_a$
Guadinyl (Arg)	+1	12.48
$\epsilon$ -Amino (Lys)	+1	10.53
$\alpha$ -Amino (N-terminal)	+1	9.6
Imidazole (His)	+1	6.0
$\alpha$ -Carboxyl (C-terminal)	-1	2.3
Carboxyl (Asp)	-1	3.86
Carboxyl (Glu)	-1	4.25
-SH (Cys)	-1	8.33
-OH (Tyr)	-1	10.07

Values are for free amino acids at 25 °C.

Each one will be summarized in terms of the separation principles, procedures, and characteristic features.

### 13.3.1

#### Denaturing 2-DE for the Separation of Polypeptides

##### 13.3.1.1 Principle

The resolution of proteins in 2-DE can be optimized when each dimension separates proteins/polypeptides according to independent parameters. The electric properties of proteins/polypeptides are decided by the dissociable groups in the proteins. The dissociable groups can be divided into two categories – those that provide one negative charge after full ionization and those that provide one positive charge, as shown in Table 13.1. Since the dissociable groups are separated from each other by the peptide bonds in the polypeptide chain backbone, it can be assumed that each group dissociates independently (i.e., the  $pK_a$  value of one group would not be affected by the dissociation state of the other groups in neighboring amino acid residues). Therefore, the following equation can be used to estimate the net charge of a polypeptide;

$$Z = - \sum (a_i K_{ai}) / ([H^+] + K_{ai}) + \sum (b_j [H^+]) / ([H^+] + K_{aj}) \quad (13.1)$$

where  $Z$  represents the net charge generated from the sum of the dissociable (acidic;  $-\text{COOH}$  and  $-\text{NH}_3^+$ ) groups in the polypeptide,  $[H^+]$  represents the concentration of  $H^+$ ,  $a_i$  and  $K_{ai}$  represent the groups  $i$  that provide a  $-1$  charge and its dissociation constant, respectively, and  $b_j$  and  $K_{aj}$  represent the groups  $j$  that provide a  $+1$  charge and its dissociation constant, respectively. Setting the net charge  $Z = 0$ , the  $pI$  ( $-\log [H^+]$  at  $Z = 0$ ) of a protein can be estimated from Eq. (13.1). Since polypeptides are defined by their specific amino acid sequence, the compositions of the dissociable groups are different between polypeptides, so they have different  $pI$  values. On the other hand, the molecular mass or the size of a polypeptide is decided by its chain

length or the number of amino acids, which is not related to the composition of the dissociable groups. For a polypeptide with known amino acid sequence, its molecular mass can be calculated by summing up all the residual masses of the component amino acids. In O'Farrell's technique, *E. coli* proteins were treated with 8 M urea, nonionic detergent NP-40, and 2-mercaptoethanol, which would ensure the solubilization of most proteins and their dissociation into single, denatured polypeptide chains [12]. The denaturants were also included in the first-dimension gels for IEF, so the polypeptides were separated according to their *pI* differences. After IEF, the IEF rod gels were equilibrated with an SDS-containing solution to form dodecylsulfate (DS) complexes of the polypeptides, each IEF gel was set on an SDS-containing slab gel [9] and then the DS–polypeptide complexes were separated according to their size differences. The size separation becomes possible because approximately 1.4 g DS<sup>−</sup> can bind per gram of polypeptide with disulfide bonds cleaved by reduction [16], which means all polypeptides have negative charges approximately proportional to their molecular mass and their native charge states become negligible. The high-resolution of this 2-DE technique is assured by strictly keeping the conditions to dissociate the proteins into single polypeptide chains in the first dimension and keeping the conditions to form DS–polypeptide complexes in the second dimension.

A major problem in the IEF step of this technique is the flattening of the pH gradient at the basic end of the IEF gels after prolonged IEF time. This pH gradient instability in the IEF gels is caused by the flow-out of the carrier ampholytes and results in gradients only in the range of pH 4–7, so the basic polypeptides that constitute ribosomes and histones cannot be separated. For the separation of basic polypeptides, a technique called nonequilibrium pH gradient electrophoresis (NEPHGE) was developed [17] in which samples are loaded at the acidic end of the gels and run for a relatively short time. However, this approach suffers from the following disadvantages: (i) two different gels are necessary for the analysis of a sample, (ii) the basic polypeptides are separated by their mobility and separation is not reached at their *pI* positions, and (iii) reproducibility of the separation is difficult to control. The immobilized pH gradient (IPG) technique was developed in 1982 [18] to overcome the problems in IEF (i.e., pH gradient instability and narrow pH range). Weak acids and bases that have the general chemical composition (Immobilines) CH<sub>2</sub>=CH–CO–NH–R, where R represents either one of several carboxyl groups or one of several tertiary amino groups, are copolymerized within the polyacrylamide network to prepare IPG gels. The pH gradient exists prior to electrophoresis since the Immobiline concentration gradients are prepared within the mixed solution of acrylamide and *N,N'*-methylenebisacrylamide (bis) monomers before the polymerization. After a series of improvements, the IPG methodology provides stabilized pH gradients, which lead to higher resolution, wider pH range (pH 3–11), and improved reproducibility for interlaboratory comparisons, and currently it is the standard IEF technique for the 2-DE analysis of polypeptides [19]. However, it is reported that some proteins, such as membrane proteins, tend to precipitate or aggregate and are lost in the IPG-IEF step. Agarose gel columns that include carrier ampholytes are recommended for IEF of such samples [20].

### 13.3.1.2 Procedures

The procedures for denaturing 2-DE using IPG gels in IEF have been described in detail [21, 22], so the important points within the procedures, characteristic to this 2-DE method, are briefly summarized.

- 1) **Sample preparation (for animal and prokaryotic cells).** During and after cell lysis, proteases are inactivated by the addition of protease inhibitors and insoluble components are removed by centrifugation. The ideal sample solubilization procedure for denaturing 2-DE would cleave all intra- and interpolypeptide disulfide bonds, and disrupt all noncovalent interactions between polypeptides and proteins. For these reasons, in a typical protocol [22] the sample is suspended in a lysis buffer that contains 9.5 M urea, 2% (w/v) CHAPS, 0.8% (v/v) Pharmalyte pH 3–10, 1% (w/v) dithiothreitol (DTT), and 5 mM Prefabloc protease inhibitor, so that the concentration of urea is higher than 8 M. The solution is subjected to sonication in an ice bath ( $3 \times 10$  s) for cell lysis and centrifuged (60 min, 42 000 g, 15 °C). The sample solutions are either used immediately or are stored at  $-78$  °C. High urea concentration disrupts the hydrogen bonds in proteins, CHAPS prevents the hydrophobic interactions between the denatured polypeptides that have exposed hydrophobic side-chains, and carrier ampholytes help the separation of polypeptides in IPG gel strips.
- 2) **Rehydration of IPG gel and application of the sample solution.** Dried IPG gel strips with polyester film backing are commercially available in various lengths (7, 11, 13, 18, and 24 cm) and for various pH ranges. The dried IPG strips are rehydrated in a solution that contains 8 M urea, 0.5% CHAPS, 0.2% DTT, and 0.2% Pharmalyte pH 3–10 overnight to form 0.5-mm thick IPG gel with polyacrylamide matrix of 4% T (T represents the sum of the weights of acrylamide and bis in 100 ml solution) and 3% C (C represents percent weight of bis in the sum of the weights of acrylamide and bis). For analytical purposes, typically 20  $\mu$ l of the sample solution (50–100  $\mu$ g of protein) are applied onto an 18-cm long IPG gel strip.
- 3) **IEF.** The IEF running conditions depend on the length and pH range of the IPG gel strip to be used. For analytical purposes using 18-cm long IPG 3–10, the voltage is raised stepwise, 150 V for 30 min, 300 V for 30 min, 1500 V for 1 h, and 3500 V for 4.6 h at 20 °C to keep the current maximum at 50  $\mu$ A per strip.
- 4) **Equilibration of IPG gel.** The IPG gel is put in a test tube that contains 10 ml equilibration buffer I, 6 M urea, 30% (w/v) glycerol, 2% (w/v) SDS, 1% DTT in 0.05 M Tris–HCl buffer, pH 8.8, and rocked for 15 min. The solution is poured off and replaced with 10 ml equilibration buffer II, 6 M urea, 30% (w/v) glycerol, 2% SDS, 0.0012% bromophenol blue (BPB), 4% iodoacetamide in 0.05 M Tris–HCl buffer, pH 8.8, and rocked for 15 min. Urea and glycerol help the transfer of polypeptides from the IPG gel, SDS is used to form DS–polypeptide complexes, and iodoacetamide to alkylate cysteine residues to block the reformation of disulfide bonds during and after the second dimension run.
- 5) **SDS gel electrophoresis.** SDS-containing slab gels can be cast on PAGfilm (GelBond<sup>®</sup>) and run horizontally or polymerized in vertically set cassettes consisting of two glass plates and two 1-mm thick spacers between them and



run vertically. Vertical gels are set in an apparatus which allows several to 20 slab gels to be run simultaneously. The gel composition of a typical vertical gel is 10, 12.5, or 15% T homogeneous, 2.6% C, 0.1% SDS, and 375 mM Tris-HCl, pH 8.8. The equilibrated IPG gel strip is placed on top of a SDS gel and overlaid with an agarose solution to achieve complete contact of the IPG strip on the surface of the SDS gel. Typically, SDS electrophoresis is run at 15 mA constant current per gel overnight at 15 °C, until the BPB tracking dye has migrated off the lower end of the gel. The procedures to visualize proteins and polypeptides on 2-DE gels are summarized in Section 13.3.4.

### 13.3.1.3 Specific Features

The most important feature of this technique is the ability to separate polypeptides with extremely high resolution. Owing to this high resolution, each spot on the 2-DE gel can be assumed to represent one polypeptide, so the 2-DE patterns would be better correlated to the activities of the genes. Therefore, changes in the protein composition during cell transformation, development, and differentiation have been studied by comparing the 2-DE patterns. However, because of the multiple manual steps in the 2-DE technique, the comparison is often time-consuming and the 2-DE patterns are difficult to perfectly superimpose. Two-dimensional difference gel electrophoresis (2-D DIGE) [23] is a technique to detect differences between two protein samples that requires only a single 2-DE gel. This is accomplished by tagging of the two samples with two different fluorescent dyes, running them on the same 2-DE gel, and the separated polypeptides are detected as two images utilizing the differences in the excitation and emission wavelengths of the two dyes. The images are then superimposed to detect the differences in fluorescence intensity for the overlapped spots. This technique enabled researchers to focus on the polypeptides with notable changes in quantity, which could be more directly related to the biological processes. Through the advances in MS-based polypeptide assignment techniques, which will be reviewed in Section 13.4, the importance of the 2-D DIGE technique has become more obvious.

## 13.3.2

### Nondenaturing 2-DE for the Separation of Biologically Active Proteins and Protein Complexes

#### 13.3.2.1 Principle

The technique of nondenaturing 2-DE – the combination of polyacrylamide gel IEF and disk gel electrophoresis – was attempted in 1969 [24] aiming at high resolution of proteins with their biological structures and functions maintained. Later, disk electrophoresis was replaced by pore gradient electrophoresis [25], because apparently a protein reaches its pore limit in a gradient gel and its band becomes sharp, resulting in higher resolution of proteins than in a uniform pore gel [26]. Human plasma proteins were separated into about 230 spots by a nondenaturing 2-DE technique in which polyacrylamide gel IEF was followed by polyacrylamide pore gradient gel electrophoresis [27]. In principle, nondenaturing 2-DE of proteins would

not have higher resolution than denaturing 2-DE, since independence of separation principles in the two dimensions cannot be strictly pursued. In nondenaturing IEF, the proteins would migrate in the gel holding their tertiary structures, which means the electric forces separating proteins are in equilibrium with the specific and/or nonspecific interactions between proteins. IPG gels are not successfully used for nondenaturing IEF, because they need high field strengths and long focusing times, which would enhance nonspecific interactions between proteins since proteins become more hydrophobic when they get closer to their  $pI$ . Therefore, column gels that contain carrier ampholytes are used for nondenaturing IEF. In the gradient gel electrophoresis step, proteins would migrate to their pore limits only when they have enough negative charges to drive them towards the anodic end of the gel. As shown in Eq. (13.1), proteins can have negative net charges when their  $pI$  values are lower than the buffer pH. Therefore, this method is only applicable for proteins that have  $pI$  values lower than the pH of the buffer solution employed (around pH 8). Apparent molecular masses of proteins can be estimated for acidic proteins ( $pI < 6$ ) when a pH 8.3 buffer is used [27], but the mass values of proteins that have  $pI$  values larger than 6 must be corrected as the proteins are more retarded when the  $pI$  values are closer to the buffer pH. In spite of these limitations, nondenaturing 2-DE is of importance in studies of proteins where their native structures and biological functions are to be maintained. Also, the fact that native proteins mostly have  $pI$  values in the range of 4–8 suggests the wide applicability of this technique. Since it is favorable to separate proteins at low field strengths and in short running times under nondenaturing conditions, a micro-gel (38 mm  $\times$  38 mm  $\times$  1 mm) 2-DE system was developed, which also enabled the parallel running of 8–16 gels [28]. It was shown that agarose IEF gels have much higher performance than polyacrylamide IEF gels in nondenaturing micro-gel 2-DE for the analysis of high-molecular-mass proteins (up to 2000 kDa) [29].

#### 13.3.2.2 Procedures

The procedures of nondenaturing 2-DE using agarose IEF gels have been described in detail [29, 30], so the important points characteristic to this 2-DE method are only briefly summarized.

- 1) **Sample preparation.** Since the method does not include the process of protein solubilization, the samples should be in the solution state, such as blood plasma or cellular cytosol fractions. A human plasma sample is prepared by inhibiting the initial stages of blood coagulation by the addition of 0.0025% (w/v) heparin or 0.1% (w/v) EDTA to the blood and centrifuging at 2000  $g$  for 10 min at 4 °C. A cytosol protein sample is prepared at 4 °C by suspending the cells (about  $2 \times 10^7$  cells/0.1 ml) in a buffer that contains 1.0 mM phenylmethylsulfonyl fluoride, sonicating the solution (6  $\times$  10 s), and centrifuging (5 min, 17 600  $g$ , 4 °C). Since the sample solutions are injected on top of the IEF column gels, they are supplemented with 20–40% (w/v) sucrose or glycerol to stabilize the sample layers.
- 2) **Preparation of IEF gels.** Agarose rod gels (internal diameter 1.4 mm, length 35 mm) that contain 1% (w/v) agarose, 2% (w/v) Ampholine™, pH 3.5–10, and 0.5% (w/v) Ampholine, pH 3.5–5 are prepared in a batch of 24 gels [28, 29].

- 3) **Preparation of pore gradient gels.** Polyacrylamide pore gradient (4.2–17.85% T linear gradient, 5% C for plasma samples and 8.4–17.85% T linear gradient, 5% C for cytosol samples) micro slab gels (38 mm × 38 mm × 1 mm), which contain 375 mM Tris–HCl buffer, pH 8.8, are prepared with the aid of a computer-controlled gradient maker in a batch of 16 gels.
- 4) **IEF.** The catholyte is 0.04 M NaOH and the anolyte is 0.01 M phosphoric acid, both precooled in ice water. Typically, 2  $\mu$ l of the plasma sample (about 120  $\mu$ g protein) or 4  $\mu$ l of the cytosol sample (about 100  $\mu$ g protein) is applied at the cathode end of an IEF gel, when the gel is to be stained with Coomassie Brilliant blue (CBB) and all the stained spots are to be subjected to protein assignment by mass spectrometric methods. IEF is run at 0.12 mA/gel constant current until a voltage of 300 V is reached (about 30 min for plasma samples and about 12 min for cytosol samples) and continued at 300 V for 15 min, keeping the IEF gel capillaries immersed in the anolyte solution cooled at 4 °C.
- 5) **Equilibration of IEF gels.** Each agarose IEF gel is set on top of a micro slab gel where a 100- $\mu$ l aliquot of a 10 mM Tris–76 mM glycine buffer, pH 8.3, is filled beforehand and the IEF gel is set for 10 min.
- 6) **Gradient gel electrophoresis.** The gels are set in an apparatus that allows 8–16 slab gels to run simultaneously. Typically, electrophoresis is run at 10 mA constant current per gel for 50 min in the case of plasma samples (4.2–17.85% T linear gradient, 5% C gels) and for 75 min in the case of cytosol samples (8.4–17.85% T linear gradient, 5% C gels) with an electrode buffer of 50 mM Tris–380 mM glycine, pH 8.3, which has been cooled at 4 °C. The procedures to visualize proteins and polypeptides on 2-DE gels are summarized in Section 13.3.4.

### 13.3.2.3 Specific Features

The most important feature of this technique is the ability to detect biological functions of proteins including enzyme activities [31, 32] and physiological binding between proteins [30, 32] after the second dimension separation. Most of the compiled recipes of activity staining of enzymes after starch gel electrophoresis [33] can be applied for the detection of enzymes on nondenaturing 2-DE gels. The dissociation of noncovalently bound protein complexes can be visualized by comparing the nondenaturing 2-DE pattern with a modified 2-DE pattern, which is obtained by nondenaturing IEF followed by SDS gel electrophoresis [34, 35]. Also, when some spots on a 2-DE gel have been assigned to contain multiple polypeptides by mass spectrometric analyses, suggesting that the spots may represent protein complexes or multiple proteins, they can be subjected to third-dimension SDS gel electrophoresis for the analysis of constituent polypeptides [36]. Another feature of nondenaturing 2-DE is that it can be done in a micro gel format [28], which enables the separation of a plasma sample to about 160 protein spots [37] and of an *E. coli* lysate to about 330 protein spots [38], in about 100 min run time or less. When chemical amino acid sequencing was the only way to obtain sequence information of proteins on 2-DE gels, nondenaturing 2-DE was disadvantageous because even when a spot on the gel represents only one protein, it may contain two or more heterogeneous polypeptide subunits and could not be sequenced. However, the advent of MS-based methods of polypeptide assignment,

which can assign multiple polypeptides in one spot, enforced the importance of this 2-DE technique.

