Springer Theses
Recognizing Outstanding Ph.D. Research

Oriol Güell

A Network-Based Approach to Cell Metabolism

From Structure to Flux Balances

Springer

Springer Theses

Recognizing Outstanding Ph.D. Research

Aims and Scope

The series “Springer Theses” brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student’s supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today’s younger generation of scientists.

Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

More information about this series at http://www.springer.com/series/8790

Oriol Güell

A Network-Based Approach to Cell Metabolism

From Structure to Flux Balances

Doctoral Thesis accepted by
University of Barcelona, Barcelona, Spain

Springer

Author
Dr. Oriol Güell
Department of Materials Science and
Physical Chemistry
University of Barcelona
Barcelona
Spain

Supervisors
Prof. Francesc Sagués
Department of Materials Science
and Physical Chemistry
University of Barcelona
Barcelona
Spain

Prof. M. Ángeles Serrano
Department of Condensed Matter Physics
University of Barcelona
Barcelona
Spain

and

Institute of Complex Systems
University of Barcelona
Barcelona
Spain

and

Catalan Institution for Research
and Advanced Studies (ICREA)
Barcelona
Spain

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-3-319-63999-4 ISBN 978-3-319-64000-6 (eBook)
DOI 10.1007/978-3-319-64000-6

Library of Congress Control Number: 2017946678

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

One of the principal objects of theoretical research is to find the point of view from which the subject appears in the greatest simplicity.

Josiah Willard Gibbs

Aquesta tesi doctoral està dedicada a les persones que estimo

Supervisors' Foreword

The study of cellular metabolism has been dominated by a reductionist approach focusing on the analysis of single reactions or specific biochemical pathways. However, metabolism is a complex system of molecular interactions and displays emergent properties and unexpected behaviours that cannot be predicted by assuming the underlying principles of molecular biology.

When searching for an explanation of the processes and responses observed in cellular metabolism, one can use a systems approach. In practical terms, this implies the consideration of genome-scale metabolic reconstructions, with thousands of metabolites and reactions which can be represented as complex networks. Network science in conjunction with systems biology offers then the tools for the endeavour of studying the intricate properties and functions of metabolic networks.

The doctoral thesis by Oriol Güell summarizes a series of studies of the cellular metabolism from a complex network and systems biology perspective. The first part of this book is devoted to the study of the robustness of metabolic networks. At the level of structure, failures of single and pairs of reactions are simulated to characterize the propagation of damage cascades. In addition, analysis of metabolic fluxes at steady state using the Flux Balance Analysis (FBA) technique is employed to extend the investigation of robustness from structure to phenotype. This brings in the concept of synthetic lethality, reviewed in relation to two of its different realizations: plasticity and redundancy. Taken together, the results indicate that essential backup mechanisms of different nature ensure the robustness to failures in metabolic networks.

The second part addresses two more issues. The first one is based on identifying metabolic backbones, which are the most important connections between metabolites constituting the metabolism. This analysis permits to detect evolutionary trends and adaptation fingerprints in metabolic networks. Finally, FBA solutions are contextualized in relation to the feasible flux space of phenotypes compliant with environmental constraints. Among all possible metabolic fluxes solutions, the FBA one is eccentric, meaning that high-growth phenotypes are metabolic states of low probability.

The importance of the results presented in this doctoral thesis goes beyond theoretical implications. The results reported here have potential applicability in biomedicine, for example to study the metabolism of diseased cells, and in biotechnological studies, such as the activation of specific metabolic reactions which will lead to the maximal production of a desired metabolite, like bio-based polymers.

Barcelona, Spain
May 2017

Prof. Dr. M. Ángeles Serrano
Prof. Dr. Francesc Sagués

List of Publications

Parts of this thesis have been published in:

Güell O, Sagués F, Serrano MÁ (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621

Güell O, Sagués F, Basler G, Nikoloski Z, Serrano MÁ (2012) Assessing the significance of knockout cascades in metabolic networks. J Comp Int Sci 3(1–2):45–53

Güell O, Sagués F, Serrano MÁ (2012) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol 10(5): e1003637

Güell O, Sagués F, Serrano MÁ (2014) Assessing the significance and predicting the effects of knockout cascades in metabolic networks. In: Extended Abstracts Spring 2013. Springer, pp 39–44. ISBN 978-3-319-08137-3

Güell O, Sagués F, Serrano MÁ (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In: Engineering of Chemical Complexity II. World Scientific Publishing, pp 39–56. ISBN 978-981-4616-12-6

Güell O, Massucci FA, Font-Clos F, Sagués F, Serrano MÁ (2015) Mapping high-growth phenotypes in the flux space of microbial metabolism. J R Soc Interface 12:20150543

Güell O, Sagués F, Serrano MÁ (2017) Detecting the Significant Flux Backbone of *Escherichia coli* metabolism. FEBS Lett 591(10):1437–1451

Acknowledgements

I would like to thank *Ministerio de Ciencia e Innovación* from the Spanish Government to provide me with a FPU fellowship (*Beca de Formación de Profesorado Universitario*) to perform the thesis. I also acknowledge *Universitat de Barcelona* for funding me, through travel grants, the congresses and the summer schools that I attended during the Ph.D.

I am grateful to Francesc Sagués and M. Àngels Serrano for the supervision of the doctoral thesis.

A special mention goes to my beloved parents, Ramon and Teresa, and also to Mariona, Anna and Albert. I also thank the rest of the family, grandfathers and grandmothers, *els tetes* Pere and Ferran, and their children. *Maite Zaitut*! An enormous hug to Lis and Nina, who were there with me during the writing of the thesis without complaining. You are incredible doggies!

Dani, Édgar, Sveto, Edo and Àngel, you are great friends and the trips to Madrid to visit you have always been incredible, especially in terms of gastronomy. *Eusebio y Gañán, Pit i Collons*.

Special thanks to my high school friends, Guilla, Gerard, Armand, Uri, Fido, Duñó, Víctor and Àlex. You have always been there, you are there, and you will always be there to help. You all know that *no falta tall*.

Trapote and Dani Igual, I will never forget that you always help me whenever it is possible.

I would also like to acknowledge Francesc Mas and the other colleagues and friends from the university for all your support.

Contents

Abbreviations

Organisms

E. coli	*Escherichia coli*
M. pneumoniae	*Mycoplasma pneumoniae*
S. aureus	*Staphylococcus aureus*

Methods

DF	Disparity Filter
DP	Degree Preserving randomization
FBA	Flux Balance Analysis
FBA-MBR	FBA–Maximum Biomass Rate
FFP	Feasible Flux Phenotypes
FVA	Flux Variability Analysis
HR	Hit-And-Run
K-S	Kolmogorov–Smirnov test
MB	Mass-Balanced randomization
PCA	Principal Component Analysis

Concepts

GCC	Giant Connected Component
GENRE	GENome-scale REconstruction
OMP	One-Mode Projection
PSL	Plasticity Synthetic Lethality
RSL	Redundancy Synthetic Lethality

SCC	Strongly Connected Component
SL	Synthetic Lethality

Pathways

AAM	Amino Acid Metabolism
ACM	Alternate Carbon Metabolism
AR	Anaplerotic Reactions
B	Biomass
CAC	Citric Acid Cycle
CEB	Cell Envelope Biosynthesis
GM	Glutamate Metabolism
G	Glycolysis
GG	Glycolysis/Gluconeogenesis
GM	Glutamate Metabolism
IITM	Inorganic Ion Transport and Metabolism
LM	Lipid Metabolism
MLM	Membrane Lipid Metabolism
NM	Nucleotide Metabolism
NSP	Nucleotide Salvage Pathway
OP	Oxidative Phosphorylation
PM	Pyruvate Metabolism
PPB	Purine and Pyrimidine Biosynthesis
PPP	Pentose Phosphate Pathway
PM	Pyruvate Metabolism
TE	Transport, Extracellular
TIM	Transport, Inner Membrane

Units

gDW	grams Dry Weight

Chapter 1
Cellular Metabolism at the Systems Level

This chapter reviews basic concepts of cellular metabolism. First, an overall view of the architecture of cellular metabolism is given, from the large-scale of Catabolism and Anabolism to biochemical pathways, reactions, and metabolites. Fundamental concepts of chemical kinetics and thermodynamics are mentioned, followed by a brief consideration of key ideas about regulation, control, and evolution of metabolism. Finally, the need for a systems-level approach is discussed. Aims and objectives, together with an outline of this thesis, are included at the end of the chapter.

Cellular metabolism is composed of enzyme-controlled biochemical reactions. They form a densely-connected metabolic network which is responsible of maintaining cells alive by generating chemical energy and by synthesizing important metabolic intermediates from nutrients taken from the environment. Over the years, cellular metabolism has attracted the attention of many researchers. At the end of the nineteenth century, the view of metabolism was dominated by studies of specific biochemical reactions or processes. It is worth mentioning in this respect the work of Eduard Buchner who, based on previous work by Louis Pasteur, demonstrated that cell-free biochemical extracts of yeast—known today as enzymes—could catalyse alcoholic fermentation. This put an end to vitalism-based ideas and boosted the then emerging field of biochemistry [1]. Later on, with the help of experimental techniques such as NMR spectroscopy add X-ray diffraction, the idea of the organization of reactions into sequences of consecutive transformations or pathways arose, creating the basis of modern biochemistry [2]. In principle, pathways were treated as entities with a definite function which operated independently of each other. Despite the enormous success achieved by biochemistry, studies focusing on single reactions, enzymes, or even single pathways are not sufficient to explain most experimental results on metabolism at the functional level, which require a high knowledge of the entire map of metabolic interactions and their interplay with other cellular components. Examples of these results are the identification of redundant metabolic pathways [3], or the observation of the effect known as synthetic lethality [4], which arises when a combination of mutations leads to cell death, whereas the individual mutations are not lethal.

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_1

Since metabolic phenotypes[1] and behaviour emerge from the interactions of many metabolic reactions and other cell components, understanding them at the systems level is crucial for our understanding of living cells. Metabolism is not isolated from the rest of the cell machinery. Therefore, a key challenge in biology is to integrate all the knowledge about the constituents of cells, from genes, to proteins, to metabolites, and reactions, in order to understand how they interact and how these interactions determine the behaviour of cells [5]. This implies a wide knowledge on how reactions are interconnected with metabolites to integrate a whole metabolic network. One can use this metabolic map to study, for example, how different pathways interact [6, 7]. A clear understanding of all these metabolic interactions, and their linkages and interdependencies with other biological scales like genetic networks, will allow us to decipher crucial questions, such as how cells are able to adapt to their environment, or in which way evolutionary processes led to the properties of metabolism as we currently observe them.

The study of integrated metabolic maps is difficult due to the inherent complexity of these intricate systems composed of thousands of interacting reactions. To ease the understanding of cellular metabolism as a complex system, the classical reductionist approach has given way to the so-called *systems-level approach*, which studies metabolism as a whole, taking into account the largest number of experimentally known constituents of the metabolic network, their interactions, and the linkages to other cell constituents such as enzymes, proteins, and genes. This emerging paradigm for the study of cell metabolism is at the core of an emerging interdisciplinary field called Systems Biology [8–11], which uses a holistic approach to understand the relationships between structure and function in biological systems, an impossible endeavour for studies that focus on specific reactions, enzymes, or metabolic processes. The use of this approach has provided a large amount of new validated hypotheses, like the heterogeneity of physiological metabolic fluxes in cells [12], or the phylogenetic analysis of metabolic environments that determine which components must be exogenously acquired [13]. Along with the development of *Complex Network Science* [14, 15], the systems-level approach has led to a huge increase in our understanding of how metabolic networks operate.

1.1 A Brief Introduction to Cellular Metabolism

Cellular metabolism comprises the complete set of chemical reactions at the cell level needed for life. While chemical syntheses in laboratory focus on specific sequences of chemical reactions in order to optimize processes, thousands of reactions, tightly interconnected through common metabolites, take place simultaneously in cells, forming a network that is precisely controlled by the combined action of enzymes, genes, etc., in order to secure functions. This network takes part in the growth of

[1] A phenotype is the composite of the observable characteristics of an organism, such as its morphology, development, biochemical or physiological properties.

cells, in the maintenance and construction of their structures, and in the response and adaptation of the cell to different environmental conditions or internal changes [16].

Cellular metabolism is divided in two big blocks. The first is called *Catabolism*, whose processes are related with the degradation of nutrients and intermediate substrates to provide energy and basic building blocks coming from the rupture of chemical bonds of nutrients. The second is referred to as *Anabolism*, whose processes are related fundamentally to the synthesis of complex organic molecules. Notice that Catabolism supplies Anabolism with the necessary energy and basic compounds or elements to synthesize new molecules. At a different scale, biochemical reactions have been classically classified into different biochemical pathways, which are sequences of consecutive reactions that transform certain metabolites into specific products. Pathways are traditionally associated with definite functions, like Glycolysis which breaks down glucose into other small compounds to extract chemical energy and basic building blocks for anabolic reactions in the synthesis of fatty acids or amino acids. Currently, we know that pathways are not isolated entities and, instead, they constantly interact with each other [6, 7].

Focusing on individual reactions, one must notice that they require the action of catalysts-called enzymes- to take place. Enzymes are a special class of proteins. Proteins are macromolecules composed of amino acids, which perform a large number of functions in living cells, participating for example in the responses to stimuli, the replication of DNA, and transportation of molecules. It is worth stressing that, even though biochemical reactions may be thermodynamically spontaneous, they would not take place without enzymes because the activation energy required inside cells is very large. To ensure that all reactions occur, enzymes decrease the necessary activation energy by generating feasible chemical mechanisms that allow these reactions to take place in a controlled way and in reasonable amounts of time [17]. The action of enzymes helps also to control reaction fluxes, i.e., the rates of biochemical reactions. Not all reactions in metabolism proceed with the same speed or are always on. Biochemical fluxes present a broad distribution of values [18] that reconfigure in response to internal or external changes and signals.

1.1.1 Key Compounds

Biochemical reactions are connected by their participating chemical species, the products of one reaction are the substrates of subsequent reactions, and so on. These compounds-metabolites-participate in many different cell functions, including catalytic activity of their own. Five different general categories of metabolites are described in the following paragraphs (see Fig. 1.1).

- Amino acids [19, 20] are compounds composed by amines (–NR_2), carboxylic groups (–COOH), and a different side chain for each amino acid. The polymerization of different amino acids generates short chains called peptides, or long chains called polypeptides that can be arranged in one or more biological functional way to form proteins.

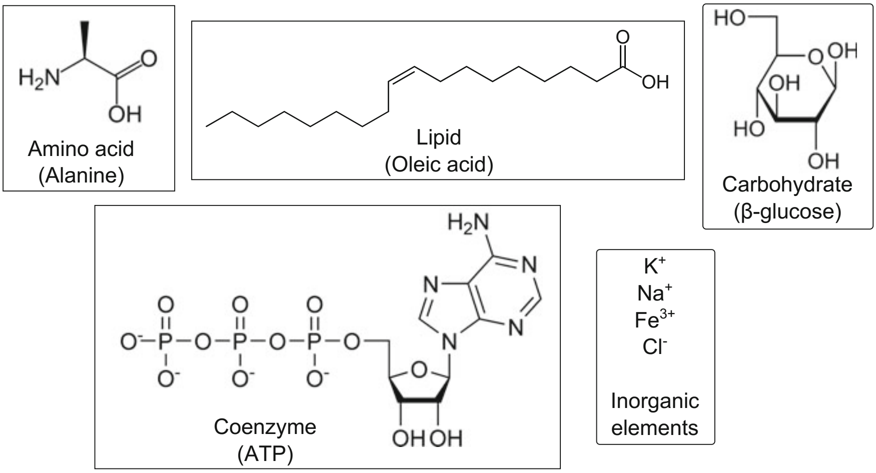


Fig. 1.1 Examples of classes of compounds that can be found in the metabolism of cells

- Lipids are amphiphilic molecules, like fats or sterols, that contain both a polar and an apolar part. This implies that they can be in contact with water (polar part), whereas at the same time are soluble in substances like oil through its hydrophobic part. The main uses of lipids are to store energy [21], signaling [22], and being constituents of membranes [23].
- Carbohydrates are large biological molecules consisting of carbon (C), hydrogen (H), and oxygen (O) arranged on a carbon backbone possibly containing in addition aldehydes (–CHO), ketones (–CO–) and hydroxyl (–OH) groups. They fulfill many roles like energy source [24], storage of energy in the form of glycogen [25], or structural functions.
- Nucleotides are organic molecules containing a nitrogenated base (an aromatic compound containing a basic[2] nitrogen), a ribose or deoxyribose sugar, and a phosphate group (–PO_4^{3-}). They are building blocks of the two nucleic acids DNA and RNA [26]. Genes are fragments of DNA that contain the hereditary information in order to code for polypeptides or for RNA chains. At the same time, RNA performs multiple vital roles in the coding, decoding, regulation, and expression of genes. Nucleotides are obtained from the phosphorylation of nucleosides, and in the form of nucleoside triphosphates, nucleotides play central roles in metabolism [27]. One of these roles is to act as coenzymes, which are important metabolic intermediates that bond loosely to enzymes so as they can perform their catalytic activity. For instance, coenzymes serve to carry energy within the cell. An important coenzyme is adenosine triphosphate (ATP). It is one of the energy currencies of the cell [28]. Many reactions depend on ATP to become thermodynamically spontaneous, taking advantage of the large content of free energy that is released when the high-energy oxygen-phosphate bond of ATP is broken. Another example

[2] In this context, basic refers to acid-base behavior.

of the importance of nucleotides as coenzymes is nicotine adenine dinucleotide NAD$^+$, a derivative of vitamin B$_3$, along with its reduced form nicotine adenine dinucleotide-hydrogen (NADH), which are in charge of balancing the quantity of reduced / oxidized species inside the cell [29, 30].

- Inorganic compounds like water (H_2O), or ionic species like potassium (K^+), sodium (K^+), chlorine (Cl^-), calcium (Ca^{2+}), etc., are simple but not less important components of metabolism. Some of them are abundant, like sodium or potassium, whereas others are present at very low concentrations (traces) [31]. They appear in the form of electrolytes, and thus their concentrations play a key role for example in fixing the osmotic pressure, pH, or the cell membrane potential [32, 33]. Some transition heavy metals like iron (Fe^{2+}/Fe^{3+}) or zinc (Zn^{2+}) are cofactors, compounds which are essential for the activity of proteins like haemoglobin [34].

1.1.2 Biochemical Reactions

Metabolites are the substrates or products of biochemical reactions in the cell. These can be classified in different categories. An important kind of metabolic reactions is a redox process, which involve the transfer of electrons from reduced species, like ammonia or hydrogen sulphide, to oxidized ones, like oxygen or nitrates. Redox reactions play fundamental roles in respiration, where glucose reacts with oxygen, the final products being carbon dioxide coming from the oxidation of glucose, and water, obtained by reduction of oxygen, along with a large quantity of free energy, which is mainly used for non-spontaneous anabolic processes.

Another type of reactions in metabolism involves the transference of entire chemical groups, like a phosphate group in a phosphorylation reaction. Other reactions involve the direct breakage of chemical bonds, like the rupture of carbon-carbon bonds in the decarboxylation of pyruvate. This is a principal process in fermentation which, in order to obtain energy and avoid pyruvate accumulation, transforms a carboxylic group in the form of carbon dioxide, generating acetaldehyde that finally gets reduced into ethanol by a redox reaction [35]. Decarboxylations are also important, for example, in the intermediate step between Glycolysis and the Citric Acid Cycle[3] to obtain acetyl-CoA, or in subsequent steps of this last pathway to generate new intermediates. Transport reactions deserve special attention, since they are responsible for the entrance of nutrients and the excretion of waste products.

An important feature of biochemical reactions is reversibility. Depending on the value of ΔG^{o},[4] reactions can be considered as reversible or irreversible [36]. More precisely, for $\Delta G^{o} \approx 0$ reactions can be considered reversible, meaning that both directions of the reaction are thermodynamically favoured; generically one would

[3] Also called Krebs Cycle or Tricarboxylic Acid Cycle (TCA Cycle).

[4] Thermodynamically speaking one should refer to ΔG, the change in Gibbs free energy (SI units J mol^{-1}). An approximate but convenient way is however to refer to ΔG^{o}, which denotes the free energy change in standard conditions of a reaction.

write $aA + bB \rightleftharpoons cC + dD$. On the contrary, if $\Delta G^o < 0$ and significantly negative, reactions are considered irreversible, and one direction is favoured $aA + bB \rightarrow cC + dD$. For $\Delta G^o > 0$, the reaction takes place mostly in the opposite direction $aA + bB \leftarrow cC + dD$.

1.1.3 Biochemical Pathways

Traditionally, sequences of consecutive biochemical reactions that transform a principal chemical into specific products are called *pathways*. In cell metabolism, there are several universal pathways that when interconnected form a complex metabolic network. Next, the central pathways of metabolism are briefly reviewed.

Glycolysis is the pathway that degrades carbohydrates. It takes place in the cytosol, and its main fuel is glucose. Basically, Glycolysis contains enzyme-catalysed chemical reactions which transform glucose into pyruvate. In its more common form, this process generates the necessary free energy in order to form two molecules of ATP along with NADH. Glycolysis contains two phases, the first one where energy must be invested, which costs two ATP molecules but that generates important intermediate compounds. On the contrary, the second phase produces energy, since four ATP molecules are generated, along with two pyruvate molecules and two NADH molecules. Therefore, Glycolysis is important not only to obtain energy but to generate important biosynthetic precursors [16]. Notice that the inverse process, which generates glucose from pyruvate is called Gluconeogenesis and corresponds to Anabolism.

Pyruvate obtained from Glycolysis can be metabolised in two different ways. The first way corresponds to anaerobic processes, when no oxygen is available. This is called fermentation, and consists in reducing pyruvate into several components like ethanol, lactate or acetate by oxidizing NADH into NAD^+. Fermentation generates two ATP molecules [16].

In case that oxygen is present, the main fate of pyruvate is to become acetyl-CoA, a chemically activated compound formed by a cofactor, called coenzyme A, and an acetyl group. Acetyl-CoA enters the Citric Acid Cycle, a route that takes simple carbon compounds and transforms them into CO_2 in order to obtain energy. The Citric Acid Cycle not only accepts acetyl-CoA from Glycolysis, but also from other routes like lipid or protein metabolism, which emphasizes the importance and centrality of this pathway (see Fig. 1.2) [16]. In eukaryotic cells, the Citric Acid Cycle occurs in the matrix of mitochondria, whereas in prokaryotic cells it takes place in the cytosol, like Glycolysis. The Citric Acid Cycle generates CO_2, guanosine-5′-triphosphate (GTP), NADH, and flavin adenine dinucleotide in hydroquinone form ($FADH_2$). GTP is transformed directly into ATP. NADH and $FADH_2$ are two reduced species that, by being oxidized, generate also ATP. This oxidation takes place in the process called Oxidative Phosphorylation.

Organisms take advantage of the processes in the electron respiratory chain called Oxidative Phosphorylation in order to oxidize the reduced species coming from the Citric Acid Cycle to generate energy. In eukaryotic cells, Oxidative Phosphorylation

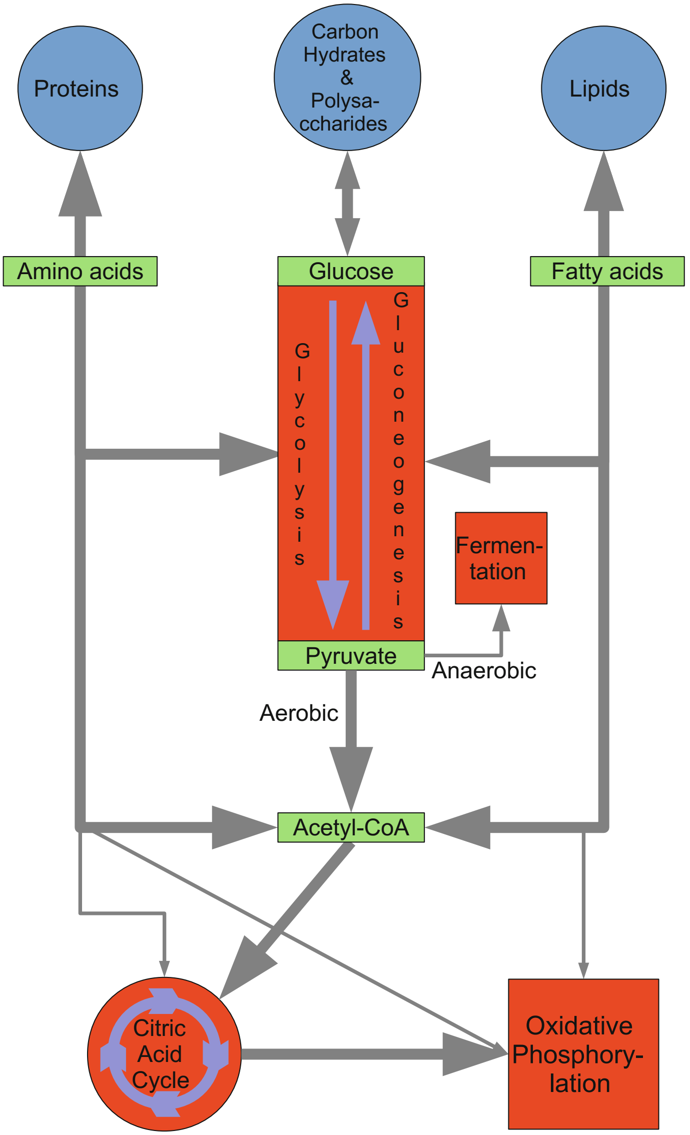


Fig. 1.2 Schematic representation of biochemical routes in central metabolism. Notice that the size of the *arrows* is only determined by aesthetics and does not contain any information about the magnitude of the fluxes through these routes. *Thin arrows* are not explained in the text but are added in this figure for completeness and to remark the interconnectivity present in cell metabolism. *Blue circles* denote major families of compounds. The Citric Acid Cycle (due to its cyclic form) is represented also with a *circle*. The other processes are represented by *squares*, *orange color* denoting pathways and *green color* denoting specific metabolites (color figure online)

takes place inside mitochondria. In prokaryotic organisms, where no mitochondria are present, it takes place across the prokaryotic cell membrane. To summarize, by coupling Glycolysis to the Citric Acid Cycle and Oxidative Phosphorylation, organisms are be able to generate up to 38 ATP molecules [16], which compared to two ATP molecules generated by fermentation, represents a great advantage in order to obtain ATP, whenever oxygen is present.

Other important pathways in cell metabolism comprise the degradation of fatty acids inside mitochondria, a process called β-oxidation, which is another source of acetyl-CoA apart from Glycolysis. Another source of acetyl-CoA comes from the degradation of amino acids, which can be synthesized by transamination [16]. Basically, transamination transforms α-ketoacids coming from the Citric Acid Cycle to generate amino acids, which emphasizes again the centrality of the Citric Acid Cycle.

There are two main routes for the synthesis of purines and pyrimidines, the building blocks of nucleic acids or coenzymes like NAD^+: the *de novo*, which refers to the synthesis from simple molecules, and the salvage pathways, where purines and pyrimidines are recycled from intermediates coming from the routes that degrade nucleotides. The *de novo* route of nucleotide synthesis has a high energetic requirement as compared to the salvage pathway. The enzymes that synthesize purines and pyrimidines perform basic, cellular activities and it is thought that are present in low, constitutive levels in all cells [37].

1.1.4 Classical Studies of Metabolism

Traditionally, metabolism has been studied using a biochemical reductionist approach focused mainly on the study of the role of biomolecules and the kinetics and on the thermodynamics of particular metabolic reactions. As an example, processes like the non-spontaneous transport across the membrane -which takes advantage of the free energy coming from a proton gradient [38] or from ATP hydrolysis- have been studied using irreversible thermodynamics. Classical questions in biochemistry that prompt new systems-level studies refer to regulation and control of metabolism, the interplay and adaptation to the environment, and the effects of evolutionary pressure.

1.1.4.1 Kinetics and Thermodynamics

Classic metabolic studies have usually focused on the kinetics of reactions. The traditional approach was to discover the chemical mechanism by which reactions take place. In this way, kinetic constants were measured for specific reactions using in vitro experimental techniques in order to obtain a velocity law.

As mentioned before, the action of enzymes decreases the necessary activation energy of a reaction, so that the reaction rate increases (otherwise it would take place more slowly or even it would take place so slowly that any progression of the reaction

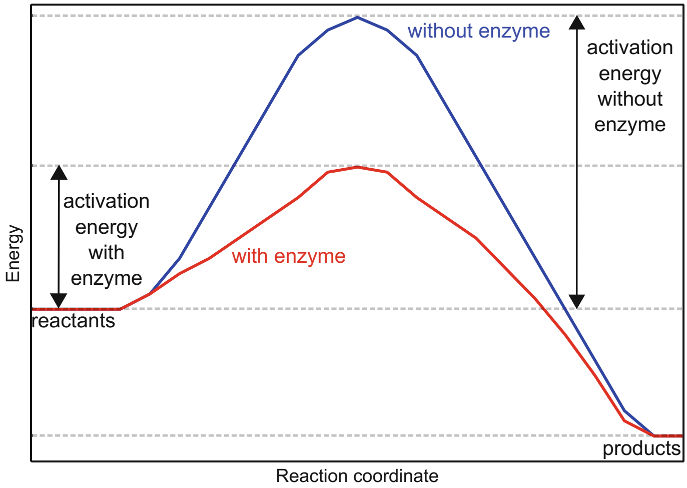


Fig. 1.3 Energy diagram showing the dependence of the energy required for the reactants in order to be transformed into products as a function of the reaction coordinate. This is an abstract coordinate that represents the progress of a reaction along the complete path

would be unnoticeable). A scheme of this decrease in the energy barrier is shown in Fig. 1.3. The best-known kinetic enzymatic mechanism in biochemistry is the famous Michaelis-Menten kinetics [39]. In fact, biochemical reactions involving a single substrate are often assumed to follow Michaelis–Menten kinetics. This model assumes that the minimal equation to describe a simple reaction with one reactant S and one product P catalysed by one enzyme E is

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} P + E \tag{1.1}$$

where k_1, k_{-1}, and k_2 are rate constants. The model relates the overall reaction rate v to the concentration of substrate $[S]$ and the concentration of enzyme $[E]$ under assumptions like steady-state conditions and low enzyme concentration. The rate v is given by the expression $v = v_{max}\frac{[S]}{K_m+[S]}$, where $v_{max} = k_2[E]$ and K_m is the substrate concentration at which the reaction rate is at half-maximum. Michaelis–Menten kinetics reaches a saturation of the velocity as a function of the substrate concentration due to the limited availability of enzyme that can bind to the substrate.

Apart from Michaelis-Menten kinetics, other mechanisms were described for reactions involving more than one substrate or even for reactions with one substrate that do not follow the Michaelis-Menten mechanism. One of these examples is cooperation, which happens when the binding of one substrate molecule to the enzyme affects the binding of subsequent substrate molecules. This effect is modelled by the Hill equation [40], which has the form $\theta = \frac{[L]^n}{K_a^n+[L]^n}$, where θ is the fraction of occupied sites and the Hill coefficient n measures how much the binding of substrate to one active site affects the binding of substrate to the other active sites. The case $n < 1$ indicates that once one substrate molecule is bound to the enzyme, its affinity for other substrate molecules decreases, whereas $n > 1$ indicates that once one substrate molecule is bound to the enzyme, its affinity for other substrate molecules increases. The case $n = 1$ indicates that the binding of one substrate does not affect the binding of other ligands. The other parameters $[L]$ and K_a are, respectively, the free unbound substrate concentration and the apparent dissociation constant derived from the law of mass action.

Other kinetic mechanisms, involving multi-substrate reactions, are the so-called ternary-complex mechanisms and ping–pong mechanisms [16]. These mechanisms describe the kinetics of an enzyme that takes two substrates, namely A and B, and turns them into two products, namely P and Q. Ternary-complex mechanisms imply that the substrates bind to the enzyme forming a ternary complex, where the reaction takes place. After this transformation, the complex dissociates, giving products P and Q. Ping–pong mechanisms consist on sequences of enzyme transformations due to interactions with the substrates. First, the enzyme binds to one substrate and one product is formed. After this process, the second substrate binds to the enzyme giving the second product.

Specific applications of thermodynamics to cell metabolism can be found for example in the description of transport of molecules across the cell membrane. On the one side, passive transport implies a movement of compounds which involves no energy supply, happening spontaneously. On the other side, active transport accounts for the movement of compounds across the cell membrane in the direction against a concentration gradient. Active transport is usually associated to the accumulation of high concentrations of molecules that the cell needs, such as ions, glucose and amino acids. If this process uses chemical energy in the form of ATP, it is termed as primary active transport. Secondary active transport involves the use of an electrochemical gradient. Examples of active transport include uptake of glucose in the human intestines [41].

Kinetics describes the rates of reactions and how fast equilibrium is reached, but it gives no information about conditions once the reaction reaches equilibrium. At the systems level, several aspects must be taken into consideration in relation to its second law. In simple terms, the second law of thermodynamics states that in a closed system entropy tends to increase. An increase in the entropy of a system implies an increase of the number of its possible reachable states. However, organisms seem to contradict this law, since biological systems are complex but ordered structures. To obey the second law and, at the same time, to generate these structures, organisms must exchange matter and energy with their surroundings (see Fig. 1.4). In this way, organisms are not in thermodynamic equilibrium, but they are dissipative systems which, to maintain their high degree of complexity and order, increase the entropy of their surroundings whereas their internal entropy is decreased. Thus, the necessary free energy required by Anabolism to generate complex molecules is obtained by coupling it to Catabolism. For example, nutrients are metabolised and small molecules like CO_2, whose entropy is much larger than that of nutrients [42, 43], are expelled as waste.

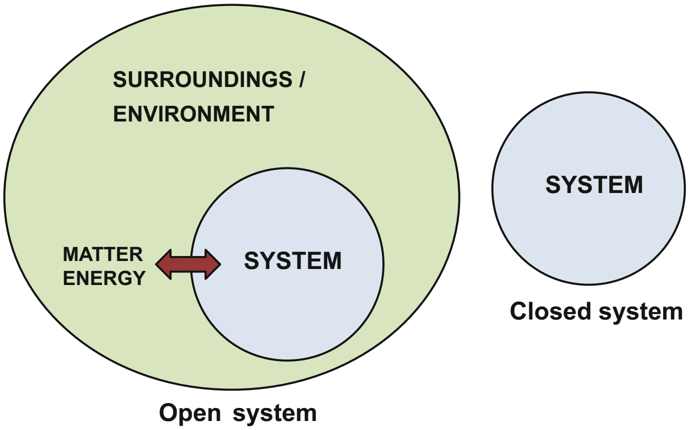


Fig. 1.4 Schematic example of an open system, with exchange of matter and energy, and a closed system, where there are no exchanges of any type

Another thermodynamic discussion concerns energy balance. The intake of energy is equal to the sum of the energy expended in the form of heat or work, and the stored energy. Energy balance states that no energy can be created or destroyed, but it can be transformed. This is indeed the first law of thermodynamics. For example, when a cell consumes nutrients, a part of the energy content of the nutrients will be diverted towards the storage as fat, or transferred inside the cell as chemical energy in the form of ATP, or immediately dissipated as heat.

1.1.4.2 Regulation and Control

The environment of organisms is constantly changing. In fact, organisms themselves modify their own surroundings by consuming nutrients and expelling waste. Therefore, organisms must be regulated in order to avoid large imbalances within themselves. Furthermore, possible internal perturbations can also lead to imbalances inside an organism. Hence, organisms have developed different regulation strategies to be able to maintain *homeostatic* states in which internal conditions remain stable [44]. Regulation requires that a system operates near steady-state conditions, which means that the temporal variation of the properties through time is practically null, except for adjustments to internal or external perturbations. This implies that concentrations of internal metabolites are maintained steady in front of variations in metabolic fluxes. This entails the regulation of enzymes by increasing or decreasing their response to signals.

A real example of homeostatic readjustment is the regulation of glucose concentration by insulin [45, 46]. When large levels of glucose are present in the blood, insulin binds to its receptors, which generates a cascade of protein kinases[5] that cause the consumption of glucose into fatty acids or glycogen. Therefore, the increase in the concentration of glucose is regulated by the control of fluxes of catabolic biochemical reactions, so as to decrease the concentration of glucose until a stable steady-state is reached.

Control has been differentiated from regulation. Metabolic control refers to the ability to change a metabolic state as a response to an external signal [47]. In this way, control can be assessed in terms of the intensity of the response to the external factor without the need of knowing how the organism is able to achieve internally this state. This implies that control is simpler than regulation, because no judgement about the function of the system is needed. For example, an enzyme may show large changes in activity due to some external signal, but these changes may have little effect on the overall flux of a certain set of reactions or pathway. Therefore, this enzyme is not involved in its control.

1.1.4.3 Evolution

Through the process of descend with modifications, organisms evolve and change in time under the driving force of survival. In cell metabolism, there are central pathways that have been conserved through evolution and that are present in practically all kinds of organisms. In fact, these pathways were in the so-called last universal ancestor, which is the most recent organism from which all organisms that now live on Earth descend [48]. Pathways like Glycolysis and the Citric Acid Cycle have been retained probably due to their optimality when producing their products and intermediates in a relatively small number of steps, which then can act as precursors for other biochemical routes. Many studies support the theory that organisms have evolved towards the maximization of the growth rate, i.e., organisms tend to reproduce as much as possible [49, 50].

There have been proposals in recent years in order to understand how metabolism might have evolved including the retention of ancestral pathways. Different mechanisms have been proposed for the evolution of metabolic pathways, for instance (1) sequential addition of old or new enzymes within short ancestral pathways, (2) duplication and then divergence of pathways, and (3) recruitment of enzymes that are already present to be assembled into a novel pathway [51]. Horizontal gene transfer is another way that organisms use to evolve, consisting on the transfer of genes between organisms. In fact, bacteria acquire resistance to antibiotics due to horizontal gene transfer [52]. This process implies modifications in the metabolic network, in the form of alterations of pathways, to generate by-passes in order to avoid the effect of the antibiotic.

[5] A protein kinase is a kind of enzyme which transfers phosphate groups from high-energy phosphate donor molecules to specific substrates. This process is called phosphorylation, not to be confused with the Oxidative Phosphorylation pathway described in Sect. 1.1.3.

Heritable epigenetic effects have also impact on evolution. Epigenetics studies the changes in gene expression that cannot be explained by changes in DNA sequences. There are two ways in which epigenetic inheritance may be different from traditional genetic inheritance. The first way corresponds to the situation where the rates of epimutation are much faster than the rates of mutation [53]. Alternatively, epimutations are more easily reversible [54]. The existence of these possibilities implies that epigenetics, and thus metabolic effects, can increase the evolvability of species.

Evolution can cause not only the gain of new metabolic functions but also the loss of functions which are not useful anymore for cells. *Mycoplasmas*, a kind of bacteria without cellular wall that act as parasites, have lost those processes and pathways that are essential for survival as independent entities, since these microorganisms obtain compounds from their hosts [55, 56].

1.2 Genome-Scale Models

A systems-level approach to the study of cell metabolism takes into account the entire set of biochemical reactions and their interactions at different levels of organization. At the core of this approach, *genome-scale metabolic networks* [10] provide high quality representations of cell metabolism which integrate biochemical information with genome annotations, physiological requirements, and constraint-based modelling refinements. These genome-scale models, after experimental validation, have predictive capacity and can be used for detailed analysis of metabolic capabilities, with applications in a range of fields like biomedicine or biotechnology [57, 58].

1.2.1 Reconstructing Metabolism

Nowadays, genome-scale metabolic models have been reconstructed and experimentally validated for different organisms like *Escherichia coli*, *Saccharomyces cerevisiae*, *Mycoplasma pneumoniae*, and *Homo sapiens*, among others. These reconstructions are called *GENome-scale metabolic REconstructions* (GENREs) (see Fig. 1.5) [10]. In GENREs, reactions are typically stoichiometrically balanced and categorized into their corresponding pathway, for example, reactions belonging to Glycolysis, Oxidative Phosphorylation, or Citric Acid Cycle. Reactions are also associated to their corresponding enzyme and metabolic gene.