### 13.3.3

#### **Blue-Native 2-DE for the Detection of Protein–Protein Interactions**

##### **13.3.3.1 Principle**

Blue-native (BN)-2-DE was developed for the analysis of mitochondrial membrane proteins by Schägger and Jagow [39] and was later extended to the analysis of soluble protein complexes [40]. Functional mitochondrial complexes are usually separated at pH 7–8 at which they are stable using nonionic detergents, but they could not be separated by native IEF because of severe aggregation. Gradient gel electrophoresis is then employed for the separation, keeping the pH of the gel buffer and electrode buffers at 7.0. The solubilized membrane samples are supplemented with CBB-G. The CBB molecules may bind mainly on the surface of the protein complexes, since the solubility of the complexes is improved and the functional activity is retained after electrophoresis for some complexes. The bound CBB molecules give total negative charges to the protein complexes and the CBB–protein complexes are separated according to the size differences in the pore gradient gels. A low concentration of CBB is added in the cathode buffer to stabilize the CBB–protein complexes during the run. A lane of the gradient gel is excised, treated with an SDS/mercaptoethanol solution, put on an SDS slab gel, and SDS gel electrophoresis is then run to obtain 2-DE patterns. On the stained 2-DE gels, each protein complex is dissociated into the constituent polypeptides and separated according to their molecular mass differences. When CBB-G dye is not used in the first dimension of gradient gel electrophoresis and it is combined with SDS gel electrophoresis (clear-native (CN)-2-DE), the spots are more overlapped than in BN-2-DE [40].

##### **13.3.3.2 Procedures**

The procedures of BN-2-DE have been described in detail [39, 40], so the important points characteristic to this 2-DE method are only briefly summarized.

- 1) **Sample preparation.** Since the method is basically a one-dimensional electrophoresis technique for the separation of proteins, the optimum solubilization conditions for the target protein complexes should be examined beforehand. The sediments of bovine heart mitochondria (200  $\mu\text{g}$  total protein) are solubilized by the addition of 40  $\mu\text{l}$  of 750 mM 6-aminocaproic acid, 50 mM Bistris, pH 7.0, and 5  $\mu\text{l}$  10% (w/v) dodecyl maltoside (a detergent), and centrifuged for 15 min at 100 000 g. Shortly before BN electrophoresis, CBB-G is added from a 5% (w/v) stock solution in 500 mM aminocaproic acid to adjust to a detergent/CBB ratio of 4: 1 (g/g). When the sample is partially purified membrane proteins and the detergent concentration is less than 0.2%, it can contain up to 200 mM NaCl and the addition of CBB is not necessary.
- 2) **Preparation of first-dimension gradient gel.** Polyacrylamide pore gradient gels (6–13% T linear gradient, 3% C or 7–16.5% T linear gradient, 3% C, 14 cm high,

16 cm wide, and 1.6 mm thick, which contain 500 mM aminocaproic acid/150 mM Bistris adjusted with HCl to pH 7.0) are prepared, sample combs are set on each gel, a sample gel solution (4% T, 3% C) is poured in the space between the separation gel top and the comb, and the sample gel is polymerized.

- 3) **BN electrophoresis.** BN electrophoresis is run with an electrode buffer of 50 mM Tricine/15 mM Bistris, pH 7.0, except that cathode buffer is supplemented with 0.02% CBB-G, at 4–7 °C at 100 V until the protein sample reaches the separation gel and then at 500 V with a current limit of 15 mA for 3–4 h. In the case of 7–16.5% T gradient gel, electrophoresis is run at 200 V overnight.
- 4) **SDS electrophoresis for the second dimension.** A 5-mm wide lane of the first-dimension gel is excised and equilibrated in 1% w/v SDS/50 mM Tris–HCl, pH 7.0, for 5 min, dipped in the equilibration buffer supplemented with 100 mM DTT for 15 min, transferred in the equilibration buffer supplemented with 55 mM iodoacetamide for 15 min, and thoroughly washed in the equilibration buffer [41]. The lane is placed on a glass plate at the usual position of stacking gels, the spacers and the second glass plate are set, the gel mold is brought into a vertical position, and a separation gel solution of SDS electrophoresis (uniform gel of 10–16% T, depends on the molecular mass distribution of the component polypeptides) is poured, leaving space for the stacking gel. After polymerization, the stacking gel is polymerized. Electrophoresis is run at room temperature at 25 °C with a current limit of 40 mA for 6 h.

#### 13.3.3.3 Specific Features

The most important feature of this technique is that it can analyze the compositions of multiple protein complexes on one 2-DE gel. The separation of proteins is done only by the one-dimensional separation in a native gradient gel and the second dimension serves for the separation of the polypeptides that constitute the protein complexes. Since gradient gel electrophoresis can separate no more than 100 protein species, the samples for BN-2-DE proteins should be prefractionated to concentrate the target protein complexes. For the same reason, the sample solubilization conditions should be optimized and sometimes BN electrophoresis may be better replaced by CN electrophoresis when the addition of CBB-G is not favorable for the sample. This technique was developed for the analysis of hydrophobic protein complexes that would easily aggregate in nondenaturing IEF, so it is complementary with nondenaturing 2-DE. Recent applications using BN or CN electrophoresis have been reviewed [42].

#### 13.3.4

##### Visualization of Proteins Separated on 2-DE Gels

As shown in Sections 13.3.1–13.3.3, the 2-DE gels obtained just after the second-dimension run can have different gel sizes and different constituents in the gels. Also, the gels can be subjected to various detection methods such as dye staining, silver staining, negative staining, and fluorescent staining. Since the methods of protein detection methods have been reviewed [43, 44], only some important points in protein visualization are briefly discussed.

#### 13.3.4.1 Fixing Before CBB, Silver, or Fluorescent Dye Staining

The polyacrylamide slab gels used for the second-dimension run would contain SDS, carrier ampholytes migrated from the IEF gels to the slab gels, and reagents used to stabilize the pH in the gels. SDS molecules attached to the polypeptide backbone and hydrophobic amino acid residues would interfere with the hydrophobic binding of organic dyes such as CBB-R and CBB-G. Carrier ampholytes in the gel would bind with the dyes, forming high-molecular-mass complexes with the dyes. These reagents should be removed from the gel with alcohols (methanol or ethanol) in rather high concentrations (30–50% v/v). Also, alcohols help the fixation of proteins with the aid of acids (5–10% v/v acetic acid). Therefore, most staining protocols – CBB, silver, or fluorescent dye staining (except zinc-imidazole staining) – include the step of protein fixation with alcohol/acid for 30 min to 1 h.

#### 13.3.4.2 CBB Staining

CBB-R and -G are the most commonly used organic dyes to stain 2-DE gels, and there are two different protocols for CBB staining: the “stain–destain” method and the “colloidal stain” method. In the former, after fixation the gels are stained in the fixation solution containing 0.1–0.2% (w/v) CBB-R for 15–30 min (gel thickness 1 mm), destained in the fixation solution with a reduced concentration (lower than a half) of alcohol for 2 h to overnight leaving a faint blue color of the gel background, and the gel is kept in the acid solution without alcohol. A high concentration of alcohol in the staining solution maintains the solubility of CBB, the lower concentration of alcohol in the destaining solution helps CBB molecules remain bound with proteins, and CBB molecules stay with the proteins in the absence of alcohol. In the latter protocol [43], the gels are fixed in an alcohol/phosphoric acid solution, washed with 2% phosphoric acid to remove alcohol, and then equilibrated in a solution containing 2% phosphoric acid, 18% (v/v) ethanol, and 15% (w/v) ammonium sulfate. The solution is then treated with 0.2% (w/v) solution of CBB-G, 1% of the gel-solution volume, and the stain is allowed to proceed for 24–72 h. Free CBB-G molecules are in a very low concentration, keeping a minimal background staining, but they diffuse into the gel and gradually accumulate on the protein spots. The detection limit of CBB staining is in the range of 50–100 ng/mm<sup>2</sup>, which means the limit is dependent on the performance (minimal spot size) of the 2-DE technique. When a protein can be separated on a micro 2-DE gel (4 cm × 4 cm after destaining) as a spot of 0.1 mm<sup>2</sup> (e.g., 0.5 mm × 0.2 mm), 10 ng is the limit [45]; when the minimal spot size is 1 mm<sup>2</sup> on a 20 cm × 20 cm gel, the detection limit is 100 ng. CBB staining is compatible with MS-based protein assignment because the bound CBB can be removed by treating the gel piece with an organic solvent.

#### 13.3.4.3 Silver Staining

Silver staining is about 20- to 100-fold more sensitive than CBB staining and there are various protocols [43]. However, the use of formaldehyde or glutaraldehyde in the prestaining step or in the silver impregnation step may cause covalent bonding of these reagents to proteins, thus hampering the MS-based assignment. A modified silver staining protocol compatible with MS has been reported, which

uses formaldehyde only at the developing step [46] and removes silver from the gel before MS [47].

#### 13.3.4.4 Reverse Staining with Zinc-Imidazole

Zinc-imidazole staining [48] is based on the selective precipitation of white imidazole–zinc complex in the gel, leaving the protein spots transparent. A 2-DE gel after SDS electrophoresis is incubated in 200 mM imidazole containing 0.1% (w/v) SDS for 15 min, the solution is then discarded, the gel is incubated in 200 mM zinc phosphate until the gel background becomes deep white, leaving the protein spots transparent (about 30 s), and staining is stopped by rinsing the gel with distilled water [49]. The sensitivity of this method is reported to be higher than CBB staining. There are, however, some drawbacks – the sensitivity for low-molecular-mass proteins and glycoproteins is relatively low, and the recognition of the transparent spots is rather difficult. Zinc-imidazole staining is highly valuable in 2-DE protein analysis because of its simplicity, speed, and MS compatibility. Higher recovery of proteins or in-gel digested peptides is expected than with other staining methods, since this method is free from the fixing processes in which polypeptides would become entangled with the gel matrices.

#### 13.3.4.5 Fluorescent Dye Staining

Proteins can be covalently derivatized with fluorophores prior to IEF as in the case of 2-D DIGE [23], but the prelabeling may cause changes to protein charge and/or protein size, which will result in heterogeneity of spot locations between the modified and nonmodified proteins. SYPRO<sup>®</sup> Ruby is a postelectrophoresis staining dye comprised of ruthenium as part of an organic complex, which allows sensitive fluorescence detection of proteins in SDS–polyacrylamide gels [50]. A 2-DE gel after SDS electrophoresis is incubated in the fixation solution for 1 h, stained with the SYPRO Ruby protein gel stain for a minimum 3 h, and then washed in 10% (v/v) methanol/7% (v/v) acetic acid solution for 10 min before scanning. The stain is MS-compatible, and the detection limit is as little as 1–2 ng/mm<sup>2</sup> and has a linear dynamic range orders of magnitude wider than that of a silver stain [51]. However, the high cost of SYPRO Ruby might limit its use in routine analysis. A scanner equipped with a light source at the specific excitation wavelength and a filter at the emission wavelength is necessary to detect the spots stained with a fluorescent dye.

#### 13.3.4.6 Quantitation

The density of the stained spots on the 2-DE gels can be quantitated with an image analyzer, which normally has the following functions; (i) spot detection, (ii) spot numbering, (iii) calculation of integrated density for each spot, (iv) annotation of each spot after calibrating with pI and molecular mass standards, (v) comparison of multiple 2-DE patterns by overlaying the spots, and so on. It must be noted that any staining technique has intrinsic limitations because the techniques are based on the selective binding of the chromophores to hydrophobic sites (CBB staining and fluorescent staining) or basic amino acid residues (silver staining), or on the affinity to SDS–polypeptide complexes (negative staining). Since different polypeptides have

different amino acid compositions, the binding of the chromophores or SDS molecules to them are not uniform. Therefore, comparisons of quantity or integrated density should be done between the spots that have been assigned to represent an identical protein/polypeptide.

### 13.4

#### **Development of Protein Assignment Techniques on 2-DE Gels and Current Status of Mass Spectrometric Techniques**

Using 2-DE techniques, hundreds or thousands of proteins/polypeptides in a complex protein system can be visualized, and information on their approximate quantity and physicochemical properties (such as *pI*s and molecular masses) can be obtained simultaneously. Also, the separated protein/polypeptides reside in very small volumes ( $1.6 \text{ cm}^3$  in micro gels and about  $40 \text{ cm}^3$  in standard gels) and can be stored for months or years. This is in contrast to liquid chromatographic techniques, in which separated proteins cannot be visualized and the protein fractions must be kept frozen in vials that will occupy much larger volumes. However, the features of 2-DE techniques become quite important only when the proteins are assigned on the 2-DE gels. With the assignment of proteins/polypeptides, the comprehensive analysis of all the spots on the 2-DE pattern would provide information necessary to reconstruct the relevant protein systems. Comparisons of 2-DE patterns before and after a biological event in a complex protein system would clarify the polypeptides/proteins with changes in quantity, *pI*, or molecular mass, which would be directly related to the biological event. Therefore, various techniques have been applied for the assignment of proteins on 2-DE gels.

#### 13.4.1

##### **Development of Protein Assignment Techniques**

Coelectrophoresis was the method employed in 1970s, but the purification of low-abundant proteins is time-consuming and enormous labor must be used for total assignment of proteins on 2-DE gels [13]. Electrophoretic transfer of proteins from polyacrylamide gels to hydrophobic membranes was devised in 1979 [52] and this method was immediately employed for the immunochemical assignment of proteins on polyacrylamide 2-DE gels. Immunoglobulin molecules are too large to diffuse in the polyacrylamide gels, but during the electrophoretic transfer the proteins come out from the small pores of a gel and bind with the surface of the large pores of a nitrocellulose or polyvinylidene difluoride membrane. After electrophoretic blotting, the membrane is incubated in a protein solution (such as bovine serum albumin) to block all the hydrophobic binding sites on the membrane, and then an antibody against a target protein is added to the solution and incubated. Now the antibody molecules can freely penetrate in the large pores of the membrane and bind with the target protein (antigen). The excess antibody is washed out, the membrane is incubated in a solution of the second antibody directed against the first antibody,



and the second antibody specifically binds with the first antibody–antigen complex. The second antibody is labeled with peroxidase, alkaline phosphatase, fluorescein, or  $^{125}\text{I}$ , and therefore the location of the specific binding site of the second antibody can be detected. The detection limit is as little as 10–100 pg in the case of enzymatic detections and the technique is widely applied for the assignment of proteins, such as human plasma proteins [53, 54]. However, purified proteins are necessary for the preparation of the first antibodies, so this method is not suited for the total analysis of samples in which most of the proteins have not been characterized. Since proteins/polypeptides are distinguished from each other only when their amino acid sequences are different, the techniques to assign uncharacterized proteins on 2-DE gels should be able to detect the structural differences.

The total amino acid sequencing of insulin by Sanger *et al.* [1, 2] prompted studies on protein purification followed by total amino acid sequencing. However, the method of Sanger for sequencing was not widely employed, because the peptide labeled at the N-terminal amino group with 1-fluoro-2,4-dinitrobenzene could not be used to determine the next amino acid. Edman [55] devised a method in which the labeled N-terminal amino acid can be cleaved and the shortened peptide is left intact. Phenyl isothiocyanate reacts with the uncharged terminal amino group of a peptide to form a phenylthiocarbamoyl derivative. Then, under mildly acidic conditions, a cyclic derivative of the terminal amino acid, a phenylthiohydantoin (PTH)-amino acid, is cleaved which can be identified by gas chromatography or liquid chromatography. The procedure can then be repeated on the shortened peptide, yielding another PTH-amino acid, which can again be identified by chromatography. Edman and Begg [56] automated the processes of Edman degradation in 1967. The necessary sample amounts for the automated sequencer were about 10–100 nmol (assuming the mass of polypeptide to be 50 000, corresponding to 0.5–5 mg) and proteins purified on a large scale by chromatographic procedures could be sequenced. The improvement of the sequencer by the use of gas-phase reagents that react with a polypeptide immobilized on a solid support [57] provides high sensitivity and polypeptides at as little as 1–5 pmol (50–250 ng) can be sequenced. N-Terminal amino acid sequencing of the polypeptide spots separated on the denaturing 2-DE gels using the gas-phase sequencer followed by protein assignment from the partial sequence became popular in 1980s and the preparation of 2-D polypeptide databases were reported in the early 1990s [58, 59]. The novelty of this protein assignment technique resided in the utilization of amino acid sequence databases to compare the obtained partial sequences with the accumulated full-length protein sequences and also on the probability-based judgment of the validity of assignment. The increasing size of protein databases, which would be further enriched by DNA sequencing, seemed to ensure the possibility of this assignment technique. However, although the chemistry of Edman degradation is well understood and the instruments are well developed, this method turned out to have the following major limitations:

- i) The polypeptide should be highly homogeneous. If a spot on a 2-DE gel is a mixture of two polypeptides, the PTH-amino acid chromatograms would always

- provide two peaks (or more, because the yield of each degradation step is not 100%) and the interpretation may become uncertain.
- ii) Polypeptides are often blocked at their N-termini by post-translational modifications (acetylation, formylation, etc.) or pyroglutamyl formation. In order to overcome this problem, internal peptides were produced by “on-membrane” or “in-gel” cleavage [60] of a polypeptide with specific proteases, separated by high-performance liquid chromatography and each purified peptide was sequenced. This approach enables the sequencing of multiple internal peptides of a given polypeptide and improves the confidence level in protein assignment, but is less sensitive and slow because of the additional steps.
  - iii) One cycle of Edman degradation needs about 1 h, so the method is too slow for the assignment of hundreds of spots on 2-DE gels, which can be produced at a rate of 10–20 gels/week in one laboratory. MS-based protein assignment techniques succeeded basic techniques developed for partial Edman sequencing and polypeptide assignment, such as in-gel digestion and probability-based assignment, but they can analyze polypeptide mixtures, have higher sensitivity, and are high-throughput. Therefore, MS is currently almost the only method employed for protein/polypeptide assignment on 2-DE gels.