Generating these representations is a difficult task and several steps are needed in the protocol [60]. First, an initial reconstruction is proposed from gene-annotation data coupled with biochemical information from databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [61], or BioCyc [62], among others. In these databases, reactions are linked with metabolic genes, enzymes, and also to functional categories like pathways. Second, the obtained reconstruction is curated by checking it against experimental evidence in the existing literature, including for

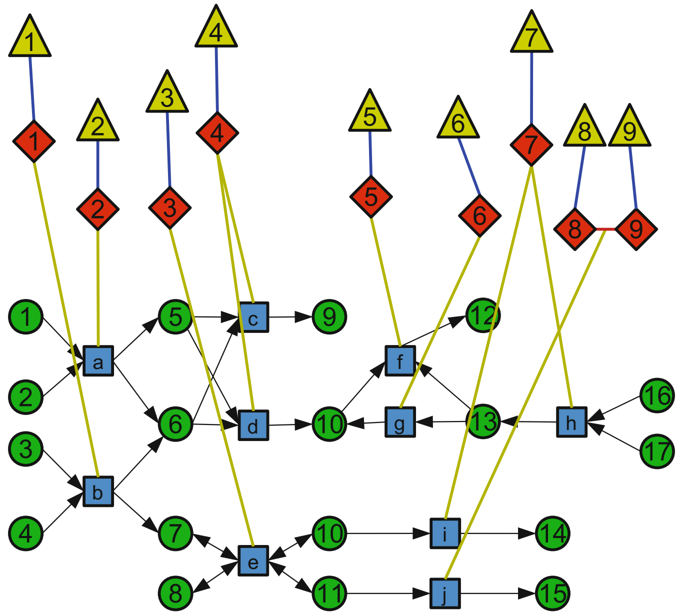


Fig. 1.5 Simplified representation of a genome-scale model. Reactions are catalysed by enzymes, whereas enzymes are codified by genes. Reactions are represented by *blue squares*, metabolites by *green circles*, enzymes by *red rhombus* and genes by *yellow triangles*. Note that enzymes 8 and 9 form a complex and the latter is the catalyst of reaction *j*. (color figure online)

instance physiological requirements. This revised reconstruction is further translated into a computational mathematical model using constraint-based approaches. Third, the reconstruction is validated by comparing the results obtained by the model with experimental evidence. After curation of inconsistencies, models are, see for instance the BiGG database [63]. Finally, one has to remember that GENREs are constantly improved in new versions as new experimental results become available.

Among all metabolic network simulation techniques for model refinement, Flux Balance Analysis (FBA) [64] is probably the most widespread. Very briefly, FBA uses constraint-based analysis to compute a metabolic phenotype, in the form of the set of fluxes of reactions, which maximizes biomass production given a set of external bounds typically referring to nutrient amounts.

Since the first GENRE reconstructed a decade ago [65], there has been a huge expansion on the construction and use of GENREs [66–70]. Their applications can be divided into four categories [57, 71, 72].

- Many advances in Biology are the result of *hypothesis-driven discoveries*. Metabolic GENREs enable the identification and confirmation of new or existing hypotheses, representing an important framework for the incorporation of cell biological data. The key to unlock the potential of GENREs for the discovery of unknown metabolic mechanisms is to ask feasible questions and to know the limitations of the used methodology, since one must always have in mind that in real living cells, many biological levels act together (metabolism, regulation, signaling, gene regulation, etc.) creating a complex system, and GENREs are after all simplified models [57].

- Many characteristic phenotypes of several organisms arise when they *interact with other species* [73, 74]. GENREs enable to analyse interactions between organisms, like for example mutualism, comensalism, parasitism, etc. [74]. It is worth mentioning in this respect the work of Bordbar et al. [75], where the authors developed a model of parasitism between a human cell and the bacterium that causes tuberculosis, *Mycobacterium tuberculosis*.
- Metabolic reconstructions serve as a framework for the contextualization of data obtained using *high-throughput techniques* [76]. A functional way to apply GENREs for contextualization of experimental data, like gene expression or ^{13}C flux data, is by imposing constraints on the fluxes of GENRE based on experimental values. If experiments suggest for instance that reactions of a particular pathway carry large fluxes, one can force the GENRE to have a minimal bound for these fluxes so as to fit the experimental observations. Then, changes in the global flux structure are studied and evaluated.
- *Metabolic engineering* involves the use of recombinant DNA technology[6] to selectively alter metabolism and improve a targeted cellular function [57, 77]. The use of GENREs for metabolic engineering has led to what has been termed as *Systems Metabolic Engineering* [78]. An example of the new advances in metabolic engineering achieved using GENREs is the modification of *Saccharomyces cerevisiae* to increase the production of industrially important intermediates of the Citric Acid Cycle [79]. Another possibility is to study gene knockouts. More precisely, in Ref. [80], the authors performed gene knockouts in *Geobacter sulfurreducens* to maximally increase its respiration rate.

1.2.2 *The Systems-Level Approach*

GENRE reconstructions and the systems-level approach have led to the development of the field called Systems Biology. It is an emerging interdisciplinary field applied to biological systems that focuses on complex interactions using a holistic approach [81]. It is not easy to have a precise and unique definition encompassing all the concepts underlying Systems Biology.

[6] Recombinant DNA molecules are DNA molecules engineered to assemble genetic material from multiple sources, creating sequences that would not otherwise be found in biological organisms.

A possible definition was stated by Ideker et al. [82]:

> Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations.

An alternative was given by Kitano et al. [9]:

> To understand complex biological systems requires the integration of experimental and computational research—in other words a systems biology approach.

These definitions share common features. On the one side, a systems-level approach considers all the components and linkages constituting the system. On the other side, the properties of the components and interactions must be integrated in a computational mathematical model. It is worth stressing the importance of the assembly of these components, i.e., how components interact between them. This can be understood with the analogy of a road-map as given in Ref. [8]. In order to understand traffic patterns, it is necessary to know not only the static road-map but also how cars interact to generate the observed final traffic patterns. Thus, to fully understand a system in a systems-level approach, one needs the diagram with all the connections of all components but also the knowledge of why, how, and to which extent components interact.

Systems Biology can therefore be defined as an approach whose aim is to study biological systems focusing on all the constituents and interactions. In this way, emergent properties which are not present at the level of the single components of the system can be discovered, and phenotype and behaviour can be related to the underlying systems architecture. Central to Systems Biology is the holistic approach. *Holism* is based on the idea that natural systems and their properties should be viewed as a whole instead on focusing on the parts that constitute the system (see Fig. 1.6). Contrarily, the focusing on single parts is called *reductionism*. Examples of traditional reductionist approaches are the study of a single protein or a single chemical mechanism, and they have dominated Biochemistry [83] and Molecular Biology [84] for decades.

Systems Biology represents a paradigm shift that requires the interplay between different disciplines, e.g., Biology, Physics, Mathematics, Chemistry, Computer Science, etc. [85, 86]. Systems Biology foments interactions from traditional computational scientists, modelling experts, and experimental researchers. Research developed to date typically requires powerful computational tools, and this particular emphasis in Systems Biology has given rise to the subfield known as Computational Systems Biology or Computational Biology [85].

Systems Biology has grown in parallel to the development of the *omics* fields. Omics are different disciplines integrating and analysing different kinds of data. Systems Biology combines the datasets obtained in these disciplines in order to achieve the maximum knowledge to model an organism (see Fig. 1.7). Examples of omics related to genes are *Genomics*, which involves sequencing an organism genome, and *Transcriptomics*, which evaluates gene transcription. In relation with proteins,

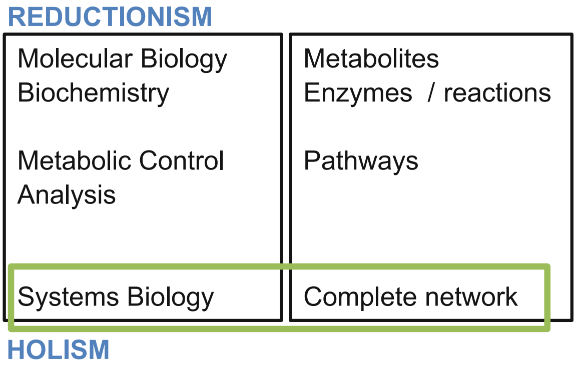


Fig. 1.6 Scheme of the different levels in the holistic versus the reductionist approaches

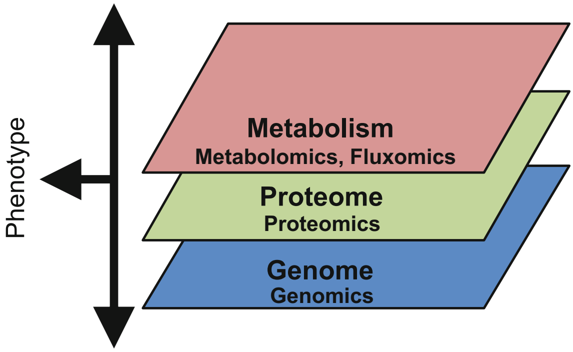


Fig. 1.7 Interplay of the different layers involved in the final phenotype of an organism

the field called *Proteomics* measures protein abundance. Regarding metabolism, *Metabolomics* deals with the study of the concentration of all the compounds present in a organism. There is another important omic field, in line with this thesis, which studies the chemical fluxes in metabolism, *Fluxomics* [87, 88]. Fluxomics provides a measure of a metabolic phenotype as the set of fluxes going through all reactions in a metabolic network.

The holistic view of metabolism including reactions, metabolites, enzymes, genes, and fluxes, represents a new paradigm that requires new tools. Complex Network Science [14, 15] has become a new promising domain for the study of biological systems. Metabolism is formed by a large amount of components and interactions and can be categorized as a complex network and, thus, many applicable techniques that belong to the complex network field are appropriate for the the study of metabolism.

In fact, some of the applications of Complex Network Science ideas to cell metabolism have led to the discovery of many unenvisaged properties such as the existence of loops [89], optimal pathway usage [90], and metabolite connectivity [91]. Other possible discoveries are the exploration of evolutionary relationships [92]. In addition, complex networks applied to metabolism serve as a tool for the identification of how evolutionary pressure has shaped the topological features of metabolic networks, such as the degree distribution [7, 93–95]. Therefore, the joint use of complex network methodologies and Systems Biology provides an excellent arena to study metabolic capabilities and the evolutionary forces that shape metabolic networks.

1.3 Aims and Objectives

This thesis aims at studying cell metabolism from a systems-level perspective, i.e., taking metabolism as a whole.

In particular, one of the questions is how metabolism responds as a whole when some of its constituents fail, i.e., when reactions or genes are non-operative by removal or mutation. It is important to mention that the aim is not to focus on the study of how to perform biochemically the perturbation or the analysis of biochemical failures at a molecular level. Instead, the investigation focus on the impact on the whole system of harmful situations and how metabolism is able to overcome them as a whole entity. In this way, one can study how different pathways reorganize to adapt to perturbations, something impossible to understand by typical molecular biology studies centred on single constituents.

Another question addressed in this thesis is focused on filtering metabolic networks in order to extract metabolic backbones providing valuable biological information. To do this, FBA and the disparity filter [96] are used. The disparity filter allows to extract backbones of the metabolic network containing the significant links. The analysis of metabolic backbones allows the identification of pathways with important roles in survival. The first role corresponds to pathways that have been present in organisms since the first stages of life, i.e., pathways central in long-term evolution. The second role corresponds to pathways more sensitive to external stimuli, i.e., pathways displaying short-term adaptation.

The last question addressed in this thesis is the assessment of FBA solutions in relation to all the feasible flux space, so as to identify whether solutions obtained with this technique describe reliably the set of possible metabolic states or, on the contrary, the FBA solution is uninformative of the entire set of metabolic phenotypes. The space of metabolic flux states can be exploited with different strategies. It can be used as a benchmark to calibrate the distance of FBA fluxes as compared to experimental measures, or to identify metabolic phenotypes unreachable by constraint-based techniques.

The main objectives of this thesis are summarized in the following bullet list:

- To study whether the structure of metabolic networks has evolved towards robustness resisting external perturbations, like gene or reaction removals or mutations.
  - To study the spreading of a cascade when a reaction or a pair of reactions fail, unveiling the interplay between multiple cascades.
  - To study the propagation of the damage to metabolism when genes fail.
  - To discuss the findings in terms of an evolutionary perspective.

- To study the effects on fluxes of individual and pairs of reactions knockouts using FBA.
  - To study the activity and essentiality of reactions.
  - To understand the mechanisms of synthetic lethality, unveiling the plasticity and the redundancy capabilities displayed by the metabolic networks of bacteria.
  - To study the dependence of plasticity and redundancy on the environment.
- To identify those pathways that perform important roles for the survival of an organism.
  - To check the efficiency of the disparity filter on metabolic networks.
  - To analyse backbones in terms of the long-term evolution of organisms
  - To extract information about the short-term adaptation of metabolism to the external environment.
- To assess the FBA solution in the entire space of metabolic solutions.
  - To demonstrate that solutions obtained using FBA as a constraint-based technique may be uninformative of typical behaviours.
  - To provide a benchmark to calibrate FBA.
  - To recover phenotypes not attainable by constraint-based techniques by using the full metabolic solution map.

1.4 Outline

After this introduction to cellular metabolism and its genome-scale models, Chap. 2 presents the general tools, methodologies, and GENREs used in this thesis.

Chapter 3 starts by considering a structural study of how metabolic networks of the bacteria *Mycoplasma pneumoniae*, *Escherichia coli*, and *Staphylococcus aureus* respond to internal perturbations, like removals of reactions or genes individually or in pairs, i.e., how the structure of the metabolic network is damaged following an internal failure which propagates as a cascade, by which the metabolic capabilities of an organism are weakened. Further, these results are linked to evolutionary explanations, i.e., how evolution has shaped and dictated the form of metabolic networks so as to respond to perturbations. This discussion is related with the robustness of organisms, in order to unveil whether the structure of the metabolic network is prepared to suffocate the advance of a damage cascade.

Chapter 4 extends the structural study of perturbations to flux distributions obtained using FBA. This study allows, on the one side, to know whether there are important reactions that must be always active in order to guarantee the survival of an organism and, on the other side, to check whether cell metabolism has developed protection mechanisms when some of its parts are unable to work. In this respect, synthetic lethal reaction pairs are analysed. These are pairs of reactions whose removal from is lethal, but metabolism is still able to survive when each reaction forming the pair is removed individually. This allows to identify two different mechanisms, plasticity and redundancy, which have helped to protect metabolism against possible reaction failures.

Chapter 5 analyses metabolic fluxes so as to extract more biological information on how organisms adapt to external environments and evolve. To perform this analysis, the disparity filter is used in order to obtain backbones as reduced versions of metabolism without losing its properties as a complex network. The structure of these backbones unveils pathways with a prominent role in the long-term evolution of the organisms and in their short-term adaptation to the environment.

Chapter 6 revises the FBA technique in relation to the whole set of feasible flux states in a metabolic network. FBA uses a strong assumption -organisms try to grow as much as possible-allowing to solve the mass action equations at steady state describing metabolism without the need of kinetic parameters. This assumption is commonly applied due to the lack of availability of kinetic constants of reactions. It is worth exploring the distribution of possible fluxes without making use of the assumption of maximal growth. This allows to perform a mapping of all the feasible flux solutions in metabolism and thus to assess the relevance of the solution obtained by FBA compared to all the other possible solutions.

General conclusions are given in Chap. 7. At the end of the thesis, there are four appendixes reviewing the basics of some specific tools used in Chaps. 3, 4, 5, and 6.

References

1. Lagerkvist U (2005) The enigma of ferment: from the philosopher's stone to the first biochemical Nobel Prize. World Scientific
2. Van Holde KE (1971) Physical biochemistry, foundations of modern biochemistry series. Prentice-Hall Inc., Englewood Cliffs
3. Segura D, Mahadevan R, Juárez K, Lovley DR (2008) Computational and experimental analysis of redundancy in the central metabolism of Geobacter sulfurreducens. PLoS Comput Biol 4(2):e36
4. Tucker CL, Fields S (2003) Lethal combinations. Nat Genet 35:204–205
5. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. Nat Rev Genet 5:101–113
6. Deville Y, Gilbert D, Van Helden J, Wodak SJ (2003) An overview of data models for the analysis of biochemical pathways. Brief Bioinform 4(3):246–259
7. Serrano MÁ, Boguñá M, Sagués F (2012) Uncovering the hidden geometry behind metabolic networks. Mol BioSyst 8:843–850
8. Kitano H (2002a) Systems biology: a brief overview. Science 295(5560):1662–1664
9. Kitano H (2002b) Computational systems biology. Nature 420(6912):206–210
10. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press
11. Alon U (2006) An introduction to systems biology: design principles of biological circuits. CRC Press
12. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427(6977):839–843

13. Borenstein E, Kupiec M, Feldman MW, Ruppin E (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci USA 105(38):14482–14487
14. Barrat A, Barthélemy D, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press
15. Newman M (2010) Networks: an introduction. Oxford University Press
16. Mathews CK, Van Holde KE, Ahern KG (2002) Bioquímica. Pearson Education
17. Lineweaver H, Burk D (1934) The determination of enzyme dissociation constants. J Am Chem Soc 56(3):658–666
18. Emmerling M et al (2002) Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. J Bacteriol 184(1):152–164
19. Steck TL (1974) The organization of proteins in the human red blood cell membrane. A Rev J Cell Biol 62(1):1–19
20. Wu G (2009) Amino acids: metabolism, functions, and nutrition. Amino Acids 37(1):1–17
21. Fitzpatrick LC (1976) Life history patterns of storage and utilization of lipids for energy in amphibians. Am Zool 16(4):725–732
22. Nishizuka Y (1995) Protein kinase C and lipid signaling for sustained cellular responses. FASEB J 9(7):484–496
23. Marrink SJ, Berendsen HJ (1994) Simulation of water transport through a lipid membrane. J Phys Chem 98(15):4155–4168
24. Bauchop T, Elsden SR (1960) The growth of micro-organisms in relation to their energy supply. J Gen Microiol 23(3):457–469
25. Good CA, Kramer H, Somogyi M (1933) The determination of glycogen. J Biol Chem 100(2):485–491
26. Neidle S (2010) Principles of nucleic acid structure. Academic Press
27. Carver JD, Allan Walker W (1995) The role of nucleotides in human nutrition. J Nutr Biochem 6(2):58–72
28. Bergman J (1999) ATP: the perfect energy currency for the cell. Creat Res Soc Q 36(1):2–9
29. Belenky P, Bogan KL, Brenner C (2007) NAD^+ metabolism in health and disease. Trends Biochem Sci 32(1):12–19
30. Pollak N, Dolle C, Ziegler M (2007) The power to reduce: pyridine nucleotides-small molecules with a multitude of functions. Biochem J 402:205–218
31. Gadd GM (1990) Heavy metal accumulation by bacteria and other microorganisms. Experientia 46(8):834–840
32. Ariño J, Ramos J, Sychrová H (2010) Alkali metal cation transport and homeostasis in yeasts. Microbiol Mol Biol Rev 74(1):95–120
33. Sychrova H (2004) Yeast as a model organism to study transport and homeostasis of alkali metal cations. Physiol Res 53:S91–98
34. Chen H, Ikeda-Saito M, Shaik S (2008) Nature of the Fe-O_2 bonding in oxy-myoglobin: effect of the protein. J Am Chem Soc 130(44):14778–14790
35. Tadege M, Dupuis I, Kuhlemeier C (1999) Ethanolic fermentation: new functions for an old pathway. Trends Plant Sci 4(8):320–325
36. Criado-Sancho M, Casas-Vázquez J (1998) Termodinámica química y de los procesos irreversibles. Pearson
37. Moffatt BA, Ashihara H (2002) Purine and pyrimidine nucleotide synthesis and metabolism. In: The arabidopsis book. American Society of Plant Physiologists, pp 1–20
38. Harvey WR, Cioffi M, Dow JA, Wolfersberger MG (1983) Potassium ion transport ATPase in insect epithelia. J Exp Biol 106(1):91–117
39. Michaelis L, Menten ML (1913) Die Kinetik der Invertinwirkung. Biochem Z 49(333–369):352
40. Hill AV (1910) The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. J Physiol (Lond) 40:4–7
41. Crane RK, Miller D, Bihler I (1961) The restrictions on possible mechanisms of intestinal active transport of sugars, pp 439–449
42. Prigogine I (1955) Thermodynamics of irreversible processes. Thomas

43. Demirel Y, Sandler SI (2002) Thermodynamics and bioenergetics. Biophys Chem 97(2):87–111
44. Desvergne B, Michalik L, Wahli W (2006) Transcriptional regulation of metabolism. Physiol Rev 86:465–514
45. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC (1985) Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 28(7):412–419
46. Lienhard GE, Slot JW, James DE, Mueckler MM (1992) How cells absorb glucose. Sci Am 266(1):86–91
47. Fell D (1997) Understanding the control of metabolism. Portland Press
48. Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature 465(7295):219–222
49. Ibarra RU, Edwards JS, Palsson BØ (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420:186–189
50. Fong S, Marciniak JY, Palsson BØ (2003) Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol 185(21):6400–6408
51. Schmidt S, Sunyaev S, Bork P, Dandekar T (2003) Metabolites: a helping hand for pathway evolution? Trends Biochem Sci 28(6):336–341
52. Kay E, Vogel TM, Bertolla F, Nalin R, Simonet P (2002) In situ transfer of antibiotic resistance genes from transgenic (transplastomic) tobacco plants to bacteria. Appl Environ Microb 68(7):3345–3351
53. Rando OJ, Verstrepen KJ (2007) Timescales of genetic and epigenetic inheritance. Cell 128(4):655–668
54. Lancaster AK, Masel J (2009) The evolution of reversible switches in the presence of irreversible mimics. Evolution 63(9):2350–2362
55. Lawrence JG (2005) Common themes in the genome strategies of pathogens. Curr Opin Genet Dev 15(6):584–588
56. Wodke JAH et al (2013) Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. Mol Syst Biol 9:653
57. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5(1)
58. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY (2012) Recent advances in reconstruction and applications of genome-scale metabolic models. Curr Opin Biotech 23(4):617–623
59. Güell O, Serrano MÁ, Sagués F (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In: Engineering of Chemical Complexity II. World Scientific Publishing, pp 39–56. ISBN 978-981-4616-12-6
60. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121
61. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucl Acids Res 28(1):27–30
62. Caspi R et al (2012) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucl Acids Res 40(D1):D742–D753
63. Schellenberger J, Park JO, Conrad TC, Palsson BØ (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinform 11:213
64. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28:245–248
65. Edwards JS, Palsson BØ (1999) Systems properties of the haemophilus influenzaerd metabolic genotype. J Biol Chem 274(25):17410–17416
66. Duarte NC, Herrgard MJ, Palsson BØ (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14:1298–1309
67. Becker SA, Palsson BØ (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5:8

68. Thiele I, Vo TD, Price ND, Palsson BØ (2005) Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. J Bacteriol 187:5818–5830
69. Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol Syst Biol 2:2006.0004
70. Jamshidi N, Palsson BØ (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26
71. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26(6):659–667
72. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol Syst Biol 9(1)
73. Fernández N, Díaz EE, Amils R, Sanz JL (2008) Analysis of microbial community during biofilm development in an anaerobic wastewater treatment reactor. Microb Ecol 56(1):121–132
74. Zomorrodi AR, Maranas CD (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Comput Biol 8(2):e1002363
75. Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. Mol Syst Biol 6(1)
76. Persidis A (1998) High-throughput screening. Nat Biotechnol 16(5):488
77. Bailey JE, Birnbaum S, Galazzo JL, Khosla C, Shanks JV (1990) Strategies and challenges in metabolic engineering. Ann NY Acad Sci 589(1):1–15
78. Park JH, Lee SY (2008) Towards systems metabolic engineering of microorganisms for amino acid production. Curr Opin Biotechnol 19(5):454–460
79. Zelle RM et al (2008) Malic acid production by Saccharomyces cerevisiae: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. Appl Environ Microbiol 74(9):2766–2777
80. Izallalen M, Mahadevan R, Burgard A, Postier B Jr, Didonato R, Sun J, Schilling CH (2008) Geobacter sulfurredonces strain engineered for increased rates of respiration. Metab Eng 10:267–275
81. Mesarović MD (1968) Systems theory and biology—view of a theoretician. Springer
82. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genom Hum G 2(1):343–372
83. Gierasch LM, Gershenson A (2009) Post-reductionist protein science, or putting Humpty Dumpty back together again. Nat Chem Biol 5(11):774–777
84. Westerhoff HV, Palsson BØ (2004) The evolution of molecular biology into systems biology. Nat Biotechnol 22(10):1249–1252
85. Kriete A, Eils R (2005) Computational systems biology. Academic Press
86. Kitano H (2005) International alliances for quantitative modeling in systems biology. Mol Syst Biol 1(1)
87. Krömer J, Quek LE, Nielsen L (2009) ^{13}C-Fluxomics: a tool for measuring metabolic phenotypes. Aust Biochem 40(3):17–20
88. Winter G, Krömer JO (2013) Fluxomics-connecting ‘omics analysis and phenotypes. Environ Microbiol 15(7):1901–1916
89. Kun A, Papp B, Szathmáry E (2008) Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. Genome Biol 9(3):R51
90. Nishikawa T, Gulbahce N, Motter AE (2008) Spontaneous reaction silencing in metabolic optimization. PLoS Comput Biol 4(12):e1000236
91. Guimerà R, Sales-Pardo M, Amaral LAN (2007) A network-based method for target selection in metabolic networks. Bioinformatics 23(13):1616–1622
92. Papp B, Teusink B, Notebaart RA (2009) A critical view of metabolic network adaptations. HFSP J 3(1):24–35

93. Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433:895–900
94. Serrano MÁ, Boguñá M, Sagués F (2011) Network-based confidence scoring system for genome-scale metabolic reconstructions. BMC Syst Biol 5:76
95. Güell O, Sagués F, Serrano MÁ (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621
96. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the mutiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106:6483–6488

Chapter 2
Methods and Data

This chapter describes the basics of the fundamental techniques used in this thesis. It is divided in three parts: (1) complex network tools applied to metabolism, (2) description of Flux Balance Analysis (FBA) -used to compute metabolic fluxes at steady state- and of Flux Variability Analysis -a variant of FBA to bound minimum and maximum fluxes for each reaction- and (3) a description of all the genome-scale metabolic reconstructions analysed in this thesis.

Nowadays, the explosion in computational power has allowed us to deal with systems of thousands or even millions of constituents and interactions, boosting the degree of our understanding on how these systems are structured and behave. Complex Network Science comprises a large amount of techniques and models which help us to study these intricate systems as a whole [1, 2]. These methodologies can be applied to any system which can be modelled as a network. Networks can be briefly defined as a set of items that interact, like for example the World Wide Web and the Internet and, in a biological context, metabolic networks [3] or protein–protein interaction networks [4]. Complex Network Science has led to an important advance in the understanding of metabolic networks [3, 5–9] which is in line with the systems-level view of metabolism in fields like Systems Biology [10, 11].

When dealing with metabolic networks, the complex network approach has to be combined with other techniques coming from Systems Biology in order to understand functional or behavioural features, for example why the inability to operate of some reactions leads to cell death, or why some reactions carry a determinate flux given a set of external nutrients. The most widespread mathematical approach used for the systems-level analysis of metabolic networks is Flux Balance Analysis (FBA) [12]. This technique is based on constraint-based analysis and optimization of an objective function, usually the biomass formation function of the cell. In this way, the fluxes through all the biochemical reactions of cell metabolism that maximize the biomass formation rate or, equivalently, the specific growth rate, can be computed. Apart from the mentioned reaction fluxes and growth rate, this technique allows to compute, for instance, the maximum yield of important compounds such as ATP or NADH [12, 13], and the effects of knockouts of genes or reactions [14, 15]. Related to FBA, other related techniques like Flux Variability Analysis (FVA) [16, 17] allow to

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_2

identify possible alternate solutions and, in conjunction with FBA, allow to perform a deep study of the flux capabilities of metabolic networks.

Complex network methodologies and constraint-based techniques applied to metabolic reconstructions represent a powerful tool for the analysis and development of new insights into metabolic functions and mechanisms that cells have developed from the earliest stages of life to the current days.

2.1 Structural Properties of Metabolic Networks as Complex Networks

Networks are discrete systems of elements that interact. These systems are represented by graphs of *nodes* (or vertices)—which represent elements—connected by *links* (or edges)—which represent interactions. The presence of a large number of nodes interacting in non-trivial connectivity patterns between order and disorder is what gives to networks their intrinsic complexity.

It is important to distinguish between complex and complicated. The main difference between these two words is better explained by a single example: solving a whole metabolic network composed of thousands of reactions is a complex problem in the sense that the large amount of interactions leads to emerging unexpected behaviours, like the effect of the removal of a biochemical reaction on other reactions, which can increase or decrease their fluxes depending on their biological activity. On the contrary, the study of a typical chemical engineering process to obtain a precise output may be a complicated problem, since one needs to draw a flowchart of all the chemical reactions and involved intermediate species that participate in the chemical synthesis. This may require a wide knowledge of the system, implying a large degree of control on all the processes, but the final behaviour of the system will be what is expected in a well-designed process.

2.1.1 Basic Representation Frameworks

Links in networks can have either a defined direction or may lack it. Therefore, when links are directed, they are depicted by arrows, specifying a source and a target. A directed link can represent, for example, a transformation between two metabolites, typically a reactant and a product with the link pointing to the product. When no specification source/target is prescribed, the interaction is mutual, like in a protein–protein interaction,[1] and links without direction are used. Associated to this, networks are classified as *directed*, *undirected*, or *semidirected*. It is worth mentioning that links can also be *bidirectional*, meaning that the interaction allows either the

[1] Protein–protein interactions refer to physical contacts established between two or more proteins as a result of biochemical events and/or electrostatic forces.

forward or backward direction at the same time and this interconnection is thus reciprocal. This is specially important in the context of metabolic networks, where reactions can be either reversible (bidirected links, meaning that both directions of the reaction are possible) or irreversible (directed links, meaning that only one direction is thermodynamically favoured). Moreover, a link can carry a weight, representing the intensity of the interaction. Therefore, networks can be *weighted* or *unweighted*. In the case of metabolic networks, weights usually correspond to fluxes of the biochemical reactions. Metabolic networks typically display a probability distribution of fluxes (or weights) that follows a power law, meaning that fluxes spanning different orders of magnitude coexist in the same metabolic state [18].

Mathematically, unweighted undirected networks are described by the adjacency matrix, a square symmetric matrix $\{a_{ij}\}$ of binary values with an entry of 1 whenever there is a link between nodes i an j and 0 otherwise. In directed networks, the matrix is instead non-symmetric. Weighted networks are encoded by the weighted adjacency matrix $\{\omega_{ij}\}$, in which the values correspond to the weight of the edge between nodes i and j.

Furthermore, networks can have different classes of nodes, leading to the so-called multipartite graphs. In multipartite graphs, links happen only between nodes in different categories. Networks with one kind of node are called *unipartite*, whereas networks with two kinds of nodes are called *bipartite* [19]. An important thing to notice is that bipartite networks can be projected into unipartite networks by performing a *one-mode projection*. To do this, one chooses a particular type of node and, in the projected reduction, places a link between two such nodes if there is at least one node of the complementary type connected to both of them.

In the real world, one can find networks combining all the mentioned properties (see Fig. 2.1). Metabolic networks are usually represented as bipartite semidirected networks, with metabolites and reactions belonging to different node categories with no direct connections between any two metabolites or any two reactions [21, 22] (see Fig. 2.1d). Although a bipartite representation is more accurate, it is sometimes preferable and always simpler to work with one-mode projections based on metabolites, which can be either directed or undirected (see Fig. 2.1a, b) depending on the reversibility of reactions, and weighted or unweighted depending on whether fluxes are taken into account. In such a projection, two metabolites get directly connected if there is at last one reaction in which they both participate (see Fig. 2.1f).

2.1.2 Degree Distribution

Nodes in networks are locally characterized by the number of their surrounding neighbours. This magnitude is called the *degree* of a node k (see Fig. 2.2). The probability of nodes having a certain degree k is written $P(k)$ and named *degree distribution*, and can be computed from the fraction of nodes in the network that has degree k.

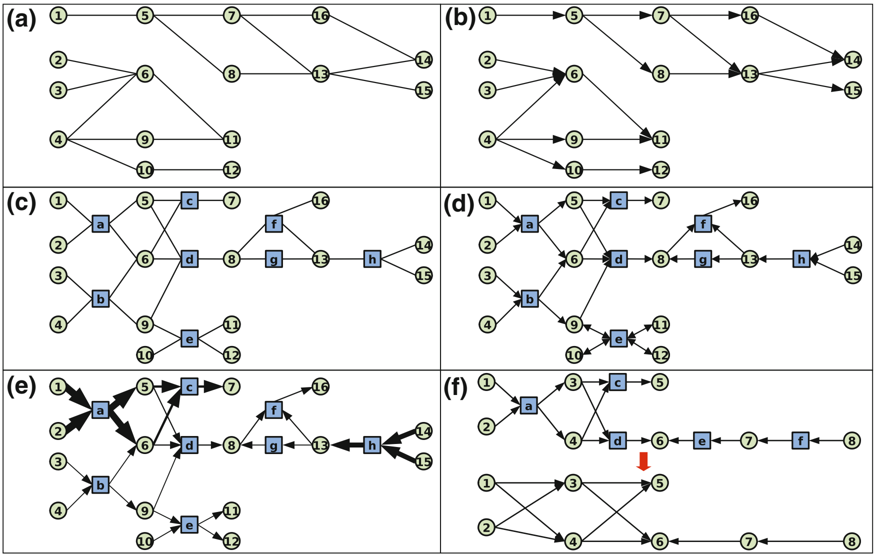


Fig. 2.1 Examples of different types of networks. **a** Undirected unipartite. **b** Directed unipartite. **c** Undirected bipartite. **d** Semidirected bipartite network. Notice that connections involving node *e* are bidirectional. **e** Semidirected weighted bipartite. The thickness of the links is proportional to their weight. **f** Example of the transformation into a one-mode projected network of metabolites from a semidirected bipartite metabolic network containing metabolites and reactions. Metabolites are represented by *circles* and reactions by *squares*. Parts of this figure have been extracted from Ref. [20] Copyright @ 2014, World Scientific Publishing

Usually, real world networks show degree distributions $P(k)$ that are highly skewed with long tails that reach values far above the mean [23]. In most cases, degree distributions follow a power-law, $P(k) \propto k^{-\gamma}$, where γ is the characteristic exponent and it has values in the range $2 < \gamma < 3$. Networks with a degree distribution described by a power-law are called *scale-free*.[2] Networks with power-law degree distributions have attracted much attention and have been studied intensively [24–26]. Notice that, usually, it is useful to work with the complementary cumulative probability distribution function $P(k' \geq k)$ in order to avoid noise effects present for large values of k.

In semidirected networks, the degrees of nodes are defined in relation to incoming (k_{in}), outgoing (k_{out}) and bidirectional (k_b) links. Correspondingly, nodes have a total degree expressed as a sum of contributions $k = k_{in} + k_{out} + k_b$. These degrees can present local correlations and so the degrees of nodes are described by the joint probability $P(k_{in}, k_{out}, k_b)$. In addition, for bipartite networks, nodes of each kind have also their own degree distribution.

[2] This name is refers to the scale-invariance that power-laws display: if $f(x) = a(x)^{\gamma}$, then $f(cx) = a(cx)^{\gamma} = c^{\gamma} f(x)$.

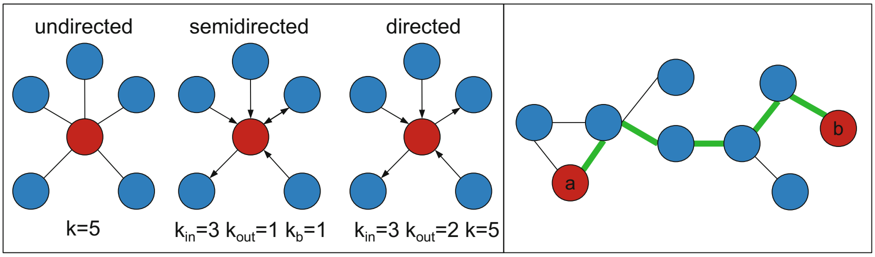


Fig. 2.2 Schematic example of a degree of a node (*left*) and a path between two nodes (*right*). *Left* Example of the degrees in a undirected, semidirected, and directed networks. *Right* Path between node a and b, highlighted in *green*. In this case, the shortest path length between nodes a and b is $\ell_{ab} = 5$ (color figure online)

Regarding specifically metabolic networks, the total degree of metabolites k_M in bipartite representations follows a power-law degree distribution $P(k_M) \propto k_M^{-\gamma}$ [1, 7]. In Ref. [3] it is found that in the organism *Escherichia coli*, the probability $P(k_{in})$ that a metabolite participates as a product and the probability $P(k_{out})$ that a metabolite participates as a reactant have both a value of γ of 2.2. Similarly, Ref. [27] shows that for the organism *Helicobacter pylori*, the exponent has a value of 2.32. The fact that metabolites display a scale-free degree distribution means that there is a high diversity in the number of reactions in which metabolites participate. The largest part of metabolites have a few connections, whereas a few metabolites, generically called hubs, have many of them. Examples of these highly-connected metabolites are ATP, H_2O, or H^+, which can participate in up to 50% of the total number of reactions for the case of H^+ in the organism *E. coli* [28]. On the contrary, reactions show a peaked distribution of total degree, the peak being located at an average degree $< k_R > \sim 4$. The bounded form of the distribution arises from the fact that reactions have a limited number of participants, typically from 2 to 12.

In Fig. 2.3a, the bipartite cumulative probability distribution function $P(k'_M \geq k_M)$ of metabolites and the bipartite probability distribution function $P(k_R)$ for reactions of the three organisms analysed in this thesis, *E. coli* [29–31], *Mycoplasma pneumoniae* [32, 33], and *Staphylococcus aureus* [34] are shown. Clearly, metabolites show a power-law degree distribution and reactions a peaked distribution, as mentioned above. In fact, all networks studied here have similar tendencies for both distribution functions, showing that metabolic networks, in spite of corresponding to quite different microorganisms, display often universal properties [3].

2.1.3 Average Path Length

Another common feature of complex networks, and in particular of metabolic networks, is the fact that any two nodes are connected by *paths* of links that are typically very short in the number of intermediate steps [7]. This is called the *small-world*

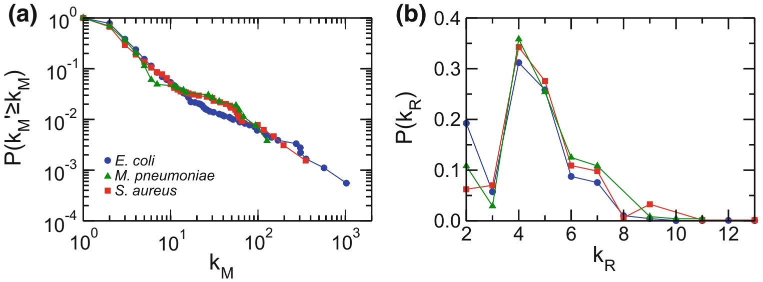


Fig. 2.3 Features of the networks of *Escherichia coli i*JO1366 (see Sect. 2.3.1), *Mycoplasma pneumoniae i*JW145 (see Sect. 2.3.2), and *Staphylococcus aureus i*SB619 (see Sect. 2.3.3). **a** Complementary cumulative probability distribution function of metabolites. **b** Degree distribution of reactions. Extracted from Ref. [35]

property. In technical terms, the distance ℓ between two nodes is defined as the number of jumps or hops along the shortest path that connects them (see Fig. 2.2). Hence, it is possible to define the average shortest path length $< \ell >$, which is the average of all the shortest distances between pairs of nodes. The small-world property is stated in the fact that $< \ell >$ increases as the logarithm of the network size N (number of nodes) [23, 25].

Small average path lengths indicate that the network contains highly-connected nodes that act as shortcuts, reducing the average number of steps needed to go from one node to another. This is crucial in many real contexts, and in particular for cell metabolism. In Ref. [3], the authors measured the average path length for 43 organisms and found a similar value for all of them, $< \ell > \sim 3.2$. This value was explained by the role of hubs, which decrease dramatically the number of steps needed to travel from one node to another. When hubs are not taken into account, longer and variable path lengths are obtained [36, 37], depending on the biological domain where organisms belong to. Typical values are 9.57, 8.50, and 7.22 for eukaryotes, archaea, and bacteria, respectively, with the differences due to evolutionary processes. Nevertheless, there remains some controversy about the small-world property in metabolic networks. In Ref. [38], it is stated that usually paths are computed by directly linking metabolites through reactions and that this is not adequate, since pathways computed in this way do not conserve their structural moieties[3] and thus they do not correspond to pathways on a traditional metabolic map. Therefore, in Ref. [38] metabolites are linked depending on the conserved structural moieties in the adjacent reactions and, as a result, it is stated that the average path length of *E. coli* metabolism is longer than it was previously thought and, consequently, the

[3] According to the IUPAC, a moiety is a part of a molecule that may include either whole functional groups or parts of functional groups as substructures.

E. coli metabolic network is not small in terms of biosynthesis and degradation of metabolites. However, it is generally accepted that metabolic networks show indeed the small-world property at the structural level. In this thesis, path lengths will be computed in Chap. 4.

2.1.4 Communities at the Mesoscale

It is thought that biological networks are composed by subsets of nodes that are functionally separable called *modules* [25, 39]. In general, this idea corresponds to the concept of communities in networks. The organization of a network into communities does not imply fragmentation. Instead, communities are subsets of a network which contain a dense interconnection pattern between nodes inside the community and lower interconnection levels with nodes outside. This can be related with the presence of a large clustering (see Sect. 2.1.6) between nodes inside the community.

Community detection [40] represents an active field in Complex Network Science motivated by the potential identification of communities with functional or operational units. Several methods, based on different exploratory techniques, have been proposed. Among the most successful community detection methods one finds, for instance, algorithms that use random walkers to partition the network into communities, like Infomap [41]. Other methods are based on the optimization of modularity. Modularity is a measure of the quality of a community structure [42]. It measures the internal connectivity of identified communities with reference to a randomized null model with the same degree distribution. Algorithms based on modularity optimization try to find the best community structure in terms of the modularity measure. Examples of successful algorithms based on this measure are SpinGlass [43] or Louvain [44]. On what follows, the three methods used in Chap. 3 of this thesis to detect communities are explained.

- *Distance hierarchical clustering*: this method starts by defining a distance between pairs of nodes in the network. Then, once the pairs of nodes have a defined distance, one groups similar nodes into communities according to this distance. There are different schemes based on distances to group nodes intro communities. The two simplest methods are single-linkage clustering, in which two sets of nodes are considered separate communities if and only if all pairs of nodes in the different sets have distance larger than a given threshold, and complete linkage clustering, in which all nodes of a community have a distance smaller than a threshold [45] (see Fig. 2.4).
- *Infomap algorithm* [41]: the main idea of this algorithm is that a random walker will tend to flow at different paces within a network, spending more time inside communities and less time to pass between them (see Fig. 2.4). The way in which the random walker moves around communities can be compared to the flow of messages between individuals. In this way, there is a strong current of messages between

individuals inside a community, and a weaker current of messages between individuals of different communities.

- *Recursive percolation*: this method has been developed in a work related to this thesis [35]. Recursive percolation identifies components in which the network is fragmented just below the percolation threshold (see Sect. 2.1.5), where the connected network disaggregates into smaller components. To find them, links are removed sequentially from lower to higher weights until the percolation transition is detected. Then, clusters are identified using a burning algorithm [46]. This procedure is applied to each component until the distribution of sizes of the obtained communities reaches some thresholds, for instance, to be similar to those given by the distance hierarchical clustering technique and Infomap. A schematic example of this process is shown in Fig. 2.4.

2.1.5 Large-Scale Connected Components

Global connectedness is one of the most fundamental properties of complex systems. The theory that describes the behaviour of network connected components is *percolation theory*. Briefly, percolation theory states that there exists a critical point, called *percolation threshold* and denoted as p_c, where a transition in the global connectedness of the network occurs, from a state where the network is formed by small isolated components to the emergence of a *giant connected component* (GCC) spanning a macroscopic fraction of the network. This means that it is always possible to find a path connecting every pairs of nodes inside the GCC.

This concept can be extended to networks with directed links. The connectivity of directed networks presents special features since the path between two nodes i and j can be different when going from i to j or vice versa. This fact leads to the existence of a bow-tie structure inside the GCC [22, 47, 48]. The main feature of the bow-tie structure of a GCC in a directed network is that one can detect the presence of a *strongly connected component* (SCC), which is a region of the network where any node is reachable from any other by a directed path. It can happen that directed networks contain more than one SCC.

Apart from the SCC, one of the other significant regions that can be found in the bow-tie structure of directed networks is called *IN component*, with nodes that can reach the SCC but that cannot be reached from the SCC. Analogously, the *OUT component* contains nodes that can be reached from the SCC but that cannot return to it. *Tubes* are sequences of nodes that connect the IN with the OUT component without going through the SCC. Finally, *tendrils* are composed by nodes that have no access to the SCC and that are not reachable from it, similarly to tubes. They go out from the IN component and come in from the OUT component. A visual scheme of the bow-tie structure of directed networks is shown in Fig. 2.5a. The bow-tie structure of *E. coli* and *Mycoplasma pneumoniae* will be explicitly considered in Chap. 5.

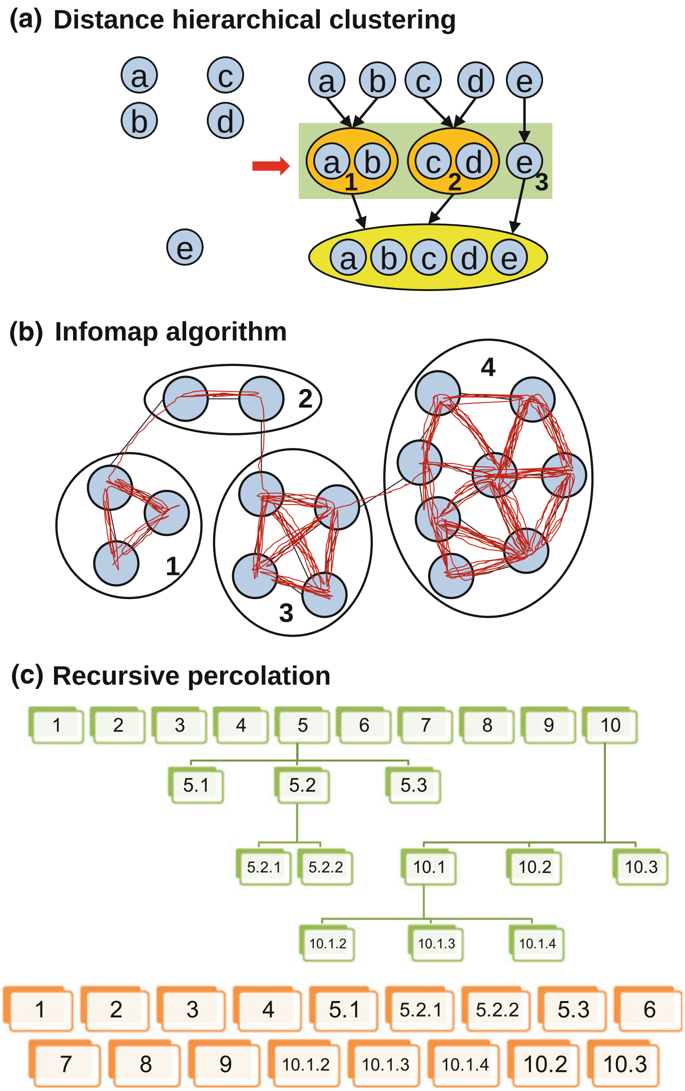


Fig. 2.4 Examples of the clustering methods. **a** Example of the distance hierarchical clustering method. Modules are formed by nodes that are nearer. Notice that with this method it is necessary to apply a threshold depending on the distances. In this example, the threshold is represented by the *green rectangle*. At this level, three communities are detected. **b** Example of the Infomap algorithm. Clusters are found with a random walker. Communities are found depending on the frequency of times that each random walker visits a set of nodes. **c** Example of the application of Recursive percolation. The first step leads to 10 clusters. Among these 10 clusters, the largest are fragmented, leading to more clusters. This partition is iterated until the distribution of sizes is similar to that in other methods. Parts of this figure have been extracted from Ref. [35] (color figure online)

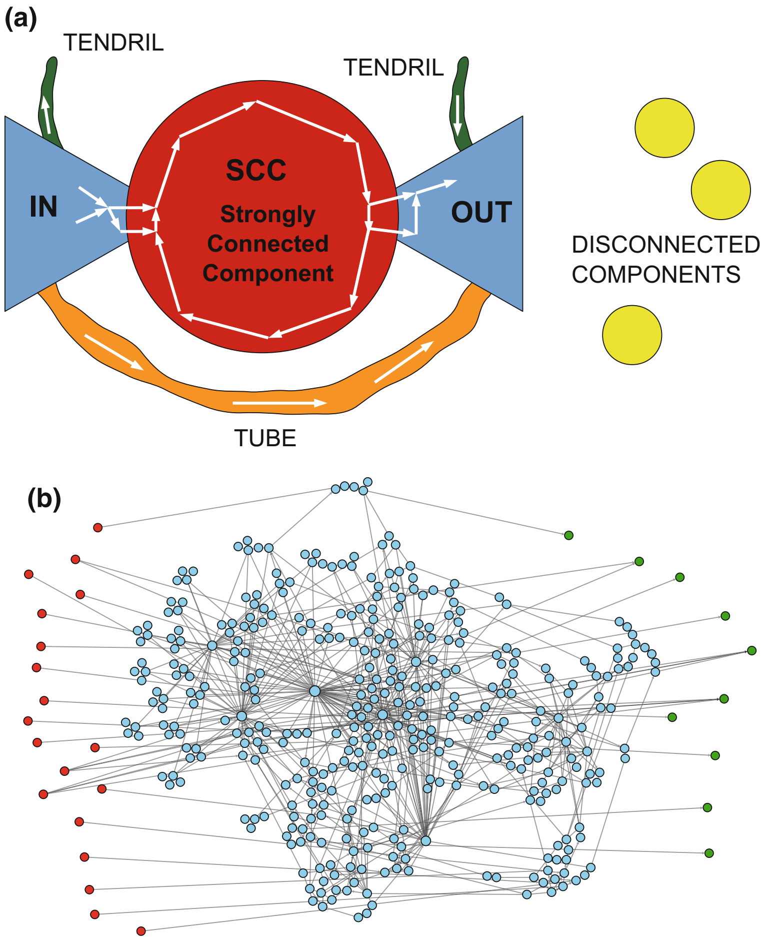


Fig. 2.5 Examples of connected components. **a** Schematic example of a bow-tie structure. **b** Example of the bow-tie structure of *Mycoplasma pneumoniae* [32], an organism studied in this thesis. *Blue nodes* compose the SCC, *red nodes* compose the IN component, and *green nodes* compose the OUT component (color figure online)

Metabolic networks show a bow-tie structure typically with a large SCC connected to non-structured IN and OUT components (see Fig. 2.5b) [47, 49]. The SCC contains the largest part of metabolites and reactions composing the network, representing thus the entire metabolic machinery of cells. IN and OUT components are formed of, respectively, nutrients and waste products directly connected to the SCC (see Fig. 2.5b).

2.1.6 Other Structural Properties of Complex Networks

Real networks exhibit also the presence of non trivial correlations in their connectivity. At the level of two nodes, it is convenient to characterize degree correlations with the average nearest neighbour degree $\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k)$, where $P(k'|k)$ is the probability of having a node with degree k' given that it is connected to a node with degree k. It basically considers the mean degree of the neighbours of a node as a function of its degree k. If $\bar{k}_{nn}(k)$ increases with k, it is said that the network is *assortative*, with nodes that connect preferentially to other nodes of similar degree. If $\bar{k}_{nn}(k)$ decreases with k, the network is named *disassortative*, with high-degree nodes attached preferentially to nodes with low degrees. Biological networks, and in particular metabolic networks, usually show a disassortative pattern [7].

Correlations among three nodes can be measured by means of the concept of *clustering*, which refers to the tendency to form triangles between the neighbours of a vertex. Watts and Strogatz [50] proposed a measure known as *clustering coefficient*, $c_i = \frac{2E_i}{k_i(k_i-1)}$, where E_i is the number of edges that exist between neighbours of the node i and k_i denotes the degree of the node i. Although this measure is helpful as a first indication for clustering, it is more informative to work with quantities which depend explicitly on the degree k. Therefore, a degree-dependent clustering coefficient $\bar{c}(k)$ is calculated as the clustering coefficient of nodes averaged for each degree class k. Metabolic networks tend to display high levels of clustering [5, 25] with $\bar{c}(k)$ having a decreasing dependence on k [6].

A final mention is deserved to structures called *motifs* [11]. Motifs are small subsets of connected nodes that are found in networks more often than expected at random. They are considered as elementary functional units, and each real network has its own set of distinct motifs. Their identification provides useful insights into the typical local connectivity patterns in the network.

2.1.7 Null Model Networks and Randomization Methods

Null models in Complex Network Science serve to study fundamental properties of complex networks and to asses the statistical significance of a property, first measuring it in the real network and then comparing the original results to the ones obtained in the randomized versions. These models can be used to prove the existence of graphs satisfying various properties, or to provide a rigorous definition of what it means for a property to hold for almost all graphs or, finally, to act as a benchmark for specific features of real networks.

One of the most known models was the graph structure proposed by Paul Erdös and Alfréd Rényi. The *Erdös–Rényi* model [51, 52] consists on generating realizations of random networks given the total number of nodes N and a total number of links L, and connecting every pair of them with probability p. This leads to a binomial degree distribution, that can be approximated by a Poisson distribution for realizations with a large number of nodes.

Another important method to construct random networks is the *Configuration model*, an algorithm to construct random networks with a degree sequence or degree distribution $P(k)$ settled a priori [53, 54]. The total number of nodes N remains constant. For each node, a random number k is drawn from the probability distribution $P(k)$ and it is assigned to the node in the form of half-edges. The network is then constructed by connecting pairs of these link ends chosen uniformly at random. These realizations, like the Erdös–Rényi networks described above, are uncorrelated and have no clusters in the thermodynamic limit $N \to \infty$.

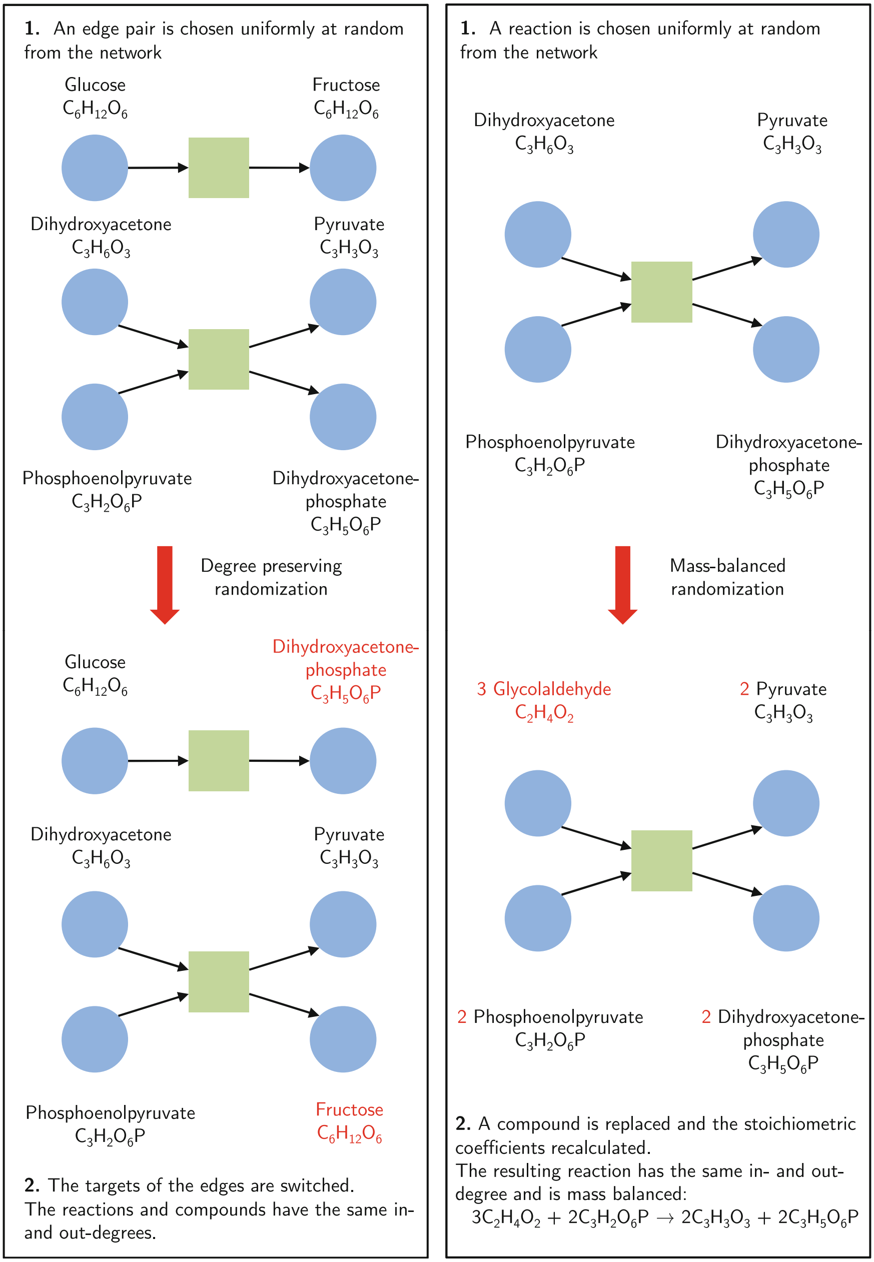
Instead of comparing real networks with null models as those described above, it is sometimes preferable to randomize a network obtained from real data by rewiring, i.e., by picking two links at random and swapping their end [55]. While randomizing, one can preserve different properties, for instance the degrees of all nodes. Two rewiring randomization methods have been used in this thesis, one that preserves the degrees of all nodes—similar to comparing with the Configuration model—called *degree-preserving* randomization, and another that generates randomized versions taking into account that new reactions must be stoichiometrically balanced, called *mass-balanced* randomization.

2.1.7.1 Degree-Preserving Randomization

In metabolic networks, the degree-preserving randomization method is similar to the Configuration model in bipartite networks. Degree-preserving randomization works by choosing two pairs of connected nodes (metabolites and reactions) of the bipartite network at random and swapping their ends, unless this would lead to a repeated metabolite in a reaction (see Fig. 2.6, left). The steps of the algorithm are:

1. Pick two links at random: $m_1 \to r_1$ and $m_2 \to r_2$ or $r_1 \to m_1$ and $r_2 \to m_2$, where m are metabolites and r reactions.
2. Swap the end of the links avoiding repeated links and self-production: ($m_1 \to r_2$ and $m_2 \to r_1$ or $r_1 \to m_2$ and $r_2 \to m_1$).
3. Repeat until L^2 swappings are performed, where L is the total number of links in the network.
4. Make several realizations of the randomized metabolic network following the three previous steps.

Reversible reactions are rewired independently of the irreversible ones in order to preserve the degrees of metabolites which correspond to reversible and irreversible reactions. This method gives networks which preserve the degrees of metabolites and reactions and it is useful, for instance, to determine the role of the degree distribution in large failure cascades in bacterial organisms, which may have evolved towards reducing the probability of having large cascades that produce metabolic damage, increasing thus robustness [56]. This method will be used in Chap. 3.



1. An edge pair is chosen uniformly at random from the network

2. The targets of the edges are switched.
The reactions and compounds have the same in- and out-degrees.

1. A reaction is chosen uniformly at random from the network

2. A compound is replaced and the stoichiometric coefficients recalculated.
The resulting reaction has the same in- and out-degree and is mass balanced:

$$3C_2H_4O_2 + 2C_3H_2O_6P \rightarrow 2C_3H_3O_3 + 2C_3H_5O_6P$$

Fig. 2.6 *Left* Scheme of the degree-preserving randomization algorithm. IN and OUT degrees are conserved, but mass balance is not satisfied. *Right* Scheme of the mass-balanced randomization. In this case metabolites are switched only if the new reaction is mass balanced; while reaction degrees are kept constant, the degrees of metabolites are not preserved. Extracted from Ref. [57] Copyright @ 2012, PACIS-JCIS

2.1.7.2 Mass-Balanced Randomization

Mass-balanced randomization generates randomized networks by rewiring the links corresponding to substrate-reaction or product-reaction relationships, while preserving atomic mass balance of the reactions [58]. Given a reaction r, its atomic mass balance is given by:

$$\sum_{e \in E_r} s_{e,r} \cdot m_e = \sum_{p \in P_r} s_{p,r} \cdot m_p \tag{2.1}$$

where E_r denotes the set of substrates and P_r the set of products in r, and m_e, m_p are the mass vectors ($m_{H_2O} = (0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$ as an example) of e and p, respectively. Finally, $s_{e,r}$, $s_{p,r}$ are their stoichiometric coefficients. For instance, consider the reaction $A \rightarrow B$, with $m_A = m_B = \mathrm{C_6H_{12}O_6}$. Then, A may be substituted by a compound C with $m_C = \mathrm{C_3H_6O_3}$ from within the network, resulting in the randomized reaction $2\ C \rightarrow B$, which satisfies Eq. 2.1 since $2\ \mathrm{C_3H_6O_3} = \mathrm{C_6H_{12}O_6}$ (see Fig. 2.6, right). In addition to substituting individual substrates or products, the method also allows more complex substitutions involving pairs of substrates or products, yielding a large number of possible substitutions.

The motivation for preserving atomic mass balance of reactions, a fundamental physico-chemical constraint, is that the resulting null model allows estimating the importance of network properties with respect to evolutionary pressure. As biological systems and their properties evolve under physical constraints and evolutionary pressure, a null model which satisfies physical principles but does not account for evolutionary pressure differs from a metabolic network only in the properties which are affected by evolutionary pressure. Thus, a property deemed statistically significant following mass-balanced randomization is beyond basic physical constraints and likely to be a result of evolutionary pressure [59]. The method preserves mass balance and reaction degrees but not the degrees of metabolites, since the stoichiometric coefficients and metabolite degrees are changed. This method will be used in Chap. 3.

2.2 Flux Balance Analysis

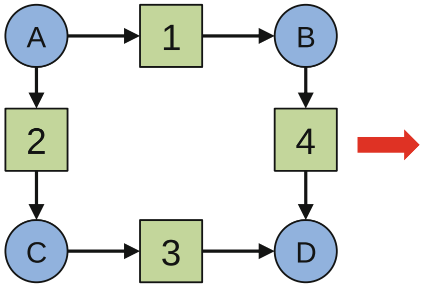
A general aim of the study of a metabolic network is to characterize and understand the configuration of fluxes of the reactions constituting the network in connection to phenotype and behaviour. The study of fluxes in metabolic networks deserves a special treatment more biochemically focused than in usual chemical kinetics schemes. With the knowledge of the kinetic constants of the reactions, it would be possible to solve the equations associated to the fluxes of reaction and the concentrations of metabolites in the metabolic network using proper mathematical methods. However, there is a lack in the availability of kinetic parameters [60] due to the difficulty in measuring them experimentally. As an alternative, computational techniques have been proposed in order to estimate fluxes through reactions of metabolic networks at steady-state.

Flux Balance Analysis is maybe the most successful and widely used approach to compute the fluxes through metabolic reactions of an organism. In addition, FBA also estimates its growth rate by maximizing the flux through the biomass reaction of the network. This technique will be used in Chaps. 4, 5, and 6.

To be more specific, metabolic reactions can be represented in terms of a *stoichiometric matrix*, this being the fundamental basis in FBA and other modelling approaches [12, 17, 61, 62]. To construct a stoichiometric matrix [63–65], one must first write the typical kinetic equations which describe the temporal variation of the concentration of metabolites, which are derived from the mass conservation principle,

$$\frac{dc_i}{dt} = \sum_{j=1}^{N_R} S_{i,j} \nu_j \tag{2.2}$$

The concentration of metabolite i is denoted by c_i, N_R is the total number of reactions, $S_{i,j}$ is the stoichiometric coefficient of metabolite i in reaction j, and ν_j stands for the flux of reaction j. Note that, typically, reaction fluxes have units of mmol gDW^{-1} h^{-1}, where gDW means grams Dry Weight. Notice that the values of the **S** matrix correspond to the stoichiometric coefficients of each metabolite in each reaction. Thus, each row represents a metabolite, whereas each column represents a reaction. Therefore, if a metabolite i does not participate in a reaction j, its stoichiometric coefficient will be 0, $S_{ij} = 0$. Otherwise, if the metabolite is a reactant, the stoichiometric coefficient will be negative, $S_{ij} < 0$, and if it is a product, it will be positive, $S_{ij} > 0$ (see Fig. 2.7).



$$\frac{dc_A}{dt} = -\nu_1 - \nu_2$$

$$\frac{dc_B}{dt} = +\nu_1 - \nu_4$$

$$\frac{dc_C}{dt} = +\nu_2 - \nu_3$$

$$\frac{dc_D}{dt} = +\nu_3 + \nu_4$$

$$\Rightarrow \frac{d\vec{c}}{dt} = \mathbf{S} \cdot \vec{\nu}$$

$$\vec{c} = \begin{pmatrix} c_A \\ c_B \\ c_C \\ c_D \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad \vec{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{pmatrix}$$

Fig. 2.7 Equations derived from mass-balance associated to a simple metabolic network. Matrix **S** is the so-called stoichiometric matrix, $\vec{\nu}$ is a vector containing all the fluxes of the metabolic network, and $\vec{c}$ denotes the vector with concentrations of metabolites

Metabolic networks are open-systems, which implies that some metabolites can leave or enter the organism. Therefore, it is not possible to arrive to a thermodynamic equilibrium state. However, it is possible to attain a non-equilibrium steady state, where the concentrations of metabolites do not change with time, forcing the system to exchange metabolites with the environment. This steady-state condition simplifies the system of coupled differential Equation 2.2 derived from mass balance into an ordinary linear system of equations, which can be written as a product of the stoichiometric matrix **S** by the vector of fluxes $\vec{\nu}$,

$$\mathbf{S} \cdot \vec{\nu} = \vec{0} \tag{2.3}$$

This is the typical form of the equation to be solved by the FBA technique. As mentioned before, it is important to notice that no kinetic parameters [66, 67] appear explicitly in Eq. 2.3 and, thus, they are not needed in relation to FBA applications.

It is important to precise that, apart from the intrinsic constraints imposed by the steady-state condition, other bounds of the form $\alpha_i \leq \nu_i \leq \beta_i$ may be imposed on the values of the fluxes to render the whole scheme both chemically and biologically realistic. These upper and lower bounds may depend on the thermodynamics of reactions, more precisely on their reversibility. If reactions are reversible, fluxes can have positive or negative fluxes, whereas for the case of irreversible reactions, reactions must have only positive fluxes. Further, since the steady-state condition forces the system to exchange metabolites with the environment, constraints on *exchange* fluxes are imposed for metabolites that can either enter or leave the organism. These exchange fluxes are taken positive from the system to the environment. Notice that fluxes obtained using FBA will depend on the particular chosen external medium.

In metabolic networks, there are usually more reactions than metabolites. The system of Eq. 2.3 is thus underdetermined, i.e., there are multiple solutions even after imposing the mentioned constraints. Therefore, a biological objective function is introduced to restrict the solution space to a single biologically meaningful solution. Technically, this means that FBA selects the state in the solution space that maximizes the value of the objective function (see Fig. 2.8). This objective function depends on the biological information that one wants to extract, but usually one chooses to optimize biomass formation adjusted to be equivalent to maximize the specific growth rate of the organism. To do this, a *biomass reaction* is added to the network which simulates the biomass production. Other possible objective functions are ATP or NADH production or yield.

Often, other auxiliary reactions are needed apart from exchange and the biomass formation reactions. The first category includes physiological requirements, like the *ATP maintenance* reaction, which is a reaction which consumes ATP in order to simulate biological energetic costs for the organism which are not associated to growth. A second category are the so-called *sink* reactions, which are reactions that have not been identified yet and that consume some metabolites to avoid accumulation. A generic sink reaction has the simple form $A \rightarrow \emptyset$.

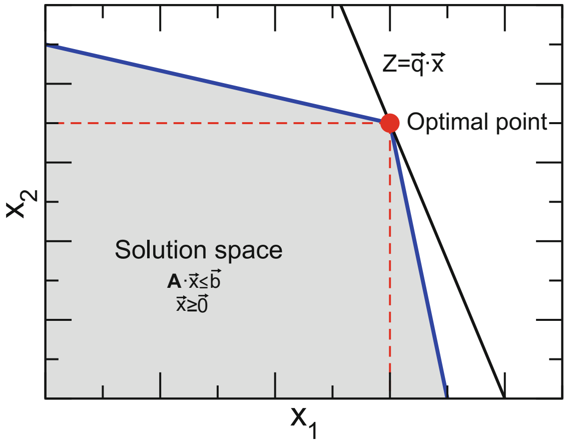


Fig. 2.8 Example of the optimization of an objective function on a system of two variables

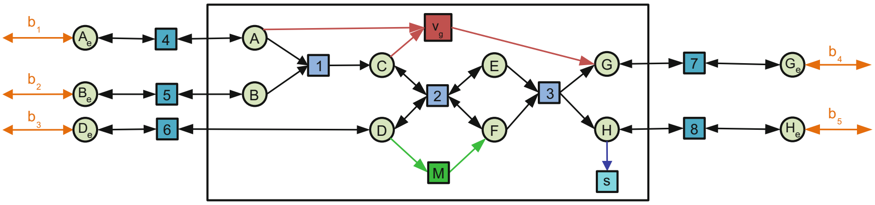


Fig. 2.9 Example of a FBA calculation in a metabolic network. Reactions are denoted by *squares* and metabolites by *circles*. The biomass production reaction (*red square*) is labelled as ν_g. Exchanges fluxes for interactions with the environment (*orange arrows*) are denoted with *b* labels. A sink reaction (*cyan square*) is shown with a *s* label. The ATP maintenance reaction (*green square*) is also shown denoted with a *M* label. Parts of this figure have been extracted from Ref. [20] Copyright @ 2014, World Scientific Publishing (color figure online)

In this way, a consistent system of equations representing the whole cell metabolism is obtained and one tries to find a solution that optimizes the value of an objective function (see Fig. 2.9 for a schematic picture of a FBA computation). If no solution exists for optimization of biomass production in a particular medium condition, one can assume that the system is not able to grow and therefore one can conclude that the organism is not able to survive in this medium.

The mathematical notation to denote a standard FBA problem choosing to optimize the specific growth rate is

$$\begin{array}{ll} \textbf{maximize} & \nu_g \\ \textbf{subject to} & \mathbf{S}\cdot\vec{\nu}=\vec{0} \\ \textbf{and} & \vec{\alpha}\le\vec{\nu}\le\vec{\beta} \end{array}$$

where $\vec{\alpha}$ and $\vec{\beta}$ represent the vectors determining the lower and upper bounds of the reactions, and ν_g denotes the specific growth rate.

The software used to perform these calculations in this thesis is GNU Linear Programming Kit (GLPK) [68–71], through its associated solver GLPSol. This solver uses a dual *simplex* algorithm to compute the solutions. It is a variant of the normal simplex algorithm [72]. The latter is an iterative algorithm which is based on finding first feasible solutions and then finding the most optimal solution based on these feasible solutions. On the contrary, dual simplex works by first finding optimal solutions and then finding a feasible solution, again, if it exists.

2.2.1 Formulation of the Biomass Reaction

FBA problems are usually solved by maximizing the flux through the biomass reaction [29, 31, 33]. This typically gives a particular flux state of the metabolic network compatible with the constraints. However, the solution obtained by FBA is often not unique. In some cases, the metabolic network is able to achieve the same specific growth rate by using alternate reactions and pathways. Therefore, phenotypically different solutions that optimize the specific growth rate are possible, implying that FBA solutions can be degenerate [12].

Technically, the biomass reactions is modelled as a reaction, $aA + bB + cC + dD... \longrightarrow xX + yY + zZ$, which produces and consumes some specific metabolites (see Fig. 2.9). These metabolites are known biosynthetic precursors present in the metabolic network under consideration. The key point is given by their stoichiometric coefficients in the biomass reaction, which are experimentally measured proportions in the biomass of the organism measured in dry weight conditions. The stoichiometric coefficients of the metabolites participating in the biomass reaction have units of mmol gDW^{-1}, and the biomass reaction has units of h^{-1}. It is worth stressing that this reaction simulates the growth of an organism given a set of external nutrients and that its coefficients are adjusted so that its flux is equivalent to the specific growth rate of the organism.

FBA can also maximize the biomass yield, which is the equivalent to maximize the specific growth rate but taking into account that the maximum uptake of the carbon source, for example glucose, must be set to 1 mmol gDW^{-1} h^{-1} to set the maximum amount of biomass that can be produced per 1 mol of nutrient.

2.2.2 Simulation of Different Environments

It is important to make explicit the way to simulate changes in the environment using FBA. To do this, one must tune the upper and lower bounds of the values of the exchange reactions of the metabolites that are present in the environment. As an example, suppose that one wants to model that glucose is present in the environment and that, therefore, the organism consumes it in order to obtain energy. The explicit form of the constraint of the exchange flux of glucose will be $-10 \leq \nu_{glucose}^{exchange} \leq \infty$, which means that the organism can expel as much as glucose as it wants but that it can eat glucose with a maximum uptake of 10 mmol gDW^{-1} h^{-1}.

Notice that nutrients have a negative lower bound and an unlimited upper bound, whereas waste products have a value of the lower bound of 0 and unlimited upper bound, which means that the organism cannot uptake it but, if the compound is generated inside the organism, it can be expelled to the exterior as waste. As an example, this would be the case for CO_2 in *E. coli*, which is not eaten by the organism but that is expelled due to respiration.

To summarize, an environment is simulated by choosing a set of nutrients and assigning a lower bound $-\alpha_i$ to each nutrient, which is the maximum uptake of each nutrient, and assigning a lower bound of 0 to components not present in the environment. For all external metabolites, the upper bound is set to ∞. Therefore, for nutrients one has $-\alpha \leq \nu_{nutrient}^{exchange} \leq \infty$, whereas for waste products one has $0 \leq \nu_{waste}^{exchange} \leq \infty$. The rest of reactions are modelled as told in the previous section.

2.2.2.1 Construction of Minimal Media

A minimal medium is the minimal set of metabolites which ensure the viability of an organism. The modelling of these media can be made as in Ref. [31]. Minimal media consist of a set of mineral salts, and one source of carbon, of nitrogen, of sulphur and of phosphorus, from four families representing carbon, nitrogen, phosphorus, and sulphur compounds, respectively. To construct different minimal media, the set of mineral salts is always the same—which contains, for example, magnesium sulphate, iron chloride, and calcium chloride-, but each source family is browsed while the other three sources are fixed to the standard metabolites of each kind (C*: glucose, N*: ammonia, P*: phosphate, S*: sulphate) (see Table 2.1).

Table 2.1 Examples of the construction of minimal media. Asterisks denote the standard metabolite of each kind. To construct carbon media, the sources of nitrogen, phosphorus and sulphur are set to the standard components of each kind whereas the carbon sources are varied. The same procedure applies to construct nitrogen, phosphorus and sulphur media

Variation of carbon sources					Variation of phosphorous sources				
Medium 1	C_1	N*	P*	S*	Medium 1	C*	N*	P_1	S*
Medium 2	C_2	N*	P*	S*	Medium 2	C*	N*	P_2	S*
Medium 3	C_3	N*	P*	S*	Medium 3	C*	N*	P_3	S*
Variation of nitrogen sources					Variation of sulphur sources				
Medium 1	C*	N_1	P*	S*	Medium 1	C*	N*	P*	S_1
Medium 2	C*	N_2	P*	S*	Medium 2	C*	N*	P*	S_2
Medium 3	C*	N_3	P*	S*	Medium 3	C*	N*	P*	S_3

2.2.2.2 Construction of Rich Media

Sometimes it can be useful to perform FBA computations in a medium with more components that the ones present in a minimal medium. These media containing more nutrients than a minimal medium are called rich media. One of this rich media is an amino acid-enriched medium. This medium can be constructed from a minimal medium with the standard metabolites explained in Sect. 2.2.2.1 (glucose, ammonia, phosphate, and sulfate), by adding the following set of amino acids: D-Alanine, L-Alanine, L-Arginine, L-Asparagine, L-Aspartate, D-Cysteine, L-Cysteine, L-Glutamine, L-Glutamate, Glycine, L-Histidine, L-Homoserine, L-Isoleucine, L-L-Leucine, L-Lysine, L-Methionine, L-Phenylalanine, L-Proline, D-Serine, L-Serine, L-Threonine, L-Tryptophan, L-Tyrosine, L-Valine. This set of amino acids enriches the minimal medium allowing the organism to take them as nutrients. Otherwise the organism would have to synthesize them, resulting in a more stringent environment for the organism. To simulate the presence of this set of amino acids in the medium, the exchange constraints bounds of these amino acids are set to -10 mmol/(gDW h).

Another famous rich medium is called Luria-Bertani Broth [73]. The Luria-Bertani Broth used in this thesis contains all the nutrients present in the amino acid-enriched medium, but it contains as additional compounds purines and pyrimidines, vitamins (namely biotin, pyridoxine, and thiamin), and also the nucleotide nicotinamide monocleotide [74]. The exchange constraints bounds of these compounds are usually set to -10 mmol/(gDW·h) ($\nu_{compound}^{exchange} \geq -10$) for *E. coli*.

2.2.3 *Activity and Essentialify of Genes and Reactions*

An important application of FBA is to compute the *activity* and *essentiality* of reactions in a network. These concepts can be applied either to genes or reactions, since a reaction is catalysed by an enzyme which at the same time is codified by a gene or a set of genes. Both concepts will be analysed in Chap. 4.

The concept of activity is quite simple. A reaction is said to be active when, given an external environment, the chosen reaction carries a non-zero flux. The concept of essentiality is more subtle. It refers to how a network, and thus the growth rate, is affected when one reaction is forced to be non-operative through the knockout of a reaction or of the corresponding gene.

To calculate the effect of the knockout of a reaction, the selected reaction is removed from the network, which is equivalent to force the chosen reaction to have a null flux. The new system is usually called a mutant. In terms of the notation used before, this is modelled as $\nu_i = 0$ with i the removed reaction and $\nu_i = 0$ its flux. Thus, a FBA problem with a reaction i constrained to be non-active is

$$\begin{array}{ll} \textbf{maximize} & \nu'_g \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = \vec{0} \\ \textbf{and} & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \\ & \nu_i = 0 \end{array}$$

where ν'_g denotes the growth rate of the mutant. As a consequence, the system can respond in three different ways as compared to the non-perturbed case ν_g:

1. The growth rate is unaltered $\nu'_g = \nu_g$.
2. The growth rate is decreased $0 < \nu'_g < \nu_g$, which means that the biomass formation of the organism is reduced but the organism is still alive at the expense of losing some performance.
3. The growth rate takes a null value $\nu'_g = 0$, meaning that the performed knockout is lethal for the organism. This is the signature of essentiality.

It has been shown that FBA predicts gene essentiality with an accuracy of 90% [29] in *E. coli* under glucose aerobic conditions, which means that FBA is a reliable tool to predict whether a knockout will be lethal or not in this particular condition.

2.2.4 Flux Variability Analysis

Sometimes it is useful to identify which are the minimum and maximum bounds that each reaction can take independently of the growth optimality condition. In this way, one can have an idea of the flux space for a particular environmental condition, and in particular which reactions can have a non-zero flux in a given environment, since some reactions may be active for low values of the growth rate but the same reactions must have a zero flux in order to ensure growth optimality. This may happen due to the fact that some reactions can compete with the growth reaction by consuming metabolites needed to grow and therefore this would reduce the flux through the biomass reaction. As a consequence, when one optimizes the flux through the biomass reaction, all reactions whose activation competes with the flux of the biomass reaction will have a null value in order to assure maximum growth conditions.

In addition to identifying those reactions that can compete with the growth rate, reactions whose minimum and maximum values are close indicate that they may be important for the organism since those reactions are allowed only to have a low variability in their fluxes. To know the minimum and maximum flux values of a reaction, one applies the technique called Flux Variability Analysis (FVA) [16, 17, 75].