#### 13.4.2

#### **MS-Based Assignment Techniques Utilizing Amino Acid Sequence Databases**

MS has been used as a highly sensitive and effective technique in structural analysis of organic compounds. An organic compound is converted to gas-phase ions (the ionization methods produce fragment ions of the original compound) in an ion source, and the ions are separated in a mass analyzer according to their  $m/z$  ratio and detected by an electron multiplier. However, the ionization principles for organic compounds are based on the direct transfer of the kinetic energy of electrons or particles to the target molecules and were not effectively applied for proteins and polypeptides. Matrix-assisted laser desorption/ionization (MALDI) [61, 62] and electrospray ionization (ESI) [63] are “soft” ionization methods developed in the late 1980s, which enabled production of gas-phase ions of intact proteins/polypeptides in high efficiency. MALDI is most conveniently combined with a time-of-flight (TOF) MS apparatus and the short pulses of nitrogen laser (half-width 3 ns) enabled accurate measurement of  $m/z$  values of the ions that travel in a flight tube in tens of microseconds. After high-resolution MALDI-MS apparatus became commercially available, the assignment of proteins on 2-DE gels by peptide-mass fingerprinting (PMF) was reported [64, 65]. ESI was combined with a tandem MS (MS/MS) apparatus that enabled sequence information of peptide samples [66]. Later, MALDI was also combined with MS/MS [67]. Up to now, MALDI-TOF-MS followed by PMF and ESI- or MALDI-MS/MS followed by peptide sequence search are the two major methods for the assignment of proteins on 2-DE gels. Therefore, the established principles and procedures of the two methods are summarized in this section. Since these methods commonly need processes to extract protease-digested peptides of the proteins/polypeptides in a gel spot, this section will be divided into three parts:

- Sample preparation for MS analysis.
- MALDI-TOF-MS and PMF.
- MS/MS and peptide sequence search.

#### 13.4.2.1 Sample Preparation for MS Analysis

In 2-DE analyses the proteins/polypeptides are separated in the second dimension according to the differences of mobility in polyacrylamide gels, which means the pore sizes of the polyacrylamide gels are comparable with the sizes of proteins/polypeptides and their extraction from gel matrices is rather difficult. However, the polypeptide assignment techniques that utilize sequence databases and probability-based judgment for the assignment do not need the intact polypeptide chains to be extracted – only their fragments are used for the analysis. Thus, proteins/polypeptides in a 2-DE gel piece are in-gel digested with a protease and then extracted. The reduction of disulfide bonds and the alkylation of cysteine residues before the in-gel digestion improves the recovery of the peptides. The presence of buffers and salts in the sample solution would suppress the ionization of peptides, so micro-columns (such as ZipTip<sup>®</sup> from Millipore [68]) packed with reverse-phase material (such as octadecyl silica (ODS)) are used to purify the peptides by removing the contaminants. A protocol for sample preparation for MS analysis from a CBB-stained gel spot (modified from [69, 70]) is shown below.

- 1) Excise a stained spot from a 2-DE gel (0.5–1.0 mm thick and volume less than about 4 mm<sup>3</sup>) with a hobby knife and transfer the gel piece to a 0.5-ml centrifuge tube (Tube 1) that contains a 50- $\mu$ l aliquot of 100 mM NH<sub>4</sub>HCO<sub>3</sub> solution. Leave Tube 1 for 10 min.
- 2) Add 50  $\mu$ l acetonitrile to Tube 1, leave it for 10 min, and discard the liquid (CBB color must be totally removed).
- 3) Add 20  $\mu$ l acetonitrile to Tube 1, leave it for 10 min, discard the liquid, and completely dry the gel with a vacuum centrifuge (about 10 min; the gel is shrunk to white and the gel matrices may be partly broken).
- 4) Add 20  $\mu$ l of 10 mM DTT, freshly prepared in 100 mM NH<sub>4</sub>HCO<sub>3</sub>, to Tube 1 and keep it at 56 °C for 1 h. Cool Tube 1 to room temperature and discard the liquid.
- 5) Add 20  $\mu$ l of 50 mM iodoacetamide in 100 mM NH<sub>4</sub>HCO<sub>3</sub> to Tube 1 (carbamidomethylation), keep it for 45 min at room temperature in the dark with occasional vortexing, and discard the liquid.
- 6) Wash the gel with 50  $\mu$ l of 100 mM NH<sub>4</sub>HCO<sub>3</sub> for 10 min and discard the liquid.
- 7) Dehydrate the gel by adding 20  $\mu$ l acetonitrile for 10 min, discard the liquid, and completely dry the gel with a vacuum centrifuge.
- 8) Rehydrate the gel with 10  $\mu$ l of 10 ng/ $\mu$ l porcine trypsin (Promega, sequence grade) solution in 50 mM NH<sub>4</sub>HCO<sub>3</sub>, tightly cap Tube 1, and incubate it at 37 °C for around 16 h (overnight).
- 9) Centrifuge Tube 1 for 10 s to collect all the liquid to the bottom and transfer the liquid to a 0.5-ml centrifuge tube (Tube 2) (recovery of hydrophilic peptides).

- 10) Add 10  $\mu\text{l}$  of 0.1% (v/v) trifluoroacetic acid (TFA)/50% (v/v) acetonitrile to Tube 2, sonicate for 20 min, and transfer the liquid to Tube 2 (recovery of hydrophobic peptides).
- 11) Apply Tube 2 with its lid opened to the vacuum centrifuge and reduce the volume of solution to around 2  $\mu\text{l}$  (reduce the amount of acetonitrile).
- 12) Add 20  $\mu\text{l}$  of 0.2% TFA to Tube 2 to adjust the pH to around 2–3, vortex the tube for about 30 s, and then centrifuge for 10 s (reduce the concentration of acetonitrile below 5% before applying the peptides to ZipTip).
- 13) Prepare ZipTip $_{\mu\text{-C18}}$  according to the manufacturer's instruction and aspirate the sample solution prepared in Step 12. Repeat aspirating and dispensing of the solution 5 times and wash the tip with 0.1% TFA 3 times.
- 14) Elute the peptides concentrated on ZipTip $_{\mu\text{-C18}}$  with the solution suited for the next analysis.

In the case of negative-stained 2-DE gels, excise a gel piece, put in Tube 1, and start the procedure from Step 3. Silver-stained gels must be treated to remove silver [47] then start from Step 3.

#### 13.4.2.2 MALDI-TOF-MS and PMF

MALDI-TOF-MS is robust and easy to perform, but two techniques were developed to achieve the high resolution (low ppm range) necessary for the polypeptide assignment by PMF: delayed extraction and electrostatic ion mirror or reflectron. Delayed extraction is achieved by placing a grid in front of the MALDI target plate, and applying the same high potential to both the target and the grid so that the initial desorption occurs in the absence of an electric field. Relatively light neutral molecules and ions desorbed by the laser pulse diffuse away rapidly, and after a delay period of around 100 ns, the grid voltage is reduced and the heavy and slow ions are simultaneously accelerated into the flight tube, avoiding collisions with the light molecules. A reflectron set close to the end of a flight tube provides a linear potential gradient that reverses the flight direction of ions and accelerates them back toward a reflectron detector. Ions of a single mass species receive a range of energies in the process of laser desorption, which causes a distribution of speeds in the flight tube. The fastest ions travel further in the potential gradient of reflectron than the slowest ions then reverse the direction to the reflectron detector, so it is possible to adjust the apparatus settings to have all ions of each unique mass arriving at the detector simultaneously. The feature of MALDI-MS that singly charged peptide molecules are predominantly detected also facilitates the PMF assignment.

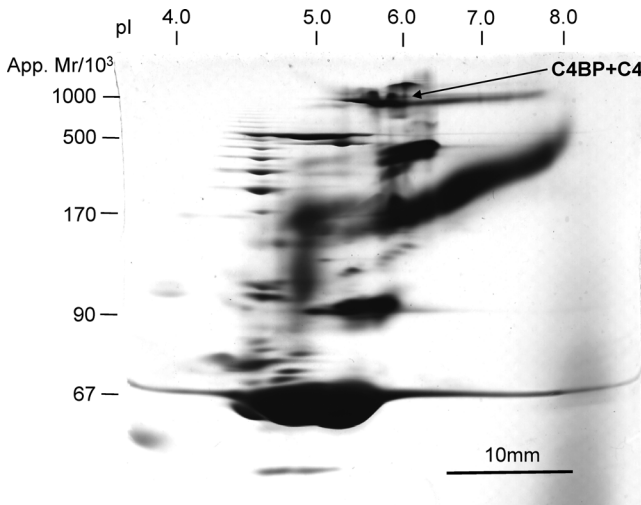
When the high-resolution apparatus became commercially available, the assignment of proteins on 2-DE gel by PMF was reported from several groups in 1993. The principle of PMF is simple; (i) the masses of peptides in-gel digested with a protease (trypsin is most often used) are measured with a high-resolution MALDI-TOF-MS apparatus at 10 ppm accuracy, (ii) one of the measured masses is compared with the peptide masses that can be calculated by virtually digesting the whole set of polypeptides in a sequence database using the specificity of the enzyme (in the case of trypsin, the carboxyl side of Arg and Lys except for -Arg-Pro- and -Lys-Pro-) and

finding the number of virtual peptides matched within a preset mass tolerance (e.g., 50 ppm), (iii) comparison for other measured peptide masses is repeated, and (iv) the polypeptide entry in the database that best matches the measured data is assigned. If the sequence database does not contain the polypeptide, the one that exhibits the closest homology is assigned. An algorithm for PMF named Mascot [71] that incorporates probability-based scoring and can be freely accessed via internet ([www.matrixscience.com](http://www.matrixscience.com)) is presently the most widely used. A protocol for MALDI-TOF-MS and PMF using Mascot is shown below.

- 1) Elute the peptides concentrated on a ZipTip<sub>μ-C18</sub> with 1 μl of 0.1% TFA/50% acetonitrile (aspirate and dispense 5 times) in a 0.5-ml centrifuge tube (Tube 3).
- 2) Add 0.3 μl of an adrenocorticotrophic hormone (ACTH) solution (3 pmol/μl 0.1% TFA, fragment 18–39, human) to Tube 3 and centrifuge for 10 s.
- 3) Add 1.0 μl of a matrix solution ( $\alpha$ -cyano-4-hydroxycinnamic acid saturated in 0.1% TFA/50% acetonitrile, kept in the dark and used within 1 week) to Tube 3 and immediately aspirate the solution (aspirate and dispense 3 times). Take 1.0 μl of the solution and spot on a stainless-steel target plate.
- 4) Leave the droplet to dry for 20 min at room temperature.
- 5) Install the target plate in the MS apparatus, and measure the peptide masses after finding the laser target spots that provide high signal-to-noise ratio and high peptides ion counts ( $m/z$  range around 700–4000, accumulating the spectra of 100–200 laser shots). In some apparatus, measurements can be automated.
- 6) Reduce the noise level of the mass spectra using the software for data analysis equipped with the MALDI-MS apparatus.
- 7) Calibrate the  $m/z$  values of each spectrum using two internal standards: an autoproteolytic product of porcine trypsin (monoisotopic mass 842.5100) and the ACTH fragment (monoisotopic mass 2464.1989). The trypsin tryptic peptide of monoisotopic mass 2211.1046 can be used instead of ACTH, but its appearance is not assured because the strength of this peak is inversely related to the protein quantity contained in the gel piece.
- 8) Detect the peptides' monoisotopic mass peaks on each spectrum using the data analysis software.
- 9) Convert the data into a peak list (ASCII file) using the software and sort the masses of the peaks by their integrated ion counts.
- 10) Open the WWW page of Mascot PMF ([http://www.matrixscience.com/cgi/search\\_form.pl?FORMVER=2&SEARCH=PMF](http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF)), input your name, e-mail address, and search title, select database (e.g., Swiss-Prot), taxonomy (e.g., *Homo sapiens*), enzyme (e.g., trypsin), allowed missed cleavages (e.g., one), fixed modifications (e.g., carbamidomethyl), variable modifications (e.g., oxidation of methionine), input peptide mass tolerance (e.g., 50 ppm), select mass values to be MH<sup>+</sup> and monoisotopic, indicate the location of the peak list in your computer, and click "start search."
- 11) The search report will appear on your display as a webpage in a few seconds, providing the list of candidate polypeptide entries and their scores. The score is  $-10 \cdot \log(P)$ , where  $P$  is the probability that the observed match is a random

event. The best match is the one with the highest score and a significant match ( $P < 0.05$ ) is typically a score around 60 (the value changes according to the size of entries in a database). The details on the search results appear as a new page by clicking the accession number of the first hit polypeptide, providing the list of matched peptides, the location of the matched peptides in the polypeptide sequence (sequence coverage), the comparisons of the values of measured and calculated masses, and so on.

MALDI-MS-PMF has proven to be easy to perform, sensitive (up to low femtomoles of polypeptides can be assigned), and high-throughput. Most of the spots detected by CBB or silver on denaturing 2-DE gels can be assigned in a relatively short period. Also, this method has been successfully applied to the assignment of proteins on nondenaturing 2-DE gels, and it could even assign heterogeneous polypeptide mixtures when the number of polypeptides is small (up to four) and their molecular masses and the contents are similar [37], as shown in Figure 13.1. However, when a



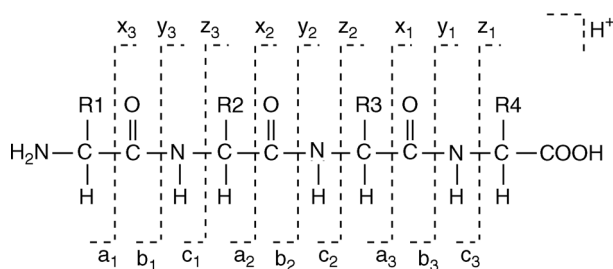
**Figure 13.1** A 2-DE pattern of nondenaturing 2-DE and MALDI-MS-PMF assignment of a protein complex. A human plasma sample was subjected to nondenaturing micro 2-DE using agarose gel IEF in the first dimension and pore gradient gel electrophoresis in the second dimension [30]. Most of the stained spots were assigned by MALDI-MS-PMF as described in Section 13.4.2.2 [29, 37] and many spots were assigned to contain multiple polypeptides. One of the spots that was assigned to contain two polypeptides is indicated by an arrow. The spot was assigned to contain C4-binding protein (C4BP)  $\alpha$ -chain (62 kDa) and complement C4 (190 kDa), whereas the spot showed an

apparent molecular mass of around 1000 kDa. C4BP is reported to be a 500- to 570-kDa complex comprised from disulfide-linked six to eight  $\alpha$ -chains and one  $\beta$ -chain (26 kDa), and each  $\alpha$ -chain has a binding site with C4b (a major fragment of C4). The PMF results strongly suggest that the spot represents a C4BP-C4b complex. The probability that the observed match is a random event was  $3.9 \times 10^{-5}$  (sequence coverage 41%) for C4BP  $\alpha$ -chain and  $3.6 \times 10^{-4}$  for C4 (sequence coverage 22%). However, the PMF method failed to assign the minor component of C4BP (i.e., the  $\beta$ -chain).

protein is composed of two (or more) polypeptides with a large difference in their molecular masses or the molar ratio of one polypeptide is much smaller than the other in a complex, the small or the minor polypeptides would provide low scores and may fail to be assigned. The small/minor polypeptides in the 2-DE spots that are predicted to represent protein complexes (from the discrepancy between the protein apparent mass and the calculated masses of the assigned polypeptides) would be visually confirmed and assigned by a third-dimension electrophoresis followed by MALDI-MS-PMF [36]. If the apparatus for MALDI-TOF/TOF is available, the small/minor peptides may be assigned by a MS/MS sequence search.

#### 13.4.2.3 MS/MS and Peptide Sequence Search

Tandem mass spectrometry (MS/MS) provides information on the amino acid sequence of in-gel digested peptides extracted from 2-DE gel spots. The peptides are ionized by MALDI or ESI and separated in the first MS (MS1) of a MS/MS apparatus. After obtaining the spectrum of the peptides, one of the peptide ions is selected, which means the other peptide ions are filtered out or shut out and only the selected peptide ion (precursor ion) can fly further. The peptide ion is introduced to a collision cell, in which collision gas is continuously filled, and it is decomposed into fragment ions by the collisions with the gas molecules (collision-induced dissociation). The degree of fragmentation can be controlled by adjusting the kinetic energy of the precursor ion or the collision gas concentration. The masses of the fragment ions are then analyzed by the second MS (MS2). Although the cleavage of the covalent bonds of the peptide ion can occur at various positions as shown in Figure 13.2 [72], the peptide bonds are the most susceptible. Therefore, N-terminal-side ions of the peptide cleaved at peptide bonds (*b*-series ions) and the counterparts C-terminal-side ions (*y*-series ions) appear as the major peaks in the MS/MS spectrum. If the mass difference between a pair of adjacent peaks of the same ion series matches with the mass of an amino acid residue in the peptide, then the order of the amino acids in the peptide can be decided. It is important to limit the degree of fragmentation so that the mass spectrum shows the two series of the peaks. Assignment by database



**Figure 13.2** Accepted nomenclature for fragment ions of peptides (exemplified by a tripeptide) [71]. Fragments will only be detected if they carry at least one charge. If this charge is retained on the N-terminal fragment, the ion is

classified as either *a*, *b*, or *c*. If the charge is retained on the C-terminal, the ion type is either *x*, *y*, or *z*. A subscript indicates the number of residues in the fragment.

searching becomes possible with the knowledge of the mass of the peptide, the nature of limited fragmentation, and the cleavage specificity of the enzyme employed to generate the peptide. In 1994–1995, three groups reported programs that match measured masses from MS/MS spectra to sequence databases [73–75]. Currently available programs for the analysis of MS/MS data are listed on the ExPASy Proteomics Server (<http://us.expasy.org/tools/>).

Several types of MS instruments can be used to provide the fragment ions of a peptide and ESI-quadrupole (Q)-TOF and MALDI-TOF-TOF are often used. In the case of ESI-Q-TOF apparatus, the peptide sample adsorbed on a ODS tip column (Section 13.4.2.1, Step 14) is eluted with a weak acid/acetonitrile solution and the eluate (1–2  $\mu$ l) is set in a gold-coated pulled glass capillary that permits a “nanospray” of the solution at a flow rate of around 20 nl/min [76]. Electrospray of the sample solution can last more than 30 min, which is enough to select several peptide peaks and accumulate their fragment ion signals to obtain the MS/MS spectra. In the case of a MALDI-TOF-TOF apparatus, the sample preparation and measurement procedures are almost as described in Sections 13.4.2.1 and 13.4.2.2. Since the same sample plate can be used for both MS analysis and MS/MS analysis, and the processes of the MS/MS analysis can be automated, MALDI-TOF-TOF is suited to high-throughput assignment of proteins/polypeptides separated on 2-DE gels.