In most applications of FVA, the biomass reaction is imposed to have a minimum value $\nu_g \geq \nu_g^{min}$ to ensure viability. Hence, one can consider that the limiting fluxes correspond to states where the organism is alive, even if the growth rate is not the maximum value that the organism can achieve. Using the mathematical notation used in Linear Programming computations, FVA for each flux of a metabolic reaction can be written as follows

$$\begin{array}{llll} \textbf{minimize} & \nu_i & \textbf{maximize} & \nu_i \\ \textbf{subject to} & \mathbf{S}\cdot\vec{\nu} = 0 & \textbf{subject to} & \mathbf{S}\cdot\vec{\nu} = 0 \\ & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} & & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \\ & \nu_g \geq \nu_g^{min} & & \nu_g \geq \nu_g^{min} \end{array}$$

However, it may happen that one is interested in capturing all the possible scenarios independently of the value of the flux of the biomass reaction, since in this way non-optimal/low-growth scenarios can be taken also into account. Therefore, FVA can be modified to compute the minimum and maximum possible values of the flux of each reaction regardless of the value of the biomass formation rate. To this end, the value of the flux of the biomass reaction is not constrained and any positive value is allowed, $\nu_g \geq 0$. Under this condition, one will obtain the maximal set of reactions that can be active in the considered medium independently of the rate of biomass formation. This variation of FVA [76, 77] will be used in Chaps. 4 and 5. Using the previous notation, this version of FVA, which we call *Biomass unconstrained Flux Variability Analysis*, can be written as

$$\begin{array}{llll} \textbf{minimize} & \nu_i & \textbf{maximize} & \nu_i \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 & \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 \\ & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} & & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \\ & \nu_g \geq 0 & & \nu_g \geq 0 \end{array}$$

2.3 Model Organisms

Information about metabolism of specific organisms [31, 33, 34, 78–85]—most single cell—are gathered in databases, like the BiGG database [86], Kyoto Encyclopedia of Genes and Genomes (KEGG) [87], BioCyc/EcoCyc/MetaCyc [88], BRENDA [89], etc. The BiGG database deserves special attention in this thesis, since it has been extensively used as it contains full reconstructions of metabolic networks for specific organisms including all the biochemical reactions and the biomass formation function in order to compute FBA solutions for different organisms.

The BiGG database provides high-quality curated information. Network reconstructions coming from this database are structured in compartments like cytosol inside cells or periplasm—the space bordered by the inner and the outer membranes in Gram-negative bacteria. Therefore, metabolites present in different compartments of the organisms are treated as different nodes. Using different compartments allows the inclusion of transport systems in both the inner and outer membrane and thus the metabolic machinery of organisms is more accurately represented. As an example, water in the periplasm will be a different metabolite than water in the cytosol. In addition, a directed bipartite representation of the metabolic network can be constructed since in the databases reactants, products, reversible, and irreversible reactions are distinguished. Further, the BiGG database specifies which enzyme catalyses each reaction and also which gene or set of genes codifies each enzyme. However, reactions are also listed which have neither associated enzymes nor genes. It may be that, for these particular reactions, enzymes have not been identified yet or that some reactions are spontaneous and they can take place without the need of an enzyme.

It is important to notice that there exist different versions for each metabolic network of each organism. This happens due to the fact that the reconstructions of metabolic networks are constantly improved and, therefore, versions are constantly updated. As an example, the first version of *E. coli* [90] contained 660 genes, 627 reactions, and 438 metabolites, while the last version of *E. coli* [31] contains 1366 genes, 2250 reactions, and 1805 metabolites.

2.3.1 *Escherichia coli*

Escherichia coli, abbreviated as *E. coli*, is the most studied prokaryotic organism and it is the bacterial model that is most frequently used in experiments due to the ease of its manipulation. More precisely, the strain studied in this thesis is K-12 MG1655. This strain colonizes the lower gut of animals. Moreover, it has been maintained as a laboratory strain with minimal genetic manipulation.

Three versions of this strain have been used in this thesis. The first one is *i*AF1260, which can be obtained either from Ref. [29] or directly from the BiGG database. This version is based on an earlier reconstruction called *i*JR904 [91], on the annotation of the genome of *E. coli* from Ref. [92], on contents from the EcoCyc (an *E. coli* version of BioCyc) database [93] and on specific biochemical characterization studies from Ref. [29]. The *i*AF1260 version contains 2077 reactions, 1669 metabolites, and 1260 genes [29] (see Table 2.2). Metabolites are located in three compartments: exterior, periplasm and cytosol. Notice that although the exterior is not a real compartment, it is treated in this way in order to be able to use the exchange reactions explained before.

The most recent version of *E. coli* is *i*JO1366 [31]. It is an update of the *i*AF1260 version. EcoCyc [94] and the KEGG database [95] were used in order to improve the *i*AF1260 version, in addition to experimental techniques [31]. It contains 2250 biochemical reactions, 1805 metabolites, and 1366 genes [31] (see Table 2.2). Like in *i*AF1260, metabolites are located in cytosol, periplasm and exterior.

A simplified version called core *E. coli* metabolic model is also used, which can be obtained either from Refs. [12, 30] or the BiGG database. It is a condensed version of the genome-scale metabolic reconstruction *i*AF1260 that contains 73 metabolic reactions in central metabolism, 72 metabolites, and 136 genes (see Table 2.2). This network is complemented with a biomass formation reaction and an ATP maintenance reaction.

2.3.2 *Mycoplasma Pneumoniae*

Mycoplasma pneumoniae, abbreviated as *M. pneumoniae*, is a human pathogen of primary atypical pneumonia that has recently been proposed as a genome-reduced model organism for bacterial and archaeal systems biology [32, 33, 96, 97]. Interest in this organism has grown recently since it lacks many anabolic processes and rescue pathways compared to more complex organisms. This in turn translates into a highly linear metabolism singularly suited to study basic metabolic functions [33]. This property will be again mentioned in Chaps. 3 and 4.

The first version of the metabolic network of *M. pneumoniae* used in this thesis was published in Ref. [32], where the authors integrated biochemical and computational studies, complementing the information using the KeGG database. Its metabolic reconstruction contains 187 reactions taking place in cytosol and in exterior, the number of metabolites is 228, and the number of genes is 140 (see Table 2.2).

The *i*JW145 version of *M. pneumoniae* is the last update [33]. This network was constructed by determining the behavior of the organism under different nutrition conditions, using literature information and experimental data. It contains 240 biochemical reactions, 266 metabolites, and 145 genes (see Table 2.2). Metabolites can be located in cytosol and exterior.

2.3.3 Staphylococcus aureus

Staphylococcus aureus, abbreviated as *S. aureus*, is found in the human respiratory tract and on the skin. It is an anaerobic bacterium which is present world-wide, and it is a common cause of skin infections, respiratory disease, and food poisoning. The strain used in this thesis is N315, a major pathogen which is able to acquire antibiotic-resistance [98].

The *i*SB619 version of *S. aureus* can be obtained either from the BiGG database or from Ref. [34]. To construct this model, the authors used the KeGG database and the Comprehensive Microbial Resource (CMR) at The Institute for Genomic Research (TIGR) website [99]. Missing functions were annotated based on reported evidence from this organism, as well as for *Bacillus subtilis* and *E. coli*. The number of reactions is 642 and the number of metabolites is 644 (see Table 2.2). Like in *M. pneumoniae*, there are only cytosol and exterior compartments.

Table 2.2 Summary of the properties of all metabolic reconstructions used in this thesis. N_R, N_M, and N_G stand for the number of reactions, metabolites, and metabolic genes respectively. Metabolites in different compartments are treated as different metabolites

Organism	N_R	N_M	N_G	Source
*E. coli i*AF1260	2077	1669	1260	Ref. [29], BiGG
*E. coli i*JO1366	2250	1805	1366	Ref. [31]
E. coli core model	73	72	136	Refs. [12, 30], BiGG
M. pneumoniae	187	228	140	Ref. [32]
*M. pneumoniae i*JW145	240	266	145	Ref. [33]
*S. aureus i*SB619	642	644	619	Ref. [34], BiGG

References

1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Newman M (2010) Networks: an introduction. Oxford University Press, New York
3. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. Nature 407:651–654
4. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42
5. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc R Soc Lond B 268:1803–1810
6. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
7. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. Nat Rev Genet 5:101–113
8. Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. BioEssays 27:176–188
9. Motter AE, Gulbahce N, Almaas E, Barabási AL (2008) Predicting synthetic rescues in metabolic networks. Mol Syst Biol 4:168
10. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, New York
11. Alon U (2006) An introduction to systems biology: design principles of biological circuits. CRC Press, Boca Raton
12. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28:245–248
13. Varma A, Palsson BØ (1993) Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165(4):477–502
14. Suthers PF, Zomorrodi A, Maranas CD (2009) Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol 5:301
15. Barve A, Rodrigues JFM, Wagner A (2012) Superessential reactions in metabolic networks. Proc Natl Acad Sci USA 1091:E1121–E1130
16. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
17. Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. BMC Bioinform 11:489
18. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427(6977):839–843
19. Guillaume JL, Latapy M (2006) Bipartite graphs as models of complex networks. Phyics A 371:795–813
20. Güell O, Serrano MÁ, Sagués F (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In: Engineering of Chemical Complexity II. World Scientific Publishing, Singapore, pp 39–56. ISBN 978-981-4616-12-6
21. Holme P, Liljeros F, Edling CR, Kim BJ (2003) Network bipartivity. Phys Rev E 68(5):056107
22. Ma HW, Zeng AP (2003a) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19:270–277
23. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
24. Barabási AL, Bonabeau E (2003) Scale-free networks. Sci Am 288(5):50–59
25. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(21):4947–4957
26. Keller EF (2005) Revisiting "scale-free" networks. BioEssays 27(10):1060–1068
27. Tanaka R (2005) Scale-rich metabolic networks. Phys Rev Lett 94(16):168101
28. Kim P et al (2007) Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. Proc Natl Acad Sci USA 104(34):13638–13642
29. Feist AM et al (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121

30. Orth JD, Fleming RM, Palsson BØ (2009) EcoSal—*Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, DC
31. Orth JD et al (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. Mol Syst Biol 7:535
32. Yus E et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
33. Wodke JAH et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol Syst Biol 9:653
34. Becker SA, Palsson BØ (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5:8
35. Güell O, Sagués F, Serrano MÁ (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621
36. Ma HW, Zeng AP (2005) Reconstruction of metabolic networks from genome information and its structural. Computational systems biology. Academic Press, New York
37. Kriete A, Eils R (2005) Computational systems biology. Academic Press, New York
38. Arita M (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101(6):1543–1547
39. Gao JT, Guimerà R, Li H, Pinto IM, Sales-Pardo M, Wai SC, Rubinstein B, Li R (2011) Modular coherence of protein dynamics in yeast cell polarity system. Proc Natl Acad Sci USA 108(18):7647–7652
40. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174
41. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105:1118–1123
42. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
43. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110
44. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 10:P10008
45. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
46. Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, Englewood Cliffs
47. Ma HW, Zeng AP (2003b) The connectivity stucture, giant strong component and centrality of metabolic networks. Bioinformatics 19:1423–1430
48. Boguñá M, Ángeles M (2005) Generalized percolation in random directed networks. Phys Rev E 72:016106
49. Serrano MÁ, De Los P (2008) Structural efficiency of percolated landscapes in flow networks. PLoS ONE 3:e3654
50. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
51. Erdös P, Rényi A (1959) On random graphs I. Publ Math Debr 6:290–297
52. Erdös P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
53. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithm 6:161–179
54. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:026118
55. Milo R et al (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827
56. Smart AG, Amaral LAN, Ottino J (2008) Cascading failure and robustness in metabolic networks. Proc Natl Acad Sci USA 105:13223–13228

57. Güell O, Sagués F, Basler G, Nikoloski Z, Serrano MÁ (2012) Assessing the significance of knockout cascades in metabolic networks. J Comp Int Sci 3(1–2):45–53
58. Basler G, Ebenhöh O, Selbig J, Nikoloski Z (2011) Mass-balanced randomization of metabolic networks. Bioinformatics 27:1397–1403
59. Basler G, Grimbs S, Ebenhöh O, Selbig J, Nikoloski Z (2012) Evolutionary significance of metabolic network properties. J R Soc Interface 9:1168–1176
60. Costa RS, Machado D, Rocha I, Ferreira EC (2011) Critical perspective on the consequences of the limited availability of kinetic data in metabolic dynamic modelling. IET Syst Biol 5(3):157–163
61. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17(2):53–60
62. Price N, Reed J, Papin J, Wiback S, Palsson BØ (2003) Network-based analysis of metabolic regulation in the human red blood cell. J Theor Biol 225(2):185–194
63. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5(1):320
64. Terzer M, Maynard ND, Covert MW, Stelling J (2009) Genome-scale metabolic networks. Wiley Interdiscip. Rev. Syst. Biol. Med. 1(3):285–297
65. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol Syst Biol 9(1):661
66. Schilling CH, Palsson BØ (1998) The underlying pathway structure of biochemical reaction networks. Proc Natl Acad Sci USA 95:4193–4198
67. Schilling CH, Edwards JS, Letscher D, Palsson BØ (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnol Bioeng 71:286–306
68. Makhorin A (2001) GNU linear programming kit. Moscow Aviation Institute, Moscow
69. Ceron R (2006) The GNU linear programming kit, Part 1: introduction to linear optimization. IBM, Raleigh
70. Ceron R (2006b) The GNU linear programming kit, Part 2: intermediate problems in linear programming. IBM, Raleigh
71. Ceron R (2006c) The GNU linear programming kit, Part 3: advanced problems and elegant solutions. IBM, Raleigh
72. Murty KG (1983) Linear programming, vol 57. Wiley, New York
73. Sezonov G, Joseleau-Petit D, D'Ari R (2007) Escherichia coli physiology in Luria-Bertani broth. J Bacteriol 189:8746–8749
74. Wunderlich Z, Mirny LA (2006) Using the topology of metabolic networks to predict viability of mutant straints. Biophys J 91:2304–2311
75. Müller AC, Bockmayr A (2013) Fast thermodynamically constrained flux variability analysis. Bioinformatics 29:903–909
76. Reed JL, Palsson BØ (2004) Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14(9):1797–1805
77. Güell O, Sagués F, Serrano MÁ (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol 10(5):e1003637
78. Duarte NC, Herrgard MJ, Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14:1298–1309
79. Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T (2004) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol Syst Biol 2:2006
80. Duarte NC, Becker SS, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104(6):1777–1782

81. Jamshidi N, Palsson BØ (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26
82. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2008) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143
83. Senger RS, Papoutsakis ET (2008) Genome-scale model for *Clostridium acetobutylicum*: part I. Metabolic network resolution and analysis. Biotechnol Bioeng 101(5):1036–1052
84. Raghunathan A, Reed J, Shin S, Palsson BØ, Daefler S (2009) Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. BMC Syst Biol 3(1):38
85. Thiele I et al (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31(5):419–425
86. Schellenberger J, Park JO, Conrad TC, Palsson BØ (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinform 11:213
87. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
88. Caspi R et al (2012) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 40(D1):D742–D753
89. Schomburg I et al (2012) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res 41:D764–D772
90. Edwards JS, Ibarra RU, Palsson BØ (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97:5528–5533
91. Reed JL, Vo TD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54
92. Riley M et al (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. Nucleic Acids Res 34(1):1–9
93. Keseler IM et al (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 33(suppl 1):D334–D337
94. Keseler IM et al (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res 37(suppl 1):D464–D470
95. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(suppl 1):D355–D360
96. Kühner S et al (2009) Proteome organization in a genome-reduced bacterium. Science 326:1235–1240
97. Güell M et al (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326:1268–1271
98. Kuroda M et al (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357(9264):1225–1240
99. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. Nucleic Acids Res 29(1):123–125

Chapter 3
Structural Knockout Cascades in Metabolic Networks

This chapter presents the analysis of the response of metabolic networks of model organisms to different forms of structural stress, including removals of individual and pairs of reactions and knockouts of single or co-expressed genes. Local metabolite motifs can be used as predictors of failure cascade sizes caused by individual failures, and for amplification effects in cascades caused by multiple failures. Correlation between gene essentiality and damages produced by single gene knockouts is detected, which points out that genes controlling high-damage reactions tend to be expressed independently of each other. This study is carried out for three characteristic organisms: *Mycoplasma pneumoniae*, *Escherichia coli*, and *Staphylococcus aureus*.

The architecture of complex networks is imprinted with universal features that affect their resilience and condition their behaviour [1–3]. Most relevant, the scale-free connectivity (see Chap. 2, Sect. 2.1.2) of many natural and man-made networks explains their fragility in front of attacks to the most connected nodes, while they are able to deal with accidental failures of single components [4, 5]. A manifestation of this fragile yet robust nature of complex networks is that the failure cascade triggered by a local shock rarely propagates to the whole system [6–9]. At the same time, it is worth to remember that network studies have mainly focused on single node failures, and that systemic responses to more globalized forms of structural and functional stress still remain to be explored.

In a more biological context, metabolic networks are among the best probed in terms of robustness in front of a variety of *in silico* perturbation experiments. They have been found to comply with the design principles of error-tolerant scale-free networks [10], and recent progress in network dynamics is also starting to portray the concept of stress-induced network rearrangements [11, 12]. The exploration of single biochemical reaction inactivations has shown that when a reaction is forced to be non-operative, a cascade of consequent failures propagates to a variable extent trough the whole network, and that the structural organization of metabolic networks reduce the likelihood of large damaging cascades [13]. At the same time, many individual mutations affecting enzyme-coding genes seem to have very little effect on cell growth [14, 15]. By contrast, the impact of multiple failures could go beyond

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_3

the mere accumulation of individual effects, producing amplified damage due to peculiar biochemical interweaving or gene epistatic interactions [16].

The analysis presented in this chapter considers the removal of single and pairs of biochemical reactions and the knockout of individual genes and clusters of co-expressed genes in three bacteria, *Mycoplasma pneumoniae*, *Escherichia coli*, and *Staphylococcus aureus*. To simulate the effect of reaction knockouts, a cascading failure algorithm [13] is used and the significance of the obtained results is assessed using two different null models called degree-preserving randomization (DP) (see Chap. 2, Sect. 2.1.7.1) and mass-balanced randomization (MB) (see Chap. 2, Sect. 2.1.7.2). One finds that, for the three organisms, the sizes of cascade distributions span a broad range of values, with many short propagations but a few that spread at the systems level. *M. pneumoniae* exhibits similar network responses to *E. coli* and *S. aureus*, although its increased linearity and reduced redundancy [17] threaten its robustness against individual reaction removals. For all three organisms, the impact of failure cascades can be predicted in terms of local network motifs. In this way, targets prone to introduce structural vulnerability can be readily detected prior to experimental testing without expensive computations, even for large and complex organisms. This chapter also reports the effects of single and multiple gene knockouts in *M. pneumoniae* by coupling, through enzyme activity, its metabolic network to the experimentally measured gene co-expression network. One observes that genes related to high-damage reactions are essential for the organism and that their expression tends to be isolated from that of other genes. This hints at the interplay between metabolism and genome, apparently evolved to favour the robustness of this organism by avoiding the potentially catastrophic effect of coupling the co-expression of structurally vulnerable metabolic genes. At the same time, one finds that this enables the organisms the ability to perform more efficient metabolic regulation at the expense of losing some of the maximum attainable robustness determined by physico-chemical constraints.

The contents of this chapter correspond to Ref. [18], to Ref. [19] Copyright @ 2012, PACIS-JCIS (reproduced with kind permission from PACIS-JCIS), and to Ref. [20].

3.1 Cascading Failure Algorithm

It is important to start by explaining how the cascading failure algorithm works after a reaction or a set of them are inactivated. First of all, the metabolic networks of *M. pneumoniae*, *E. coli* (*i*AF1260), and *S. aureus* (*i*SB619) (see Chap. 2, Sect. 2.3) are modelled as a bipartite semidirected network (see Chap. 2, Sect. 2.1.1), with two specific criteria:

1. All biochemical reactions in the genome-scale metabolic reconstructions (GENREs) are considered except exchange, sink, biomass formation, and ATP maintenance reactions.

2. All metabolites involved in the reactions included in the network representation are considered. In particular, hubs participating in a huge number of reactions are not excluded. Hubs stay neutral with respect to structural cascades and do not contribute to propagate them. Due to their large number of connections, they are highly unlikely to become non-viable as a consequence of single or double cascades reaching them.

The cascading failure algorithm [13] is based on the states of the nodes on the network, *i.e.*, nodes can be viable or non-viable. Non-viable nodes spread the perturbation, whereas viable do not. To define viability, two aspects are considered since a bipartite representation of the metabolic network is used. The first one refers to metabolites, and consists on the fact that each viable metabolite must have at least one outgoing and one incoming connections so as to prevent accumulation or depletion of the metabolite. For reactions, the criterion is that all metabolites participating in a reaction must be viable.

The algorithm works by removing one or more reactions, and then checking the viability of its surrounding metabolites. If they are viable, the cascade stops, otherwise the cascade is spread into other reactions and metabolites until all remaining nodes satisfy the mentioned criteria (see Fig. 3.1). When the cascade stops, the corresponding damage is quantified as the number of reactions turned non-operational.

Reversible reactions (see Chap.1, Sect. 1.1.2) deserve a special treatment in this algorithm. They are decoupled in two half-nodes, the forward and the reverse direction. A cascade propagating to a metabolite of a reversible reaction fixes it in the forward or reverse direction depending on whether the single incoming or outgoing link left to the affected metabolite is connected to the forward or reverse half of the reaction (see Fig. 3.1, step 3). In all cases, when any metabolite of a reversible reaction has this reaction as the single one producing and consuming it, the reaction must be removed to satisfy the viability criterion.

3.2 Impact of Reaction Failures

This first result is the distribution of damages for cascades triggered by individual and by pairs of reactions in the metabolic networks of *M. pneumoniae*, *E. coli*, and *S. aureus*. Later, local network motifs responsible for the propagation of cascades are identified, and a local predictor for damage is proposed.

3.2.1 Impact of Individual Reactions Failures

Although close to 50% of all individual reaction failures in the three organisms considered do propagate cascades, most cascades are indeed small (59% of the cascades in *M. pneumoniae*, 38% in *S. aureus*, and 55% in *E. coli* propagate to only one or two

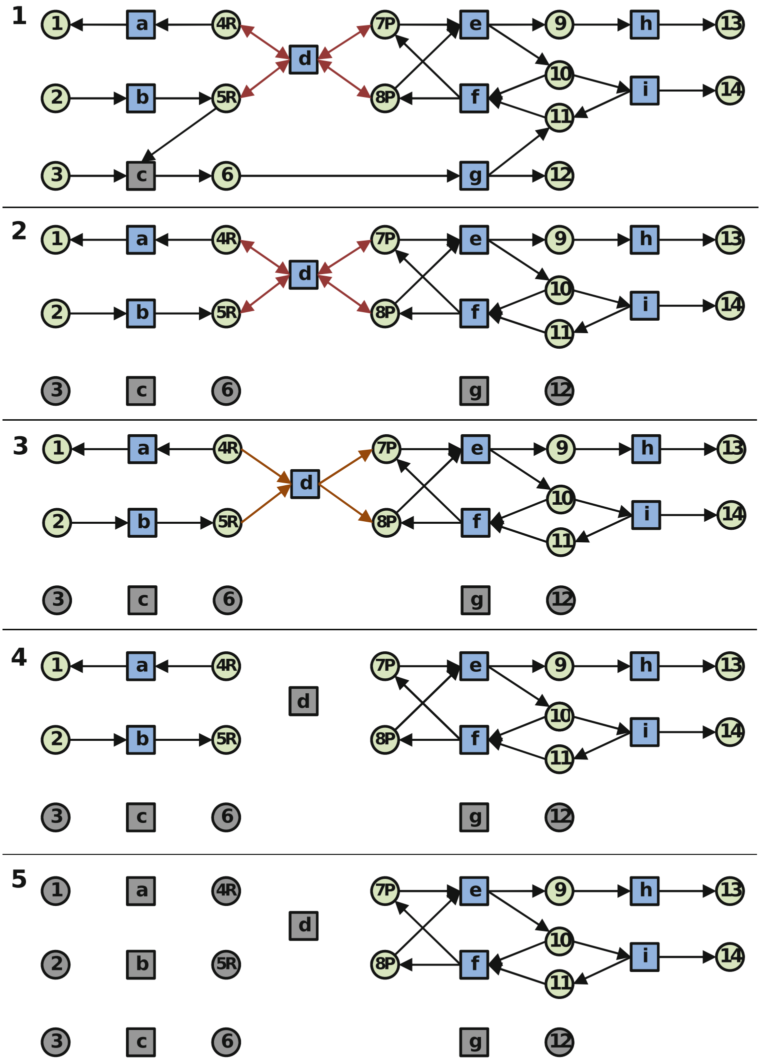


Fig. 3.1 Example of how the cascading failure algorithm is applied to a metabolic network. (1) For clarity, metabolites 4 and 5 are labelled with R and 7 and 8 with P depending on whether they are reactants or products of the reversible reaction denoted *d* (for simplicity, only a reversible reaction is considered in this illustration, the rest being assumed to be irreversible). The cascade starts when reaction *c* fails. (2) Therefore, metabolites 3 and 6 become non-viable. Because metabolite 6 is connected to reaction *g*, the later becomes non-viable, turning also metabolite 12 non-viable. Notice that metabolite 11 loses one IN connection, but it is still viable, meaning that one of the waves of the cascade stops here. However the other wave keeps spreading. (3) Metabolite 4 causes the reversible reaction *d* to remain viable only towards the production of metabolites 7 and 8. (4) Consequently, metabolite 4 becomes non-viable, and so its associated reactions also become non-viable. (5) The cascade spreads until all metabolites and reactions affected by the cascade remain viable. Finally, note that metabolites 1, 2, 3, 13, and 14, which initially have no incoming or outgoing connections, are not considered non-viable by the algorithm. Extracted from Ref. [19] Copyright @ 2012, PACIS-JCIS

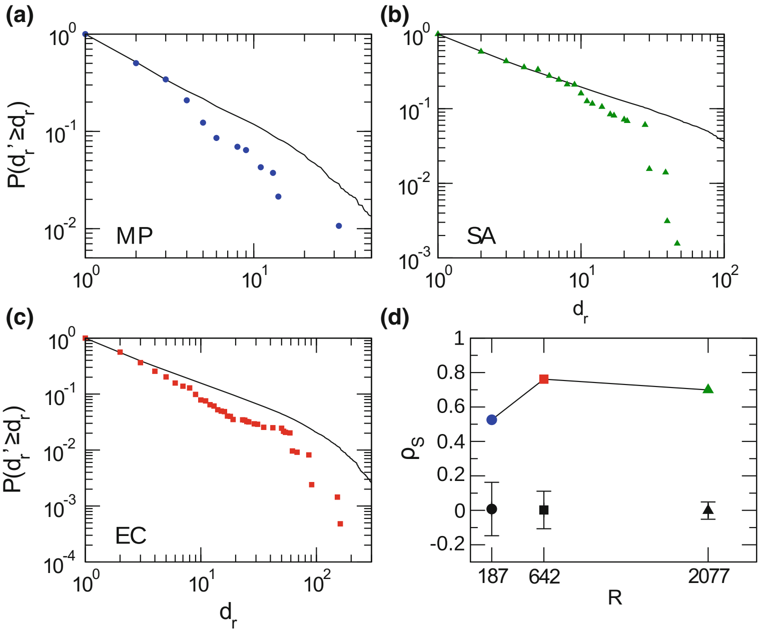


Fig. 3.2 Damage in cascades triggered by individual reactions. **a**–**c** Cumulative probability distribution functions of damages in *M. pneumoniae*, *E. coli*, and *S. aureus*. Results are compared with damages produced in DP randomized versions of the metabolic networks in order to discount structural effects. In each case, the *solid black curve* is the average over 100 realizations. Results for *S. aureus* and an older version of *E. coli* were already presented in Ref. [13]. The results of the Kolmogorov-Smirnov tests are given in terms of the K-S statistic and its associated significance level (K-S statistic/associated significance level) (see Appendix A): 0.095/0.07, 0.086/0.0002, and 0.079/1.4 $\cdot 10^{-11}$ for *M. pneumoniae*, *S. aureus*, and *E. coli* respectively. With a significance value of $\alpha = 0.05$, distributions of damages can be considered not consistent with those for randomized variants, except for *M. pneumoniae*. **d** Spearman's rank correlation coefficient ρ_S between predictors and damages, plotted against metabolic network size (number of reactions R). Results are compared to random reshuffling of the predictor value associated to reactions (100 realizations for each organism). Average Spearman's rank correlation coefficients for the randomizations appear in *black*, and *error bars* delimit the maximum and the minimum values obtained. Extracted from Ref. [18]

reactions). However, the removal of some particular reactions may trigger relatively far reaching damages. This is shown in Fig. 3.2a–c, that display the cumulative probability distributions $P(d_r' \geq d_r)$ that the failure of a reaction r attains at least $d_r - 1$ other reactions in each metabolic network. All species show similar broad distributions, although the crossover in the tail of the distribution from power-law-like to exponential-like is not evident in *M. pneumoniae* probably due to its limited redundancy. In order to assess the significance of cascades, the computed distributions are compared with those corresponding to DP randomized variants of the metabolic networks taken as null models (see Chap. 2, Sect. 2.1.7.1).

To check consistency, Kolmogorov-Smirnov (K-S) tests [21] (see Appendix A) are performed measuring the maximum absolute difference between the null model and the empirical distributions (see caption of Fig. 3.2 for specific values). This difference is transformed into a significance level directly compared to a chosen threshold, typically $\alpha = 0.05$. If the significance associated to the K-S test statistic is equal or smaller than α, the compared distributions cannot be considered consistent. Both *E. coli* and *S. aureus* display values much below the threshold, meaning that the empirical distributions are not determined just by the connectivity imposed by the degrees of metabolites. Comparing both distributions, the metabolic organization of the organisms appears to have evolved towards reducing the likelihood of large failure cascades (probably lethal for the organisms) or, equivalently, towards increased structural robustness, as previously seen for *S. aureus* and for an older version of the metabolic network of *E. coli* in [13]. In contrast, the value of the associated significance level for *M. pneumoniae* is very similar to the threshold. As a consequence, one cannot say that the difference between cascade size distributions in the original network and in the randomized counterparts is statistically significant, even though the probability for large cascades is still smaller in the original metabolic network. This can be explained by the increased linearity and limited redundancy of *M. pneumoniae* metabolic network structure, according to available data [17].

Along with structure, biochemical insight contributes to explain why some reactions trigger larger cascades. For *M. pneumoniae*, the most vulnerable reactions can be classified into four groups related to vital functions. One group is associated to metabolites phosphoenolpyruvate and protein L-histidine, each solely produced by one generating reaction and both of them directly related to phosphorylation processes, vital for instance in the synthesis of ATP. The second group relates to formate, which has a prominent role in the energy metabolism on many bacteria. The third group involves reactions where the important metabolite is thioredoxin, an antioxidant protein essential to reduce oxidized metabolites, along NADP$^+$. Finally, the failure of reactions in the fourth group trigger large cascades that affect the synthesis of fatty acids by turning acyl carrier proteins non-viable.

Prediction of the size of the cascades is possible by looking to the local information corresponding to the triggering reaction. An expression for the predictor P_r for the damage spreading from the triggering reaction r which is surrounded by m metabolites is:

$$
\begin{aligned}
P_r = \sum_{m \in r} \Big[& (k_i + k_b)\delta^0_{k_o}(\delta^1_{k_b} + \delta^0_{k_b})(\delta^1_{k'_o - k_o} + \delta^1_{k'_b - k_b}) \\
& + (k_o + k_b)\delta^0_{k_i}(\delta^1_{k_b} + \delta^0_{k_b})(\delta^1_{k'_i - k_i} + \delta^1_{k'_b - k_b}) \\
& - \delta^0_{k_i}\delta^0_{k_o}\delta^1_{k_b}\delta^1_{k'_b - k_b} \Big] .
\end{aligned} \tag{3.1}
$$

Degrees k_i, k_o and k_b refer respectively to the number of incoming, outgoing and bidirectional links of metabolite m (reactant or product) associated to the triggering reaction r after discounting the links used to propagate the cascade, and k'_i, k'_o and

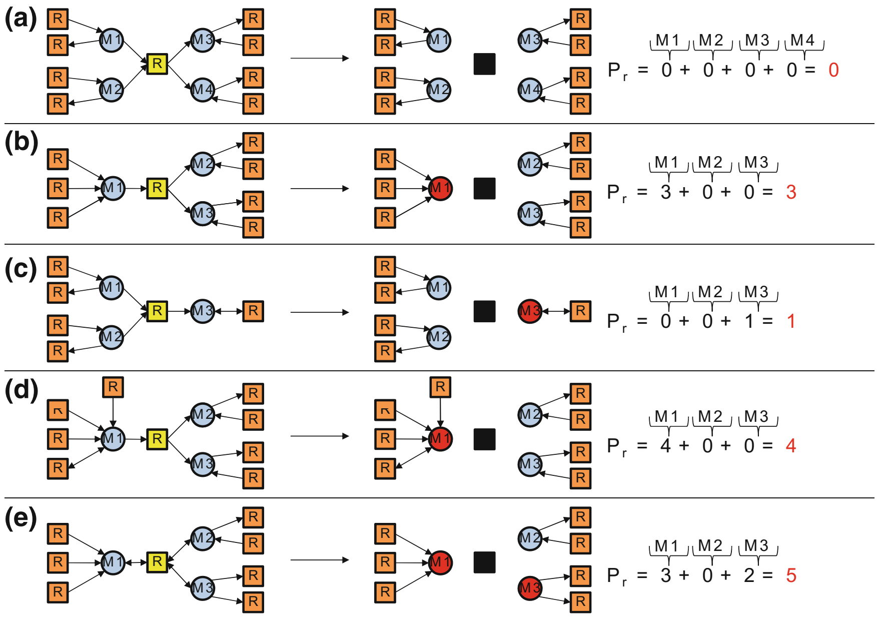


Fig. 3.3 Examples of application of Eq. 3.1 to several configurations of metabolites and reactions. Triggering reactions are coloured *yellow*, whereas metabolites which spread the cascade are coloured *red*. For clarity, the contribution of each metabolite to the value of P_r is also given. Extracted from Ref. [18] (color figure online)

k'_b denote the original values before the cascade is triggered. δ_a^b are used for the Kronecker's delta function. Basically, this predictor identifies metabolites susceptible to propagate the cascade, which are those having originally just one IN or just one OUT link, which is the one connecting them to the triggering reaction, or those connected to the triggering reaction by a *bidirectional* link and lacking in or out connections. The contribution to the predictor of one of those metabolites counts the number of connections of this metabolite with the rest of reactions, which can then be considered susceptible to become non-viable and propagate the cascade (see Fig. 3.3 for illustrations showing how the measure works for some particular cases).

Propagator motifs are represented by branched metabolites with just one in or out connection that happens to be attached to the triggering reaction. The higher the branching ratio of these metabolites, the higher the likelihood that the reaction propagates a large cascade, and thus to become a target for structural vulnerability in the network. To give an example, the two most vulnerable reactions in *M. pneumoniae* produce phosphoenolpyruvate, a compound involved in Glycolysis and Gluconeogenesis that acts as a source of energy. It happens to be a highly-branched cascade propagator motif connected to two reversible reactions and, as a product, to eight irreversible reactions (see Fig. 3.4 for a categorization of cascade propagator motifs in bipartite networks).

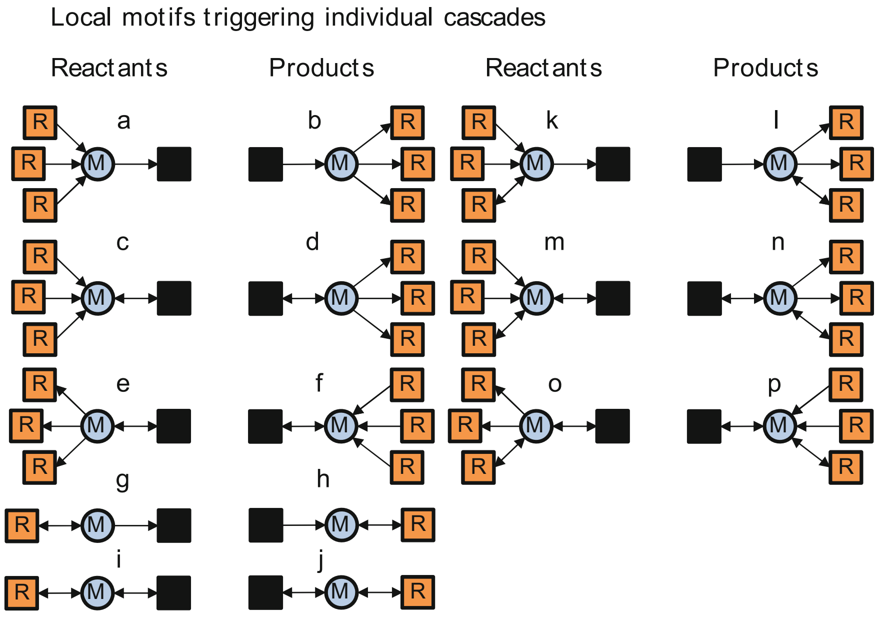


Fig. 3.4 Motifs of cascade propagation after failure of individual reactions. Cases **a**–**j** result into cascades with d_r larger than 1, while cases **k**–**p** correspond to potential transmitters in the sense that they may or may not spread the cascade. Extracted from Ref. [18]

To check the predictive power of our predictor P_r, Spearman's rank correlation coefficient ρ_S between predictors and damages are measured for each organism (see Appendix B). Basically, Spearman's correlation [22] is the Pearson correlation coefficient between two ranks, here given by the positions in ordered lists of reactions according to predictor values P_r and damages d_r. A high ranking position by predictor value is expected to correlate with vulnerable reactions at the top of the damage ranking. For all three organisms, very high values of the correlation coefficient are found, which are statistically significant (see Fig. 3.2d). This evidences the ability of this predictor, calculated on the basis of local information, to rank reactions by damage without directly computing the effect of the failure.

3.2.2 Non-linear Effects Triggered by Pairs of Reactions Cascades

As expected, the simultaneous failure of two reactions leads to higher damages compared to single reaction failures as shown in Fig. 3.5. The graphs display the cumulative probability distributions $P(d'_{rr'} \geq d_{rr'})$ calculated from all possible pairs of reactions initiating the cascades. It is worth stressing that the order of initiation is

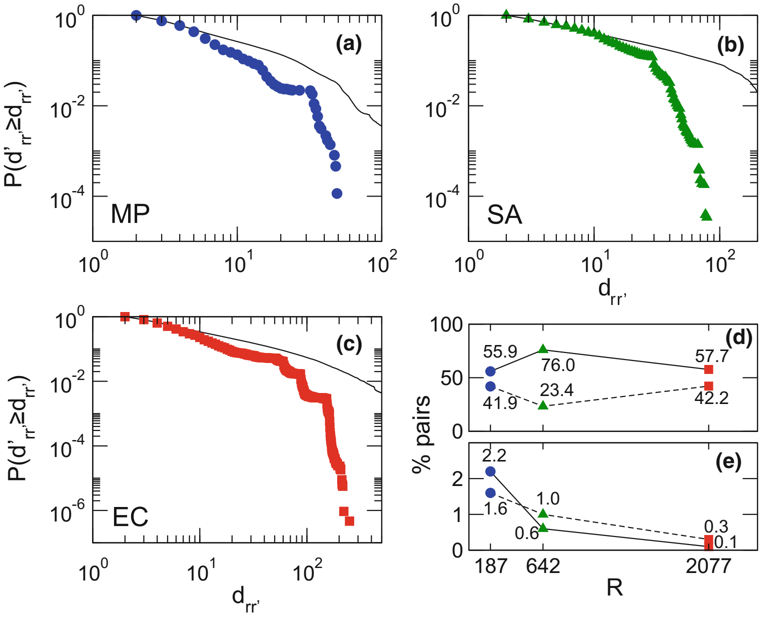


Fig. 3.5 Damage in cascades triggered by pairs of reactions. **a**–**c** Cumulative probability distribution functions of damages in *M. pneumoniae*, *E. coli*, and *S. aureus*. Results are compared with damages produced in DP randomized versions of the metabolic networks in order to discount structural effects. In each case, the *solid black curve* is the average over 100 realizations. Results of the Kolmogorov-Smirnov tests (K-S statistic/associated significance level) (see Appendix A): 0.15/0, 0.14/0, and 0.13/0 for *M. pneumoniae*, *S. aureus*, and *E. coli* respectively. Taking $\alpha = 0.05$, distributions of damages can be considered not consistent with those for randomized variants. **d** Most frequent double cascades output. *Solid line* interference without amplification. It is related with cases **b** and **c** in Fig. 3.6. *Dashed line* no interference, which is related with case **a** in Fig. 3.6. **e** Non-linear effects in double cascades. *Solid line* overlap. It is related with cases **c** and **e** in Fig. 3.6. *Dashed line* amplification. Amplification is related with cases **d** and **e** in Fig. 3.6. Extracted from Ref. [18]

irrelevant. Notice that the exponential cut-off is still present, and becomes prominent even for *M. pneumoniae*. Again, metabolic robustness is assessed by comparing cascades in the original networks with those in DP randomized counterparts using K-S tests (see caption of Fig. 3.5 for specific values). One finds that, for all three organisms including *M. pneumoniae*, the probability for large cascades triggered by pairs of reactions is significantly smaller in the original metabolic networks as compared to those in the randomized variants, suggesting that the organization of metabolic networks has evolved towards protecting metabolism against multiple reaction failures.

It can also be observed that cascades caused by individual reactions combine in different ways when two reactions fail simultaneously (see Fig. 3.6). The crucial concept here is that of the pattern of interference of the respective areas of influence

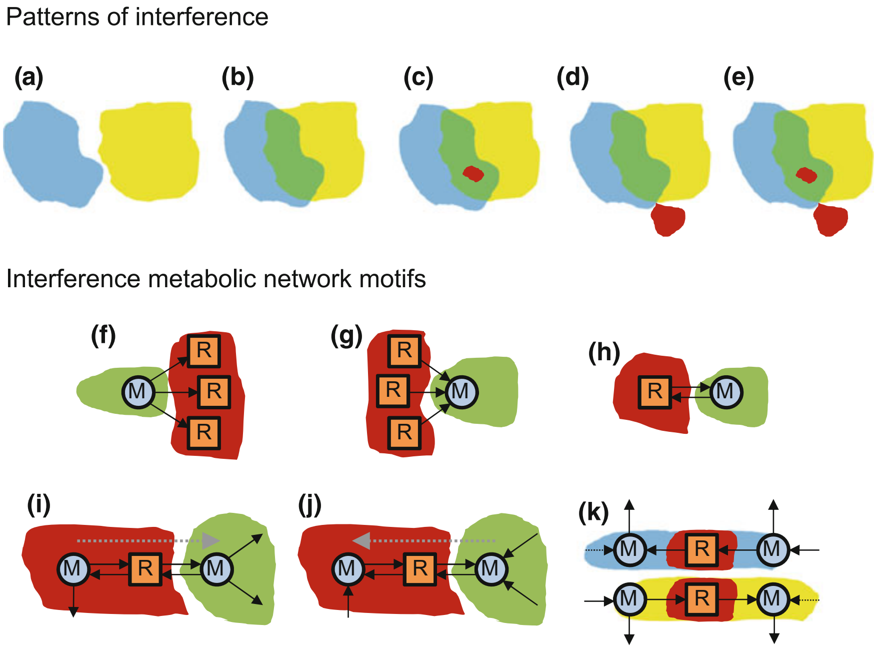


Fig. 3.6 Cascade propagator network motifs and typology of double cascades. **a**–**e** Illustration of possible interference patterns between individual cascades: additive, interference without overlap or amplification, interference with overlap and without amplification, interference without overlap and with amplification, interference with overlap and amplification, respectively. *Blue* and *yellow* stand for single cascades, *green* for interference, and *red* for overlap and amplification, depending on whether the red zone is in the interference zone (*green*) or not. **f**–**k** Metabolic network motifs in the interference of two individual cascades that induce amplification. Cases **f**–**g** Motif caused by a metabolite which loses its only generating reaction and at the same time it is the reactant of several reactions. These reactions are going to be become non-viable. Case **g** is equivalent to f but inverting the sense of the links. Case **h** Metabolite which has been left with one connection to a reversible reaction. This reversible reaction has zero net flux and becomes non-viable. Cases **i** and **j** This motif appears when a modified metabolite is lead with only one incoming connection coming from a reversible direction. This fixes the reversible reaction towards the production of this metabolite. If this step turns a metabolite of the reversible reaction non-viable, the reversible reaction becomes non-viable. Therefore, this motif is a potential trigger of amplification. Case **j** is equivalent to case **i** when the senses of the reactions are inverted. Case **k** The individual cascades fix the sense of a reversible reaction oppositely, one cascade forwards (k top) and the other backwards (k bottom) (note that the pictures illustrate the effects of both cascades individually). After superimposing the effects of the two cascades, one can see that this reversible reaction becomes non-viable. Thus, metabolites surrounding the reaction may become non-viable as well, depending on their degrees. It is also a potential trigger, as in cases **i** and **j**. Extracted from Ref. [18]

of the two individually considered cascades. By that, one refers to all metabolites and reactions altered,[1] removed or not, by each single cascade. If there is no interference,

[1] Reactions altered but not removed are reversible reactions that become directed by effect of the cascade.

the total damage $d_{rr'}$ is additive and equal to the sum of the two single damages d_r and $d_{r'}$. Otherwise, different situations are possible leading to a combined damage that can be equal, larger or smaller than the single added values. The latter case is a univocal signature of cascade overlapping $o_{rr'}$, pointing to the existence of a common subset of reactions that fail in both cascades (the most extreme realization is when one cascade is totally contained in the other). More interesting is the situation when, irrespectively of the presence or absence of overlap, a non-linearly amplified damage is detected, involving a number $a_{rr'}$ of new reactions that break down under simultaneous black outs. For all cascades,

$$d_{rr'} = d_r + d_{r'} - o_{rr'} + a_{rr'} \tag{3.2}$$

Interference without amplification is the most common situation, followed by the absence of interference (see Fig. 3.5d). In contrast, overlap and amplification happen for a very small fraction of all double cascades, and their occurrence decreases with the size of the organism (see Fig. 3.5e). In particular, the reduced incidence of amplification represents a new signature that organizational principles at play ensure the robustness of the organisms, despite increasing complexity and interweaving.

However, amplification may have a very large impact when it occurs. For instance, pyruvate (a product of glucose metabolism and a key intersection in several metabolic pathways) provides energy by fermentation. This process reduces pyruvate into lactate, a reaction that does not trigger any black out cascade when it fails, so $d_r = 1$. At the same time, pyruvate can also be decarboxylated to produce acetyl groups, the building blocks of a large number of molecules that are synthesized in cells. The failure of the first reaction in such pathway triggers a cascade of length $d_{r'} = 3$. In contrast, the simultaneous failure of both the fermentation and the reduction of pyruvate induces a large cascade of size $d_{rr'} = 36$, most likely lethal. As a biological explanation, one could argue that both processes are strongly interdependent to maintain the oxidation-reduction balance when fermentation is in action.

Collateral effects offer the clue to understand this amplification phenomenon. In parallel to rendering non-operational some reactions and their corresponding metabolites, a cascade can reduce the connectivity and increase the branching ratio of other viable metabolites in its influence area. When stricken by the propagation front of a second cascade, these metabolites are susceptible of becoming non-viable, further spreading the failure wave. In this way, interference is a necessary but not a sufficient condition for amplification, and a large amplification can be possible even when there is no overlap and the interference between the individual cascades is small. To predict which pairs will trigger amplification, one must focus on metabolites in the interference of the influence areas of the two individual cascades. Those metabolites that remain viable after each individual cascade but become non-viable when the two effects are superposed will produce amplification, propagating the double cascade to new reactions. In Fig. 3.6f–k, the connectivity structure of all interference cascade propagator motifs responsible for amplification is provided.

3.3 Impact of Gene Knockouts in Metabolic Structure

Reaction failures are usually associated to the disruption of an enzyme due to knockout, inhibition, or deleterious mutation of the corresponding gene. In *M. pneumoniae*, enzyme multi-functionality and gene essentiality are higher as compared to other prokaryotic bacteria, so gene malfunctioning can potentially produce an acuter stress response at the level of metabolism. To address this issue, the metabolic network of *M. pneumoniae* is coupled to its gene co-expression network through the activity of enzymes, and knockouts of individual genes and clusters of co-expressed genes are performed. Inherent to this analysis is the potential occurrence of individual, double, or multiple cascades simultaneously. Multiple knockouts are algorithmically handled as an obvious extension of the previously considered situation of pair cascades.

The genome of *M. pneumoniae* [23] comprises 688 genes, 140 of which have a metabolic function. Except for one spontaneous reaction and 20 reactions with unknown regulation, these metabolic genes codify 142 enzymes that catalyse reactions in the metabolic network of this organism.

3.3.1 Metabolic Effects of Individual Mutations

Individual metabolic gene knockouts or mutations inhibit the production of catalytic enzymes and induce black outs of reactions propagating in the metabolic network as a failure cascade (see Fig. 3.7). From existing data, 71% of the 140 metabolic genes in *M. pneumoniae* have a one-to-one relation with reactions, and 21% of the genes regulate multiple reactions. Seldom the same reaction may be individually regulated by different enzymes produced by different genes, which happens for only four non-damaging reactions. More often, several genes are necessary to regulate the activity of a single reaction through an enzymatic complex. Twelve complexes codified by 26% of genes regulate the activity of 10% of metabolic reactions in *M. pneumoniae*. The removal of any of the genes involved in a complex is expected to cause the inactivation of the reaction controlled by the complex, which in principle may increase vulnerability. However, it can be observed that almost all complexes are associated to low damage reactions, which indicates a certain degree of structural robustness.

To study the metabolic effects of individual gene mutations, the knockout of all reactions associated to the gene under consideration are simulated. As explained, most often this corresponds to one single reaction but sometimes multiple reactions are removed simultaneously. The first observation is that metabolic genes affecting vulnerable reactions trigger large failure cascades. More interestingly, genes with large associated damages in metabolism turn out to be essential or conditionally essential for *M. pneumoniae* (see Table 3.1), with a unique exception discussed below. The classification given in Ref. [17] is used, where essentiality is defined according to the measured metabolic map and the definition of a minimal medium which allows

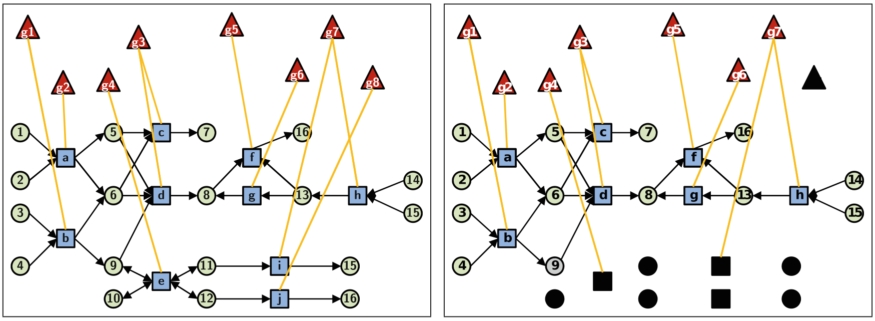


Fig. 3.7 *Left* Scheme of genes connected to reactions. Direct connections between genes and reactions are shown, but notice that connections between genes and reactions can be done only due to the existence of enzymes. *Right* effect of how performing a knockout of a gene, labelled as *g8*, spreads a cascade in the metabolic network. *Squares* denote reactions, *circles* denote metabolites and *triangles* genes. *Black* nodes denote nodes that have become non-operational, whereas *gray* nodes are viable nodes that have reduced their connectivity due to the effect of the cascade. Parts of this Figure have been extracted from Ref. [24] Copyright @ 2014, World Scientific Publishing

M. pneumoniae to grow. Essential genes are those that are required for the survival of the organism, meaning that the products of the reactions that they control are essential for life and cannot be produced by alternative pathways, while conditional means that essentiality depends on the media composition available.

In fact, all conditionally essential genes with the potential of producing high damage in the metabolism of *M. pneumoniae* have been found to have an essential orthologue (essentiality determined by loss-of-function experiments) in *Mycoplasma genitalium* [25], a comparable genome-reduced bacterium. The only exception to essentiality in Table 3.1 is gene MPN062, considered as non-essential in Ref. [17], while in this study it triggers a large failure cascade and so it can be classified as a vulnerable target for metabolic structure. Its damaging potential can be explained by the fact that each of the four reactions controlled by the gene has a contribution that, although not extremely high individually, adds to the total damage and interferes to produce amplification. Therefore, MPN062 can be proposed as an important gene for metabolic function in *M. pneumoniae*, a conjecture that is supported by the essentiality of its orthologue in *M. genitalium* [25].

Another interesting case is essential gene MPN429, whose knockout triggers the largest cascade in *M. pneumoniae*. Each of the four affected reactions in the Glycolysis pathway is not able to propagate a cascade individually. However, when they all are removed simultaneously, the strongest amplification effect is observed. The biochemical explanation is that the non-linear interaction of the cascades stops the production of phosphoenolpyruvate, which disrupts the synthesis of ATP, a circumstance particularly harming to the organism.

The obtained results of the study of structural cascades to predict gene essentiality in *M. pneumoniae* is in agreement with the gene essentiality computed in Ref. [26].

Table 3.1 Largest structural damages produced in metabolism by gene knockouts and correspondence with gene essentiality as given in Ref. [23]. Damage in metabolic structure caused by gene knockout (third column) is measured in number of deleted reactions. In the fourth column, the number of reactions regulated by the corresponding gene is given, and in parentheses the damage associated to each of these reactions is also given. Genes in monocomponent clusters are highlighted in boldface, and braces are used to denote genes that form complexes. Note that the complex at the end of the list is not detected by any of the three clustering procedures. Finally, gene MPN062 is the only one in the table annotated as non-essential although it is associated to a large failure cascade. Extracted from Ref. [18]

Gene	Essentiality	Damage	Reactions
MPN429	yes	49	4 (1,1,1,1)
MPN606	yes	32	1 (32)
MPN628	yes	32	1 (32)
MPN017	yes	25	3 (14,1,9)
MPN303	yes	18	8 (1,1,1,1,8,1,2,3)
MPN062	*no*	17	4 (6,3,2,3)
MPN576	cond	16	2 (13,2)
MPN005	yes	13	1 (13)
MPN336	yes	13	3 (4,3,6)
MPN354	yes	13	1 (13)
MPN627	yes	11	1(11)
MPN066	yes	9	4 (1,1,2,5)
MPN240	cond	9	1 (9)
MPN299	cond	9	1 (9)
MPN322	cond	9	4 (1,1,2,1)
MPN323	cond	9	4 (1,1,2,1)
MPN324	cond	9	4 (1,1,2,1)
MPN034	yes	7	4 (1,1,2,3)
MPN378	yes	7	4 (1,1,2,3)

3.3.2 *Metabolic Effects of Knocking Out Gene Co-expression Clusters*

Groups of co-expressed genes in *M. pneumoniae* can be identified from gene expression data under different conditions [27–29], which reveals a complex gene regulatory machinery [23]. The functional deactivation of these clusters might be produced by the failure of common regulatory elements and important damage could be transmitted to metabolism.

In this subsection, results on the effects on the metabolic structure of *M. pneumoniae* by suppressing gene co-expression clusters are shown. Information about gene expression is provided in Ref. [23]. Correlations in the expression of genes were measured from tilling arrays under 62 different environmental con-

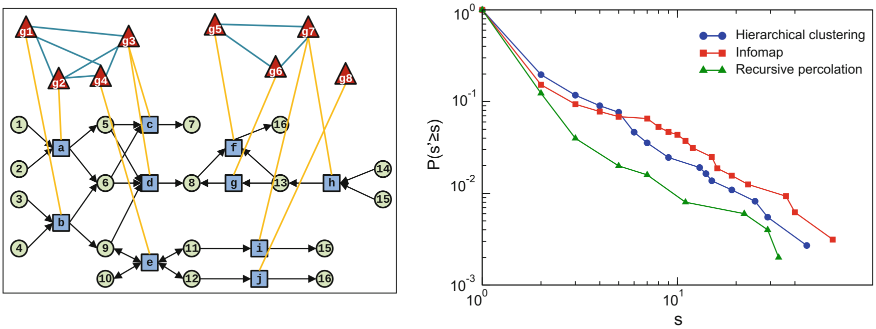


Fig. 3.8 Pictures of co-expressed genes and distribution of sizes of gene clusters. *Left* Groups of co-expressed clusters regulating a metabolic network. *Right* Distribution of sizes of the clusters obtained using distance hierarchical clustering (*blue*), Infomap (*red*) and recursive percolation (*green*). Parts of this figure have been extracted from Ref. [24] Copyright @ 2014, World Scientific Publishing, and from Ref. [18]

ditions. This matrix of correlations between the expression levels of pairs of genes gives a fully connected network where the link between two genes carries a weight ranging from -1 to 1. This gene correlation matrix can be coupled to the metabolic network of *M. pneumoniae* through the activity of enzymes to produce a multilevel network representation.

To detect gene co-expression clusters, three different strategies -distance hierarchical clustering, Infomap, and recursive percolation (see Chap. 2, Sect. 2.1.4)- are applied to the gene expression correlation matrix in order to discount biases introduced by the specifications of the community detection method. The distributions of sizes of the obtained clusters with the three strategies are shown in the right panel of Fig. 3.8, where it is indeed possible to see a certain degree of similarity between the distribution of sizes of the clusters obtained using the three different methods.

The comparative analysis of the detected clusters of genes showed that, although the partitions found by each algorithm may differ in their composition and in the maximum size of the clusters, there are preserved commonalities independently of the method. One of them is that all methods are able to detect seven of the twelve complexes, since the related genes always appear classified in the same cluster. Another remark is that, as explained in the previous paragraph, the three detection methods result in qualitatively similar power-law-like cluster size distributions (see Fig. 3.8, right), with most clusters having small size while some are relatively big. Interestingly, genes related to high damage spreading reactions are secluded into mono-component clusters. To be more precise, eight of the nineteen genes in Table 3.1 are recognized by all three methods as having an expression profile that is not correlated to other gene activity levels. This is surprising since, in principle, high-damage genes might be expected to be co-regulated with other genes, as influencing big parts of metabolism usually requires coordinated gene activity. The fact that these genes appear isolated pinpoints them as potentially important metabolic

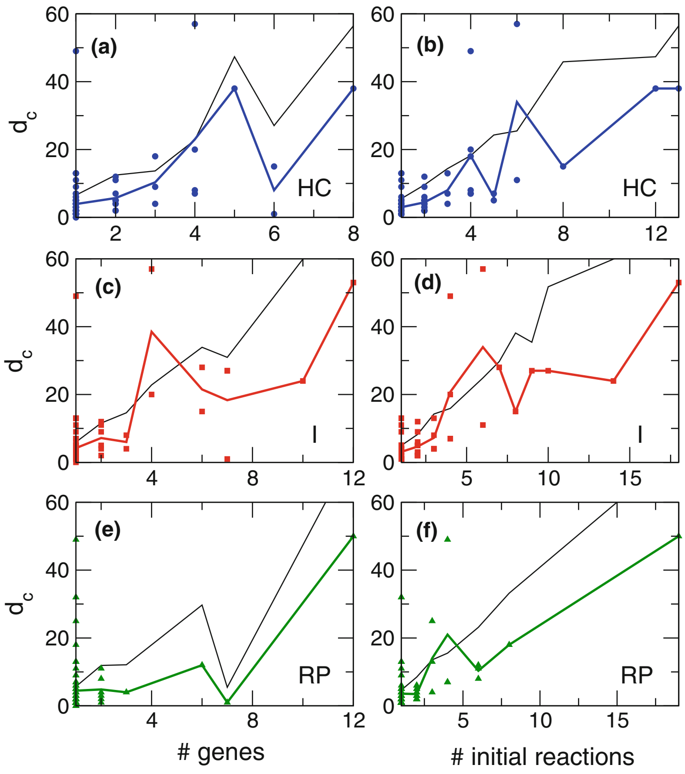


Fig. 3.9 Damages as a function of the number of metabolic genes and reaction failures in gene co-expression cluster knockouts. Clusters are defined according to three different methods: Hierarchical Clustering (HC), Infomap (I), and Recursive Percolation (RP). Results are compared with damages produced in randomized versions of the metabolic networks in order to discount structural effects. In each case, the *solid black curve* is the average over 100 realizations. Extracted from Ref. [18]

regulator targets, since the alteration of only one gene may affect a large number of metabolic reactions. In any case, the lack of co-regulation of genes related to high damage spreading reactions is again an indication that the structural organization of the organism has evolved towards protecting the system against multiple failures.

Taking averages for equally sized clusters, it can be found that knockouts of co-expression clusters produce a damage on metabolic structure that increases with the number of affected metabolic genes, except when most metabolic genes in a cluster codify an enzymatic complex regulating one reaction (see Fig. 3.9, left panels). The damage produced by the failure of the cluster also increases with the number of associated reactions (right panels in Fig. 3.9). In order to discount structural effects, these results are compared with those measured on DP randomized versions of the metabolic network of the genome-reduced bacterium. As evidenced in Fig. 3.9, all cluster detection methods identify clusters that produce lower damages in the real metabolic network of *M. pneumoniae* as compared to the randomized network. This supports the idea that the regulatory machinery that controls the coupled-to-metabolism co-expression of genes has evolved towards robustness.

Finally, since the three cluster detection methods propose different forms of aggregating metabolic genes, it is relevant to consider whether cluster composition is relevant for failure propagation. As a null model, one can consider randomization restricted not to the network itself but to the specific gene metabolic composition, while maintaining the total number of metabolic genes in each cluster. It can be observed that such a reshuffling of metabolic genes in clusters has no relevant effect on the damages measured on the metabolic network (see Fig. 3.10). This means that, surprisingly, the composition of the clusters is not as statistically relevant for metabolic vulnerability as the distribution of the cluster sizes itself. This feature, together with the large detected amount of mono-component clusters, point out to the existence of multiple levels of regulation, depending on experimental conditions and, at the same time, explains why genes controlling high damage spreading reactions operate preferentially under functional isolation as a metabolism protection mechanism.

3.4 Robustness Versus Regulation in Metabolic Networks

The null model used in the first part of this chapter, called degree-preserving randomization, does not account for the most basic physico-chemical constraints and may lead, in the case of metabolic networks, to consideration of reactions which are not mass (*i.e.*, stoichiometrically) balanced (which do not preserve the same type and number of atoms on the substrate and product sides). As a result, the randomized networks may not be chemically feasible. As an alternative, the null model called mass-balanced randomization [30] accounts for this issue (see Chap. 2, Sect. 2.1.7.2). It is worth stressing that this method preserves the degrees of reactions but not the degrees of metabolites (see Fig. 3.11).

In this section, cascades originated by single reaction and pair of reaction failures in the original networks of the three bacteria are compared with those obtained from two null models: DP, already used in the previous section, and MB randomization. As in the first part, K-S tests are used to statistically assess whether the null models are relevant to explain the resulting damage distributions in the original networks. The analysis reinforces the importance of choosing an appropriate null model according to the question at hand, since the null model ultimately affects the interpretation of the findings [19].

First, cascades triggered by individual removal of reactions are studied, each cascade having its associated damage d_r. When comparing the cumulative distributions $P(d_r' \geq d_r)$ of the damage d_r produced by individual removal of reactions between the original and randomized networks (see Fig. 3.12, left panels), it can be observed that the distributions of the original networks lie in between the distributions of the two null models. To check whether or not the cumulative probability distributions

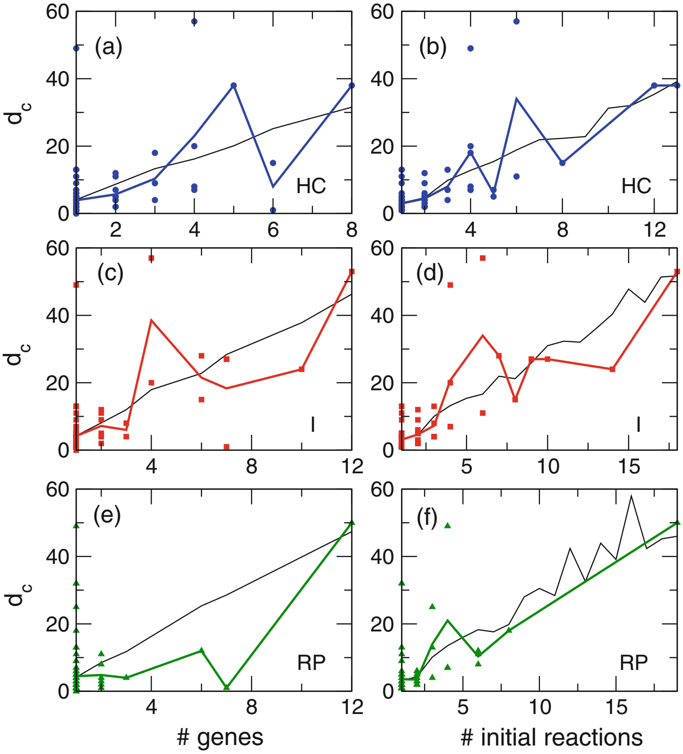


Fig. 3.10 Damage distributions as a function of the number of genes and reaction failures, similar to Fig. 3.9, but now randomizing the specific genetic contents of each cluster while maintaining the total number of metabolic genes in each cluster. The size of the clusters are defined according to three different methods: Hierarchical Clustering (HC), Infomap (I), and Recursive Percolation (RP). Extracted from Ref. [18]

are significantly different in the original networks and in their randomized variants, K-S tests are performed (see Table 3.2), taking as the standard significance level $\alpha = 0.05$. The compared distributions are considered significantly different from the null models because their associated significance is smaller than 0.05, except for *M. pneumoniae*, whose distribution can be considered consistent with the DP model as seen in Sect. 3.2, probably due to its linearity. Both for *E. coli* and *S. aureus*, damages are smaller compared to their DP randomizations but larger when compared to their MB randomizations. Thus, the robustness of the analysed networks cannot be explained by the distribution of degrees or by basic physical constraints. For the DP null model, this finding indicates that robustness is positively influenced by factors other than the degrees. The results from the MB null model suggest that, for all three organisms, evolutionary pressure leads to larger cascades of non-viable reactions as compared to those imposed by physico-chemical constraints, and thus lower robustness.

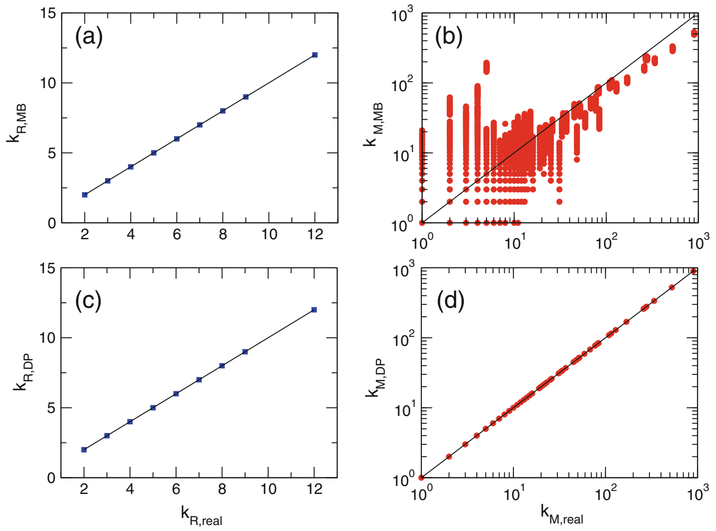


Fig. 3.11 Comparison of the degrees of reactions and metabolites obtained by the two null models applied to *E. coli* network. In this representation, each point is a reaction or a metabolite with coordinates $(k_{real}, k_{randomized})$, where k_{real} is the metabolite/reaction degree in the original network, and $k_{randomized}$ the corresponding degree in a randomized network. Points fall in the diagonal if degrees are preserved in the randomized networks. **a** and **b** MB randomization. This method gives networks in which the degrees of the reactions are preserved. However, degrees of metabolites are not conserved. **c** and **d** DP randomization. This method gives networks with preserved degrees of reactions. Degrees of metabolites are also preserved with DP randomization, however at the expense of violating mass balances of reactions. Extracted from Ref. [19] Copyright @ 2012, PACIS-JCIS

After performing single reaction removals, the same analysis for the removal of each possible pair of reactions is done. Similar to the single reaction case, the cumulative probability distributions $P(d'_{rr'} \geq d_{rr'})$ of the damage $d_{rr'}$ resulting from the knockout of two reactions is determined (see Fig. 3.12). K-S tests with a standard significance level $\alpha = 0.05$ are again applied (see Table 3.2), finding that the distributions of the original networks are significantly different from those of both randomization methods. All organisms display in this case similar results, the distributions of the original networks lie again between the distributions of the two null models, and all of them can be considered inconsistent with both null models. Consequently, the observations for individual failures also hold for the failure of reaction pairs: robustness is positively influenced by factors other than degrees, but negatively influenced by evolutionary pressure.

The cascade algorithm produces larger damages in the original networks as compared to those in MB randomized networks, but smaller cascades as compared to those in DP randomized counterparts. A possible explanation is offered by the difference in global properties of the networks obtained from the two randomization methods [31]. DP randomization decreases the average path length and increases the clustering coefficient of the randomized network, increasing its small-world property.

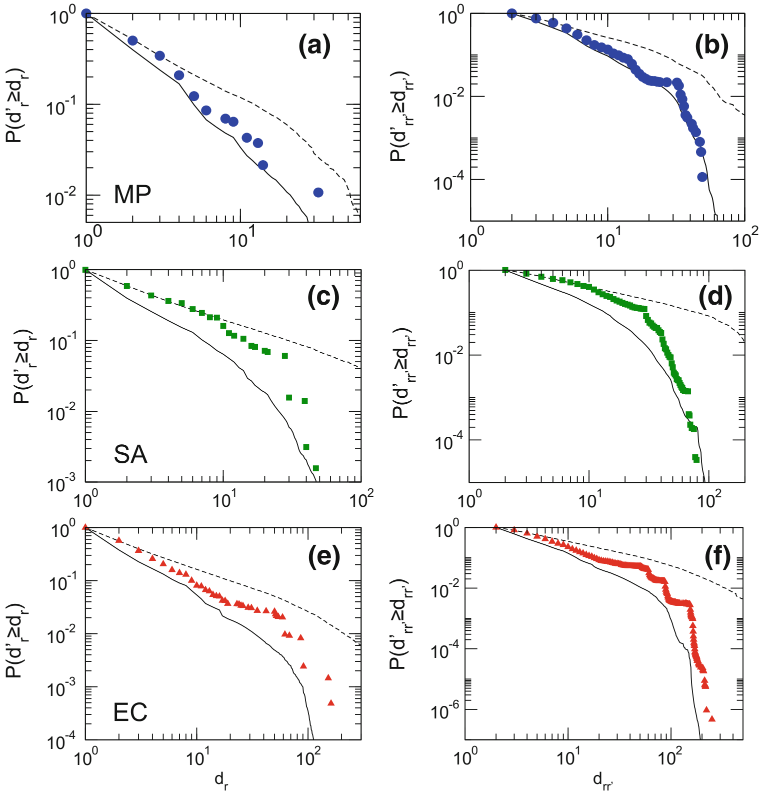


Fig. 3.12 Distributions of damage caused by removal of reactions. **a**, **c**, **e** Cumulative probability distributions for *M. pneumoniae* (*blue*), *S. aureus* (*green*), and *E. coli* (*red*). Averaged distributions over 100 randomizations of the original networks are shown for DP (*dashed line*) and MB randomization (*continuous line*). **b**, **d**, **f** Damages caused by pairs of removal of reactions. Parts of this Figure have been extracted from Ref. [19] Copyright @ 2012, PACIS-JCIS (color figure online)

Consequently, such networks are more interconnected and, thus, a cascade may in principle propagate further in the network. The opposite holds for MB randomization, which increases the average path length while decreasing the clustering coefficient of the randomized network so that the spread of the damage is less likely. Although the average path length does not resemble the length of metabolic inter-conversion, the small-world property may still affect the impact of removal of reactions due to its functional importance.

It can also be pointed out that the principle of cascade propagation relies on violation of a structural precondition for a steady-state, namely that all metabolites can be produced and consumed in order to avoid their depletion or accumulation. However, the steady-state assumption is only meaningful for networks which satisfy fundamental physical principles. Therefore, the use of MB randomization, which guarantees preservation of mass balance, allows to discern whether the measured

Table 3.2 Kolmogorov-Smirnov tests for comparing single reaction (SR) and pairs of reactions (PR) failure cascades in the three metabolic networks with both randomization methods, MB and DP. The values of the K-S statistic / associated significance level are given. Parts of this Table have been extracted from Ref. [19] Copyright @ 2012, PACIS-JCIS

Organism	SR		PR	
	MB	DP	MB	DP
M. pneumoniae	0.10/0.03	0.095/0.07	0.15/0	0.15/0
S. aureus	0.19/0	0.086/0.0002	0.27/0	0.14/0
E. coli	0.19/0	$0.079/1.4 \cdot 10^{-11}$	0.21/0	0.13/0

property is a result of basic physical principles, or, instead, whether it is affected by evolutionary pressure. Since the size of cascades in MB randomized networks is significantly lower than those in real networks, evolutionary pressure may indeed lead to larger cascades.

Consequently, this finding indicates that evolutionary pressure may favour lower robustness of metabolic networks with respect to the failure of reactions, seemingly contradicting the general requirement of robustness in biological systems. On the one hand, this finding may be a result of the evolutionary versatility of metabolic networks, which favours organisms that are able to evolve quickly, i.e., by few modifications to their metabolic networks. On the other hand, it is worth stressing that a cascade may not only be interpreted as the harmful spreading of failure, but also as the ability to regulate metabolism by activating/deactivating reactions, e.g., by transcriptional regulation [32]. Thus, large cascades, favoured by evolutionary pressure, may point at the evolutionary requirement of regulating large parts of metabolism through the regulation of small sets of enzyme-coding genes. The ability to regulate the activity of metabolic reactions by deactivating competing reactions is a well-known principle of metabolism. These results thus indicate that evolutionary pressure may favour the ability of efficient metabolic regulation at the expense of robustness to reaction or gene knockouts, pointing at the necessary integration of trade-offs from various cellular functions.

3.5 Conclusions

Results obtained in this chapter demonstrate that when *E. coli* and *S. aureus* are subjected to reaction failures, their metabolic networks have a structure that minimizes the number of large cascades. In this way, the largest part of reaction failures lead to small cascades, resulting in a small damage for the metabolic network. Hence, one can conclude that these organisms have a robust metabolic network against reaction failures. *M. pneumoniae* exhibits network responses that are qualitatively comparable to *E. coli* or *S. aureus*, although it is found that it less robust against individual reaction removals with reactions more prone to trigger large metabolic failure cascades identified as key participants in the regulation of energy and fatty acid synthesis.

The concept of cascade amplification has been for the first time formulated and interpreted as a signature of the subtle non-linearities underlying the structure of complex networks. Specific scenarios in *M. pneumoniae* have been discussed. In addition, there is a motivation to assess the predicting power of the used formalism. In this sense, a predictor of damage propagation for single cascades, and structural motifs underlying amplified failure patterns in situations of concurrent spreading have been proposed.

On what respects to the analysis of single gene knockouts, it reveals its potentiality in capturing most of the scenarios of experimentally determined lethality for *M. pneumoniae*. Moreover, when clustered and knocked together new trends of the complex genomic regulation of the metabolism emerge. First, the distribution of cluster sizes seems to matter more than the actual composition of the clusters. This is connected to the fact that the regulation of high-damage genes tends to appear isolated from that of other genes, a kind of functional switch in metabolic networks that at the same time acts as a kind of genetic firewall.

The introduction of a randomization model that generates new realizations of the network which are mass balanced indicates that evolutionary pressure favours the ability of efficient metabolic regulation at the expense of robustness to gene knockouts. This is explained because it favours organisms to evolve quickly by little modifying their metabolic networks, and because a failure cascade can be interpreted as an ability to regulate metabolism by activating/deactivating reactions, apart from being interpreted as a harmful spreading of a failure.

3.6 Summary

- The metabolic networks of three bacteria, *M. pneumoniae*, *E. coli*, and *S. aureus*, have been found to be robust against reaction failures, although *M. pneumoniae* is less robust against individual reaction removals due to its simplicity [18].
- A predictor of damage propagation for cascades produced by single reaction failures and the structural motifs underlying amplified failure patterns have been proposed. It has been checked that the predictor successfully predicts damage without the need of computing cascades [18].
- The concept of cascade amplification has been formulated and interpreted as a signature of the subtle non-linearities underlying the structure of complex networks [18].
- The study of structural stress at the level of metabolic genes reveals its potentiality in capturing most of the scenarios of experimentally determined lethality for *M. pneumoniae* [18].

- The distribution of gene cluster sizes seems to matter more than the actual composition of the clusters in relation to failure propagation in the metabolic network [18].
- The studied organisms show a trade-off between robustness and efficient regulation of their metabolic networks [19].

References

1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Dorogovtsev SN, Goltsev AV, Mendes JFF (2008) Critical phenomena in complex networks. Rev Mod Phys 80:1275–1335
3. Barrat A, Barthélemy D, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press, Cambridge
4. Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the internet to random breakdown. Phys Rev Lett 85:4626
5. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
6. Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci USA 99:5766–5771
7. Moreno Y, Gómez JB, Pacheco AF (2002) Instability of scale-free networks under nodebreaking avalanches. Europhys. Lett. 58:630–636
8. Motter AE (2002) Cascade-based attacks on complex networks. Phys Rev E 66:065102(R)
9. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464:1025–1028
10. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. Nat Rev Genet 5:101–113
11. Szalay MS, Kovacs IA, Korcsmaros T, Bode C, Csermely P (2007) Stress-induced rearrangements of cellular networks: consequences for protection and drug design. FEBS Lett 581:3675–3680
12. Motter AE, Gulbahce N, Almaas E, Barabási AL (2008) Predicting synthetic rescues in metabolic networks. Mol Syst Biol 4:168
13. Smart AG, Amaral LAN, Ottino J (2008) Cascading failure and robustness in metabolic networks. Proc Natl Acad Sci USA 105:13223–13228
14. Edwards JS, Palsson BØ (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol. Prog. 16:927–939
15. Segrè D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99:15112–15117
16. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7:501
17. Yus E et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
18. Güell O, Sagués F, Serrano MÁ (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621
19. Güell O, Sagués F, Basler G, Nikoloski Z, Serrano MÁ (2012) Assessing the significance of knockout cascades in metabolic networks. J Comp Int Sci 3(1–2):45–53
20. Güell O, Sagués F, Serrano MÁ (2014) Assessing the significance and predicting the effects of knockout cascades in metabolic networks. In: Extended Abstracts Spring 2013. Springer, pp 39–44. ISBN 978-3-319-08137-3

21. Smirnov NV (1948) Tables for estimating the goodness of fit of empirical distributions. Ann Math Stat 19:279
22. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101
23. Güell M et al (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326:1268–1271
24. Güell O, Serrano MÁ, Sagués F (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In: Engineering of chemical complexity II. World Scientific, pp 39–56. ISBN 978-981-4616-12-6
25. Glass JI et al (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425–430
26. Wodke JAH et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol Syst Biol 9:653
27. Borate BR et al (2009) Comparison of threshold selection matrices for microarray gene co-expression matrices. BMC Res Notes 2:240
28. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:1544–6115
29. Khanin R, Wit E (2005) Construction of malaria gene expression network using partial correlations. In: Methods of microarray data analysis V, pp 1544–6115
30. Basler G, Ebenhöh O, Selbig J, Nikoloski Z (2011) Mass-balanced randomization of metabolic networks. Bioinformatics 27:1397–1403
31. Basler G, Grimbs S, Ebenhöh O, Selbig J, Nikoloski Z (2012) Evolutionary significance of metabolic network properties. J R Soc Interface 9:1168–1176
32. DeRisi JL, Iyer V, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278(5338):680–686

Chapter 4
Effects of Reaction Knockouts on Steady States of Metabolism

The activity and essentiality of metabolic reactions of two model organisms, *Escherichia coli* and *Mycoplasma pneumoniae*, are studied using Flux Balance Analysis in different environments. In particular, synthetic lethal pairs correspond to combinations of active and active or inactive non-essential reactions whose simultaneous deletion causes cell death. Lethal knockouts of pairs of reactions separate in two different groups depending on whether the pair of reactions works as a backup or as a parallel use mechanism, the first corresponding to essential plasticity and the second to essential redundancy. Within this perspective, functional plasticity and redundancy are essential mechanisms underlying the ability to survive of metabolic networks.