### 13.5

#### Conclusions

The advent of sensitive and high-throughput techniques for protein assignment using MS reinforced the utility of 2-DE in the analysis of complex protein systems. Protein complexes, native proteins, and polypeptides can be separated on 2-DE gels, in the small volumes of which hundreds of protein/polypeptide samples are put in order according to the differences in two properties, such as *pI* and molecular mass. Almost all the stained proteins/polypeptides can be assigned by MALDI-MS and PMF and/or by MS/MS and a peptide sequence search. The combined techniques of 2-DE and MS are now applied in one way to the analysis of structural details of proteins such as post-translational modifications and in another way to the global analysis of protein–protein interactions, aiming at the total reconstruction of biological structures and functions of living organisms.

#### References

- 1 Sanger, F. and Tuppy, H. (1951) *Biochemical Journal*, **49**, 481–490.
- 2 Sanger, F. and Thompson, E.O.P. (1953) *Biochemical Journal*, **53**, 366–374.
- 3 Ornstein, L. (1964) *Annals of the New York Academy of Sciences*, **121**, 321–349.
- 4 Davis, B.J. (1964) *Annals of the New York Academy of Sciences*, **121**, 404–427.
- 5 Shapiro, A.L., Vinuela, E., and Maizel, J.V. (1967) *Biochemical and Biophysical Research Communications*, **28**, 815–820.
- 6 Weber, K. and Osborn, M. (1969) *Journal of Biological Chemistry*, **244**, 4406–4412.



- 7 Vesterberg, O. (1969) *Acta Chemica Scandinavica*, **23**, 2653–2666.
- 8 Awdeh, Z.L., Williamson, A.R., and Askonas, B.A. (1968) *Nature*, **219**, 66–67.
- 9 Laemmli, U.K. (1970) *Nature*, **227**, 680–685.
- 10 Mesyanzhinov, V.V., Leiman, P.G., Kostyuchenko, V.A., Kurochkina, L.P., Miroshnikov, K.A., Sykilinda, N.N., and Shneider, M.M. (2004) *Biochemistry*, **69**, 1190–1202.
- 11 Kaltschmidt, E. and Wittmann, H.G. (1970) *Proceeding of the National Academy of Sciences of the United States of America*, **67**, 1276–1282.
- 12 O'Farrell, P.H. (1975) *Journal of Biological Chemistry*, **250**, 4007–4021.
- 13 Neidhardt, F.C. and Phillips, T.A. (1984) *Two-Dimensional Gel Electrophoresis of Proteins* (eds J.E. Celis and R. Bravo), Academic Press, New York, pp. 417–444.
- 14 Stein, L.D. (2004) *Nature*, **431**, 915–916.
- 15 Manabe, T. (2003) *Journal of Chromatography B*, **787**, 29–41.
- 16 Pitt-Rivers, R. and Impiombato, F.S.A. (1968) *Biochemical Journal*, **109**, 825–830.
- 17 O'Farrell, P.Z., Goodman, H.M., and O'Farrell, P.H. (1977) *Cell*, **12**, 1133–1142.
- 18 Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Görg, A., Westermeier, R., and Postel, W. (1982) *Journal of Biochemical and Biophysical Methods*, **6**, 317–339.
- 19 Görg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000) *Electrophoresis*, **21**, 1037–1053.
- 20 Altenhofer, P., Schierhorn, A., and Fricke, B. (2006) *Electrophoresis*, **27**, 4096–4111.
- 21 Rabilloud, T. and Chevallet, M. (2000) *Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods* (ed. T. Rabilloud), Springer, Berlin, pp. 9–29.
- 22 Görg, A. and Weiss, W. (2000) *Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods* (ed. T. Rabilloud), Springer, Berlin, pp. 57–97.
- 23 Ünlü, M., Morgan, M.E., and Minden, J.S. (1997) *Electrophoresis*, **18**, 2071–2077.
- 24 Dale, G. and Latner, A.L. (1969) *Clinica Chimica Acta*, **24**, 61–68.
- 25 Emes, A.V., Latner, A.L., and Martin, J.A. (1975) *Clinica Chimica Acta*, **64**, 69–78.
- 26 Slater, G.G. (1968) *Analytical Biochemistry*, **24**, 215–217.
- 27 Manabe, T., Tachi, K., Kojima, K., and Okuyama, T. (1979) *Journal of Biochemistry*, **85**, 649–659.
- 28 Manabe, T., Hayama, E., and Okuyama, T. (1982) *Clinical Chemistry*, **28**, 824–827.
- 29 Jin, Y. and Manabe, T. (2009) *Electrophoresis*, **30**, 939–948.
- 30 Jin, Y. and Manabe, T. (2009) *Electrophoresis*, **30**, 931–938.
- 31 Kadofuku, T., Sato, T., Manabe, T., and Okuyama, T. (1983) *Electrophoresis*, **4**, 427–431.
- 32 Manabe, T., Takahashi, Y., and Okuyama, T. (1983) *Electrophoresis*, **4**, 359–362.
- 33 Shaw, C.R. and Prasad, R. (1970) *Biochemical Genetics*, **4**, 297–320.
- 34 Manabe, T., Mizuma, H., and Watanabe, K. (1999) *Electrophoresis*, **20**, 830–835.
- 35 Manabe, T. and Jin, Y. (2008) *Electrophoresis*, **29**, 2672–2688.
- 36 Manabe, T. and Jin, Y. (2007) *Electrophoresis*, **28**, 2065–2079.
- 37 Manabe, T., Jin, Y., and Tani, O. (2007) *Electrophoresis*, **28**, 843–863.
- 38 Manabe, T. and Jin, Y. (2010) *Electrophoresis*, **31**, 2740–2748.
- 39 Schägger, H. and von Jagow, G. (1991) *Analytical Biochemistry*, **199**, 223–231.
- 40 Schägger, H., Cramer, W.A., and von Jagow, G. (1994) *Analytical Biochemistry*, **217**, 220–230.
- 41 Lasserre, J.-P., Beyne, E., Pyndiah, S., Lapallierie, D., Claverol, S., and Bonneu, M. (2006) *Electrophoresis*, **27**, 3306–3321.
- 42 Wittig, I. and Schägger, H. (2008) *Proteomics*, **8**, 3974–3990.
- 43 Rabilloud, T. and Charmont, S. (2000) *Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods* (ed. T. Rabilloud), Springer, Berlin, pp. 107–126.
- 44 Görg, A., Weiss, W., and Dunn, M.J. (2004) *Proteomics*, **4**, 3665–3685.
- 45 Manabe, T. and Okuyama, T. (1983) *Journal of Chromatography*, **264**, 435–443.

- 46 Yan, J.X., Wait, R., Berkelman, T., Harry, R.A., Westbrook, J.A., Wheeler, C.H., and Dunn, M.J. (2000) *Electrophoresis*, **21**, 3666–3672.
- 47 Gharahdaghi, F., Weinberg, C.R., Meagher, D.A., Imai, B.S., and Mische, S.M. (1999) *Electrophoresis*, **20**, 601–605.
- 48 Ortiz, M.L., Calero, M., Fernandez-Patron, C., Patron, C.F., Castellanos, L., and Mendez, E. (1992) *FEBS Letters*, **296**, 300–304.
- 49 Castellanos-Serra, L., Proenza, W., Huerta, V., Moritz, R.L., and Simpson, R.J. (1999) *Electrophoresis*, **20**, 732–737.
- 50 Berggren, K., Chernokalskaya, E., Steinberg, T.H., Kemper, C., Lopez, M.F., Diwu, Z., Haugland, R.P., and Patton, W.F. (2000) *Electrophoresis*, **21**, 2509–2521.
- 51 Lopez, M.F., Berggren, K., Chernokalskaya, E., Lazarev, A., Robinson, M., and Patton, W.F. (2000) *Electrophoresis*, **21**, 3673–3683.
- 52 Towbin, H., Staehelin, T., and Gordon, J. (1979) *Proceeding of the National Academy of Sciences of the United States of America*, **76**, 4350–4354.
- 53 Anderson, N.L., Nance, S.L., Pearson, T.W., and Anderson, N.G. (1982) *Electrophoresis*, **3**, 135–142.
- 54 Manabe, T., Takahashi, Y., Higuchi, N., and Okuyama, T. (1985) *Electrophoresis*, **6**, 462–467.
- 55 Edman, P. (1950) *Acta Chemica Scandinavica*, **4**, 283–293.
- 56 Edman, P. and Begg, G. (1967) *European Journal of Biochemistry*, **1**, 80–91.
- 57 Hewick, R.M., Hunkapiller, M.W., Hood, L.E., and Dreyer, W.J. (1981) *Journal of Biological Chemistry*, **256**, 7990–7997.
- 58 Hanash, S.M., Strahler, J.R., Neel, J.V., Hailat, N., Melhem, R., Keim, D., Zhu, X.X., Wagner, D., Gage, D.A., and Watson, J.T. (1991) *Proceeding of the National Academy of Sciences of the United States of America*, **88**, 5709–5713.
- 59 Baker, C.S., Corbett, J.M., May, A.J., Yacoub, M.H., and Dunn, M.J. (1992) *Electrophoresis*, **13**, 723–726.
- 60 Rosenfeld, J., Capdevielle, J., Guillemot, J.C., and Ferrara, P. (1992) *Analytical Biochemistry*, **203**, 173–179.
- 61 Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., and Yohida, T. (1988) *Rapid Communications in Mass Spectrometry*, **2**, 151–153.
- 62 Karas, M. and Hillenkamp, F. (1988) *Analytical Chemistry*, **60**, 2299–2301.
- 63 Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989) *Science*, **246**, 64–71.
- 64 Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) *Current Biology*, **3**, 327–332.
- 65 Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993) *Proceeding of the National Academy of Sciences of the United States of America*, **90**, 5011–5015.
- 66 Shevchenko, A., Chernushevich, I., Ens, W., Standing, K.G., Thomson, B., Wilm, M., and Mann, M. (1997) *Rapid Communications in Mass Spectrometry*, **11**, 1015–1024.
- 67 Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., and Standing, K.G. (2000) *Analytical Chemistry*, **72**, 2132–2141.
- 68 Bagshaw, R.D., Callahan, J.W., and Mahuran, D.J. (2000) *Analytical Biochemistry*, **284**, 432–435.
- 69 Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) *Nature*, **379**, 466–469.
- 70 Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) *Analytical Chemistry*, **68**, 850–858.
- 71 Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) *Electrophoresis*, **20**, 3551–3567.
- 72 Biemann, K. (1990) *Methods in Enzymology*, **193**, 886–887.
- 73 Mann, M. and Wilm, M. (1994) *Analytical Chemistry*, **66**, 4390–4399.
- 74 Yates, J.R. 3rd, Eng, J.K., McCormack, A.L., and Schieltz, D. (1995) *Analytical Chemistry*, **67**, 1426–1436.
- 75 Clauser, K.R., Hall, S.C., Smith, D.M., Webb, J.W., Andrews, L.E., Tran, H.M., Epstein, L.B., and Burlingame, A.L. (1995) *Proceeding of the National Academy of Sciences of the United States of America*, **92**, 5072–5076.
- 76 Wilm, M. and Mann, M. (1996) *Analytical Chemistry*, **68**, 1–8.

## 14

# Bioinformatics Tools for Detecting Post-Translational Modifications in Mass Spectrometry Data

*Patricia M. Palagi, Erik Arhné, Markus Müller, and Frédérique Lisacek*

### 14.1

#### Introduction

Proteins may be chemically modified during and after translation, one of the later steps in protein biosynthesis. Known as co- and post-translational modifications (simplified by one single acronym PTMs), they may alter the chemical nature of an amino acid (e.g., deamination, the conversion of an arginine into a citrulline) or the function of a protein by attaching a biochemically functional group such as acetate, phosphate, various lipids, and carbohydrates to an amino acid, or by making structural changes, like the formation of disulfide bridges between cysteines or the cleavage of polypeptide chains. Some examples of their biological effects include the regulation by phosphorylation of intracellular signal transduction pathways that mediate development, immune system function, and intracellular communication [1]; cell adhesion, targeting, and signaling due to glycosylation [2]; attachment of fatty acids for membrane anchoring [3], to name just a few. The knowledge and analysis of proteins and their PTMs are extremely important for the understanding of biological pathways and the study of various diseases (e.g., cancer, diabetes, and neurodegenerative diseases).

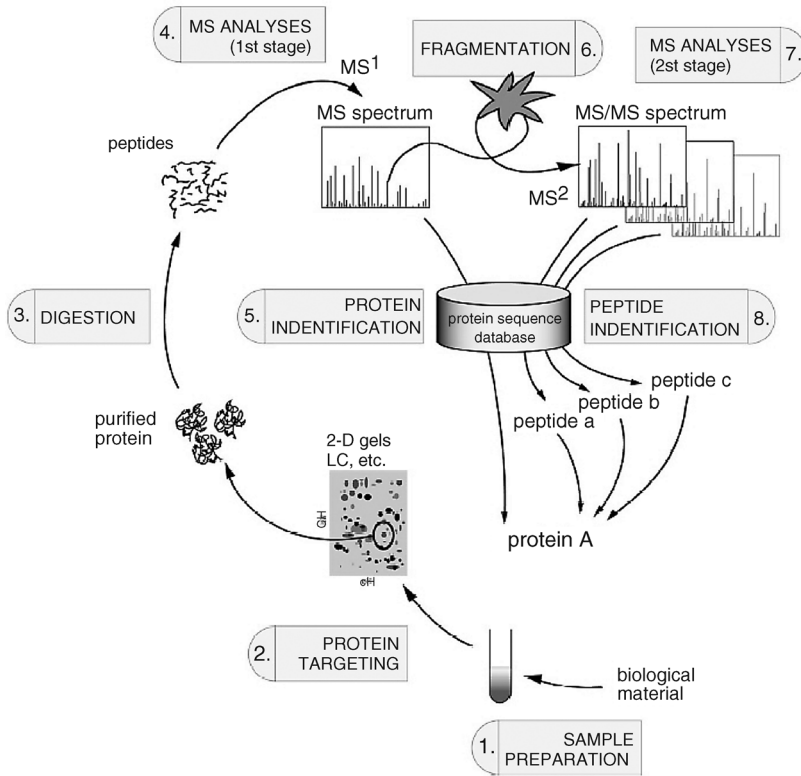
Proteomics and its analytical methods have greatly improved the study of PTMs. The word “proteome,” introduced in 1994 to picture the PROTEin complement of a genOME [4], describes the ensemble of protein forms expressed in a biological sample at a given point in time and in a given situation. Proteomics, defined as the study of proteomes, has four main objectives:

- i) Identify and catalog all proteins in a proteome.
- ii) Analyze differential protein expression associated with different cell or tissue states depending on a disease, a treatment, a developmental stage, and so on.
- iii) Characterize proteins with the specificity of their function in the studied tissue, their cellular localization, variants, PTMs, and so on.
- iv) Describe and understand protein interaction networks.

Proteomics is in constant evolution, and relies on different but complementary analytical methods for protein separation and analysis, and on bioinformatics. The most common technologies for separating complex protein mixtures are capillary and gel (one- and two-dimensional) electrophoresis, liquid chromatography, and high-performance liquid chromatography. In the last decade, mass spectrometry (MS) – a technology that measures the mass-to-charge ( $m/z$ ) ratio of molecules – has become the key method for protein identification, quantitation, and characterization in proteomics projects [5]. There are currently more than 20 different mass spectrometers commercially available, differing in design and functionality, but having common components.

Schematically, mass spectrometers are composed of an ionization source, a mass analyzer that measures the  $m/z$  of the ionized analytes, and a detector that counts the number of ions appearing at each  $m/z$  value. Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are the two most commonly used techniques to ionize proteins and peptides in MS-based proteomics analyses. MALDI usually generates protonated proteins or peptides carrying a single charge, whilst ESI generally generates a mixture of singly and multiply charged ions. The various types of mass analyzers differ in their accuracy, resolution, dynamic range of  $m/z$ , and ability to generate fragments of peptides (tandem MS (MS/MS) spectra). The most popular are time-of-flight (TOF), quadrupoles, quadrupole ion-traps, Fourier transform ion cyclotron resonance, and, the most recently added to the list, Orbitrap [6]. There are many possible combinations of ionization sources and analyzers. Nevertheless, MALDI is generally combined with TOF analyzers, resulting in MS spectra of intact peptides, whilst ESI has been mostly combined with ion traps and quadrupoles, generating MS/MS spectra (collision-induced spectra) of fragmented selected peptides. (See also Chapter 1 of this volume for a more comprehensive overview of MS of amino acids and proteins.)

In a simple bottom-up proteomics analysis workflow (Figure 14.1), a protein mixture is separated by two-dimensional electrophoresis, for example. Each spot of interest is excised from the gel and digested with a proteolytic enzyme (e.g., trypsin). The resulting peptides undergo MS analysis and the generated spectra are matched against the theoretically digested list of peptides from a protein sequence database. This bioinformatics procedure of identification of proteins is called peptide mass fingerprinting (PMF). If desired, the peptides of interest can also be fragmented by tandem MS/MS and the resulting fragment masses are also correlated with theoretical peptide sequences by a MS/MS protein identification algorithm (Figure 14.2). Protein identification using MS/MS spectra is less ambiguous than that achieved by PMF, because, in addition to the peptide masses, information about the peptide sequence can be derived from the peak pattern in the MS/MS spectrum. Although available tools interpreting peptide fragmentation are mainly searching sequence databases for protein identification, they can gear into searching for PTMs. However, the latter task is not straightforward – it requires more time as well as critical manual intervention and evaluation. A range of bioinformatics resources helps scientists detect or predict potential PTMs in experimental spectra.