The previous chapter reported the study of structural perturbations modelled by the removal of a reaction or a set of them and the application of ta viability criterion at the structural level. This chapter goes from structure to function by using the technique called Flux Balance Analysis (FBA) (see Chap. 2, Sect. 2.2) to implement reaction knockouts. A FBA analysis goes beyond the structural characterization of a cascade triggered by a reaction knockout in the sense that FBA intrinsically assigns zero fluxes to all the reactions in the network that turn out to be non-viable, *i.e.*, that are not able to maintain a balanced steady state in a certain environmental condition. In addition, using FBA one can compute how the environment affects the fluxes of reactions in metabolic networks. In particular, FBA allows to compute the activities and essentialities of reactions at steady state (see Chap. 2, Sect. 2.2.3), and to study the concept of synthetic lethality and how it is related to concepts such as *plasticity* and *redundancy*.

The computation of the activity of reactions using FBA has permitted a better understanding of how metabolism adapts to environmental changes by means of modifications in the biochemical fluxes [1, 2]. Beyond the concept of activity, the study of essentiality can help to understand how metabolism adapts to an internal failure, analysing the adaptation of the fluxes when one reaction is forced to be non-operative. In fact, the concept of essentiality has been studied extensively, from single reaction failures [3–5] to multiple failures [6, 7].

Plasticity and redundancy are large-scale strategies that offer the organism the ability to exhibit no or only mild phenotypic variation in front of environmental

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_4

changes or upon malfunction of some of its parts. In particular, these mechanisms protect metabolism against the effects of single enzyme-coding gene mutations or reaction failures, the final outcome being that most metabolic genes result to be not essential for cell viability. However, some mutants fail when an additional gene is knocked out, so that specific pair combinations of non-essential metabolic genes or reactions become essential for biomass formation. As an example, double mutants defective in the two different phosphoribosylglycinamide transformylases present in *Escherichia coli*-with catalytic action in purine biosynthesis and thus important as crucial components of DNA, RNA or ATP- require exogenously added purine for growth, while single knockout mutants do not result in purine auxotrophy [8].

These synthetic lethal (SL) combinations [9–12] have recently attracted attention because of their utility for identifying functional associations between gene functions and, in the context of human genome, for the prospects of new targets in drug development. However, non-viable synthetic lethal mutants are difficult to characterize experimentally despite the high-throughput techniques developed recently [13]. We are still far from a comprehensive empirical identification of all SL metabolic gene or reaction pairs in a particular organism [6], even more when considering different growth conditions. Metabolic screening based on computational methods becomes then a powerful complementary technique particularly suited for an exhaustive *in silico* prediction of SL pairs in high-quality genome-scale metabolic reconstructions.

This chapter unveils how functional plasticity and redundancy are essential systems-level mechanisms underlying the viability of metabolic networks. In previous works on cellular metabolism [2, 14], plasticity was some times associated to changes in the fluxes of reactions when an organism is shifted from one growth condition to another. Instead, here functional plasticity is discussed as the ability of reorganizing metabolic fluxes to maintain viability in response to reaction failures when the environment remains unchanged. On the other hand, functional redundancy applies to the simultaneous use of alternative fluxes in a given medium, even if some can completely or partially compensate for the other [15]. An exhaustive computational screening of SL reaction pairs is performed in *E. coli* in glucose minimal medium and it is found that SL reaction pairs divide in two different groups depending on whether the SL interaction works as a backup or as a parallel use mechanism, the first corresponding to essential plasticity and the second to essential redundancy. When comparing the metabolisms of *E. coli* and *Mycoplasma pneumoniae*, one can find that the two organisms exhibit a large difference in the relative importance of plasticity and redundancy. In *E. coli*, the analysis of how pathways are entangled through SL pairs supports the view that redundancy SL pairs preferentially affect a single function or pathway [9]. In contrast -and in agreement with reported SL genetic interactions in yeast [16]-essential plasticity, which is the dominant class in *E. coli*, tends to be inter-pathway but concentrated and unveils cell envelope biosynthesis as an essential backup for membrane lipid metabolism. Finally, different environmental conditions are tested to explore the interplay between these two mechanisms in coessential reaction pairs. Knockouts of genes are not considered because approaching directly pairs of reactions without the scaffold of enzymes and genes allows to determine in a clean and systematic way the minimal combinations of reactions that turn out to be essential for an organism.

The contents of this chapter correspond to Ref. [17] Copyright @ 2014, World Scientific Publishing, and to Ref. [18].

4.1 Activity and Essentiality of Single Reactions of *E. coli* Across Media

This section summarizes the results of the study of how the activity and essentiality of reactions in the *i*JO1366 version of *E. coli* (see Chap. 2, Sect. 2.3.1) depend on the nutrient composition of the environment [17] Copyright @ 2014, World Scientific Publishing. To this end, the activity and essentiality of all active reactions in a set of minimal media are computed, and then, depending on their behaviour on each environment, each reaction is classified according to four general categories. This study also allows to identify reactions as eventual candidates to form part of SL pairs.

A total number of 555 minimal media can be constructed as proposed in Chap. 2, Sect. 2.2.2.1, with a final number of 333 which allow growth for the *i*JO1366 version of *E. coli*. In addition to these minimal media, 10,000 random media are also analysed, of which 3707 give a non-zero growth. To construct these random media, one considers all metabolites present in the extracellular environment of *E. coli*. Then, one chooses the number of these metabolites that can act as nutrients. In this case, 90% of the total number of external metabolites are allowed to act as nutrients. Once the number of nutrients is selected, one chooses at random metabolites until one reaches the selected number of nutrients, and the lower bound of the exchange reactions of each metabolite is changed to a value of -10 mmol $\text{gDW}^{-1}\text{h}^{-1}$.

4.1.1 Quantifying Activity and Essentiality

The activity and essentiality of each reaction are computed in every medium, with the obvious constraint that essentiality is computed only if the reaction is active. On what follows, an explanation to compute the accumulated values of the activity and essentiality is given.

The activity $a_{i,j}$ of a reaction i in a medium j is defined as

$$a_{i,j} \equiv \begin{cases} 1 \text{ if } \nu_{i,j} > 0 \\ 0 \text{ if } \nu_{i,j} = 0 \end{cases} \tag{4.1}$$

where $\nu_{i,j}$ denotes the flux of reaction i in medium j. To obtain a representative value of the activity, FBA calculations are performed in both minimal and random media. In addition, the activity is normalized by the number of media in which the calculations have been performed. Therefore, the activity of a reaction i for a given set of media n_{media} will be obtained according to

$$a_i \equiv \frac{1}{n_{media}} \sum_{j=1}^{n_{media}} a_{i,j} \tag{4.2}$$

with $0 \leq a_i \leq 1$.

Essentiality is defined on the subset of active reactions. To compute the essentiality of a particular reaction, the FBA growth rate is examined after removing the corresponding reaction. An expression of the essentiality of a reaction i in a medium j is given as

$$e_{i,j} \equiv \begin{cases} 0 \text{ if } \nu'_{g,j} > 0 \\ 1 \text{ if } \nu'_{g,j} = 0 \end{cases} \tag{4.3}$$

where $\nu_{g,j'}$ denotes the flux of the reaction of production of biomass in medium j when reaction i is constrained to have zero flux. Again, the results are averaged on several media and normalized by dividing by the number of media. In this way, the bounds of essentiality of a reaction lay between 0 and the corresponding activity of the reaction, $0 \leq e_i \leq a_i$,

$$e_i \equiv \frac{1}{n_{media}} \sum_{j=1}^{n_{media}} e_{i,j} \tag{4.4}$$

Another useful magnitude to be used later on is the ratio of media where reaction i is essential with respect to the number of media where it is active. This measure is trivially computed according to $p_i = \frac{e_i}{a_i}$.

4.1.2 *Characterization of the Reactions*

After computing FBA on all environments and for all mutants, essentiality *vs* activity is plotted for all reactions. All points must fall on the diagonal or under it. This plot is shown in Fig. 4.1 for both minimal and random media.

Reactions can be classified into four categories:

1. **Essential whenever active reactions**: $0 < a_i = e_i$. They are essential in all media where they are active. These reactions lay on the diagonal of the aforementioned plot.
2. **Always active reactions**: $a_i = 1$, $0 < e_i < a_i$. They are always active and sometimes essential. These reactions are located in the opposite y axis.
3. **Never essential reactions**: $0 < a_i < 1$, $e_i = 0$. They are never essential but sometimes active. These reaction are located in the x axis.

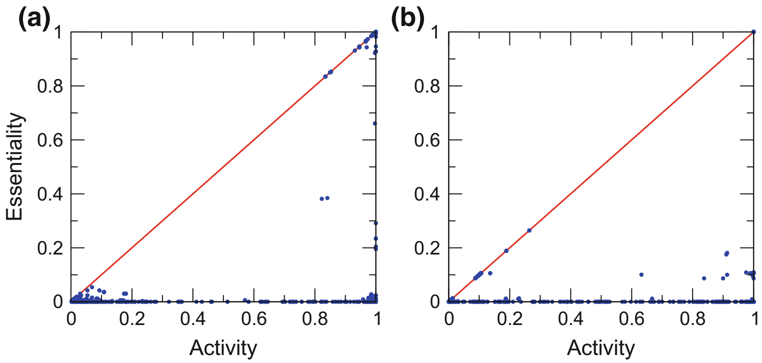


Fig. 4.1 Representation of essentiality versus activity. **a** Minimal media. **b** Random media. In both pictures the four different categories can be clearly differentiated. *Diagonal*: *essential whenever active reactions*. Opposite *y* axis: *always active reactions*. *x* axis: *never essential reactions*. Inside *triangle*: *partially essential reactions*. Extracted from Ref. [17] Copyright @ 2014, World Scientific Publishing

4. **Partially essential reactions**: $0 < a_i < 1$, $0 < e_i < a_i$. They are essential only a fraction of times when they are active. These reactions are located inside the triangle formed by the diagonal, and the *y* and *x* axes.

To understand the obtained results, the study focuses on the different subnetworks obtained by filtering the complete original network according to the four basic explained categories. More precisely, these subnetworks are obtained by maintaining in the network only those reactions within the respective mentioned categories. Once these subnetworks are obtained, the number of connected components in the subnetwork are computed in order to know whether the selected subnetwork is fragmented or not. In particular, the giant connected component (GCC) and the strongly connected component (SCC) (see Chap. 2, Sect. 2.1.5) of the subnetworks are computed. This study is done in order to detect whether reactions within a specific type are responsible for the percolation state of the network.

In Table 4.1, the statistics of active and essential reactions are summarized together with values of the sizes of the connected components of the subnetworks. Results correspond to the set of minimal media. A precise discussion of such statistics is provided on what follows. Notice first that there are several reactions which are strictly never active (902). This may be explained by the fact that these computations have been done in minimal media, which may only activate a few number of reactions needed to survive. In addition, it can be seen that the complete network, which corresponds to values of activity $0 \leq a_i \leq 1$ and essentiality $0 \leq e_i \leq a_i$, is constituted by a single GCC and that, in addition, it has a large SCC, a typical situation in metabolic networks.

Table 4.1 Connected components and number of reactions N_R in each subnetwork

Category	CC	N_R
Essential whenever active reactions	**Total**	**665**
$a_i > 0$	**GCC**	**611(91.9)**
$e_i = a_i$	**SCC**	**409 (66.4)**
Essential and active in some media	Total	458
$0 < a_i < 1$	GCC	409 (89.3)
$e_i = a_i$	SCC	200 (48.8)
Essential and active in all media	Total	207
$a_i = 1$	GCC	198 (95.7)
$e_i = a_i$	SCC	174 (86.1)
Always active reactions	**Total**	**37**
$a_i = 1$	**GCC**	**34(91.9)**
$0 < e_i < a_i$	**SCC**	**29(85.3)**
Never essential reactions	**Total**	**494**
$0 < a_i < 1$	**GCC**	**494**
$e_i = 0$	**SCC**	**476(96.4)**
Partially essential reactions	**Total**	**152**
$0 < a_i < 1$	**GCC**	**145(95.4)**
$0 < e_i < a_i$	**SCC**	**129(90.0)**
All reactions	Total	2250
$0 \leq a_i \leq 1$	GCC	2250
$0 \leq e_i \leq a_i$	SCC	2076 (92.0)
Never active	Total	902
$a_i = 0$		
$e_i = 0$		

Values in parentheses correspond to percentages. GCC percentages are computed by dividing the number of reactions in GCC relative to the total number of reactions in each category, whereas SCC percentages are computed by dividing the number of reactions in SCC relative to the number of reactions in the GCC subnetwork. Categories in bold correspond to the four basic categories mentioned in the text. Extracted from Ref. [17] Copyright @ 2014, World Scientific Publishing

4.1.2.1 Essential Whenever Active Reactions

A histogram of the values of the essentiality of the set of reactions *essential whenever active reactions* is shown in Fig. 4.2a, b. A bimodal distribution is clearly displayed, with peaks at extreme values, $a_i = e_i \simeq 0$ and the other at $a_i = e_i \simeq 1$. This means that there is a core of reactions that are always active and essential, as pointed out in Ref. [2], and there is another set of reactions that are active very few times. This histogram coincides with the classification of the dependence of essentiality on the environment given in Ref. [5]. The peak at values of activity $\sim$0 corresponds to *environment-specific* essential reactions, whereas the peak at values of activity ~ 1 corresponds to *environment-general* essential reactions. The first

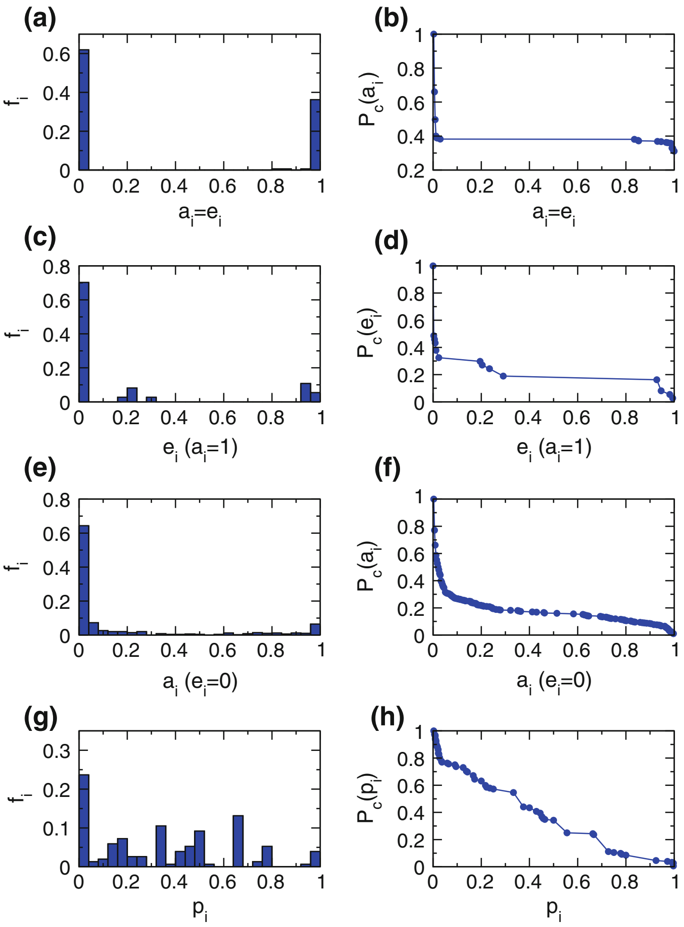


Fig. 4.2 Histograms (fraction) and complementary cumulative probability distribution function of activity or essentiality (depending on the category) for minimal media. **a**, **b** Essential whenever active reactions. **c**, **d** Always active reactions. **e**, **f** Never essential reactions. **g**, **h** Partially essential reactions. Extracted from Ref. [17] Copyright @ 2014, World Scientific Publishing

region includes reactions whose deletion abolishes growth in specific environments, whereas the second one corresponds to reactions whose deletion suppresses growth in all environments.

A deeper characterization of this set of reactions is made in Table 4.1, which shows that this subnetwork has a large GCC with nearly 90% of the subnetwork. If reactions with activity-essentiality index of 1 are excluded from this subnetwork, another subset is obtained which has also a large GCC (89.3% of the total 458 reactions). This means that reactions with $a_i = e_i = 1$ are not responsible for the percolation state of the subnetwork of *essential whenever active reactions*, which points out to a large degree of redundancy.

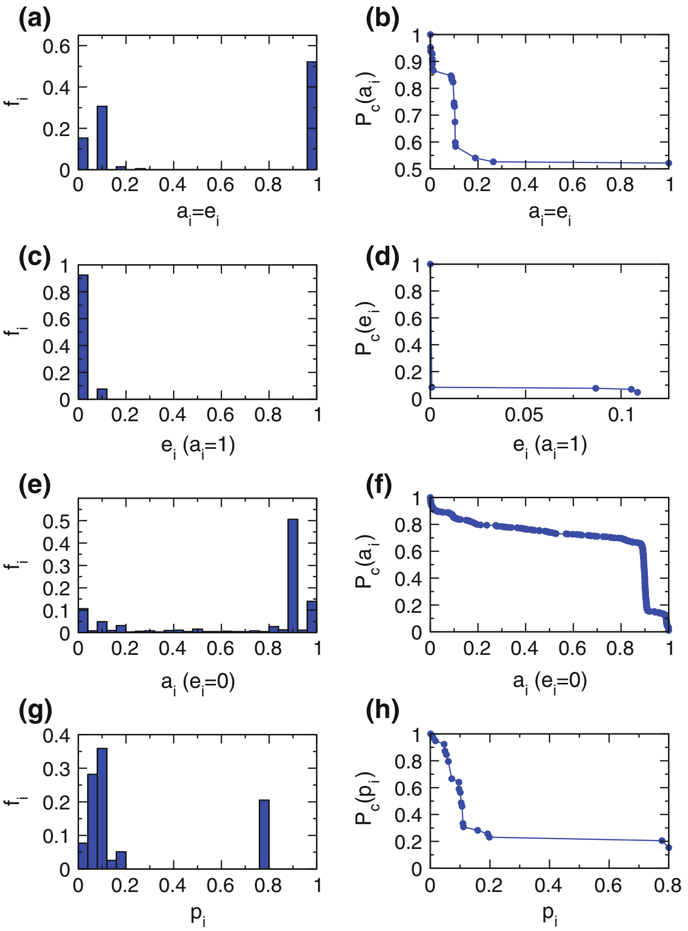


Fig. 4.3 Histograms (fraction) and complementary cumulative probability distribution function of activity/essentiality (depending on the category) for random media. **a**, **b** Essential whenever active reactions. **c**, **d** Always active reactions. **e**, **f** Never essential reactions. **g**, **h** Partially essential reactions. Extracted from Ref. [17] Copyright @ 2014, World Scientific Publishing

An illustrative example of a particular reaction in this subcategory is *Potassium transport* (Ktex). This is a reaction which supplies the organism with potassium. This mineral salt is an important metabolite which influences the osmotic pressure through the cell membrane and also secures the propagation of electric impulses. Since these are important processes for organisms, this reaction is always active in order to secure that these processes are done properly and that the organism has a non-zero growth.

For random media (see Fig. 4.3a, b), a similar behaviour is obtained, with larger probabilities at the extrema, but an extra peak is obtained for low values of activity and essentiality. This means that there are some reactions which are not as specific as *environment-specific* reactions because they are active and essential in more than one medium, loosing in this way their specificity. This makes sense for random media, since they contain many metabolites that may activate many reactions and, in this way, they lose the specificity of a minimal medium, which triggers only the reactions that allow an organism to grow on it.

4.1.2.2 Always Active Reactions

The set of reactions called *always active reactions* contains reactions with $a_i = 1$ and $0 < e_i < a_i$. In Fig. 4.2c, d one can see that, in this case, there is a large peak at values of $e_i = 0$, meaning that the largest part of reactions with $a_i = 1$ have a value of $e_i = 0$. This means that, although these reactions are always active, they are not essential. One may be tempted to think that reactions with very low values of essentiality are useless and hence they could be removed from the network. Nevertheless, there are two reasons that justify their consideration.

- The first one is that these reactions may improve the life conditions of the organism. These reactions, in spite of being non-essential, might be active in order to increase the growth of the organism. As a matter of fact, to survive to hard conditions, an organism which is able to reproduce fast and efficiently will, with large probability, survive to unfriendly life conditions.
- The second one is more subtle. These reactions could form SL pairs. As an example, two reactions regulated by the genes *tktA* and *tktB*, which are in the peak at $e_i = 0$ and $a_i = 1$, form a synthetic lethal pair and the removal of these reactions would abolish growth by impeding the synthesis of nucleotides, nucleic acids, and aromatic amino acids. Briefly, the reactions regulated by these mentioned genes, called TKT1 and TKT2 and both with a complete name of *Transketolase*, are reactions which belong to the Pentose Phosphate Pathway. This pathway generates NADPH and pentoses phosphate, the latter being a precursor used in the synthesis of nucleotides, nucleic acids and aromatic amino acids. Both reactions are always active to ensure a sufficient production of these mentioned products, and when one of these reactions is knocked out, the other reaction is in charge to restore this function.

In Fig. 4.3 one can see that, as in *essential whenever active reactions*, this category of *always active reactions* form a subnetwork with a GCC which is almost the full subnetwork with also a large SCC.

Note that for random media (see Fig. 4.3c, d), a similar trend to minimal media is obtained.

4.1.2.3 Never Essential Reactions

Never essential reactions have values of activity and essentiality which satisfy $e_i = 0$ and $0 < a_i < 1$. The histogram of the values of the activities for these reactions is shown in Fig. 4.2e, f. A similar histogram to that corresponding to *always active reactions* is recovered again. This means that, not surprisingly, the largest part of *never essential reactions* are not much active. The individual removal of these reactions will leave the growth rate unaltered or only reduced. The existence of these reactions could be explained again in terms of improving the growth of the organism and, again, for the possibility of participating in SL pairs.

An example of a reaction of this kind is *Manganese transport in via permease*(no H+), MN2tpp, a reaction which pumps manganese into the organism. Its non-essentiality comes from the fact that there exists an alternative reaction called *Manganese*(Mn+2)*transport in via proton symport (periplasm)*, MNt2pp, which also pumps manganese into the organism, but the latter uses a proton gradient to perform the transport.

In Table 4.1 one can see again the same trend as the other categories of reactions. This subnetwork contains a GCC that is almost the full subnetwork with a large SCC.

For random media, different results are obtained in this case (see Fig.4.3e, f). The largest peak is located at large values of activity, which means that there is a large set of reactions which are mainly active but never essential. The peak located slightly above 0.8 could appear due to the fact that the random media are in fact rich media. Hence, it is possible that a common set of metabolites activate the same reactions in many media. These reactions are responsible for the increase of the value of the flux of the biomass reaction.

4.1.2.4 Partially Essential Reactions

Partially essential reactions contain reactions with activity and essentiality values of $0 < a_i < 1$ and $0 < e_i < a_i$. Since these reactions have both values of essentiality and activity different from zero, the histogram is represented in terms of $\frac{e_i}{a_i}$ as shown in Fig. 4.3g, h. The distribution is rather homogeneous, meaning that these two quantities may be largely uncorrelated, their ratio spanning the whole range of allowed values.

Table 4.1 shows again a large GCC containing a large SCC. Notice that this trend has been maintained for all categories of reactions.

Again, different results are obtained for random media (see Fig. 4.3g, h). In this case, homogeneously distributed values as for minimal media are not obtained. Instead, the behaviour resembles that of the *essential whenever active* subset, they are concentrated at low values and at a value of $\frac{e_i}{a_i} = 0.8$. This means that reactions are essential in fewer environments as compared to those in which they are active, showing again that activity does not imply essentiality.

4.2 SL Pairs and Plasticity and Redundancy of Metabolism

In metabolism, *synthetic lethality* arises when the individual failures of two reactions are not essential for cell growth but, contrarily, their simultaneous removal causes cell death [9–12, 19, 20].

Synthetic lethality has been originally proposed in relation to genes [9–13]. Its definition is that two genes are synthetic lethal when their individual knockout does not lead to the death of the organism but when both genes are removed simultaneously the organism is not able to overcome which leads to the death of organism (see Fig. 4.4). Genes code for enzymes, and enzymes determine the kinetics of reactions and thus whether reactions take place in a feasible amount of time. Therefore, as for essentiality of individual genes and reactions, it is possible to extend the concept of synthetic lethality to reactions.

FBA is a powerful technique particularly suited for an exhaustive *in silico* prediction of SL pairs [6, 21]. Using FBA, a reaction pair deletion is annotated as non-viable, and so as a synthetic lethal, if the double mutant shows a no-growth phenotype.

This section presents the study of plasticity and redundancy of metabolism by directly computing the effects of double reaction knockouts, excluding those reactions that are individually essential in order to identify SL pairs. On what follows,

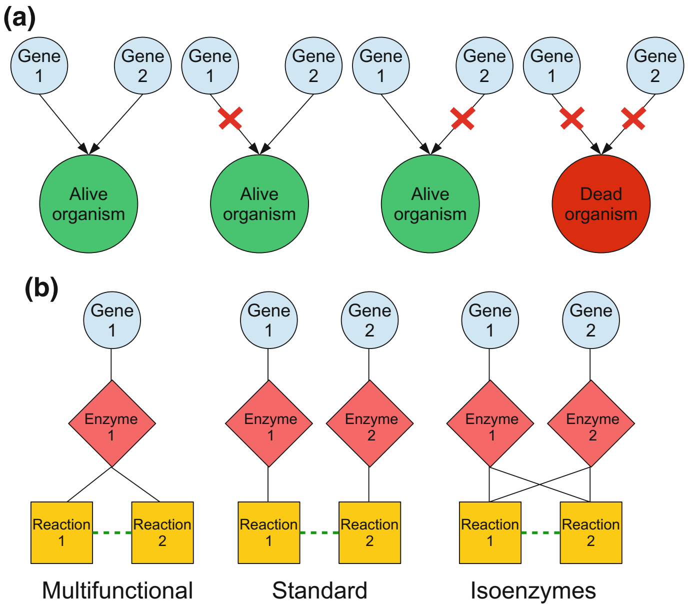


Fig. 4.4 Synthetic lethality schemes. **a** Simplified scheme to illustrate the concept of synthetic lethality of two genes. **b** Different possible organization of genes, enzymes and reactions in SL pairs. Parts of this figure have been extracted from Ref. [18]

a detailed analysis of the classification of identified SL reaction pairs into plasticity and redundancy subtypes in the *i*JO1366 version of *E. coli* and in the *i*JW145 version of *M. pneumoniae* (see Chap. 2, Sect. 2.3.2) is presented.

4.2.1 Classification of SL Pairs

Some considerations are needed in relation to the space of reactions to be considered in forming potential SL pairs, the set of reactions that can be active but not essential in glucose minimal medium (see Chap. 2, Sect. 2.2.2.1). Different from the analysis in the previous section, in this section the study is primarily focused in one medium, not in a set of environments. In addition, the space of reactions to be considered is preliminary reduced using a method that we call "Biomass unconstrained Flux Variability Analysis", where Flux Variability Analysis (FVA) is applied irrespective of the level of attainable growth (see Chap. 2, Sect. 2.2.4). The final ensemble, formed of 1176 reactions in *E. coli* and 66 in *M. pneumoniae*, is a subset of the original reconstruction that includes but that is not limited to the set of FBA active reactions under maximum growth constraint [22, 23].

An important remark is worth mentioning at this point. Some FBA computationally predicted SL pairs can be inconsistent with experimental data since they may contain at least one gene reported as essential *in vivo*. For *E. coli*, results are checked with essentiality information given in Ref. [24]. Given the lack of direct evidence, results for *M. pneumoniae* are compared to a genome-wide transposon study in *Mycoplasma genitalium* given in Ref. [25]. Since a functional ortholog in *M. genitalium* can be assigned to 128 metabolic genes in *i*JW145 (of a total of 145 genes), the essentiality of that ortholog can be associated to the corresponding gene in *M. pneumoniae*. The other 17 genes are assumed, similarly to Ref. [26], to be not essential for growth due to their absence in *M. genitalium* and the high similarity of the metabolic networks of both organisms [27]. Three cases may occur when FBA *in silico* results are compared to experimental essentiality:

- Both reactions in the *in silico* SL pair involve non-essential genes. In this case, the pair can be considered a potential synthetic lethal (see Fig. 4.5a).
- One reaction involves a non-essential gene whereas the other is regulated by an essential one. In this case, if the essential gene regulates more than one reaction, one can consider that the *in silico* prediction is not an inconsistency (see Fig. 4.5c), since the essentiality might refer to the rest of regulated reactions. Otherwise, the pair is considered as inconsistent with experimental data (see Fig. 4.5b).
- Both reactions are regulated by essential genes. With the same argument as before, for the case that both reactions have associated genes which regulate more than one reaction, one can still consider the pair to be a potential synthetic lethal (see Fig. 4.5d). The other possible combinations are considered inconsistent with empirical evidence (see Fig. 4.5e).

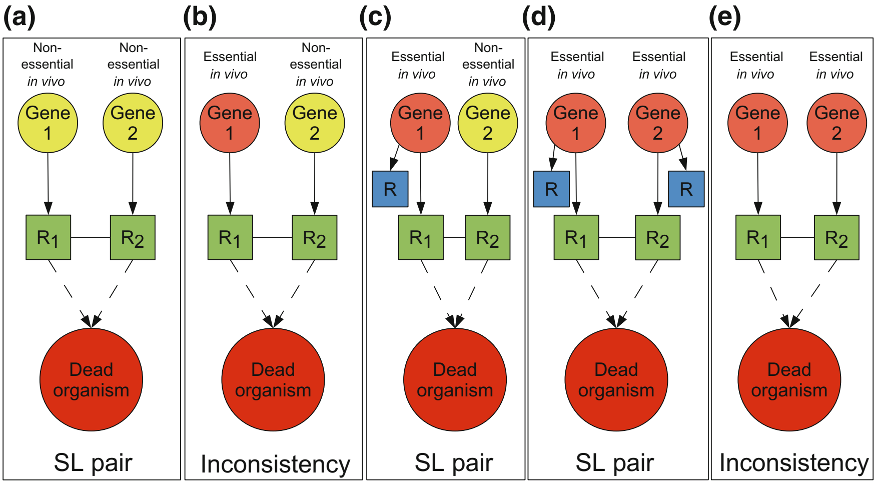


Fig. 4.5 Schematic representation of the identification synthetic lethal inconsistencies

Detected SL pairs associated to isoenzymes (see Fig. 4.4b) and multifunctional enzymes (see Fig. 4.4b) are also classified as inconsistencies. Isoenzymes (also known as isozymes) are enzymes that differ in amino acid sequence but that catalyse the same chemical reaction. In this way, a reaction can be catalysed by two different enzymes in case that one of them becomes non-operative. Multifunctional enzymes are those that can catalyse more than one reaction at the same time. They are very important for organisms, since they are responsible of the catalysis of more than one reaction and their failure may cause important damage to organisms, since many reactions can become non-operative.

4.2.2 *Classification of SL Reactions Pairs into Plasticity and Redundancy*

Of all reaction pair deletions in *E. coli*, 0.04% are *in silico* synthetic lethals and can be separated in two different subtypes. In the biggest group, having a relative size of 91%, one of the paired reactions is active in the medium under evaluation while the second reaction has no associated flux. The rest of SL reaction pairs are formed by two active reactions. Moreover, in accordance with results in Ref. [6], it is found that inconsistencies correspond to 4% of all identified *in silico* SL pairs in *E. coli*.

Active-inactive coessential reaction pairs are referred to as *plasticity synthetic lethal* (PSL) pairs (see Fig. 4.6a). 219 PSL reaction pairs are found in *E. coli*, 86% of all diagnosed SL pairs in the *i*JO1366 version of *E. coli* (see Fig. 4.7). Coessential inactive and active reactions in these pairs have zero and non-zero FBA flux respectively. When the active reaction is removed from the metabolic network, fluxes

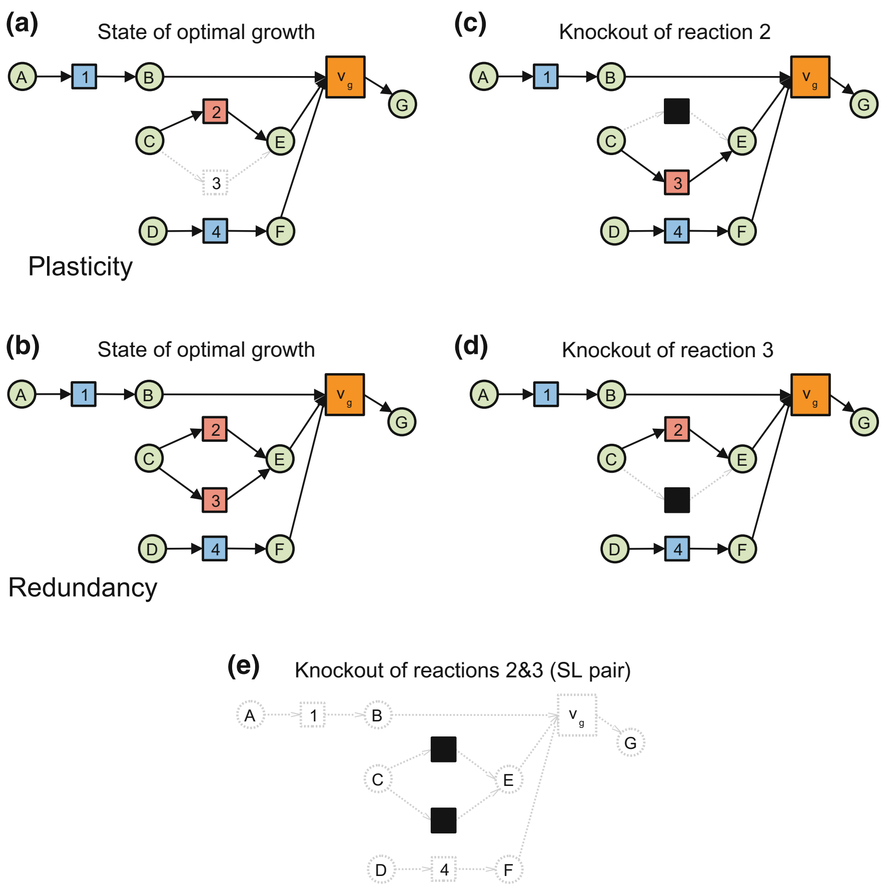


Fig. 4.6 Schematic representation of plasticity and redundancy synthetic lethality subtypes in metabolic networks. Metabolites are represented by *circles* and reactions by *squares*. Coloured reactions with *black arrows* represent active reactions, whereas *gray discontinuous lines* are used for inactive reactions and metabolites and *black* for knockouts. The biomass production reaction is represented as a *larger square* with an associated flux ν_g. When it turns to inactive, meaning that it has no associated flux, the organism is not able to grow. For simplicity, SL reaction pairs are illustrated in this figure as having a common metabolite, although this is not necessarily always the case. **a** Initial configuration of a plasticity synthetic lethality reaction pair (reaction 2 active and reaction 3 inactive). **b** Initial configuration of a redundancy synthetic lethality reaction pair (both reactions 2 and 3 active). **c** Final configuration after knockout of reaction 2 in **a** or **b**. **d** Final configuration after knockout of reaction 3 in **a** or **b**. **e** Final configuration after simultaneous knockout of reactions 2 and 3 in **a** or **b**. Extracted from Ref. [18]

reorganize such that the zero-flux reaction in the pair turns on as a backup of the removed reaction to ensure viability of the organism, even though the growth is generally lowered. In contrast, the level of growth is unperturbed when the inactive reaction is removed. As an example, the SL pair valine-pyruvate aminotransferase

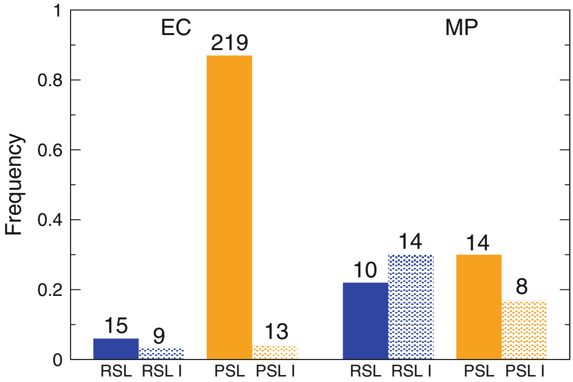


Fig. 4.7 Histogram for the four different categories of SL pairs in *E. coli* (*left*) and *M. pneumoniae* (*right*). RSL (*blue*) redundancy synthetic lethal pairs, RSL I (*blue points*) redundancy synthetic lethal pairs showing inconsistencies, PSL (*orange*) plasticity synthetic lethal pairs, PSL I (*orange points*) plasticity synthetic lethal pairs showing inconsistencies. Extracted from Ref. [18] (color figure online)

and valine transaminase form a PSL pair, the second reaction being the backup of the first, whose simultaneous knockout produces auxotrophic mutants requiring isoleucine to grow [28].

While the single activation of one of the reactions in a PSL pair is enough to ensure viability in front of single reaction disruptions, the parallel use of both coessential reactions may happen in other cases. *Redundancy synthetic lethal* (RSL) pairs are those in which both reactions are active and used in parallel (see Fig. 4.6b). Of all SL reaction pairs in the *i*JO1366 version of *E. coli*, one finds that 15 (6%) are RSL (see Fig. 4.7). Indeed, for 13 of the 15 RSL pairs the simultaneous use of both reactions increases fitness as compared to the situation when only one of the reactions is active (fitness is here understood as the maximal FBA biomass production rate for the organism). For the remaining two pairs growth remains unchanged. As an illustrative example of parallel use, oxygen transport combines with reactions in the ATP forming phase of Glycolysis to form RSL reaction pairs. If Oxydative Phosphorylation is blocked by the absence of oxygen and no alternative anaerobic process like Glycolysis is used, the energy metabolism of *E. coli* collapses and so the whole organism.

It is interesting to compute the shortest path length (see Chap. 2, Sect. 2.1.3) between reactions in SL pairs. It is found that network distances between reaction counterparts is slightly shorter in RSL pairs than in PSL pairs. Indeed, not all reactions in RSL or PSL pairs are directly connected through common metabolites. Direct connections happen for 60 and 38% of pairs respectively, while the rest can be separated by up to four other intermediate reactions so that the average shortest paths are 3.33 and 3.80, respectively (the average shortest path of the whole metabolic network is 5.02). Both essential plasticity and redundancy display overlap in reactions and associated genes. In the 15 RSL pairs, one can identify 17 different reactions controlled by 15 genes or gene complexes. The 219 PSL pairs involve 108 different reactions controlled by 61 genes or gene complexes.

Although this analysis refers to reactions, specific signatures of enzyme activity may be worth stressing in connection with the analysis of coessential reaction pairs. For some of the identified SL pairs, direct experimental evidence is reported in the literature [28, 29]. Other experimental results support the buffering activity of reactions in some SL pairs, like in the aerobic/anaerobic synthesis of Heme [30, 31] and in the oxidative/non-oxidative working phases of the Pentose Phosphate Pathway [32]. Enzymatic degeneracy can be responsible for explaining two of the *in silico* detected RSL reaction pairs in *E. coli*. One RSL reaction pair, which produces isopentenyl diphosphate and its isomer dimethylallyl diphosphate -biosynthetic precursors of terpenes in *E. coli* that have the potential to serve as a basis for advanced biofuels [33]— is catalysed by a single enzyme encoded by an essential gene (one-to-many enzyme multifunctionality (see Fig. 4.6f)). Conversely, isoenzymes are encoded by different genes but can catalyse the same biochemical reactions. This many-to-one relationship ensures that single deletion mutants lacking any of the genes encoding one of the isoenzymes can still be viable (see Fig. 4.6f). This case happens in one RSL reaction pair catalysed by isoenzymes encoded by non-essential genes associated to transketolase activity in the Pentose Phosphate Pathway [20].