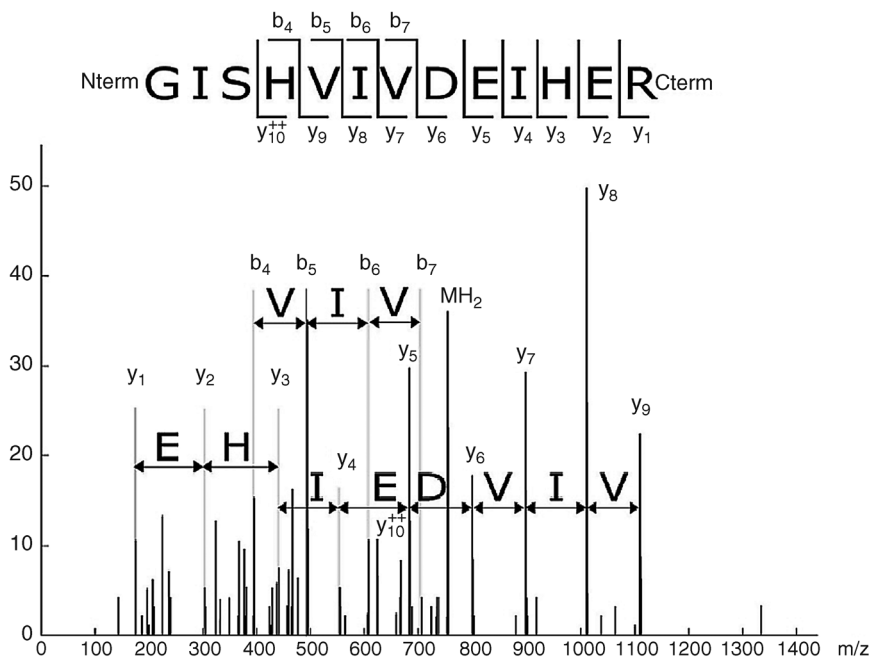
**Figure 14.1** Typical workflow for “bottom-up” protein identification. The first steps (1) and (2) are preparing the sample and targeting proteins to analyze. Then the proteins are digested (3) and the resulting peptides undergo MS analysis (4) yielding an MS spectrum that can be identified by an MS identification algorithm (5). The analysis can go further

by isolating peptides, fragmenting them (6), and measuring the resulting fragment masses (second stage of MS) (7), yielding an MS/MS spectrum for each isolated peptide. The obtained MS/MS spectra are then correlated with theoretical peptide sequences using an MS/MS identification algorithm (8).

## 14.2

### PTM Discovery with MS

The basic principle of discovering PTMs in MS spectra relies on the mass difference between the modified and unmodified states of an amino acid. The precise mass of a modification can be predicted from its atomic composition and as such can determine the mass of a modified amino acid. For example, the acetylation of serine will increase its mass by 42.0367 mass units. As a consequence, the mass of a peptide or protein carrying an acetylated serine will be 42.0367 mass units heavier than that of peptide or protein bearing an unmodified serine. The description of known amino



**Figure 14.2** Example of match between theoretical (expected y and b ions) and experimental masses of fragments (MS/MS spectrum). For instance, 440.3 is the theoretical mass of  $y_3$  = 'HER' which also matches the sum of masses of  $y_2$  = 'ER' (303.2) and  $y_1$  = 'R-C-term' (174.1) as visible in the experimental spectrum.

acid modifications can be found in the UniMod [7] and RESID [8] databases – both are described later in this chapter.

There are currently three main types of bioinformatics tools for the analysis of PTMs from MS data: (i) search for known mass differences between expectedly modified and unmodified peptides in MS or MS/MS data, (ii) discover unexpected mass differences in MS or MS/MS data, and (iii) predict possible modifications from MS or sequence data. These three topics are the subject of the following sections. Popular and maintained bioinformatics resources for the study of PTMs are summarised in Table 14.1.

### 14.2.1

#### Detecting PTMs in MS and MS/MS Data

In the PMF analysis, peak lists extracted from an experimental spectrum are compared with theoretical masses computed from protein sequences stored in databases that were digested *in silico* using the same cleavage specificity of the protease employed in the experiment (e.g., K and R for trypsin). Practically, the procedure counts matches of experimental and theoretical peptide masses. A protein is selected as a candidate when it contains a threshold-dependent number of peptide hits. Other criteria, such as peptide coverage of the sequence, mass error, and so on,

**Table 14.1** Popular resources for PTM discovery available on the web.

Type of resource	Name	Homepage
Databases	UniMod	<a href="http://www.unimod.org">www.unimod.org</a>
	RESID	<a href="http://www.ebi.ac.uk/RESID">www.ebi.ac.uk/RESID</a>
	UniProtKB	<a href="http://www.uniprot.org">www.uniprot.org</a>
	NeXtProt	<a href="http://www.nextprot.org/db/">http://www.nextprot.org/db/</a>
	PhosphoSite	<a href="http://www.phosphosite.org">www.phosphosite.org</a>
	Phosida	<a href="http://www.phosida.com">www.phosida.com</a>
	GlycoSuiteDB	<a href="http://glycosuitedb.expasy.org">glycosuitedb.expasy.org</a>
	dbOGAP	<a href="http://cbsb.lombardi.georgetown.edu/OGAP.html">cbsb.lombardi.georgetown.edu/OGAP.html</a>
Identification tools	UbiProt	<a href="http://ubiprot.org/ru/">ubiprot.org/ru/</a>
	Mascot	<a href="http://www.matrixscience.com">www.matrixscience.com</a>
	Phenyx	<a href="http://phenyx.vital-it.ch/pwi">phenyx.vital-it.ch/pwi</a>
	OMSSA	<a href="http://pubchem.ncbi.nlm.nih.gov/omssa/">pubchem.ncbi.nlm.nih.gov/omssa/</a>
	SEQUEST	<a href="http://fields.scripps.edu/sequest/">fields.scripps.edu/sequest/</a>
	X!Tandem	<a href="http://www.thegpm.org/TANDEM/">www.thegpm.org/TANDEM/</a>
Open Modification tools	SpectraST	<a href="http://www.peptideatlas.org/spectrast/">http://www.peptideatlas.org/spectrast/</a>
	GutenTag	Only available for download: <a href="http://fields.scripps.edu/downloads.php">http://fields.scripps.edu/downloads.php</a>
	Inspect	<a href="http://proteomics.ucsd.edu/LiveSearch">proteomics.ucsd.edu/LiveSearch</a>
	Modiro	<a href="http://www.modiro.com:8080/licenseserver/home.seam">http://www.modiro.com:8080/licenseserver/home.seam</a>
Prediction tools	QuickMod	<a href="http://web.expasy.org/quickmod/">http://web.expasy.org/quickmod/</a>
	MOD <sup>i</sup>	<a href="http://prix.uos.ac.kr/modi/search.jsp">http://prix.uos.ac.kr/modi/search.jsp</a>
	FindMod	<a href="http://web.expasy.org/findmod/">http://web.expasy.org/findmod/</a>
	GlycoMod	<a href="http://web.expasy.org/glycomod/">http://web.expasy.org/glycomod/</a>
	NetOGlyc	<a href="http://www.cbs.dtu.dk/services/NetOGlyc/">www.cbs.dtu.dk/services/NetOGlyc/</a>
	OGlcNAcScan	<a href="http://cbsb.lombardi.georgetown.edu/OGAP.html">cbsb.lombardi.georgetown.edu/OGAP.html</a>
	NetPhos	<a href="http://www.cbs.dtu.dk/services/NetPhos/">www.cbs.dtu.dk/services/NetPhos/</a>
	NetAcet	<a href="http://www.cbs.dtu.dk/services/NetAcet/">www.cbs.dtu.dk/services/NetAcet/</a>
	N-Ace	<a href="http://n-ace.mbc.nctu.edu.tw/">n-ace.mbc.nctu.edu.tw/</a>
	Sulfinator	<a href="http://web.expasy.org/sulfinator/">http://web.expasy.org/sulfinator/</a>
	PrePS	<a href="http://mendel.imp.ac.at/PrePS/">http://mendel.imp.ac.at/PrePS/</a>
	NMT	<a href="http://mendel.imp.ac.at/myristate/SUPLpredictor.htm">http://mendel.imp.ac.at/myristate/SUPLpredictor.htm</a>

contribute to defining a score. The candidate proteins are then sorted according to their scores. The top-ranked proteins are considered as potential identifications of the spectrum – a higher score indicating a higher likelihood that the corresponding protein is the target one.

A variety of similar tools for PMF analysis currently exist, differing in their scoring schemes and the parameters taken into account [9], which include the possible amino acid modifications. All PMF tools allow the user to select, from a list of known or suspected amino acid modifications, those that will be searched against the MS data. Two types of modifications can be applied: fixed or variable. A fixed modification is chosen when an amino acid is definitely modified. For example, protein samples

separated with sodium dodecylsulfate–polyacrylamide gel electrophoresis will usually produce cysteines carrying a carboxymethylation on cysteine (“CAM”). When choosing this option with tolerance of 1, the program will generate the theoretical peptide with all cysteines modified and peptides with all but one cysteine modified. On the other hand, a variable modification is chosen when residues may or may not be modified. For example, for methionine oxidation (“MSO”) with variable option and a tolerance of 2, the program will generate the theoretical peptides with zero, one, or two methionines modified. These programs support every combination of possible modifications occurring on the same locus except *N*- or *O*-linked glycosylation or other complex modifications like glycan phosphatidylinositol anchors.

Mascot [10] and Aldente [11] are two examples of PMF tools. However, in the high throughput era in which proteomics entered several years ago, protein identification with PMF has lost momentum to the benefit of MS/MS identification [4].

MS/MS identification tools work on the same principle of PMF tools (i.e., the matching of experimental MS/MS spectrum against theoretical spectra computed from protein sequences stored in databases that were previously *in silico* digested). Nevertheless, it is better adapted to searching in larger databases as a more specific and sensitive identification method when working with complex mixtures of peptides. Indeed, the sequence information given by peptide fragmentation in MS/MS spectra increases the chances of finding true-positive hits in database searches and the identification of PTMs.

Most classical MS/MS search programs split the identification process into two stages: the first stage is aimed at building a list of candidate proteins from confidently identified spectra and the second stage is aimed at matching unidentified spectra against this list with more combinatorial parameters (e.g., taking into account a larger number of modification types or increasing the mass error tolerance). The main idea behind this strategy is to increase the sequence coverage by loosening constraints while searching for less usual PTMs. Popular programs of this category are Mascot [10], SEQUEST [13], X!Tandem [14], OMSSA [15a] and Phenyx [15] – a software platform developed by GeneBio in collaboration with the Proteome Informatics Group.

Note that the improvement of instrument accuracy in the recent years led to the development of spectral library searches that bypass the costly step of processing millions of amino acid sequences while simply matching new spectra with libraries of annotated spectra [16a].

PMF and MS/MS tools all include the option of looking for a selection of PTMs as the search is performed. In the first case, more mass differences will be found between the experimental and the theoretical masses of the peptides. In the second case, a wider range of mass shifts will be considered for matching the spectra of fragmented ions. However, in both cases values of differences or shifts are known or possibly user-defined.

#### 14.2.2

##### **Discovering PTMs in MS or MS/MS Data**

The advantage of classical search tools described in Section 14.2.1 is the speed with which larger datasets can be processed in a reasonable amount of time. It is very



useful nowadays for the analysis of high-throughput proteomics projects with thousands of spectra to be processed. The drawback is its conceptual limitation to identify only spectra with expected modifications. *Open-modification search* tools, also known as the tag approach or blind search, address this problem. This search strategy is also based on detecting matches between experimental spectra and sequences entries of a database; however, in contrast to classical search tools, they implement algorithms optimized to match spectra with unexpected mass shifts such as PTMs. Four programs are known to a large community: GutenTag [16], Modiro [17], Inspect [18], and Popitam [19]. More recent and faster software include Modi [20a] and QuickMod [20b]. For a more detailed review, see [20].

Open modification search algorithms are designed to take into account any type of PTM that would allow a better match between the spectrum peak pattern and a candidate peptide. However, to avoid the search time explosion problem, these tools include an initial database reduction step, which is often performed by a MS/MS identification strategy such as those described in Section 14.2.1. In fact, MS/MS search tools and open modification tools are complementary in the search for PTMs in MS data.

### 14.2.3

#### PTM Prediction Tools

##### 14.2.3.1 From MS Data

The principle of PTM prediction tools with MS data is to locate mass differences in the spectra and compare these shifts against a collection of known PTM mass values allowing a certain error margin. For example, FindMod [11], a tool developed by the Swiss Institute of Bioinformatics (SIB), can predict more than 70 potential PTMs, and single amino acid substitutions in peptides, by comparing experimental MS data with theoretical peptides calculated from a given UniProt protein sequence or from a user-entered sequence. GlycoMod [21], a more specific tool, can predict from experimental masses possible glycopeptides and free or derivatized oligosaccharide structures that can occur on proteins. In contrast to the tools seen in the previous section, the prediction tools do not need to go through the protein identification step, only through the prediction of PTMs.

##### 14.2.3.2 From Sequence Data

All the tools reviewed so far are based on the analysis of high-throughput proteomics experiments with MS. It is notorious that new genome sequencers, such as the Illumina Genome Analyzer system based on the Solexa technology, are capable of decoding whole organism genomes in a few days. Even though such huge amounts of nucleic acid sequences are available today, it is unfortunately very difficult to predict from these data or from the protein sequence alone the type of modification a residue would carry. Nevertheless, some protein sequences may indicate whether a modification can occur in a specific amino acid, with a motif-like signal. At least three servers have specialised in hosting a collection of tools for PTM prediction from sequence data. The oldest is the ExPASy server [22], known since 1993 as a Web portal for biological knowledge on protein sequences. An example of tools listed in the ExPASy server is the Sulfinator tool [23], based on hidden Markov models (HMM),

predicts tyrosine sulfation sites in proteins of any kind of organism. The Center for Biological Sequence Analysis in Denmark has also traditionally developed artificial neural network and HMM algorithms such as NetOGlyc [24], NetPhos [25], and NetAcet [26] that respectively predict mucin-type GalNac O-glycosylation sites in mammalian proteins; serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins; and N-acetyltransferase A substrates in mammalian and yeast proteins. The prediction tools work basically on the same principle. The user enters an amino acid sequence and the tool will output the loci in the sequence where the predicted PTM would appear. The third server is located in the Research Institute of Molecular Pathology in Vienna and specialised in lipid modifications such GPI anchor or prenylation (see [26a] for review). All of the above and other more recent initiatives are listed in Table 14.1

Note that in all cases, the accuracy of the prediction is hardly ever higher than 70% and that the inferred information needs to be considered with caution.

### 14.3

#### Database Resources for PTM Analysis

The results of protein and peptide identification with MS and MS/MS data, as well as the evidence of experimentally found PTMs, have been constantly annotated in public and Internet-ready databases, such as the UniProt Knowledgebase (UniProtKB) [12]. UniProtKB is a protein-centered database and one of the most comprehensive resources on protein information. It was developed by the UniProt consortium (SIB, European Bioinformatics Institute, and the Protein Information Resource group at the Georgetown University Medical Center and National Biomedical Research Foundation). UniProtKB is composed of two sections: UniProtKB/Swiss-Prot, which contains nonredundant manually annotated records, and UniProtKB/TrEMBL, which contains the computer-annotated translations of coding sequences. Together, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL cover all proteins identified so far whether characterized or only inferred from nucleotide sequences. Swiss-Prot gives a high-quality synthetic view of the knowledge that is available on a particular protein, such as nomenclature, the sequence and its variants, function, implication on different diseases, links to other specialized databases, and so on. A dedicated group of curators extract, from the scientific literature, information that is summarized and integrated into the knowledge base. UniProtKB Release 2011\_07 (July 2011) contained 530 264 manually annotated and reviewed entries (in the UniProtKB/Swiss-Prot part) and 16 014 672 computer-annotated, unreviewed entries (in the UniProtKB/TrEMBL part). Note that a new imitative for the in-depth annotation of human proteins was launched under the name of neXtProt by the founder of Swiss-Prot in 2010 (cf: Table 14.1). This database contains reviewed information on PTMs specifically in human proteins.

PTMs are described in three different parts in the UniProtKB/Swiss-Prot entries. In the “Keywords” subsection of the “Ontologies” section, PTMs are listed using a set of controlled vocabulary making a search process very easy (e.g., when looking for proteins that have a “disulfide bond” or a “formylation”). The list of controlled

vocabulary is given at the UniProt website as well as the list of all proteins linked to a particular keyword. The “Sequence annotation (Features)” section of UniProtKB/Swiss-Prot entries provide information on the location and types of PTMs annotated with a specific feature key: “LIPID” for the addition of lipids, “DISULFID” for disulfide bonds, “CROSSLNK” for other cross-links (e.g., isoglutamyl cysteine thioester), and “MOD\_RES” for the addition of all other small groups (e.g., acetyl, methyl or phosphate groups). Glycoproteins, annotated as “CARBOHYD,” are further classified according to the identity of the atom of the amino acid that binds the carbohydrate chain (i.e., “N-linked” when bound to nitrogen, “O-linked” when bound to oxygen, and “C-linked” when bound to carbon).

If the site of the PTM is not known or if the consequences of the modifications require further comments, the information is stored in the “Post-translational modification” subsection of the “General annotation (Comments)” section.

The annotated information on PTMs in the UniProtKB/Swiss-Prot entries can be obtained by:

- Using some specific high-quality prediction tools (e.g., for N-glycosylation, sulfation, and myristoylation). In this case the reliability code “Potential” is added to the annotation (see section 14.2.3.2 for the discussion on prediction tools).
- Propagation from already annotated orthologous proteins. In this case they will carry the reliability tag “By similarity”.
- Using results of published low- or high-throughput proteomics studies. In this case the PTM has been experimentally proven (the most reliable level) and there is a Ref tag to the published article where the experience was described.

A UniProtKB entry of protein “Proteasome subunit alpha type-3” (accession number P25788) is exemplified in Figure 14.3 and the PTM annotations are

**P25788 (PSA3\_HUMAN)** Reviewed, UniProtKB/Swiss-Prot  
 Last modified: June 28, 2011. Version: 135. History...

Clusters with 100%, 90%, 50% identity | Documents (4) | Third-party data

Names · Attributes · General annotation · Ontologies · Interactions · Alt products · Sequence annotation · Sequences · References · Cross-refs

Entry info · Documents · Customize order

**Sequence annotation (Features)**

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
<b>Molecule processing</b>					
<input type="checkbox"/> Initiator methionine	1	1	Removed		
<input type="checkbox"/> Chain	2 – 255	254	Proteasome subunit alpha type-3		PRO_0000124091
<b>Amino acid modifications</b>					
<input type="checkbox"/> Modified residue	2	1	N-acetylserine (Ref:9) (Ref:20)		
<input type="checkbox"/> Modified residue	57	1	N6-acetyllysine (Ref:32)		
<input type="checkbox"/> Modified residue	110	1	N6-acetyllysine (Ref:32)		
<input type="checkbox"/> Modified residue	161	1	Phosphotyrosine (Ref:16)		
<input type="checkbox"/> Modified residue	206	1	N6-acetyllysine (Ref:32)		
<input type="checkbox"/> Modified residue	230	1	N6-acetyllysine (Ref:32)		
<input type="checkbox"/> Modified residue	238	1	N6-acetyllysine (Ref:32)		
<input type="checkbox"/> Modified residue	243	1	Phosphoserine (Ref:22)		
<input type="checkbox"/> Modified residue	250	1	Phosphoserine (Ref:3) (Ref:13) (Ref:17) (Ref:18) (Ref:20) (Ref:21) (Ref:22) (Ref:24) (Ref:25) (Ref:26) (Ref:27) (Ref:28) (Ref:29) (Ref:31)		

**Figure 14.3** UniProtKB/Swiss-Prot entry for “Proteasome subunit alpha type-3” (accession number P2788). The screenshot shows the “Sequence annotation (Features)” section of

this protein, which has been annotated as carrying four PTMs: one N-acetylserine, one phosphotyrosine, and two phosphoserines.

highlighted. Note that four PTMs have been experimentally evidenced: one acetylated serine, one phosphorylated tyrosine, and two phosphorylated serines. In the References section of this UniProt entry (not shown here), there is a link to the article where this experimental evidence was described and the keyword “MASS SPECTROMETRY”. The later information is useful to help query, for example, all phosphorylated proteins annotated in UniProtKB that have MS experimental evidence. This query can be summarized in the following sentence:

```
scope:PHOSPHORYLATION and scope: "MASS SPECTROMETRY"
AND annotation: (type:mod_res phospho*
confidence:experimental)
```

which can be typed in the query window of the UniProtKB interface.

In UniProtKB Release 2011\_07, the most represented PTMs classes are, in decreasing order, MOD\_RES (with 185'346 sites annotated), CARBOHYD (103'995), DISULFID (101' 650), LIPID (10985), and CROSSLNK (6092). In the MOD\_RES class, phosphorylated proteins are the most abundant (49'392 entries) followed by the acetylated proteins (20'951 entries).

RESID, a database of protein modifications from the European Institute of Bioinformatics, is also publicly available, and contains PTM-centered information such as chemical formula and atomic mass, literature citations, cross-references, structure diagrams, and molecular models [8]. Each RESID entry describes a unique chemical modification (e.g., *N*-acetyl glycine) with a cross-reference to all UniProtKB entries annotated with that PTM. Among many possible queries, RESID allows a search for modifications containing a certain mass value. For example, by searching the mass range 41 to 43 daltons (input “41:43”) in the CWeightp field (physical weight) of the RESID query tool, 21 different types of acetylation are listed in Release 66.00 30-Jun-2011.

UniMod [7] is a participative public database on chemical modifications of proteins; any scientist can add or modify a PTM record provided that he/she has been previously registered. Each UniMod entry describes a chemical modification with its chemical formula, and the monoisotopic and average mass values. It also lists the possible protein sites where the modification has been experimentally proven and the list of references.

More specific databases focusing on critical specific PTMs (phosphorylation and glycosylation) are also available. Phosphorylation is a PTM that regulates an amazing number of biological processes, concentrating a whole field of research and literature (for review, see [27]), and as such has deserved many databases. PhosphoSite [28] and Phosida [29] are two examples of phosphorylation site databases. GlycoSuiteDB is a curated database on glycan structures while dbOGAP holds information on O-glycosylated proteins. Ubiquitinated proteins also have their own database in UbiProt. See Table 14.1 for relevant links; an overview of other tools and databases is given in [27].

## 14.4

### Conclusions

MS currently plays a central role in the analysis of amino acids modifications; however, because of the intrinsic complexity of some PTMs, it cannot yet be considered as the ultimate solution. Mass spectrometers give the difference between modified and unmodified peptides, but some modifications cannot be discriminated because they produce the same mass shift [30]. For example, some types of glycosylation (glucose, mannose, galactose) have the same monoisotopic compound mass (162.058 Da). Some modifications such as phosphate and sulfate also have very similar masses (79.663 and 79.957 Da, respectively). With the exception of glycines, all the other standard amino acids have D- and L-isomers. L-Amino acids are the most common; however, D-amino acids are also often found in mammals and human cells [31, 32]. Distinguishing this mirror effect of D- and L-amino acids isomers remains a challenging task for MS techniques because this PTM does not induce any change in the molecular mass.

Many PTMs have been the subject of numerous successful proteomic studies. Among these, phosphopeptides [33], implicated in a multitude of biological processes, have benefited from specific MS improvements developed to facilitate their analysis. Another class of PTMs is considerably less frequent and less studied. Nonenzymatic glycation (Maillard reaction) are among the less frequently studied PTMs, although they are notoriously implicated in various pathological processes [34]. The analysis of glycopeptides by MS methods also remains a challenge in terms of quantification, assignment, and interpretation of spectra due to the complexity and heterogeneity of their molecular structures. From this perspective, PTMs remain an exciting and fertile area of proteomics and MS research, and chemical proteomics has its key role to play [35]. Currently, less than 4% of all proteins in UniProtKB/Swiss-Prot have been annotated with at least one PTM (in 2006, there was 1% [36]). The protein modification databases UniMod and RESID contain approximately 500 different modification entries each. Estimates of the PTM occurrence vary, but we can safely assume that there is an important gap between what is currently known and what remains to be discovered. New powerful mass spectrometers, capable of MS<sup>3</sup> to differentiate fragments of fragments, as well as steady improvements in MS resolution can discriminate very small mass (Da) differences. Further developments of related bioinformatics tools will greatly enhance throughput in PTM analysis of MS-based proteomics studies.

### References

- Hubbard, S.R. and Till, J.H. (2000) Protein tyrosine kinase structure and function. *Annual Review of Biochemistry*, **69**, 373–398.
- Ohtsubo, K. and Marth, J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell*, **126**, 855–867.
- Kannicht, C. (2002) *Posttranslational Modification of Proteins – Tools for Functional Proteomics*, Humana, Totowa, NJ.

- 4 Wilkins, M.R. and Appel, R.D. (2007) Ten years of the proteome, in *Proteome Research: Concepts, Technology and Application* (eds M.R. Wilkins, R.D. Appel, K.L. Williams, and D.F. Hochstrasser), Springer, Berlin.
- 5 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
- 6 Hernandez, P., Binz, P.-A., and Wilkins, M.R. (2007) Protein identification in proteome projects, in *Proteome Research: Concepts, Technology and Application* (eds M.R. Wilkins, R.D. Appel, K.L. Williams, and D.F. Hochstrasser), Springer, Berlin.
- 7 Creasy, D.M. and Cottrell, J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.
- 8 Garavelli, J.S. (2004) The RESID database of protein modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
- 9 Palagi, P.M., Hernandez, P., Walther, D., and Appel, R.D. (2006) Proteome informatics I: bioinformatics tools for processing experimental data. *Proteomics*, **6**, 5435–5444.
- 10 Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using MS data. *Electrophoresis*, **20**, 3551–3567.
- 11 Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005) Protein identification and analysis tools on the ExPASy server, in *The Proteomics Protocols Handbook* (ed. J.M. Walker), Humana, Totowa, NJ.
- 12 The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, **37**, D169–D174.
- 13 Eng, J.K., McCormack, A.L., and Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.
- 14 Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
- 15 Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, **3**, 1454–1463.
- 15a Tharakan, R., Martens, L., Van Eyk, J.E., and Graham, D.R. (2008) OMSSAGUI: An open-source user interface component to configure and run the OMSSA search engine, *Proteomics*. Jun; **8** (12), 2376–2378.
- 16 Tabb, D.L., Saraf, A., and Yates, J.R. III (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*, **75**, 6415–6421.
- 16a Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., Stein, S.E., and Aebersold, R. (2008) Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods*, **5** (10), 873–875.
- 17 Schaefer, H., Chamrad, D.C., Marcus, K., Reidegeld, K.A., Blüggel, M., and Meyer, H.E. (2005) Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. *Proteomics*, **5**, 846–852.
- 18 Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P.A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, **77**, 4626–4639.
- 19 Hernandez, P., Gras, R., Frey, J., and Appel, R.D. (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, **3**, 870–878.
- 20 Ahrné, E., Müller, M., and Lisacek, F. (2010) Unrestricted identification of modified proteins using MS/MS. *Proteomics*, **10**, 671–686.
- 20a Na, S. and Paek, E. (2009) Prediction of novel modifications by unrestrictive search of tandem mass spectra. *J. Proteome Res.*, **8** (10), 4418–1427.

- 20b Ahrné, E., Nikitin, F., Lisacek, F., and Müller, M. (2011) QuickMod: A tool for open modification spectrum library searches *J. Proteome Res.*, **10** (7), 2913–2921.
- 21 Cooper, C.A., Gasteiger, E., and Packer, N.H. (2001) GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, **1**, 340–349.
- 22 Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, **31**, 3784–3788.
- 23 Monigatti, F., Gasteiger, E., Bairoch, A., and Jung, E. (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, **18**, 769–770.
- 24 Julenius, K., Molgaard, A., Gupta, R., and Brunak, S. (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **15**, 153–164.
- 25 Blom, N., Gammeltoft, S., and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, **294**, 1351–1362.
- 26 Kiemer, L., Bendtsen, J.D., and Blom, N. (2005) NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, **21**, 1269–1270.
- 26a Eisenhaber, B. and Eisenhaber, F. (2010) Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods Mol Biol.*, **609**, 365–384.
- 27 Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007) Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, **4**, 798–806.
- 28 Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
- 29 Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology*, **8**, R250.
- 30 Packer, N.H. and Harrison, M.J. (1998) Glycobiology and proteomics: is mass spectrometry the Holy Grail? *Electrophoresis*, **19**, 1872–1882.
- 31 Sachon, E., Clodic, G., Galanth, C., Amiche, M., Ollivaux, C., Soyez, D., and Bolbach, G. (2009) D-Amino acid detection in peptides by MALDI-TOF-TOF. *Analytical Chemistry*, **81**, 4389–4396.
- 32 Kirschner, D.L. and Green, T.K. (2009) Separation and sensitive detection of D-amino acids in biological matrices. *Journal of Separation Science*, **32**, 2305–2318.
- 33 Boersema, P.J., Mohammed, S., and Heck, A.J.R. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. *Journal of Mass Spectrometry*, **44**, 861–878.
- 34 Capote, F.P. and Sanchez, J.-C. (2009) Strategies for proteomic analysis of non-enzymatically glycosylated proteins. *Mass Spectrometry Reviews*, **28**, 135–146.
- 35 Tate, E.W. (2008) Recent advances in chemical proteomics: exploring the post-translational proteome. *Journal of Chemical Biology*, **1**, 17–26.
- 36 Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, **34**, D187–D191.





## Index

### a

- accurate mass tags (AMTs) 9
- acetylation of lysine side-chains 25
- N*-acetylglucosaminyltransferase 22
- Achromobacter lyticus* 30
- acidification 26
- $\alpha$ -conotoxin BuIA 407
- $\alpha$ -crystallin 39
- active nuclei in NMR 98
- acylated peptides 15
- acylation 24
- agarose rod gels 446
- aggregation-suppressing arginine 278
- Ala peptide models
  - conformational energy maps for 158
- alcohol solution 279
- Allicin formation 385
- Alzheimer's  $\beta$ -amyloid (A $\beta$ ) aggregation 218, 219
- Alzheimer's disease 127, 218, 255
- amine coupling 232
- amino acids
  - chemical shift 99, 100
  - energy levels, and spin states 98, 99
  - historical significance 101
  - labile protons 110–112
  - main NMR parameters 99–101
  - metabolomics 112, 113
  - NMR, active nuclei in 98
  - nuclear magnetic resonance 97
  - nuclear Overhauser effect (NOE) 100
  - properties and structures 103–105
  - random coil chemical shift 102–105
  - residual dipolar couplings (RDCs) 101
  - residue masses 13–15
  - scalar coupling constants 100
  - spin systems 105–110
  - structure 101, 102
- amylin fibers 253
- amyloidogenic proteins 255
- $\beta$ -amyloid (A $\beta$ ) peptide 127
  - Alzheimer's aggregation and electrochemical detection method 218, 219
  - atomic force microscopy (AFM) images 220
  - kinetic study 219, 220
  - label-free electrochemical detection 219–221
  - label-free electrochemical monitoring 218–221
- angular momentum 98
- antibody–analyte interactions 230
- antifreeze proteins 62
- antigen–antibody reaction 213
- antimicrobial peptides (AMP) 121, 238
  - interaction 238
- antiprotozoan effects 121
- apical membrane antigen-1 (AMA1) 161
  - enhanced affinity for 162
- aqueous acidic arginine solution 279
- aqueous solution, role in hydrating protein surface 345
- arginine effect on fragment ion formation 17
- aromatic amino acids in aqueous solution, optical properties 177
- asparagine 13, 22, 27, 104, 137
- aspartic acid 13, 19, 22, 27, 137, 188
- L*-aspartyl-phenylalanine methyl ester 125
- aspartyl residue, NMR  $^1\text{H}$  one-dimensional spectra 112
- asymmetric unit 61, 66, 67
- atomic force microscopy (AFM) imaging 220, 249
  - amplitude resonance curve 253
  - amylin fibrils grow on 254

- of amyloid peptides reconstituted in membrane bilayers 257
  - analysis of transthyretin aggregation by 256
  - cytoplasmic gap junction 259
  - force measurements 269
  - imaging problems 261
  - for information obtained 254–255
  - issues
    - accuracy of surface tracking 266–268
    - artifacts related to too low free amplitude 262–266
    - imaging force 263–264
    - repetitive stress 264–265
    - resolution 262–263
    - step artifacts 268
    - transient force and bandwidth 266
  - liquid imaging 269–272
    - maps morphology 249, 250
    - not free of imaging artifacts 261
    - phase resonance curve 253
    - principle and basic modes, of operation 250–251
    - major disadvantage 251
    - optical resolution 250
    - tapping mode 251
    - tapping restrains 251
  - properties of tapping mode 261
  - sample preparation, for bioimaging 272
  - adhesion 272–273
  - chemical binding 274
  - physical entrapment 273–274
  - of striated domains induced in supported DPCC bilayers 258
  - tapping mode AFM probe 252
  - tool to image single biomolecules 261
  - Autolab PGSTAT 100 system 216
  - Azotobacter vinelandii* 369
- b**
- belladonna mottle virus (BDMV) 306
  - Biacore technology 227
  - bilayer model membrane system 237
  - bimolecular interaction 241
  - binding capacity 243
  - bioactive peptides
    - aspartame 125, 126
    - cyclopeptides 128
    - opioids 126, 127
    - transmembrane helices 127, 128
  - bioimaging, sample preparation for 272–274
  - biological functions, of protein molecules 277
  - biological macromolecules, separation of 288
  - Biological Magnetic Resonance Data Bank 101
  - biomolecular interactions 225
  - biomolecules 181, 226, 227, 233, 249, 261, 345, 347
  - biopolymerization 253
  - Blue-native (BN)
    - 2-DE 448
    - proteins 449
    - electrophoresis 449
  - bovine heart mitochondria 448
  - bovine pancreatic trypsin inhibitor (BPTI) 398
    - $\alpha$ -helical and  $\beta$ -sheet regions of 398
  - Bragg's law 75–77
    - in reciprocal space 77–78
  - brain microtubules 277, 278
  - Bravais lattices 64, 66
  - B-values 94
  - b/y-type fragment ions 20
- c**
- CamCoil approach 105
  - Canavan disease 113
  - carbamylation 26
  - carboxymethylation 468
  - carpet-like disruption 255
  - cartesian coordinates 94
  - caspases 368
  - catalytically active thiol 367
  - catholyte 447
  - CBB molecules 450
  - CBB–protein complexes 448
  - CBB staining, detection limit of 450
  - CCCC zinc finger 370
  - c-centered orthorhombic cell 64, 65
  - cell culture, amino acids in 426
  - central nervous system (CNS) 156
  - cerebrospinal fluid (CSF) 31, 156
  - charge-coupled devices (CCD) 52, 82, 88
  - charged peptide ions 420
  - chelators 370
  - chemical immobilization 274
  - chemical mass 1
  - chemical shift perturbation (CSP)
    - experiment 138
  - chemical shifts 99, 100
    - random coil 101
    - values 106–109
    - use 113
  - 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) 236
  - chromatography, concepts in 169
    - chemical and physical factors