Finally, a comparative study shows that coessential reaction pairs are 50 times more abundant in a much simpler genome-reduced organisms of increased linearity and reduced complexity such as *M. pneumoniae*. To perform the computations, the medium given in Table S5 of the Supplementary Information of Ref. [26] is used. Constraints corresponding to the category called defined medium have been used, adding also D-ribose. 2% of all potential candidate reaction pairs in *M. pneumoniae* are synthetic lethals versus solely the 0.04% in *E. coli*. Inconsistencies are also much more abundant relatively to *E. coli* and the balance of RSL vs PSL reaction pairs is also different (see Fig. 4.7). Parallel use happens as frequently as the backup mechanism in coessential reactions, with 42% of all synthetic lethals being RSL pairs and 58% being PSL pairs. As compared to results reported in Ref. [26] for the synthetic lethality of genes, the used methodology detects the same 29 SL gene pairs and 15 new SL gene pairs. Since the 8 different genes in these pairs form two different complexes of four and three genes and one gene remains isolated, the 15 SL gene pairs reduce to just 2 SL reaction pairs (in the RSL and RSL I categories) sharing one of the reactions. The three reactions involved in the pairs are uptake of G3P (glycerol 3-phosphate), G3P oxidation to dihydroxyacetone phosphate, and uptake of orthophosphate. As reported in Ref. [26], two independent routes through third-party pathways connect Glycolysis to Lipid Biosynthesis. The first two reactions above, R1 and R2, are involved in one of the routes, while the last reaction R3 influences the flux through the other route. When R1 and R3 or R2 and R3 are removed from *i*JW145 model, the organism collapses due to the simultaneous failure of both routes.

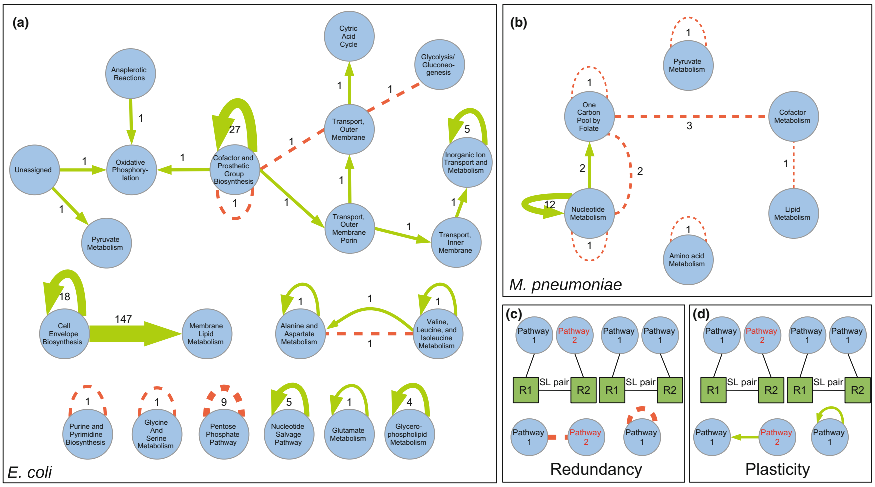

(a)
Cytric Acid Cycle
Glycolysis/ Gluconeo- genesis
Anaplerotic Reactions
1
1
1
Transport, Outer Membrane
27
5
1
Unassigned
1
Oxidative Phosphory- lation
1
Cofactor and Prosthetic Group Biosynthesis
Inorganic Ion Transport and Metabolism
1
1
1
1
1
Pyruvate Metabolism
Transport, Outer Membrane Porin
1
Transport, Inner Membrane
18
147
Cell Envelope Biosynthesis
Membrane Lipid Metabolism
1
1
1
Alanine and Aspartate Metabolism
Valine, Leucine, and Isoleucine Metabolism
1
1
1
9
5
1
4
Purine and Pyrimidine Biosynthesis
Glycine And Serine Metabolism
Pentose Phosphate Pathway
Nucleotide Salvage Pathway
Glutamate Metabolism
Glycero- phospholipid Metabolism
E. coli
(b)
1
Pyruvate Metabolism
1
One Carbon Pool by Folate
3
Cofactor Metabolism
2
2
1
12
Nucleotide Metabolism
Lipid Metabolism
1
1
Amino acid Metabolism
M. pneumoniae
(c)
Pathway 1
Pathway 2
Pathway 1
Pathway 1
R1
SL pair
R2
R1
SL pair
R2
Pathway 1
Pathway 2
Pathway 1
Redundancy
(d)
Pathway 1
Pathway 2
Pathway 1
Pathway 1
R1
SL pair
R2
R1
SL pair
R2
Pathway 1
Pathway 2
Pathway 1
Plasticity

Fig. 4.8 Metabolic pathways entanglement through essential plasticity and redundancy in *E. coli* and *M. pneumoniae*. *Nodes* represent pathways and two pathways are joined by a link whenever there exists a SL pair containing one reaction in each pathway. Links corresponding to plasticity SL pairs are represented by *green continuous arrows* pointing from backup to active. Redundancy SL pairs are represented by *discontinuous red lines*. Labels correspond to the number of pairs which generate this combination of pathways, being thicker those links with more associated pairs. Self-loops correspond to SL pairs with both reactions in the same associated pathway. **a** Pathways entanglement in *E. coli*. **b** The same for *M. pneumoniae*. **c** Scheme of how pathways entanglement is derived from RSL pairs. **d** The same for PSL pairs. Extracted from Ref. [18] (color figure online)

4.2.3 Pathways Entanglement

To investigate further the role of essential plasticity and redundancy in the global organization of metabolic networks, one can study the entanglement of biochemical pathways [34] through synthetic lethality. To do this, it is necessary to annotate all reactions in synthetic lethal pairs in terms of the standard metabolic pathway classification and to count the frequencies of dual pathways combinations both for plasticity and redundancy subtypes. In Fig. 4.8, a visual summary of pathways entanglement through essential plasticity and redundancy is given. A graph representation is used, where pathways are linked whenever they participate together in a SL interaction (discontinuous lines represent redundancy SL interactions (see Fig. 4.8c) and continuous arrows stand for plasticity SL interactions (see Fig. 4.8d). The frequency of a given pathway combination in RSL or PSL pairs defines the weight of the corresponding link.

In *E. coli* (see Fig. 4.8a), one can observe that the synthetic lethality entanglement of pathways is in general very low, with the exception of the entanglement between Cell Envelope Biosysthesis and Membrane Lipid Metabolism. Redundancy SL pairs are basically intra-pathway, with only 3 of 15 being inter-pathway. Of all intra-pathway RSL pairs, 75% concentrate in the Pentose Phosphate pathway. Interestingly, the distribution of PSL reaction pairs avoids that of RSL pairs and, in contrast, tends to be inter-pathway. Of all PSL pairs, 67% include zero-flux reactions in Cell Envelope Biosysthesis and active reactions in the Membrane Lipid Metabolism, which unveils Cell Envelope Biosysthesis as an essential backup for Membrane Lipid Metabolism. Intra-pathway plasticity coessentiality amounts to 29% of PSL pairs and is concentrated in Cofactor and Prosthetic Group and Cell Envelope Biosynthesis.

In *M. pneumoniae* (see Fig. 4.8b) pathways entanglement through coessentiality of reactions is very low as in *E. coli*. Redundancy SL pairs can be intra-pathway (4 of 10) or inter-pathway (6 of 10) and PSL pairs are basically intra-pathway (12 of 14). Redundancy SL pairs denote the parallel use of reactions in Folate Metabolism and reactions in Nucleotide and Cofactor Metabolism. These two pathways, Folate and Nucleotide Metabolism, are also linked by two PSL pairs with non-essential reactions in Folate Metabolism and essential reaction backups in Nucleotide Metabolism. Nucleotide Metabolism is also the pathway that concentrates most PSL pairs. Both RSL and PSL reaction pairs unveil Nucleotide and Folate Metabolism as the most entangled pathways. Taken together, these results indicate that Folate and Nucleotide Metabolic pathways preserve most rescue routes for reaction deletion events, in accordance with results in Ref. [26]. The fact that the proportion of plasticity SL pairs is considerably decreased in *M. pneumoniae* as compared to *E. coli* could be indicative that, even if both plasticity and redundancy serve an important function in achieving viability, essential plasticity is a more sophisticated mechanism that requires a higher degree of functional organization, using at the same time less resources for maximum growth. At the same time, this can also be explained by the relative unchanging environmental conditions of *M. pneumoniae* in the lung, that could have induced the elimination of pathways not required in that medium [26]. This suggests that the adaptability of *M. pneumoniae* is very much reduced and its behaviour could not be resilient to environmental changes.

4.2.4 Sensitivity to Differences on Environmental Conditions

The last part of this section presents the analysis of plasticity and redundancy depending on the growth condition under evaluation. Environmental specificity of genes and reactions has been explored experimentally [24, 35, 36] and *in silico* [5] for different organisms and for random viable metabolic network samples, and it has also been extended to multiple knockouts in yeast [14, 21] and *E. coli* [37].

To investigate the sensitivity of SL reaction pairs in *E. coli* to changes in minimal medium composition, the study focuses on the 234 SL pairs detected in glucose minimal medium and checks their classification over the 333 minimal media constructed as in the previous Sect. 4.1. Figure 4.9a shows the SL reaction pairs ranked by the fraction of media in which the pairs are synthetically lethal. For most pairs, coessentiality is not specific of an environment and only a minimal number of pairs shows environmental specificity. In particular, 53% coessential pairs are lethal in all media and 95% are lethal in more than 95% of environments. For each SL pair, one can

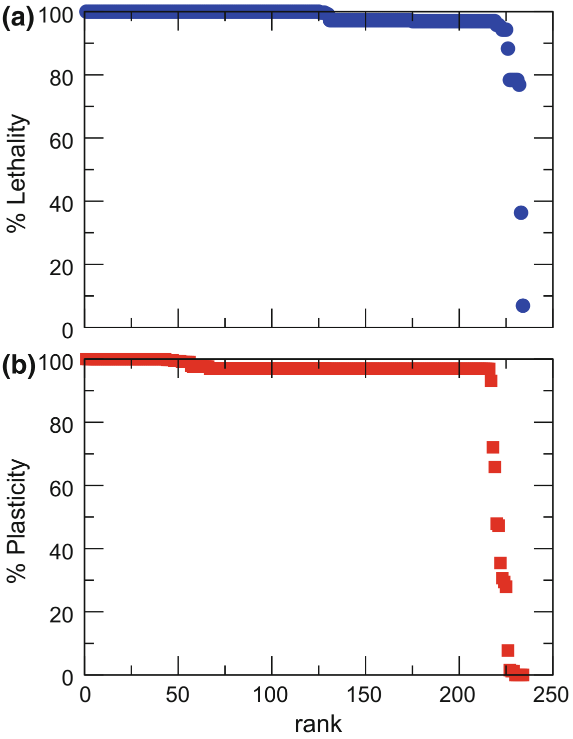


Fig. 4.9 Synthetic lethal reaction pairs in minimal media. **a** Synthetic lethal reaction pairs ranked by the fraction of minimal media for which the pair is synthetically lethal. **b** Synthetic lethal reaction pairs ranked by the fraction of minimal media in which the SL pairs are classified as essential plasticity and, complementary, as essential redundancy, provided that the pairs remain synthetically lethal. Extracted from Ref. [18]

count the number of media in which the SL pair is classified in the plasticity subtype as compared to the total number of media in which the pair is predicted to be coessential. Results are shown in Fig. 4.9b. Nearly all SL pairs, 93%, are in the plasticity subclass for more than 93% of the media, while 12 pairs display a switching behaviour between plasticity and redundancy. Noticeably, these pairs are intra-pathway and share common metabolites. Of them, three pairs contribute to biosynthesis of amino acids (Valine, Leucine, and Isoleucine Metabolism and Glycine and Serine Metabolism) and five pairs belong to the Pentose Phosphate Pathway and are related to the production of carbon backbones used in the synthesis of aromatic amino acids. Finally, five reaction pairs maintain in the redundancy subclass across all conditions in which are coessential.

The behaviour of *E. coli* can be explored in an amino acid-enriched medium (see Sect. 2.2.2.2). Comparing with glucose minimal medium, the first observation is that 223 of the 234 SL pairs detected in glucose minimal medium are also found to be lethal in amino acid-enriched medium, which means that 11 pairs are rescued. Of the 11 RSL pairs in amino acid-enriched medium, eight are conserved and three switch from plasticity in the minimal to redundancy in the amino acid-enriched medium. On the other hand, 208 of the 212 PSL pairs are conserved and four change from redundancy in the minimal to plasticity in the amino acid-enriched medium.

Noticeably, only in one of the 208 conserved PSL pairs the pattern of activity changes from the reductase reaction producing dimethylallyl diphosphate to the isomerization of the less reactive isopentenyl pyrophosphate. In addition, a new set of 12195 lethal reaction pairs occurs, all of them involving however one essential reaction in glucose minimal medium that in amino acid-enriched medium becomes non-essential and instead takes part in SL pairs. Apart from those, no other new SL pairs are found.

In addition, this study also considers a rich medium. To construct this rich medium, a Luria-Bertani Broth (see Sect. 2.2.2.2) has been taken into consideration. In this rich medium, 13 new rescues are found when compared to the minimal medium (two new rescues as compared to the amino acid-enriched medium) and only three SL pairs change their plasticity/redundancy category.

Plasticity and redundancy are still conserved when the growth maximization requirement is loosen. To implement the relaxation of the growth maximization requirement, again the glucose minimal medium is taken as a reference and the biomass production or the basic nutrients uptake rates are limited. In the first case, a FVA calculation is performed fixing the growth of the biomass to 30% of the maximal growth in glucose minimal medium and the exchange bounds of all nutrient uptakes are obtained. In comparison to the reference values, one can observe that the only metabolites which lower their maximal uptakes are the mineral salts (approximately reduced also to a 30%), while the uptake rates for the rest of compounds remained with the same bounds. Then, it is possible to perform FBA calculations in this overconstrained condition and compute SL pairs and their classification in RSL and PSL. If growth is relaxed in *E. coli* to 30% of its maximum value in glucose minimal medium by doing this, *in silico* essentiality of individual reactions does not change but activation of reactions increases. It is found, however, that the effect of this reorganization is indeed mild for plasticity and redundancy. All SL pairs are conserved and 82% of them maintain their PSL or RSL classification. The absolute number of RSL pairs increases from 15 to 50 since four RSL pairs in the reference condition given by glucose minimal medium change to plasticity in the overconstrained medium, and at the same time 39 PSL pairs change to RSL. On the other hand, 180 SL pairs of 219 in the reference medium remain as PSL pairs in the overconstrained condition. However, the pattern of activity in the pair has switched in 14% of the PSL pairs in this case, which indicates that the specific selection of the active reaction in a PSL pair can have an impact in the level of attainable growth.

If instead of limiting the uptake of mineral salts, the uptake rates of basic nutrients providing sources of carbon, nitrogen, phosphorus and sulphur are overconstrained, the effect is even softer and indeed negligible as compared to the reference medium. To do this overcostraining, it is necessary to first apply FVA setting the value of biomass growth to the maximum in glucose minimal medium in order to determine an upper uptake limit. Then, the maximum rate uptake of glucose and of the other three basic compounds is constrained to 30% of the maximum possible values while keeping the reference values for the mineral salts. FBA is then applied in the resulting overconstrained medium and SL pairs and their classification in RSL and PSL are computed. The number of active reactions only increases in three, the essentiality of individual reactions and SL pairs is conserved, and 99% of them maintain their PSL or RSL classification with only three SL pairs that switch class and only one PSL pair that changes the active reaction.

In both overconstrained modifications of the glucose minimal medium, the number of active reactions changed from 412 to 490 in the mineral salts overconstrained medium and to 415 in the basic nutrients overconstrained medium. It is important to stress that, in both cases, the essentiality of individual reactions and all SL pairs were conserved (except for two new RSL inconsistencies in the basic nutrients overconstrained medium).

4.3 Conclusions

The first part of this chapter presents the results of a study of the activity and the essentiality of single reactions of *E. coli* in different environments. Reactions can be divided in four categories depending on their values of essentiality and activity. By doing this, one recovers *environment-specific* and *environment-general* reactions as given in Ref. [5]. These correspond to the bimodal behaviour in the category called *essential whenever active reactions*. Given their importance, these reactions can be selected as drug targets since they are fundamental constituents of the metabolism of *E. coli*. Another important feature that can be observed is the fact that some reactions, in spite of being never essential, are always active, which may favor an increase of the growth rate of the organism and the robustness of metabolism through redundancy. The categories of reactions which show this behaviour are *always active reactions* and *never essential reactions*. The last feature that one can extract from the category *partially essential reactions* is that active reactions are not necessarily essential. Therefore, in general extrapolating activity to essentiality is not correct.

Beyond the essentiality of single reactions, SL pairs are complex functional combinations of reactions (or genes) that denote at the same time both vulnerability in front of double deletions and robustness in front of the failures of any of the two counterparts. Working at the level of reactions, synthetic lethality is meditated by two different mechanisms, essential plasticity and essential redundancy, depending on whether one reaction is active for maximum growth in the medium under consideration and the second inactive, or in contrast both reactions have non-zero flux. Plasticity sets up as a sophisticated backup mechanism (mainly inter-pathway in *E. coli*) that is able to reorganize metabolic fluxes turning on inactive reactions when coessential counterparts are removed in order to maintain viability in a specific medium. Redundancy corresponds to a simultaneous use of different flux channels (mainly intra-pathway in *E. coli*) that ensures viability and besides increases fitness. Apparently, it could seem extremely improbable that the removal of an inactive reaction together with a non-essential active one, like in PSL pairs, could have any lethal effect on an organism. However, it is found that this situation is indeed overwhelmingly dominant in *E. coli* as compared to redundancy synthetic lethality, and it is still relatively frequent even in a less complex organism like *M. pneumoniae*.

Synthetic lethal mutations have been assumed to affect a single function or pathway [9], which reinforces the idea that pathways act as autonomous self-contained functional subsystems. In contrast, other investigations in yeast [16] report that synthetic-lethal genetic interactions are approximately three and a half times as likely to span pairs of pathways than to occur within pathways. In this chapter, it is found that RSL pairs in *E. coli* are predominantly intra-pathway while PSL pairs, more abundant, tend to be inter-pathway although concentrated in the entanglement of just two pathways, Cell Envelope Biosynthesis and Membrane Lipid Metabolism. The comparative study here shows that although pathways entanglement through coessentiality of reactions is low in both organisms, RSL pairs in *M. pneumoniae* can be intra-pathway or inter-pathway, linking Folate Metabolism and Nucleotide and Cofactor Metabolism, and PSL pairs are basically intra-pathway and located in Nucleotide Metabolism. Taken together, these results indicate that Folate and Nucleotide Metabolic pathways preserve most rescue routes for reaction deletion events, in accordance with results in Ref. [26]. The fact that the proportion of PSL pairs is considerably decreased in *M. pneumoniae* as compared to *E. coli* could be indicative that, even if both plasticity and redundancy serve an important function in achieving viability, essential plasticity is a more sophisticated mechanism that requires a higher degree of functional organization, using at the same time less resources for maximum growth. At the same time, this can also be explained by the relative unchanging environmental conditions of *M. pneumoniae* in the lung, that could have induced the elimination of pathways not required in that medium [26]. This suggests that the adaptability of *M. pneumoniae* is very much reduced and its behaviour could not be resilient to environmental changes.

It has also been found that SL reaction pairs and their subdivision in plasticity and redundancy are highly conserved independently of the composition of the minimal medium that acts as environmental condition for growth, and even when this environment is enriched with non-essential compounds or overconstrained to decrease the maximum biomass production. These environment unspecific SL pairs can thus be selected as potential drug targets operative regardless of the chemical environment of the cell.

4.4 Summary

- There exists a set of reactions, and thus enzymes and genes, that must be always active in order to ensure the viability of an organism [17] Copyright @ 2014, World Scientific Publishing.
- Non-essential reactions deserve special attention for two causes: their role as growth enhancers and for their potential participation in synthetic lethal pairs [17] Copyright @ 2014, World Scientific Publishing.
- Synthetic lethality is meditated by two different mechanisms, essential plasticity and essential redundancy, depending on whether one reaction is active for max-

imum growth in the medium under consideration and the second inactive, or in contrast both reactions have non-zero flux [18].

- Plasticity sets up as a sophisticated backup mechanism that is able to reorganize metabolic fluxes turning on inactive reactions when coessential counterparts fail in order to maintain viability in a specific medium [18].
- Redundancy corresponds to a simultaneous use of different flux channels that ensures viability and besides increases fitness [18].
- Plasticity and redundancy are highly conserved independently of the composition of the minimal medium that acts as environmental condition for growth, and even when this environment is enriched with non-essential compounds or overconstrained to decrease the maximum biomass production [18].

References

1. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427(6977):839–843
2. Almaas E, Oltvai ZN, Barabási AL (2005) The activity reaction core and plasticity of metabolic networks. PLoS Comput Biol 1:0557–0563
3. Baba T et al (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2(1)
4. Joyce AR et al (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. J Bacteriol 188(23):8259–8271
5. Barve A, Rodrigues JFM, Wagner A (2012) Superessential reactions in metabolic networks. Proc Natl Acad Sci USA 1091:E1121–E1130
6. Suthers PF, Zomorrodi A, Maranas CD (2009) Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol 5:301
7. Wang Z, Zhang J (2009) Abundant indispensable redundancies in cellular metabolic networks. Genome Biol Evol 1:23–33
8. Nygaard P, Smith JM (1993a) Evidence for a novel glycinamide ribonucleotide transformylase in *Escherichia coli*. J Bacteriol 175:3591–3597
9. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. Science 291:1001–1004
10. Tucker CL, Fields S (2003) Lethal combinations. Nat Genet 35:204–205
11. Masel J, Siegal ML (2009) Robustness: mechanisms and consequences. Trends Genet 25:395–403
12. Nijman SMB (2011) Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Lett 585:1–6
13. Kaelin WG (2005) The concept of synthetic lethality in the context of anticancer therapy. Nat Rev Cancer 5:689–698
14. Harrison R, Papp B, Pál C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci USA 104:2307–2312
15. Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. BioEssays 27:176–188
16. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23:561–566
17. Güell O, Serrano MÁ, Sagués F (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In *Engineering of Chemical Complexity II*. World Scientific Publishing, Singapore ISBN 978-981-4616-12-6

18. Güell O, Sagués F, Serrano MÁ (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol 10(5):e1003637
19. Novick P, Osmond BC, Botstein D (1989) Suppressors of yeast actin mutations. Genetics 121:659–674
20. Zhao G, Winkler ME (1995) An *Escherichia coli* K-12 tktA tktB mutant deficient in transketolase activity requires pyridoxine (vitamin B_6) as well as the aromatic amino acids and vitamins for growth. J Bacteriol 176:883–891
21. Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nat Genet 38:993–998
22. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
23. Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. BMC Bioinform 11:489
24. Orth JD et al (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism - 2011. Mol Syst Biol 7:535
25. Glass JI et al (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425–430
26. Wodke JAH et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol Syst Biol 9:653
27. Yus E et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
28. Whalen WA, Berg CM (1982) Analysis of avtA: Mu d1(ap lac) mutant: metabolic role of transaminase C. J Bacteriol 150:739–746
29. Nygaard P, Smith JM (1993b) Evidence for a novel glycinamide ribonucleotide transformylase in *Escherichia coli*. J Bacteriol 175:3591–3597
30. Troup B, Hungerer C, Jahn D (1995) Cloning and characterization of the *Escherichia coli* hemN gene encoding the oxygen-independent coproporphyrinogen III oxidase. J Bacteriol 177:3326–3331
31. Rompf A et al (1998) Regulation of *Pseudomonas aeruginosa* hemF and hemN by the dual action of the redox response regulators Anr and Dnr. Mol Microbiol 29:985–997
32. Jiao Z, Baba T, Mori H, Shimizu K (2003) Analysis of metabolic and physiological responses to gnd knockout in *Escherichia coli* by using C-13 tracer experiment and enzyme activity measurement. FEMS Microbiol Lett 220:295–301
33. Rude MA, Schirmer A (2009) New microbial fuels: a biotech perspective. Curr Opin Microbiol 12:274–281
34. Serrano MÁ, Boguñá M, Sagués F (2012) Uncovering the hidden geometry behind metabolic networks. Mol BioSyst 8:843–850
35. Giaever G et al (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418:387–391
36. Steinmetz LM et al (2002) Systematic screen for human disease genes in yeast. Nat Genet 31:400–404
37. Nakahigashi K et al (2009) Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. Mol Syst Biol 5:306

Chapter 5
Detection of Evolution and Adaptation Fingerprints in Metabolic Networks

Metabolic fluxes present an heterogeneity that can be exploited to construct metabolic backbones as reduced versions of metabolic networks. These backbones can be analysed to extract important biological information. In this chapter, the disparity filter is applied to two organisms, *Escherichia coli* and *Mycoplasma pneumoniae*. Backbones offer information about long-term evolution since they contain the core of ancestral pathways related with energy obtainment optimized by evolution to maximize growth. At the same time, backbones unveil short-term adaptation capabilities to variable external stimuli.

The analysis of metabolic networks is a difficult task which requires a mixed use of tools that belong to Systems Biology, such as Flux Balance Analysis (FBA) (see Chap. 2, Sect. 2.2), and tools that belong to complex network science, such as modelling of metabolic networks as bipartite semidirected networks (see Chap. 2, Sect. 2.1.1.). The combination of these approaches has enabled a huge step further towards the elucidation of important biological information hidden in the complexity of genome-scale metabolic reconstructions.

A useful tool in the endeavour of extracting useful biological information is the concept of backbone. Backbones maintain relevant biological information while displaying a substantially decreased number of interconnections and, hence, can provide accurate but reduced versions of the whole system. In particular, the work by Almaas et al. [1] introduced a filtering technique that selects the reaction that dominates the production or consumption of each metabolite such that a high-flux backbone can be retrieved. Although this method recovers pathways, the obtained backbones present a linear structure with very little interconnectivity and lack many of the features of real metabolic networks [2, 3].

Filtering approaches have also interested researchers working on networks in a more general context. A filtering method for weighted networks based on the disparity measure [4, 5] was developed in Ref. [6]. This approach exploits the heterogeneity present in the intensity of interactions in real networks both at the global and local levels [7] to extract the dominant set of connections for each element. Typically, the obtained disparity backbones preserve almost all nodes in the initial network and a large fraction of the total weight, while reducing considerably the number of links that pass the filter. At the same time, disparity backbones preserve the heterogeneity

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_5

and cut-off of the degree distribution, the level of clustering, and the bow-tie structure (see Chap. 2, Sect. 2.1.5), and other characteristic features of the original networks [6]. Hence, the complex features of the original networks are preserved.

In this chapter, FBA is used to determine reaction fluxes and the disparity filter (see Appendix D) [6] is applied to extract the metabolic backbones of two organisms: *Escherichia coli* and *Mycoplasma pneumoniae*. These backbones are investigated for fingerprints of evolution and adaptation. One finds that the metabolic backbones of both organisms in minimal medium are mainly composed of a core of reactions belonging to ancient pathways. This means that the significant fluxes in these bacterial metabolic backbones are associated to reactions which have been present from the earliest stages of their life and still remain at present significant for biomass production. At the same time, external conditions modify the structure of the backbones, which allows to identify pathways that are more sensitive to changes in the environment and so prone to short-term adaptation.

The contents of this chapter correspond to Ref. [8].

5.1 Identification of the Disparity Backbones of Metabolic Networks

FBA is used to compute the fluxes of the reactions composing the metabolic networks. These fluxes are treated as weights by the disparity filter. In this chapter, the *i*JO1366 version of *E. coli* K-12 MG1655 and the *i*JW145 version of *M. pneumoniae* are used (see Chap. 2, Sect. 2.3). FBA calculations are performed in glucose minimal medium with a maximum uptake of glucose limited to 10 mmol $\text{gDW}^{-1}\ \text{h}^{-1}$ for *E. coli* and 7.37 mmol $\text{gDW}^{-1}\ \text{h}^{-1}$ for *M. pneumoniae* (D-ribose is added to enrich the medium for *M. pneumoniae*). Once the fluxes are computed, the disparity filter is applied to the incoming and outgoing connections of each metabolite, such that only those links to reactions which concentrate a significant amount of flux are selected for the backbone (see Appendix D). The connectivity structure (see Chap. 2, Sect. 2.1.5) of the obtained backbones is analysed from an evolutionary perspective, and additional media are considered to analyse environmental sensitivity (see Chap. 2, Sect. 2.2.2).

An important feature of flux solutions obtained using FBA is the heterogeneity of the flux distributions. In the same state, fluxes of reactions can span several orders of magnitude [1, 9]. To check this statement, the probability distribution functions of the obtained fluxes are shown (disregarding zero-flux reactions) in the insets of Fig. 5.1b, c, confirming that, indeed, fluxes show an heterogeneous distribution at the global level. The set of metabolites in non-zero flux reactions is considerably reduced from the original total number, from 1805 to 445 metabolites in *E. coli*, and from 266 to 227 metabolites in *M. pneumoniae*. To characterize the existence of such heterogeneity also at the local level, the disparity measure [1, 6] is calculated for every metabolite (see Appendix D). Figure 5.1b, c display the disparity values for all metabolites as a function of their incoming and outgoing degree in *E. coli* and *M. pneumoniae*, respectively. The shadowed areas correspond to values compatible with a random distribution of fluxes among the reactions producing or consuming a metabolite and help to discount local heterogeneities produced by random fluctuations (see caption of Fig. 5.1). As shown, most metabolites present flux disparity values that cannot be explained by random fluctuations meaning that the local distribution of the fluxes of reactions associated to metabolites is significantly heterogeneous. One concludes then that the disparity filter will be able to efficiently extract a backbone with the most relevant connections for both organisms, while preserving the characteristic features of metabolism as a complex network.

Briefly, the disparity filter works by comparing weights of links with a random assignment. The filter preserves a link in the backbone if the probability that its normalized weight α_{ij} is compatible with the random assignment (p-value) is smaller than a chosen threshold α which determines the filtering intensity (see Appendix D). One proceeds to filter the metabolic networks with fluxes of reactions as weights of the connections between metabolites and reactions. For each metabolite i, the α_{ij} of each connection between metabolite i and its neighbouring reactions j is computed and the obtained p-value is compared with the significance level α. The disparity filter can be adjusted by tuning this threshold to observe how the metabolic networks of both *E. coli* and *M. pneumoniae* are reduced as α is decreased from 1 to 0, both of them included, $\alpha = 1$ meaning the complete network. Notice that, after applying the filter, one recovers a bipartite representation of the metabolic backbone. To avoid working with stoichiometrically non-balanced reactions, the filtered bipartite representation is transformed into a one-mode projection of metabolites placing a directed link between two metabolites if there is a reaction whose flux is simultaneously relevant for the consumption of one metabolite and for the production of the other [1]. In this one-mode projected backbone, one computes how many links E, nodes N and total weight W remain. These magnitudes are normalized by dividing them by the corresponding values in the original network, E_T, N_T, and W_T.

Figure 5.1d, e show the dependencies N/N_T versus E/E_T, and W/W_T versus E/E_T in the associated insets, for the one-mode metabolic projections of the backbones of both *E. coli* and *M. pneumoniae*. While the filter can reduce considerably the fraction of links, the corresponding fraction of nodes is maintained at almost the original value. In addition, the total weight in the backbone only starts to drop appreciably after more than 50% of the links are removed. One takes the critical value α_c as the point where the fraction of nodes starts to decay (see Fig. 5.1d, e). This critical value can be seen as an optimal point which reduces greatly the number of links in the network preserving at the same time most nodes and so as much biochemical and structural information as possible. The values are $\alpha_c = 0.21$ for *E. coli* and $\alpha_c = 0.37$ for *M. pneumoniae*.

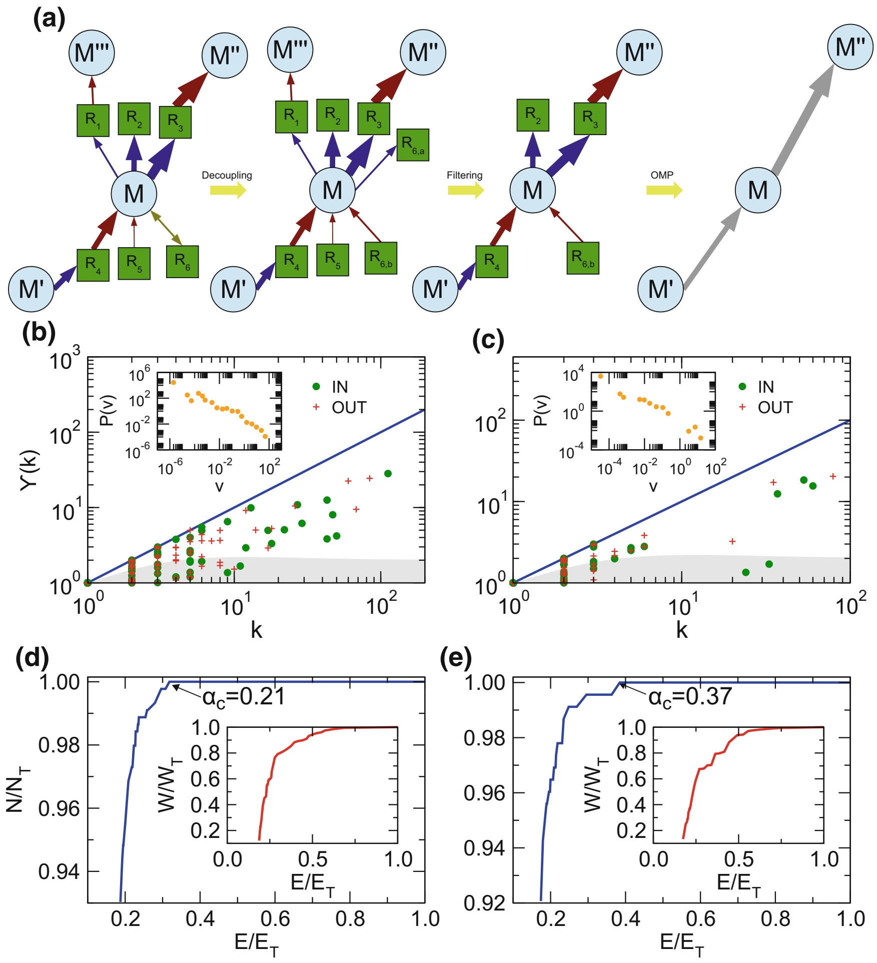


Fig. 5.1 Scheme of the application of the disparity filter and measures of the heterogeneity of reaction fluxes in *E. coli* and *M. pneumoniae*. **a** Scheme of the filtering method. *Blue* nodes are metabolites and *green* squares denote reactions. Incoming connections to metabolites are represented by *red arrows*, outgoing connections with *blue arrows*, and bidirectional connections with *dark yellow arrows*. OMP denotes one-mode projection. **b** Disparity measure as a function of incoming and outgoing degrees (k) in *E. coli*. The *shadowed area* corresponds to the average plus 2 standard deviations given by the null model, meaning that points which lie outside this are can be considered heterogeneous [6]. *Inset* global distribution of fluxes of *E. coli*. **c** Disparity measure as a function of IN and OUT degrees (k) for *M. pneumoniae*. Again, the *shadowed area* corresponds to the average plus 2 standard deviations given by the null model. *Inset* global distribution of fluxes of *M. pneumoniae*. **d** Fraction of nodes as a function of the fraction of links in *E. coli*. *Inset* remaining weight as a function of the fraction of links in the network. **e** Fraction of nodes as a function of the fraction of links in *M. pneumoniae*. *Inset* remaining weight as a function of the fraction of links in the network. Extracted from Ref. [8] (colour figure online)

5.2 Evolutionary Signatures in the Backbones of Metabolites

The metabolic backbones of both *E. coli* and *M. pneumoniae* are constructed using the identified critical values for the significance level. The backbones retain all the 445 and 227 metabolites present in active reactions respectively. Next, one analyses their structure in terms of connectedness. Metabolic networks have been found to display typical large-scale connectivity patterns of directed complex networks, called the bow-tie structure, with most reactions in a interconnected core, named the strongly connected component (SCC), together with in (IN) and out (OUT) components formed mainly by nodes directly connected to the SCC component [2, 10] (see Chap. 2, Sect. 2.1.5). This is the case of the original metabolic networks of both organisms, whose SCCs contain the largest part of the metabolites and reactions of the network, and whose IN and OUT components are formed, respectively, by nutrients and waste metabolites.

Metabolites in the backbone of *E. coli* are arranged in a connected component of 178 nodes and several disconnected small components (51). Three different SCCs can be identified in the connected part of the backbone, each with 25, 10, and 6% of the nodes in the connected component (see Fig. 5.2a). The two smallest SCCs are in the OUT component of the largest SCC. For the three of them, the IN and OUT components and tendrils are recovered. Metabolites corresponding to central compounds of metabolism are identified in these SCCs: protons, water, ATP, glutamate, phosphate, NAD^+, diphosphate, ADP and FAD^+. These metabolites are highly-connected metabolites even in the metabolic backbone, helping to preserve the same structural features of the complete metabolic network.

Since links in the metabolic backbone denote reactions transforming metabolites, it is interesting to annotate links with the pathway associated to the corresponding reaction. In this way, it is possible to count the composition of the three SCCs in terms of pathways. Starting with the largest SCC (see Fig. 5.2a), one finds that the major contributions are Oxidative Phosphorylation (26%), Citric Acid Cycle (16%), Glycolysis/Gluconeogenesis (15%), Pentose Phosphate Pathway (9%), and Glutamate Metabolism (9%) (see Fig. 5.2c). It has been demonstrated that these routes are ancient pathways that have been conserved through evolution. More precisely, Glycolysis and Pentose Phosphate Pathway take place without the need of enzymes in a mimetic Archean ocean [12]. Concerning the Citric Acid Cycle, it is also an ancient pathway that has evolved in order to achieve maximum ATP efficiency [13] by being coupled to Oxidative Phosphorylation and Glycolysis [14], in addition to help the organism to decrease their quantity of reactive oxygen species by modulation of their participating metabolites [15]. Another pathway significantly present in the largest SCC is Glutamate Metabolism. Glutamate has been reported to be one of the oldest amino acids used in the earliest stages of life [16].