- of mobile and stationary phase 179
- dispersion effects of solutes 171
- extra-column band broadening 172
- modes and exploited molecular properties
- of target compounds 178
- peak zone band broadening 171
- plate number, defined 171
- reduced mobile phase velocity 171
- reduced plate height 172
- resolution 170, 173
- retention factor 170
- retention times 170
- separation by 168, 169, 173, 176, 178, 191, 196
- solvent pH and ionic strength for 288–295
- van Deemter–Knox equation 171, 172
- clear-native (CN)
  - 2-DE 448
  - electrophoresis 449
- cleavage
  - A $\beta$  peptides originate from 127
  - b/y cleavage 18
  - of covalent bonds of the peptide ion 459
  - at C-terminal side of aspartic acid 18
  - endogenous cleavage site 26
  - on membrane or in-gel cleavage 454
  - nonspecific cleavages 30
  - proteolytic cleavage sites 25
  - require protonation 16
- clopidogrel 378
- Clostridium pasteurianum* 369
- c-mannosylation 22
- coelectrophoresis 441, 452
- collision energies 20
- collision-induced dissociation (CID) 6, 459
- competitive inhibition assay 233
- connexon pores 260
- conotoxins, M-superfamily branch of 406
- core-shell nanoparticle 218
- core-shell nanoparticle 215
- correlation spectroscopy (COSY)
  - experiments 109, 131
- cosolvents
  - applications 277
  - for proteins 280
  - classification of 283
  - destabilize protein 280
  - effect 279
  - systems 279
- coumarin solubility, as a function of cosolvent (amino acid) concentration 326
- coupling constant. *See* scalar coupling constants
- Cpn10 oligomers, crystallographic structures 122
- cryocrystallography 61
- cryoelectron microscopy 59
- cryoprotectants 61
- crystallization 56
  - conditions 58, 59
  - experiments 58
  - crystals 55
    - growth 57
    - method of 58
    - lattice (*See* lattice)
    - nucleation 58
    - quality 59
    - structures 60
- C–S–S–C dihedral angle 399
- cubic crystal system 64
- culture media 278
- cyclic peptides 128, 159, 160, 173
- cyclophilin D 159
- cyclophilin G 159
- cyclosporin A 159
- Cys1–Cys9 disulfide 403
- Cys3–Cys11 disulfide bond 402
- Cys13–Cys33 disulfide-deficient 400
- Cys3–Cys35 disulfide-deficient analog
  - of ShK 402
- Cys2–Cys15 disulfide-deficient analogs 403
- Cys63-SH 365
- CysSSSCys-like trisulfides 386
- cysteine 363, 468
  - based thiol/disulfide, redox potential of 372
  - biochemistry 363, 364
  - containing tripeptide glutathione 367
  - framework 404–407
  - ligands, oxidation of 371
  - mutational studies 406
  - proteins/enzymes, active site of 365
  - residues 365, 403, 404, 409
  - modification of 388
  - role of 365
  - thiol(ate), nucleophilicity of 368
- cysteine-based catalytic cycles
  - comparison of 367
- cysteine perthiols 387
- cysteine, role of 361–389
  - disulfide bonds, dynamic picture of 371–374
  - dormant catalytic sites 378–379
  - peroxiredoxin/sulfiredoxin catalysis 379–384
  - sulfur 361–365
  - oxidation states 384–388

- as target for oxidants, metal ions, and drug molecules 388–389
  - S-thiolation, chemical protection and regulation 374–378
  - thiols 365–371
  - cytochrome c 39
  - cytosolic O-GlcNAc modifications 22
- d**
- dark-adapted receptor 144
  - data analysis 240–243
    - linearization analysis 240, 241
    - numerical integration analysis 241, 242
    - steady-state approximations 242, 243
  - database search programs 8, 9
  - data collection strategy 82
  - deamidation 27
  - Debye–Hückel theory of solubility 56
  - denaturation 26, 294, 299, 308, 313, 326, 340, 343, 396
  - dendrotoxins 403
  - de novo* sequencing programs 35, 36
  - 2-DE techniques 446, 452
  - detoxification systems 367
  - deuterium oxide (D<sub>2</sub>O) 37
  - diallyltrisulfide 389
  - diffraction-quality crystals 56
  - dimethylsulfoxide (DMSO) 115, 118, 431
  - dimyristoylphosphatidylcholine (DMPC) 133, 216, 236
  - dimyristoylphosphatidylglycerol (DMPG) 236
    - liposomes 236
  - dipole–dipole couplings 101
  - direct detection 233
  - direct immunoassay 233
  - displacement assay 233
  - dissociable groups, pK<sub>a</sub> values of 442
  - dissociation rate constant 241
  - distance geometry methods 134
  - disulfide bonds 25, 395, 400, 408
    - applications of 408–409
    - conformations of 400
    - conservation of 403
    - contribution of
      - protein stability 396–397
      - non-native disulfide connectivities 407–408
      - presence/absence of 403
      - in protein dynamics 401–403
      - protein engineering 408
      - and protein topology
        - conservation/evolution of 403–404
        - conservation of 404
    - roles of 396, 399
      - probing 396
      - in protein folding 397–399
      - in protein structure 399–401
  - disulfide bridge-based PEGylation 408
  - disulfide bridges 128, 363, 369, 396, 409, 463
  - disulfide connectivity 404–407
  - disulfide-containing proteins
    - dormant catalytic sites in 380
  - disulfide linkage, structure of 399
  - disulfide oxidoreductases 366, 373
  - disulfide radical cation 385
  - disulfide-S-oxides 388
  - dithiothreitol (DTT) 396, 444
  - DNA-binding protein 369
  - DNA–DNA hybridizations 211
  - DNA sequence 419
    - data 441
    - databases 40
  - DS–polypeptide complexes 443
  - dynorphin A 118
    - partial NOESY spectrum 119
- e**
- echoviruses 305
  - Edman degradation 453
  - electrochemical biosensing systems
    - for analyzing functional peptides 211
  - electrochemical impedance spectroscopy 215
  - electrochemical LSPR-based label-free detection
    - E-LSPR 215
    - hybrid bilayer membrane
      - E-LSPR substrate fabrication, and formation 215–217
    - of melittin 215
    - peptide toxin, membrane-based sensors measurements 217, 218
  - electromagnetic radiation 52, 67, 70
  - electron density equation 80–81, 84
  - electron density maps 91
  - electron microscopy 59, 249, 348
  - electrospray ionization (ESI) 4, 168, 464
  - electrospray ionized tryptic peptides, low-energy CID 16
  - electrostatic stabilization, of virus 305
  - enantiomers 65, 66
  - endorphin 123
  - β-endorphin 126
    - NOESY spectra 127
  - end-point immunoassays. *See* enzyme-linked immunosorbent assays
  - energy levels 98, 99

enkephalins 117, 119, 122, 126  
 enzyme-linked immunosorbent  
 assays 225  
 equilibrium constants 161, 242  
*Escherichia coli* 56, 401  
 – malate synthase G (MSG) 137, 139  
 – periplasmic expression, of foreign  
 proteins in 278  
 ethanolamine hydrochloride (EAH) 232  
 ethoxylate 58  
 1-ethyl-3-(3-dimethylaminopropyl)  
 carbodiimide hydrochloride (EDC) 232  
 Ewald's sphere 83  
 ExPASy server 469  
*ex vivo* protein modifications 26–28

## f

FAB ionization 5  
 fabricated microfluidic LSPR chip  
 – photograph 212  
 $\alpha$ -factor receptor 120  
 familial amyloidotic polyneuropathy  
 (FAP) 254–255  
 fast atom bombardment (FAB) 4  
 fiber growth 253  
 fibrillar aggregates 255  
 fluorescence spectroscopy 396  
 fluorescent confocal microscopy 249  
 fluorescent dye staining 450  
 N-fold rotation 65  
 foot-and-mouth disease (FMD) virus 307  
 formulation  
 – of biological products 297–300  
 – composition for viruses 309  
 Förster resonance energy 249  
 Fourier transform ion cyclotron resonance  
 (FT-ICR) 5, 464  
 Fourier transform mass spectrometer 168  
 Fpg DNA binding surface 142  
 fragment ions 11  
 free energy 327, 337  
 freezing protein crystals 61  
 Friedel's law 80–81, 83

## g

gas-phase sequencer 453  
 gel isoelectric focusing 440  
 gel-permeation chromatography  
 (HP-GPC) 178–179  
 gene expression 431  
 GenPept database 8  
 $\gamma$ -glutamyl-cysteinyl-glycine (GSH) 367  
 glass-forming stabilizers 345  
 globular aggregates 255

globular proteins 39, 272, 338, 401, 402  
 glutathione peroxidase (GPx) 378  
 glutathione redox 368  
 glutathione-S-transferases (GSTs) 365  
 glutathione system 388  
 glyceraldehyde-3-phosphate dehydrogenase  
 (GAPDH) 363  
 glycoapture 23  
 glycopeptides 20, 22, 183, 469, 473  
 GlycoSuiteDB 472  
 glycosylation 20, 24, 26, 35, 156, 468,  
 471, 473  
 – categories 22  
 gonadotropin releasing hormone 155.  
*See also* luteinizing-hormone releasing  
 hormone  
 G-protein-coupled receptors 120, 160  
 GraphPad Software 240  
*Grifola frondosa* 30  
 gyromagnetic ratio 99

## h

hanging drop crystallization 57  
 $\alpha$ -helix secondary structure 52  
 hemoglobin 53  
 herpes simplex virus (HSV) 302  
 heteronuclear experiments 113  
 heteronuclear single quantum correlation  
 spectrum (HSQC) 131  
 hexafluoroacetone hydrate (HFA) 122  
 HIC. *See* hydrophobic interaction  
 chromatography (HIC)  
 hidden Markov models (HMM) 469  
 high-performance liquid chromatography  
 (HPLC) 9, 21, 29, 32, 167, 168, 177, 198,  
 202, 203  
 – separation modes, in peptide and protein  
 analysis 177–178  
 -- gel-permeation chromatography  
 (HP-GPC) 178–179  
 -- HP-AC 188–189  
 -- HP-ANPC 183–184  
 -- HP-HIC 184–187  
 -- HP-HILIC 181–183  
 -- HP-IEX 187–188  
 -- HP-NPC 181  
 -- reversed-phase chromatography (HP-  
 RPC) 168, 179–180  
 high-performance reversed-phase  
 chromatography (HP-RPC), methods 189  
 – analytical method 190–191  
 -- column efficiency *N*, optimization of  
 191–192  
 -- retention factor, optimization of 192–196

- fractionation 198
  - quality of fractionation, analysis 198
  - scaling up to preparative chromatography 196–198
  - steps 190
  - high-resolution electrospray mass spectra, advantage of 10
  - HIV virus, coat of 389
  - HPA sensor chip 234
  - human plasma proteins, differential precipitation of 287
  - human Prdx, crystal structure of 381
  - human serum albumin (HSA) 378
    - peroxidase-like catalysis 379
  - hybrid bilayer membrane (HBM)-immobilized surface 216
  - hydrated water 280
  - hydration property, of salt ions 279
  - hydrodynamic radius of PEG 342
  - hydrolytic enzymes 368
  - hydrophilic interaction chromatography (HP-HILIC) 168
  - hydrophilic surface 272–273
  - hydrophobic interaction chromatography (HIC) 168, 184, 278, 289, 291
    - and RP-HPLC, using hydrophobic ligands 288
  - hydrophobic surface 273
  - N-hydroxysuccinimide (NHS) esters 232
- i**
- ICAT. *See* isotope-coded affinity tags
  - ice crystal 62
  - ICK fold, structures and sequences of peptides 405
  - IEF, IPG gels 444
  - I $\kappa$ B–NF $\kappa$ B complex 373
  - Illumina Genome Analyzer system 469
  - immobilization 249
  - immobilized DMPG liposomes 238
  - immobilized pHgradient (IPG) technique 443
  - immunosuppressant drug
    - cyclosporine 163
  - impedance spectroscopy 218
  - inactivation of viruses 300, 309–310
    - formalin-induced 305
    - heat-induced 305
    - rate 305
  - indirect competitive inhibition
    - immunoassay 233, 234
  - indirect competitive inhibition SPR immunoassay 235
  - infectious bronchitis virus (IBV) 302
  - insulin 371
    - peptide chains of 395
    - real-time monitoring 213
  - integrated fluidics cartridges (IFCs) 229
  - interactions
    - process 214
    - of water/cosolvent with macromolecule 330
  - interconverting resonances
    - phenomenon 115
  - internal fragment ions 18
  - intracellular signal transduction pathways
    - phosphorylation of 463
  - intrinsically unstructured proteins 145
  - inversion symmetry 65
  - in vivo* protein modifications 21–26
  - ion channel activity 257–259
  - ion-exchange chromatography 440
  - ion-trap mass spectrometer 168
  - IPG gels 443–446
  - IR spectroscopy, applications 97
  - isoaspartic acid 27
  - isobaric tag for relative and absolute quantification (iTRAQ) 427, 428
  - isoelectric point 233
  - isomorphous replacement 85–88
  - isotope-coded affinity tags (ICAT) 2, 427
  - isotope dilution 2
  - isotope spacing method 10
- j**
- J*-coupling 100
- k**
- Karplus equation 100
  - Kretschmann configuration 211
- l**
- Langmuir binding model 240, 241
  - Larmor frequency 99
  - lattice 62, 73–75, 83
    - Bravais 65
    - Fourier transform equation from 79
    - symmetry of 63
    - systems 64
    - two-dimensional 63
  - Leu-enkephalin amide complex
    - in DMSO 124
    - schematic model 124
  - light microscopy 51–52
  - linear harmonic oscillator 251
  - linearization analysis 240, 241
  - N/O-linked glycosylation 468
  - O-linked glycosylation 26

- lipid bilayer membrane systems 234
- lipid modifications 470
- liquid chromatography 30, 31, 230, 420, 440, 453, 454, 464
- liquid chromatography (LC)
  - liquid phase of 420
  - MS chromatographic signal intensities 423
- liquid imaging 269–272
- liver enzyme rhodanese 387
- localized surface plasmon resonance (LSPR) system 211, 214
  - for analyzing functional peptides 211
  - microfluidics biosensor
    - insulin–anti-insulin antibody reaction on chip detection 213, 214
    - insulin peptide hormone detection 211–215
    - microfluidic chip fabrication and measurement 212, 213
    - and micro total analysis systems 211, 212
- low-energy CID
  - of nonmobile peptide ions 19
  - of peptide ions 19
  - spectra of phosphopeptides 20
- luteinizing-hormone releasing hormone 155. *See also* gonadotropin releasing hormone
- Lysobacter enzymogenes* 30
- m**
- macromolecular solutes 279
- macromolecules 55, 133, 277, 278, 279, 288, 300, 314, 316, 321, 322, 325, 332, 348, 363
- magnetic field 5, 67, 99, 100, 101, 135
- magnetic moments 98
- MALDI-MS
  - apparatus 454, 457
  - feature of 456
- MALDI-MS-PMF 458
  - 2-DE pattern of 458
  - of protein complex 458
- MALDI-PSD spectra 19
- MALDI-TOF-MS 454–457
- MALDI-TOF-TOF, apparatus 460
- mass accuracy 7
  - high, role of 8
- mass defect 1, 4
- masses of positively charged fragment ions 16
- mass mapping 32
- mass spectrometry (MS) 439
  - components of 4–6
  - mass spectrometry (MS)-based proteomics 419
    - quantification in 420–423
    - label-free approaches 423–425
    - SIL, in quantitative proteomics 425–430
  - mass spectrometry data
    - detecting post-translational modifications
      - bioinformatics tools for 463–472
      - PTM
        - database, analysis of 470–472
        - detection 466–468
        - discovering 468–469
        - discovery 465
        - prediction tools 469–470
  - mass-to-charge (m/z) ratios 4
  - matrix-assisted laser desorption ionization (MALDI) 4, 6, 19, 26, 454, 464
  - Matthews coefficient 61
  - melanocortin-4 receptor (MC4R) peptide agonists 408
  - melittin
    - calibration curves for 218
    - C-terminal positive residues 238
    - electrochemical LSPR-based label-free detection 215
    - interactions with HBM 217
  - membrane-based sensor, fabrication 216
  - membrane binding 255
    - of antimicrobial peptides by SPR 238
    - and lysis 255
    - of small oligomers of amyloid
      - implicated in neurodegenerative diseases 255
  - membrane interactions
    - of antimicrobial lytic peptides 255
    - of proteins 255
    - studies protocols 236–240
      - antimicrobial peptides membrane binding 238–240
      - bilayer systems formation 236, 237
      - liposome preparation 236
      - membrane system analyte binding 237, 238
  - metabolites 112
  - metal ions 86
  - metallothionein (MT) 369
    - S-glutathiolation of 371
  - methionine 14, 26, 30, 90, 105, 362, 365, 386, 457
  - methionine oxidation (MSO) 468
  - N-methylated peptides 155–163
  - N-methylation
    - antimalarial peptide 161, 162
    - conformational effects 157–159

- cyclic peptides 159, 160
  - effects on bioactive peptides 159–162
  - somatostatin analogs 160, 161
  - thyrotropin-releasing hormone 159
  - 2-methyl-2,4-pentanediol (MPD) 58
  - as protein precipitant 287
  - metrix algorithm 134
  - micro total analysis systems ( $\mu$ TASs) 211
  - microtubule-associated proteins (MAPs) 278
  - microtubule proteins 278
  - Miller indices 73–75
  - mirror symmetry 65
  - $m/N$  translation 65
  - mobile proton 15, 18
  - molecular dynamic simulations 249
  - molecular replacement 83–85
  - monoisotopic mass 1
  - Monte Carlo algorithms 136
  - Monte Carlo calculation 125
  - mounting crystals, for X-ray analysis 61–62
  - MRM-based assay 429
  - mRNA transcript abundance, analysis of 420
  - MS acquisition parameters 425
  - MS-based protein 450
  - MS-based proteomics methods 420
  - MS instruments, types of 460
  - MS/MS analysis 460
  - MS/MS spectra 5
  - MS/MS spectrum 466
  - multidimensional HPLC 198–200
    - design of an effective scheme 203–206
    - fractionation of complex peptide, and protein mixtures 202
    - operational strategies for 202–203
    - purification of peptides, and proteins by 200–201
  - multidimensional liquid chromatography 167
  - multidomain proteins 59
  - multiple-reaction monitoring (MRM) approach 429
  - multiwavelength anomalous dispersion (MAD) 85, 88
    - Harker diagram to determine phase of protein 89
    - requiring incorporation, of heavy atom 90
  - myoglobin 52, 53, 85, 91
  - myristoylation 24, 471
  - $m/z$  ratio 5
- n**
- NADH peroxidase 384
  - nanoscale LC-MS analyses
    - quantifying peptides in 422, 423
  - natural peptides 116, 128
  - neurodegenerative disease 255
  - neurotoxicity 255
  - NMR  $^1\text{H}$  spectrum 114
  - NMR relaxation parameters 402
  - noble metal nanostructure phenomena 211
  - NOESY spectroscopy 118, 132, 142
    - cross-peaks in 133
  - nominal mass 3
  - nondenaturing 2-DE 458
    - procedures of 446
  - nonequilibrium pH gradient electrophoresis (NEPHGE) 443
  - nonisomorphism 86
  - normal hydrogen electrode (NHE) 368
  - nuclear magnetic resonance (NMR) 37, 97, 159
    - active nuclei in 98
    - isotope filtering/editing experiment 142
    - parameters 99–101
    - titration 141
  - nuclear Overhauser effects (NOEs) 100, 162
  - nuclear spin 98
  - numerical integration analysis 240, 241
- o**
- obtustatin 403
  - octadecyl silica (ODS) 454
  - octyl-glucoside (HPA) 237
  - O'Farrell's technique 443
  - oligomerization
    - pathways 221
    - of transthyretin 255
  - one-dimensional polyacrylamide gel electrophoresis (1-DE)
    - techniques of 440
  - OOIBase32 software 212
  - open-modification search tools 469
  - Orbitrap instruments 5
  - organic solvents 326–327, 340–341
    - enhance unfolding at elevated temperatures 285
    - to induce protein crystallization 287
    - polar 181
    - for reversed-phase chromatography 278
    - used as cosolvents 284
    - used to induce protein crystallization 287
    - used to weaken hydrophobic binding 291
  - orthorhombic cells 64
  - orthorhombic crystal system 64
  - oscillating current 5
  - osmolytes 325, 339–340
  - oxidation
    - of methionine 26



- of tryptophan 27
- oxidative damage 39, 375, 377
- P**
- palmitoyloleoylphosphatidylcholine (POPC) 236
- paraffin's protons 97
- parallel reaction model 242
- paramagnetic relaxation effect (PRE) 139
- partial methylation of lysine side-chains 27
- Patterson maps 86–88
- peak absorbance intensity 217
- pentaerythritol propoxylate 58
- PepNovo 36
- peptide hormones 25
- peptide ions undergoing low-energy CID 18
- peptide-mass fingerprinting (PMF) 454, 464
  - analysis 466, 467
  - tools 468
- peptide–membrane interaction 255
- peptide protonation, sites of 15
- peptide–receptor interaction 119
- peptides
  - analysis, HPLC separation modes in 177–178
  - bioactive conformation 116
  - bioactive peptides 116, 117
  - biophysical properties 173
  - chemical shift 99, 100
  - chemical structure of 173
  - choice of solvent 117–124
  - conformational properties 176
  - conformational transitions in proteins
    - oligopeptides models for 114–116
  - energy levels, and spin states 98, 99
  - ensemble calculations 125
  - fragment ions of 459
  - historical significance 113, 114
  - main NMR parameters 99–101
  - membranes 120, 121
  - NMR, active nuclei in 98
  - nuclear magnetic resonance 97
  - nuclear Overhauser effect (NOE) 100
  - optical properties 176–177
  - receptor cavities 122–124
  - residual dipolar couplings (RDCs) 101
  - scalar coupling constants 100
  - sensorgrams 239
  - transport fluids 118–120
- peroxiredoxins (Prdxs) 363
- perthiol, overview of 362, 378, 386, 387, 388, 389
- pharmaceutical proteins 278
- phase separation 184, 287, 311, 340
- phenylmethylsulfonyl fluoride 446
- phosphate-buffered saline (PBS) 216
- phosphofructokinase (PFK) 346
- phosphorylation 20, 23, 24, 420, 470, 472
- pH scouting 232
- point groups 65
- polarizing microscope 59, 60
- poliovirus 305
- polyacrylamide IEF gels 446
- polyacrylamide pore gradient 447
- polyamino acids 115
- polydimethylsiloxane (PDMS) 211
- poly(ethylene glycol) (PEG) 56, 155, 408
- polypeptide chain, mass properties of 21
- polysulfides 389
- pore-forming peptide toxins 215
- postsourc decay (PSD) 11
- post-translational modifications (PTMs) 21
  - analysis of 466
  - database resources for PTM analysis 470–472
  - detecting PTMs in MS and MS/MS data 466–468
  - discovery with MS 465
    - popular resources 467
  - prediction tools 469–470
- potential mass and time tags (PMTs) 9
- Prdx enzymes 379, 389
  - catalytic cycle of 368
  - redox control network 382
  - sulfenic acid reduction 383
  - thiosulfinate 385
- Prdx-like functions 383
- Prdx proteins
  - chaperone activity of 381
- precipitation 287
- preferential hydration 329
- probe oscillation, physics of 251
  - amplitude 252
  - amplitude resonance curve 252
  - sinusoidal periodic motion 252
  - tapping mode AFM probe 252
  - trajectory 252
- PROCHECK program 94
- proteases 37
- proteasome subunit alpha type-3
  - protein, UniProtKB entry of 471
  - UniProtKB/Swiss-Prot entry for 471
- protection effect 130
- protein aggregation 253
- protein analysis
  - aim of 439, 441
  - amino acid sequence, determination of 439
  - HPLC separation modes in 117

- protein cysteine residue (PrSH) 374
  - Protein Data Bank (PDB) 91, 137, 159, 397
  - protein disulfide isomerase (PDI) 368, 399
  - protein–DNA-specific binding 261
  - protein G, B1 domain 116
  - protein–glutathione disulfides (PrSSG) 373
  - protein–ligand complex 140
  - protein/polypeptide assignment 439–460
  - protein–protein interactions 37, 38, 273, 430
    - identification of 420
    - with quantitative proteomics 430–433
  - protein–protein lattice 61
  - protein–protein/protein–ligand interactions 236
  - proteins
    - analysis, HPLC separation modes in 177–178
    - biophysical properties 173
    - chemical shift 99, 100
    - chemical structure of 173
    - conformational properties 176
    - crystallization 59
    - crystals 60, 61
    - diffractometric methods, alternative to validation of 129
    - electrophoretic transfer of 452
    - energy levels, and spin states 98, 99
    - folding models 398
    - folds 133, 396, 398, 409
    - identification, bottom-up, workflow for 465
    - isolation 167
    - main NMR parameters 99–101
    - microenvironment 365
    - NMR, active nuclei in 98
    - nuclear magnetic resonance 97
    - nuclear Overhauser effect (NOE) 100
    - oligomerization 253–255
    - optical properties 176–177
    - principles, effects of cosolvents on 280
    - protein spectra 129, 130
    - purification 56, 169, 453
    - refolding 296–297
    - residual dipolar couplings (RDCs) 101
    - sample 59
    - scalar coupling constants 100
    - spectra 129, 130
    - stability 396
    - stabilizing salts 302
    - structure refinement 91–93
    - structure validation 93–94
    - *in vivo* 277
    - Wüthrich’s protocol 130–144
  - proteins identification 32
    - from MS/MS spectra of peptides 32–35
  - protein–solvent interactions
    - in frozen and freeze-dried systems 342
    - freeze-dried system 345–348
    - frozen systems 342–345
  - protein–water interactions 280
  - proteolysis 25, 27
  - proteolytic digestion 21
  - proteolytic peptides, mass of 25
  - proteomics 21, 463
    - bottom-up
      - analysis of peptide mixtures 30–32
      - enzymatic digestion for 29–30
    - quantitative mass spectrometry-based 419–433
    - shotgun *vs.* targeted 28–29
    - top-down *vs.* bottom-up 28
  - ProteOn XPR36 Protein Interaction Array System 228
  - protonation, of backbone amide 16
  - PROXIMO algorithm 39
  - P2Y<sub>12</sub> receptor antagonist 388
  - pyroglutamic acid 27
- q**
- Q-TOF hybrid 6
  - Q-trap instruments 6
  - quadrupole mass filters 5
  - quantitative errors 423
  - quantitative proteomics
    - application of 432
    - comparing experimental workflows 424
    - experiments 433
  - quantum mechanics 99
  - quantum number 98, 99
- r**
- racemic mixture 66
  - Ramachandran plot 94
  - rate constants 213, 214, 240–242
  - reactive oxygen species (ROS) 373, 375, 377, 381
  - receptor–ligand interactions 227
  - reciprocal lattice 83
  - reciprocal space 73–75
  - recombinant proteins 278
    - culture media, increase periplasmic expression of 278
    - refolding of 278
  - redox chameleon 363
  - refolding technologies 278
  - refractive index 226, 230
  - residual dipolar couplings (RDCs) 101, 133, 134
    - use 134

- residue mass 11, 12, 21
  - for amino acid 13
- resolution
  - defined 6
  - and high mass accuracy 7
  - types of 8
- resonance assignments 136
- resonance units (RUs) 227
- restrained molecular dynamics (RMD)
  - calculations 125
- restricted access materials (RAMs) 167
- reversed-phase chromatography 179–180
- R-factor 92
  - minimization 92, 93
- Rhodococcus* sp. 371
- ribbon isomer 407
- ribonuclease A (RNase A) 398
- ribonuclease S-complex 39
- ring current shift 130
- $R_{\text{merge}}$  value 83
- rotational nuclear Overhauser effect
  - spectroscopy (ROESY) experiment 117, 118
- S**
- Saccharomyces cerevisiae* 373
- salting-in effect 305
- salting-out salts 288
- sandwich assay 233
- sandwich immunoassay 233
- scalar coupling constants 100
- Scorodocarpus borneensis* 386
- scorpion toxin ChTx, amino acid
  - sequences of 406
- SDS-containing slab gels 444
- SDS-containing solution 443
- SDS gel electrophoresis 445, 447, 448, 449
- SDS–polypeptide complexes 451
- secretory processes
  - potent inhibitory effects on 160
- selected reaction monitoring (SRM)
  - experiment 5
- selenium–selenium bond 409
- sensor chip 228, 229
  - surface 229
- sensorgrams 227, 240
  - SPR schematic presentation 228
- sequence-dependent effect 102
- S-glutathiolation 368, 374, 375
- ShK toxin 155, 400
- shotgun proteomics 40
  - analytical workflow 421
  - bottom-up 35
- SILAC. *See* stable isotope labeling with amino acids in cell culture (SILAC)
- SIL by amino acids in cell culture 427
- SIL-enriched atoms 425
- silica nanoparticles 215
- silver staining 450, 451
- simple collinear optical system 216
- simple membrane-based sensor 217
- small unilamellar vesicles (SUVs) 236
- sodium dodecylsulfate–polyacrylamide gel
  - electrophoresis (SDS–PAGE) 21, 440, 468
- solid-phase extraction (SPE) 167
- solubility 59
  - change of glycine in different organic solvents 327
  - of various amino acids in salt solution 324
- solute–solute interactions 56
- solvent accessibility surface (SAS) 39
- solvent applications 280
  - ability of organic solvents, to denature proteins 285
  - actin depolymerization 283
  - alteration in protein–protein interactions 283
  - demonstrating instability, under unnatural environments 281
  - for DNA 310
  - forces and mechanism of cosolvent/water interaction with 314
  - isolation and purification of DNA 310–312
  - mechanism 314
  - physical mechanism 315
    - excluded volume effects 318–322, 321
    - hydration 315
    - ion hydration 315–316
    - Jones–Dole viscosity B coefficient 317
    - molecular crowding effect 320
    - osmolyte hydration 317
    - protein hydration 317–318
  - preferential interaction 328–333
    - organic solvents 340–341
    - osmolytes 339–340
    - PEG 341–342
    - salts 333–339
  - stability of DNA, in a cosolvent system 312–314
  - thermodynamic interaction 322
    - group interaction 322–323
    - organic solvents 326–327
    - osmolytes 325–326
    - salts 323–325
- effects of salts, Hofmeister series 280, 281
- enzymes in organic solvents 284

- folding of SNase in methanol 285
  - glycerol and sucrose, biochemical research on proteins 281
  - H/D exchange methodology
    - exchange of water with organic solvents 286
    - timecourse of H/D exchange reaction 286
  - helical polymer, assembly reactions in presence of 283
  - lyophilized protein, and folding kinetics 287
  - order of salts in altering protein solubility 281
  - organic solvents extensively used as 284
  - protein folding/enzyme reaction 284
  - research on biological samples 281
  - strong salting-out salt 283
  - tubulin losing structure 281–282
    - microtubule assembly, critical monomer concentration of 282
    - for viruses 300–301
      - decrease in herpesvirus titer following 306
      - formulation composition for freeze-dried viruses 307
      - inactivation of 309–310
      - for isolation of viruses 301–302
      - for purification of viruses 301–302
      - stabilization
        - and destabilization by salts 303–304
        - and formulation of viruses 302, 305–306, 308–309
  - solvent molecules 92, 279
  - space groups 66
  - spatially anisotropic dipolar couplings 133
  - spectral counting 425
  - spin states 98, 99
  - spin systems 110, 111
  - square-wave voltammetry (SWV) 219
  - Srx proteins 382
  - stable isotope labeling with amino acids in cell culture (SILAC) 1, 2, 426, 428, 429, 431
  - steady-state approximations 242, 243
  - steric hindrance 27
  - Streptococcus faecalis* 384
  - sucrose, potassium phosphate, and glutamate (SPG) 307
  - sulfenic acids 384
  - sulfenic acid switch 389
  - sulfur-based radicals 385
    - with biological systems 385
  - sulfur chemotype
    - biochemical flexibility of 362
    - diversity of 363
  - sulfur chemotypes
    - schematic overview of 364
  - SUMO1-modified proteins 24
  - superoxide dismutases (SODs) 137
  - superoxide radical anion 375
  - surface plasmon resonance (SPR) spectroscopy 215
    - angle 226
    - application
      - in immunosensor design 230–234
      - in membrane interactions 234–244
    - assay design 233, 234
    - assay development 232–234
    - based immunosensors 230
    - based optical biosensors 225, 226
    - in biosciences 225
    - biosensors, principle of operation 226–228
    - data analysis 240–243
    - detection system 230
    - direct immunoassay 233
    - flow system 229, 230
    - immobilization of analyte, to specific chip surface 232, 233
    - immunoassay technique
      - schematic view 227
    - indirect competitive inhibition immunoassay 233, 234
    - instrument description 228–230
    - membrane interaction studies protocols 236–240
    - sandwich immunoassay 233
    - sensor surface 228, 229
  - surface tension 62
  - Swiss Institute of Bioinformatics (SIB) 469
  - Swiss-Prot entries 470, 471
  - synaptic cleft 120
  - synchrotron radiation 88
  - synthetic peptides 167
  - SYPRO Ruby protein gel stain 451
- t**
- tandem mass tags (TMT) 427
  - target voltage-gated calcium channels 155
  - Temussi model 126
  - N-terminal amino acid 18, 453
  - N-terminal carbamidomethylated cysteine 27
  - tetrapeptide unit 156
  - thermal energy 98
  - thermodynamic interaction 322
  - thiol 366
    - modification 368
  - thiol-based reducing agent (Trx) 378
  - thiols, chemical protection of 376
  - thiol-specific oxidants 389

- thiol-specific reagents, chemical
    - structures of 378
  - thiol/thiolate, nucleophilic character of 368
  - thioredoxin (Trx) 372
  - thiosulfates 375, 377
  - thiosulfonate 377, 386
  - thiyl radicals, characterization of 385
  - THRASH algorithm 10
  - thrombospondin repeat (TSR) domains 403
  - thyrotropin-releasing hormone 159
  - time-of-flight hybrid (TOF) 454, 464
    - mass analyzer 5
  - total correlation spectroscopy (TOCSY) 109
  - transfer free energy, from water to ethanol 323
  - translational symmetry 65
    - glide plane 65
  - transmembrane pores 255
    - formation 255
  - trans-proteomic pipeline (TPP) 32
  - transverse relaxation-optimized spectroscopy (TROSY) 129, 131, 135
  - triose phosphate isomerase (TIM) 138
  - tripeptide thyrotropin-releasing hormone 159
  - Tris-HCl buffer 444
  - trisulfides 386
  - trNOE effect 143, 144
  - Trx reductase 379
  - trypsin 27
  - twinned crystals 60
  - two-dimensional difference gel electrophoresis (2-D DIGE) 445
  - two-dimensional gel electrophoresis (2-DE) 439–460
    - blue-native, protein–protein interactions
      - principle of 448
      - procedures of 448–449
      - specific features 449
      - current status of 441–442
    - MS-based assignment techniques
      - amino acid sequence databases 454–455
      - MALDI-TOF-MS 456–459
      - peptide sequence search 459–460
      - sample preparation for 455–456
    - nondenaturing, biologically active proteins separation
      - feature of 447–448
      - principle 445–446
      - procedures of 446–447
      - polypeptides, separation of 442
      - principle 442–444
      - procedures for 444–445
    - specific features 445
    - protein analysis, and development of 439–441
    - protein assignment techniques, development of 452–454
    - proteins separated, visualization of
      - CBB, silver/fluorescent dye staining 450
      - CBB staining 450
      - fluorescent dye staining 451
      - quantitation 451–452
      - silver staining 450–451
      - zinc-imidazole staining 451
  - tyrosine kinase interacting protein (Tip) 140
- u**
- ubiquitin 39
  - ubiquitin-like modifiers 24
  - UniProtKB 470–472
  - unit cell 62, 63
    - centering 64
    - defined 64
    - three-dimensional 64
  - UV absorbance measurements 433
  - UV/Vis spectrophotometer 212
- v**
- van der Waals repulsion 133
  - vapor diffusion 57
  - Veber–Hirschmann peptide 160
  - viruses, destabilized by freezing and drying 306
  - voltammetric techniques 221
- w**
- waves
    - mathematical representation of 67–70
    - cosine function 68
    - geometric law of cosines 68
    - mathematical relationship, in exponential form 69
    - vector representation, and mathematically expression 69, 70
    - plot of two waves 68
  - wax-impregnated spectroscopic graphic electrode 219
  - WHATCHECK program 94
  - Wüthrich's protocol 130–144
    - chemical shifts 132, 133
    - complexes interactions 140–142
    - conformational constraints 132–134
    - deuteration 142
    - interactions 138–144
    - isotope editing/filtering 142, 143

- malate synthase G(MSG) 137, 138
- model building 134
- NOEs 133
- RDCs 133, 134
- recent developments 134–136
- recording NMR spectra 131
- sample preparation 131
- sequential assignment 131, 132
- structures 136–144
- superoxide dismutases (SODs) 137
- trNOE effect 143, 144

**x**

- X-ray crystallography 52, 53, 97, 440
- limitations 54
- X-ray crystal structure 55

- X-ray diffraction 52, 55, 67–70, 75, 77, 83, 118, 137
- from a crystal 75–77
- intensity 83
- X-ray diffractometry 37
- X-ray methods, validation of 129
- X-ray scattering 55, 61, 67–70, 69, 70, 76, 77, 79, 85
- X-rays interaction, with matter 70–73

**z**

- zinc fingers 369–371, 389
- zinc-imidazole staining 450, 451
- zinc/sulfur complex, oxidation of 371
- zinc/sulfur proteins 369
- Z-score 94