Links in the other two SCCs correspond also to reactions belonging to ancestral pathways. The second largest SCC contains links that belong mainly to Purine and Pyrimidine Biosynthesis (91%). Purines and pyrimidines serve as activated

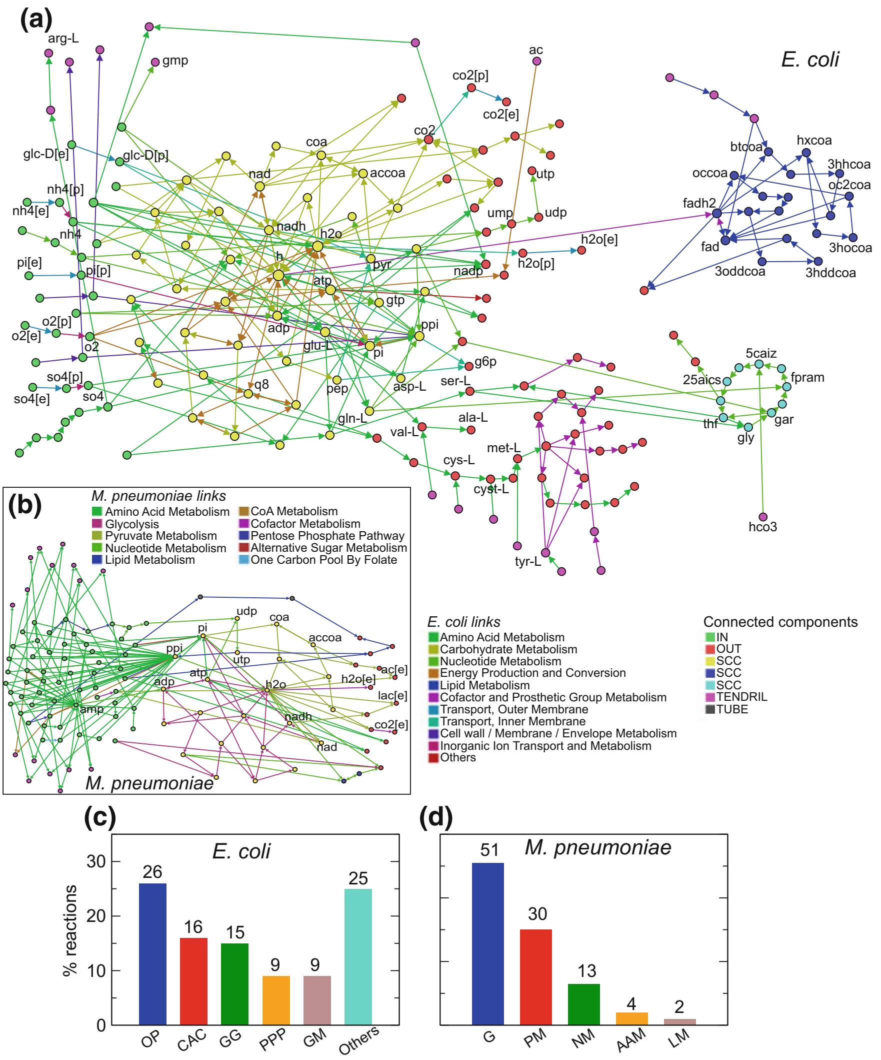


Fig. 5.2 SCCs of the backbone of metabolites and corresponding pathways. **a** Connected component in the metabolic backbone of *E. coli*. The *colors* of the nodes depend on the component each node belongs to. The *color* of the links, and its association given in the legend, depends on the functional categories given in Ref. [11], where each category contains pathways that realize similar tasks. **b** Connected component of the metabolic backbone of *M. pneumoniae*. The *color* of the nodes denote again the component each node belongs to. The *color* of the links, and its association given in the legend, depends on the pathway each reaction belongs to **c** Percentage of links in pathways for the largest SCC in the metabolic backbone of *E. coli*. **d** The same for *M. pneumoniae*. Extracted from Ref. [8] (colour figure online)

precursors of RNA and DNA, glycogen, etc. [17, 18], and it has been found that the synthesis of purines and pyrimidines was the first pathway involving enzyme-based metabolism [19]. Interestingly, the other contribution to this SCC is Glycine and Serine Metabolism. Glycine is a precursor of purines and pyrimidines. Pathways related to the third SCC are Membrane Lipid Metabolism (97%) and Cofactor and Prosthetic Group Biosynthesis (3%). Membrane Lipid Metabolism supplies the necessary lipids to generate the cell membrane needing the participation of the cofactor $FAD^+/FADH_2$. It has been shown that the pathways involved in lipid metabolism exhibit differences between different lineages in organisms [20], whereas pathways related to central metabolism are more conserved and are transversal [20].

When considering α values smaller than the critical one, implying that the filter is more restrictive and more heterogeneity is needed to overcome it, we observe that the smallest SCCs discussed above disappears. More precisely, it happens for a value of $\alpha = 0.19$. Decreasing even more the significance level to $\alpha = 0.15$ the SCC containing reactions in the Purine and Pyrimidine Biosynthesis pathway retains the 30% of the nodes for $\alpha_c = 0.21$, whereas the largest SCC still contains a 86%, showing the large resistance of this large core to lose nodes. At a value of $\alpha = 0.14$, the second SCC finally disappears and there only remains a single SCC, still preserving 82% of the nodes in it for $\alpha_c = 0.21$. Hence, energy metabolism shows a large resistance to get fragmented even though the filter becomes progressively more and more restrictive.

To contrast the obtained results in *E. coli*, the same analysis in *M. pneumoniae* is performed. Its critical value α_c is 0.37 (see Fig. 5.1). The connected component of its metabolic backbone is shown in Fig. 5.2b. It contains two SCCs, one of them being irrelevant with only two nodes (see Fig. 5.2b). The relevant SCC contains 21% of the nodes in the connected component, and the largest part of its links are related also with energy metabolism as in *E. coli*. The dominant pathways in this core are Glycolysis and Pyruvate Metabolism (see Fig. 5.2d). Along Glycolysis, Pyruvate Metabolism is also an ancestral pathway that was present in the earliest stages of life [21], when no oxygen was present in the early atmosphere.

5.3 The Metabolic Backbones of *E. Coli* Encode Its Short-Term Adaptation Capabilities

The previous section analyses the metabolic backbone of *E. coli* in glucose minimal medium in terms of the long-term evolution of the organism. In this section, the study is focused on how changes in the environment modify this backbone, which exposes short-term adaptation capabilities. First, FBA fluxes that maximize the growth rate of *E. coli* in the rich medium Luria–Bertani (LB) Broth [22, 23] are calculated. Afterwards, the disparity filter is applied to extract the metabolic backbone in this new environment, that is obtained for a significance level threshold $\alpha_c = 0.4$. This value is noticeably larger than $\alpha_c = 0.21$ identified for the glucose

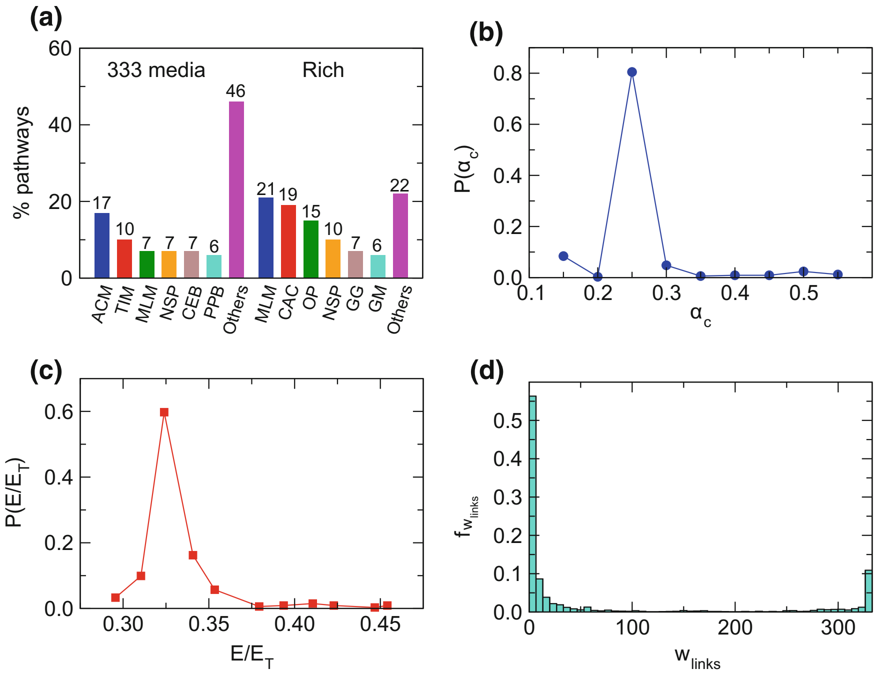


Fig. 5.3 Dependence of the distribution of pathways in the metabolic backbone of *E. coli* with the composition of environment. **a** Histogram of the fraction of links belonging to each pathway (x axis) for the 333 minimal media (*left*) and in the rich medium (*right*). **b** Probability distribution function of α_c for all minimal media. **c** Probability distribution function of the fraction of links in the metabolic backbones for all minimal media. **d** Histogram of weights of links in the metabolic superbackbone. Extracted from Ref. [8]

minimal medium. Interestingly, this rich medium activates 400 reactions, 11 less than in glucose minimal medium. Of them, 279 are active in both media, of which 247 have a larger flux in LB Broth. An analysis of the connected components in the metabolic backbone of *E. coli* in rich medium is also performed. One finds that it contains a large connected component with 449 metabolites and 60 small disconnected components. The connected component contains also three SCCs. However, two of them are tiny with only two nodes, whereas the largest one encloses 34% of the nodes in the connected component. Interestingly, the pathway contributing more reactions to this large SCC is Membrane Lipid Metabolism (see Fig. 5.3a). This fact is in accordance with Ref. [24], where the authors found that the expression of the genes which synthesize fatty acids was generally elevated in rich medium. Another important difference is the loss of prominence of Oxidative Phosphorylation and the Pentose Phosphate Pathways.

Next, the set of minimal media given in Ref. [11] (see Chap. 2, Sect. 2.2.2.1) are considered, where different carbon, nitrogen, phosphorus and sulphur sources are alternated. For each minimal medium, α_c is scanned as in Fig. 5.1b, c. In Fig. 5.3b, c, one plots, respectively, the probability distribution functions of the tuned α_c values and of the fraction of links remaining in the metabolic backbones for all media. One finds that there is a characteristic value of these magnitudes with no outliers, meaning that the flux structure is very similar across media in spite of the difference in the composition of nutrients. The presence of these characteristic values of α_c and the retained fraction of links in the metabolic backbones motivates to merge all of them into a single merged metabolic backbone. The links in this superbackbone correspond to reactions that passed the filter in any of the external media considered and are annotated with a weight that corresponds to the number of media in which the corresponding metabolic backbone contains the link. The histogram of the distribution of these weights is shown in Fig. 5.3d, characterized by a clear bimodal behaviour. One peak corresponds to links being common to all media, and the other corresponds to the most common situation of links specific to a few media.

An analysis of connectedness shows that this superbackbone contains a large connected component and 11 disconnected components. The connected component is composed by a large SCC with 43% of its nodes, in addition to three small SCCs containing only two nodes each. A pathway composition analysis in the large SCC indicates that, again, one obtains significantly different results from the glucose minimal medium (see Fig. 5.3a). The most prominent pathway is Alternate Carbon Metabolism, in agreement with Ref. [25], where the authors found that Alternate Carbon Metabolism is related to genes whose expression depends on external stimuli, particularly on alteration of carbon sources. It is also in agreement with results in Ref. [26], where the authors hypothesize that Alternate Carbon Metabolism can adapt to different nutritional environments, and also with results in Ref. [27], where Alternate Carbon Metabolism is found to be an important intermediate pathway in the network of pathways. The second most abundant pathway corresponds to Transport, Inner Membrane, which again is in agreement with Refs. [25, 27]. It is a transversal pathway which is in charge of the transport of metabolites between periplasm and cytosol. Finally, if one retains links present at least in 25% of the minimal media, the network fragments into 40 components with the largest one containing five SCCs, which indicates that links with small weight, i.e. links specific for a few media, have an important role in providing global connectivity to the superbackbone.

5.4 Conclusions

Identifying high-flux routes in metabolic networks has been useful in order to, for example, identify principal chains of metabolic transformations [1, 28, 29]. In this chapter, one goes beyond the mere identification of high-flux routes with metabolic pathways. Using a high-flux fluctuation analysis, it is possible to identify ancestral pathways and, on the other hand, pathways with capabilities to adapt to short-term external changes. At the core of the high-flux fluctuation analysis, a filtering tool which needs no *a priori* assumptions for the connectivity of the filtered subnetworks is used, but that produces reduced versions which are globally connected and retain the characteristic complex features of the original network. This procedure allows to extract a metabolic backbone which contains all relevant connections given a set of external nutrients, recovering both intra- and inter-pathway connections which can be understood as the superhighways of metabolism. Further, an evolutionary explanation can also be given for this identification of both intra- and inter-pathway connections since the cooperation between reaction inside and outside pathways implies that the overall performance of a cell will be improved due to a better and more efficient utilization of the available resources. This fact reinforces the idea that pathways are not isolated identities performing their tasks independently of others [27].

As stated in Ref. [30], properties that originate from evolutionary pressure should not be observed in random networks. Due to the fact that the disparity filter identifies links that deviate from a random null model, it allows to identify those reactions for which evolutionary pressure has had a large incidence. Since FBA flux solutions are used, in this chapter the effect of evolutionary pressure is understood to favor the maximization of the growth of the organism [31–33]. The evolutionary analysis of the metabolic backbones of the two considered organisms in minimal medium shows that their SCCs are composed by reactions that belong to ancient pathways. In *E. coli*, each SCC has different and definite metabolic functions. In both *E. coli* and *M. pneumoniae*, the largest SCC contains pathways related to energy metabolism, meaning that these organisms have evolved towards maximum efficiency in obtaining chemical energy, something very important in case of nutrient scarcity. A smaller SCC is responsible for the synthesis of purines and pyrimidines, vital for DNA / RNA synthesis. The third SCC corresponds to the metabolism of lipids, the most important constituents that compose the cell membrane. Two findings relating the two small SCCs deserve also special attention. Firstly, the two small SCCs are located in the OUT component of the large SCC. Secondly, as the filter becomes more restrictive, the small SCCs fragment, while the large SCC still maintains a large part of links and nodes. These features could be explained in terms of the functional requirements of the small SCCs. On the one side, they need chemical energy to perform their tasks and, on the other side, they need also basic building blocks. These tasks are performed in the large SCC by, for example, Glycolysis/Gluconeogenesis or the Citric Acid Cycle. Therefore, it suggests that those SCCs were added to the OUT component of the large SCC in later steps of evolution. A simpler organism, *M. pneumoniae*, has no other relevant SCCs apart from energy metabolism, as a result of its parasitism, which has led to the loss of many metabolic functions [34]. More precisely, in *M. pneumoniae* the Citric Acid Cycle and Oxidative Phosphorylation do not take place [34, 35], meaning that it must rely on organic acid fermentation to obtain energy. Moreover, changes in the growth rate greatly affect the fluxes through Glycolysis and Pyruvate Metabolism [34].

The study of the dependence on the environment of the *E. coli* metabolic backbone allows to identify short-time adaptation capabilities. Regarding rich medium, one observes that the critical value of α is substantially different than the one in glucose minimal medium, suggesting that this enriched medium modifies significantly the flux structure compared to the glucose minimal medium. The bacterium in rich medium displays less active reactions than in glucose minimal medium since, in minimal medium, many reactions must be active in order to synthesize biosynthetic precursors that in the rich medium can be obtained from the environment, in agreement with Ref. [24]. The pathway called Membrane Lipid Metabolism achieves a high relevancy, being the most abundant pathway in the largest SCC of the rich medium metabolic backbone. This happens because the instantaneous response of *E. coli* to this rich medium, which induces a large increase in the growth rate of the organism due to nutrient abundance, is to synthesize as much as membrane lipids as possible, since fast-growing cells must synthesize membrane components more rapidly to satisfy the high lipid demand to generate new cells [24]. The analysis of the adaptation of *E. coli* to 333 different minimal media shows that the distribution of fluxes is practically independent on the composition of the nutrients present in these environments, allowing to extract characteristic features that describe the backbones of the metabolic network independently of the environment. This permits the construction of a merged backbone that comprises all the links composing the metabolic backbone in each media. This leads to the identification of pathways whose associated reactions are more sensitive to changes in the environment, unveiling Alternate Carbon Metabolism as the pathway with more capabilities to respond to external stimuli, in accordance with previously reported results [25, 26].

The use of filtering methods usually imply a drastic reduction of the complexity of metabolic maps, which weakens the validity of potentially inferred conclusions. The application of the disparity filter based on a high-flux fluctuation analysis to produce metabolic backbones enables to reduce the system while maintaining all relevant interactions and so it becomes a useful tool to unveil sound biological information. For instance, the investigation of *E. coli* and *M. pneumoniae* revealed metabolic backbones in minimal medium mainly composed of a core of reactions belonging to ancient pathways, for which the effects of evolutionary pressure are higher, and unveiled pathways with high capacity to respond to external stimuli.

5.5 Summary

- The disparity filter is very efficient in order to compute metabolic backbones as reduced versions of metabolism which retain its complexity [8].
- The study of the bow-tie structure of the backbones in a glucose minimal medium reveals that pathways related with energy obtainment have an important evolutionary role in *E. coli* and *M. pneumoniae* [8].
- The study of the backbone of *E. coli* in rich medium identifies the pathway Membrane Lipid Metabolism as relevant for growth in the nutritionally rich medium, due to the necessity of large amounts of lipids to generate the cell membrane [8].
- The analysis of the superbackbone, constructed by merging all the backbones corresponding to different minimal media, recognizes the pathway Alternate Carbon Metabolism as the most relevant pathway to respond to external stimuli [8].

References

1. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427(6977):839–843
2. Ma HW, Zeng AP (2003) The connectivity stucture, giant strong component and centrality of metabolic networks. Bioinformatics 19:1423–1430
3. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. Nat Rev Genet 5:101–113
4. Herfindahl OC (1959) Copper costs and prices: 1870–1957. John Hopkins University Press, Baltimore
5. Hirschman AO (1964) The paternity of an index. Am Econ Rev 54:761–762
6. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the mutiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106:6483–6488
7. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci USA 101(11):3747–3752
8. Güell O, Sagués F, Serrano MÁ (2017) Detecting the significant flux backbone of *Escherichia coli* metabolism. FEBS Lett 591(10):1437–1451
9. Bianconi G (2008) Flux distribution of metabolic networks close to optimal biomass production. Phys Rev E 78(3):035101
10. Serrano MÁ, De Los Rios P (2008) Structural efficiency of percolated landscapes in flow networks. PLoS ONE 3:e3654
11. Orth JD et al (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. Mol Syst Biol 7:535
12. Keller MA, Turchyn AV, Ralser M (2014) Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. Mol Syst Biol 10(4):725
13. Meléndez-Hevia E, Waddell TG, Cascante M (1996) The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. J Mol Evol 43(3):293–303
14. Ebenhöh O, Heinrich R (2001) Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. Bull Math Biol 63(1):21–55
15. Mailloux RJ, Bériaut R, Lemire J, Singh R, Chénier DR, Hamel RD, Appanna VD (2007) The tricarboxylic acid cycle, an ancient metabolic network with a novel twist. PLoS ONE 2(8):e690
16. Fell DA, Wagner A (2000) The small world of metabolism. Nat Biotechnol 18:1121–1122
17. Evans DR, Guy HI (2004) Mammalian pyrimidine biosynthesis: fresh insights into an ancient pathway. J Biol Chem 279(32):33035–33038
18. Powner MW, Gerland B, Sutherland JD (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature 459(7244):239–242
19. Caetano-Anolles G, Kim HS, Mittenthal JE (2007) The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci USA 104(22):9358–9363
20. Suen S, Lu HHS, Yeang CH (2012) Evolution of domain architectures and catalytic functions of enzymes in metabolic systems. Genome Biol Evol 4(9):976–993
21. Tadege M, Dupuis I, Kuhlemeier C (1999) Ethanolic fermentation: new functions for an old pathway. Trends Plant Sci 4(8):320–325
22. Sezonov G, Joseleau-Petit D, D'Ari R (2007) *Escherichia coli* physiology in Luria-Bertani Broth. J Bacteriol 189:8746–8749
23. Güell O, Sagués F, Serrano MÁ (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol 10(5):e1003637
24. Tao H, Bausch C, Richmond C, Blattner FR, Conway T (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J Bacteriol 181(20):6425–6440
25. Lourenço A, Carneiro S, Pinto JP, Rocha M, Ferreira EC, Rocha I (2011) A study of the short and long-term regulation of *E. coli* metabolic pathways. J Integr Bioinform 8(3):195–209

26. Monk JM et al (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci USA 110(50):20338–20343
27. Serrano MÁ, Boguñá M, Sagués F (2012) Uncovering the hidden geometry behind metabolic networks. Mol BioSyst 8:843–850
28. Bourqui R et al (2009) Metabolic network visualization eliminating node redundancy and preserving metabolic pathways. BMC Syst Biol 1:29
29. Faust K, Dupont P, Callut J, van Helden J (2010) Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 26:1211–1218
30. Basler G, Ebenhöh O, Selbig J, Nikoloski Z (2011) Mass-balanced randomization of metabolic networks. Bioinformatics 27:1397–1403
31. Ibarra RU, Edwards JS, Palsson BØ (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420:186–189
32. Blank LM, Kuepfer L, Sauer U (2005) Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6(6):R49
33. Llaneras F, Picó J (2008) Stoichiometric modelling of cell metabolism. J Biosci Bioeng 105(1):1–11
34. Wodke JAH et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol Syst Biol 9:653
35. Manolukas JT, Barile MF, Chandler DK, Pollack JD (1988) Presence of anaplerotic reactions and transamination, and the absence of the tricarboxylic acid cycle in mollicutes. J Gen Microbiol 134(3):791–800

Chapter 6
Assessing FBA Optimal States in the Feasible Flux Phenotypic Space

Optimal growth solutions can be confronted with the whole set of feasible flux phenotypes (FFP), which provides a reference map that helps to assess the likelihood of optimal and high-growth states and their extent of conformity with experimental results. In addition, FFP maps are able to uncover metabolic behaviours that are unreachable using models based on optimality principles. The information content of the full FFP space of metabolic states provides with an entire map to explore and evaluate metabolic behaviour and capabilities, opening new avenues for biotechnological and biomedical applications.

The results presented in previous chapters required an extensive use of Flux Balance Analysis (FBA) (see Sect. 2.2) in order to extract backbones or to compute the effect of failures of reactions. If the removal of a reaction or a pair of reactions is not lethal for the organism, i.e., the growth rate is not zero, there can exist many flux solutions for the organism to be alive. As it has been already explained, the FBA solution is a possible solution, the one which maximizes the growth rate. One may be tempted to ask where the solutions given by FBA lay in the whole space of possible flux solutions of a metabolic network. In this way, it will be possible to know whether the state given by FBA is indeed representative of the system or, on the contrary, it is not a representative solution of the flux space, this eventually being interpreted for example due to evolutionary effects.

FBA studies, like in the previous Chap. 4, reveal that metabolism is a dynamically regulated system that reorganizes to safeguard survival [1, 2], implying that metabolic phenotypes directly respond to environmental conditions. For instance, unicellular organisms can be stimulated to proliferate by controlling the abundance of nutrients available. In rich media, cells reproduce as quickly as possible by fermenting glucose, a process which produces high specific growth rates as well as large quantities of excess carbon in the form of ethanol and organic acids [3]. To survive the scarcity of nutrients during starvation periods, Glycolysis is hypothesized to switch to oxidative metabolism, which no longer maximizes the specific growth rate, but instead the ATP yield needed for cellular processes. Cells of multicellular organisms show similar metabolic phenotypes, relying primarily on Oxidative Phosphorylation when not stimulated to proliferate and changing to non-oxidative glycolytic metabolism during

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_6

cell proliferation, even if this process -known in cancer cells as the Warburg effect [4, 5]—is much less efficient at the level of energy yield.

These metabolic phenotypes are captured by FBA. However, the identified solutions are frequently inconsistent with the biological reality since no single objective function describes successfully the variability of flux states under all environmental conditions [6, 7], and in fact the highest accuracy of FBA predictions is achieved whenever the most relevant objective function is tailored to particular environmental conditions according to the empirical evidence for a very specific metabolic phenotype. For instance, FBA requires either a rich medium or a manual limitation of the oxygen uptake to a physiological enzymatic limit to mimic the observed fermentation of glucose to formate, acetate, or ethanol typical of proliferative metabolism, while in minimal medium optimization of growth rate relies primarily on Oxidative Phosphorylation, which increases ATP production converting glucose to carbon dioxide, as in starvation metabolism. However, along optimal metabolic phenotypes, there is a whole space of possible states non-reachable by invoking optimality principles that prevent non-optimal biological states. Optimization of a biological function in the absence of *a priori* biological justification, which happens for instance under conditions for proliferative or starvation metabolism, may weaken *in silico* predictions.

In this chapter, optimal growth rate solutions are confronted to the whole set of feasible flux phenotypes (FFP) of core *Escherichia coli* metabolism in minimal medium, which provides a reference map that helps to assess the likelihood of optimal and high-growth states [8]. The whole set of feasible flux phenotypes is determined by mass-balance conditions and the bounds imposed on metabolites. Mathematically, it constitutes a convex finite polytope, and it is sampled using an algorithm called Hit-And-Run (HR) (see Appendix E) [9]. One can quantitatively and visually show that optimal growth flux phenotypes are eccentric with respect to the bulk of states, statistically represented by the feasible flux phenotypic mean, which suggests that optimal phenotypes are uninformative about the more probable states, most of them low-growth rate. Feasible flux phenotypic space is proposed as a benchmark to calibrate the deviation of optimal phenotypes from experimental observations. Finally, the analysis of the entire high-biomass production region of the feasible flux phenotypic space unveils metabolic behaviours observed experimentally but unreachable by models based on optimality principles, like FBA, which forbid aerobic fermentation -a typical pathway utilization of proliferative metabolism- in minimal medium with unlimited oxygen uptake.

The contents of this chapter correspond to Ref. [8].

6.1 Optimal Growth Is Eccentric with Respect to the Full FFP Space

As in FBA, feasible flux states of a metabolic network are those that fulfil stoichiometric mass balance constraints together with imposed upper and lower bounds on the reaction fluxes. These constraints restrict the number of solutions to a compact convex set which contains all possible flux steady states in a particular environmental condition. In glucose minimal medium (see Chap. 2, Sect. 2.2.2.1), the FFP space of the core *E. coli* model is determined by 70 potentially active reactions, including biomass formation and the ATP maintenance reaction, and 68 metabolites. Using the HR algorithm, a raw sample of 10^9 feasible states is obtained, from which a final uniform representative set of 10^6 feasible states is extracted.

Notice that the used approach is suitable for genome-scale network sizes beyond the reduced size of the core *E. coli* model. There is not any fundamental or technical bottleneck that prevents its application to complete metabolic descriptions at the cell level. In this chapter, the core *E. coli* model is used due to a matter of computational time and ease of visualization.

From the sampled set of core *E. coli* metabolic states in minimal medium of glucose bounded to 10 mmol gDW^{-1}h^{-1}, the metabolic flux profiles of each individual reaction is collected as the set of its feasible metabolic fluxes. From such profile, one can compute the probability density function $f(\nu)$ which describes the likelihood for a reaction to take on a particular flux value. In Fig. 6.1, the profiles of all reactions are shown. One can observe a variety of shapes, all of them low-variance, most displaying a maximum probability for a certain value of the flux inside the allowed range,[1] and many being clearly asymmetric. The allowed range is computed using Biomass unconstrained Flux Variability Analysis (see Chap. 2, Sect. 2.2.4).

To characterize the dispersion of the possible fluxes for each reaction, one can measure its coefficient of variation $CV(f(\nu))$ calculated as the ratio between the standard deviation of possible fluxes and their average. For all but three reversible reactions (Malate dehydrogenase, Glucose-6-Phosphate isomerase, and Glutamate dehydrogenase), the only reversible reactions having a low associated flux mean and thus a higher $CV(f(\nu))$, this metric is below one and when ranked for all reactions it steadily decreases to almost zero, Fig. 6.2a. Interestingly, it can be found that this coefficient is significantly anticorrelated with the essentiality of reactions as observed experimentally [10] (point-biserial correlation coefficient -0.29 with p-value 0.01, see Appendix C). This means that essential reactions tend to have a highly concentrated profile of feasible fluxes. Besides, and only for the glucose transferase reaction GLCpts, one finds a zero probability of having a zero flux, which indicates that this reaction is essential in glucose minimal medium as expected. The asymmetry of each profile is characterized by the distance between the more probable flux in the FFP space and the lower flux bound of the flux variability range rescaled by the flux variability range of the reaction (see Chap. 2, Sect. 2.2.4). In Fig. 6.2b, a scatter plot of values for all 68 core reactions is shown. Strikingly, the rescaled distances cluster in three regions around 0, 0.5 and 1 forming groups of sizes 38, 15 and 17 respectively. This indicates that the most probable flux is close to either the lower or upper bound or, conversely, the probability distribution function tends to be quite symmetric. Moreover, it can be also observed that an anticorrelation between

[1] Notice that none of these histograms can have more than one peak due to the convexity of the steady-state flux space.

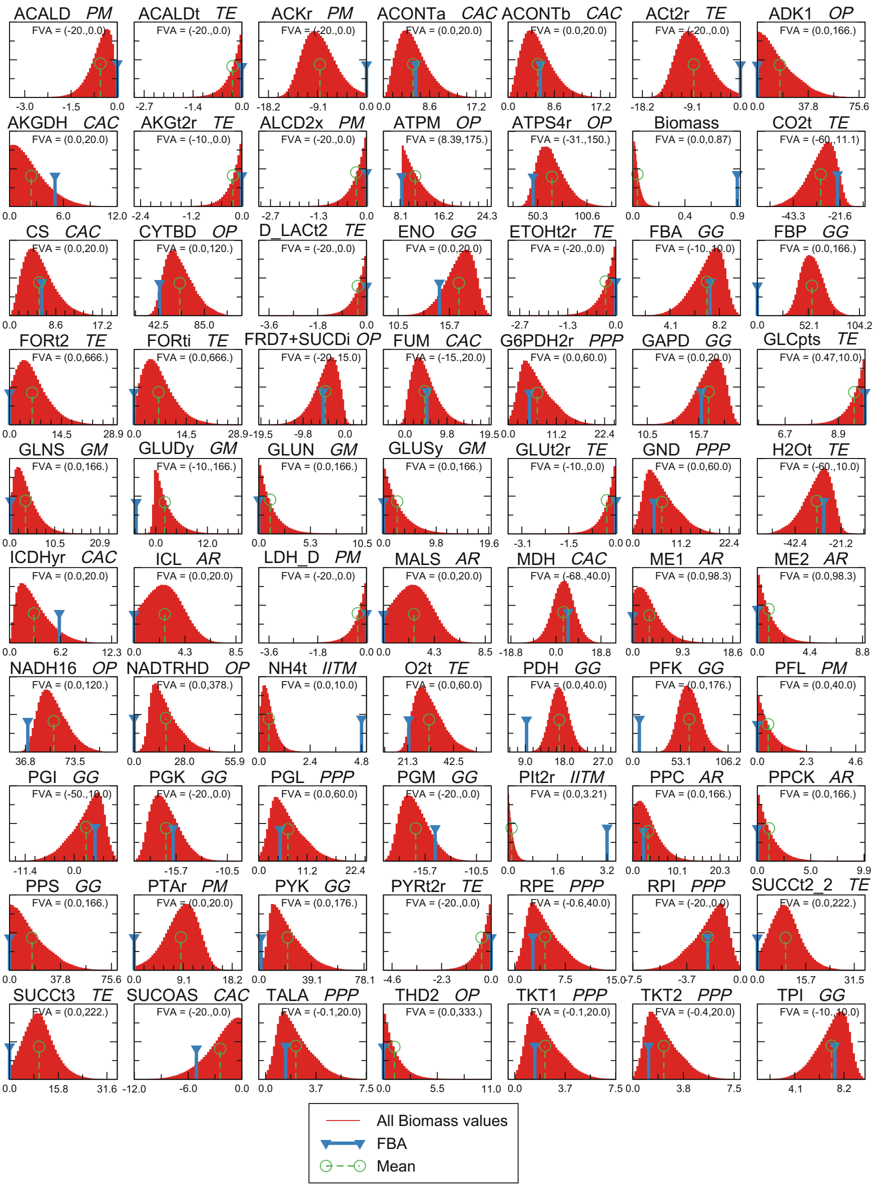


Fig. 6.1 Probability density functions of metabolic fluxes values for all reactions in core *E. coli* under glucose minimal conditions. *Each graph* shows the reaction label, the flux variability range (values inside parentheses), and each associated pathway (acronyms in *italics*). Notice that the range plotted in the axes does not coincide with the flux variability range, since in the axes an optimal x range for each reaction is chosen to distinguish the shape of each profile. In addition, in each profile the position of the FBA point (*blue marker*) and the position of the Mean (*green marker*) are also shown. Extracted from Ref. [8] (color figure online)

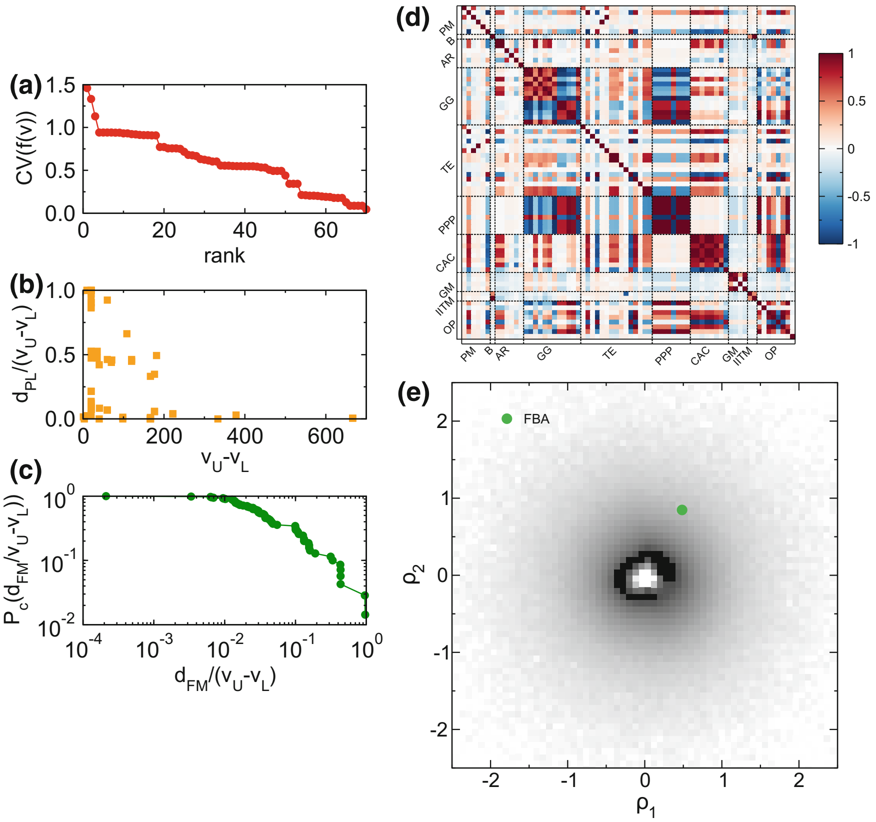


Fig. 6.2 Analysis of reaction profiles and visualization of the FFP space. **a** Coefficient of variation for all core reactions ranked by value. **b** Scatter plot of distances between the more probable flux in the FFP space and the lower flux bound rescaled by flux variability range for each reaction. **c** Complementary cumulative distribution function of distances between FBA maximal growth flux and FFP space mean flux rescaled by flux variability range for each reaction, in log-log scale. **d** Matrix of Pearson correlation coefficients measuring the degree of linear associations between feasible fluxes of reactions (acronyms of the pathways are shown in abbreviations). **e** Projection of the FFP space onto the two principal component vectors of the correlation matrix in **e**. All sampled flux phenotypes are normalized and projected along the first (ρ_1) and second (ρ_2) principal components. The plot is in polar coordinates, with the negative logarithm of the radius. The majority of points lies in a *circle* close to the origin (the *darker area*). The FBA solution (*green circle*) is, conversely, rather eccentric. Parts of this Figure have been extracted from Ref. [8]

the length of the flux range and the position of the most probable flux is present, so that the closer is this to its maximum value the shorter is the allowed range of fluxes.

In order to assess the likelihood of flux states corresponding to FBA maximization of the flux through the biomass reaction (FBA-MBR) (or equivalently of the growth

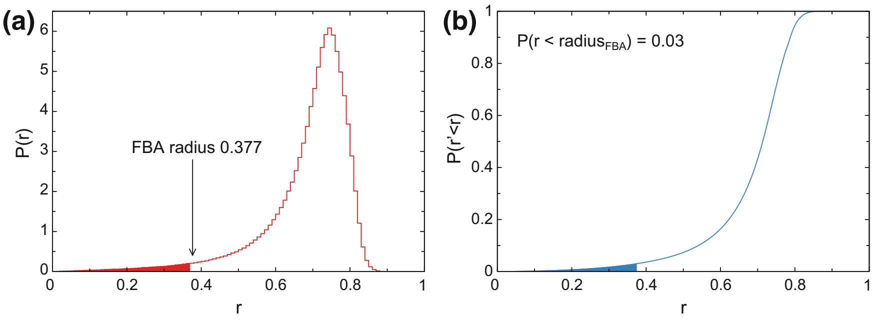


Fig. 6.3 **a** Probability distribution function of the radii of all solutions before applying the negative logarithmic transformation. The *red area* denotes the probability of having a smaller radius than the FBA solution. This fraction of area is the 3% of the total area, which means that the 97% of the solutions have a larger radius than the FBA solution. **b** Cumulative probability distribution function of the radius. The *blue* region denotes the range of solutions with a radius smaller than FBA. The probability of having a radius smaller than FBA is the y-value of the *curve* at the rightmost side of the region. Extracted from Ref. [8] (color figure online)

rate) in relation to typical[2] points within the whole FFP space, one can calculate the average flux value for each reaction, the mean, and compare it to the FBA optimal biomass production flux. The complementary cumulative distribution function of the distances between these two characteristic fluxes rescaled by the flux variability range of reactions is shown in Fig. 6.2c. A broad distribution of values can be observed over several orders of magnitude with no mean value actually very close to the FBA maximal solution except for a few reactions, typically working at maximum growth. At the other end of the spectrum, deviated reactions include for instance excretion of acetate and phosphate exchange. As a summary, one can conclude that the mean and the FBA biomass optimum are rather distant, which suggests that FBA optimal states are uninformative about phenotypes in the bulk of states in the FFP space.

To visualize neatly the eccentricity of the FBA maximum growth state with respect to the bulk of metabolic flux solutions, Principal Component Analysis [11, 12] is used in order to reduce the high-dimensionality of the full flux solution space projecting it onto a two-dimensional plane from the most informative viewpoint (see Appendix F). Reaction profiles are taken in pairs to calculate the matrix of Pearson correlation coefficients measuring their degree of linear association (see Fig. 6.2d). Note that an ordering of reactions by pathways allows to have a clear visual feedback of intra- and inter-pathway correlations taking place in the core *E. coli* metabolic network, such that clusters of highly correlated reactions appear as bigger darker squares. The two axes of our projection correspond to the two first principal components of this profile correlation matrix ρ_1 and ρ_2, which account for most of the variability in profile correlations. Each sampled metabolic flux state has been rescaled as a z-score centred around the mean and projected onto these axes, as shown in the scatter

[2] In the mathematical/computational context, typical means statistically representative in relation to the whole set of flux states contained in the FFP space.

plot Fig. 6.2e in polar coordinates, where a negative logarithmic transformation to the radial coordinate for ease of visualization has been applied. The majority of phenotypes have a radius close to zero. Since points closer to the origin are better described by the two principal components (see Appendix F), this implies that ρ_1 and ρ_2 capture the largest variability of the sampled FFP. Clearly, the FBA optimal growth solution is rather eccentric with respect to typical solutions, with an associated radius of 0.98 in this representation. In fact, 97% of states have a smaller radius than the optimal growth solution (see Fig. 6.3).

6.2 The FFP Space Gives a Standard to Calibrate the Deviation of Optimal Phenotypes from Experimental Observations

This section focuses on the relationship between primary carbon source uptake and oxygen need to illustrate the potential of the FFP space as a benchmark to calibrate the deviation of *in silico* predicted optimal phenotypes from experimental observations. Sampled FFP states of core *E. coli*, in particular FFP mean values, as a function of the upper bound uptake rate of the carbon source are compared with reported experimental data for oxygen uptakes in minimal medium with glucose, pyruvate, or succinate as a primary carbon source (see Fig. 6.4). The line of optimality representing FBA optimal growth solutions is also considered. Glucose experimental data points were used from Ref. [1], experimental results for pyruvate are reported in Ref. [13], and experimental results in Ref. [14] report the quantitative relationship between oxygen uptake rate and acetate production rate as a function of succinate uptake rate.

In all cases, FBA-MBR reproduces well experimental data points in the low carbon source uptake region [14], where *E. coli* is indeed optimizing biomass yield. However, oxygen uptake rate saturates after some critical threshold of carbon source uptake rate (which depends on the carbon source) reaching a plateau which, among other possibilities, could be explained by the existence of a physiological enzymatic limit in oxygen uptake that lessens the capacity of the respiratory system [15]. The plateau levels are 18.8±0.7 mmol gDW^{-1} h^{-1} for glucose [14], 16.8±0.4 mmol gDW^{-1} h^{-1} for pyruvate [13], and 19.49±0.78 mmol gDW^{-1} h^{-1} for succinate [14]. In this region of high carbon source uptake, FBA-MBR predicts an oxygen uptake overestimated by around 25% with respect to the values reported from experiments. While this amount is in principle large, the FFP space gives a standard that helps to calibrate it.

The eccentricity of experimental observations is measured as their distance to the FFP mean. For glucose, this value is 19.4 mmol gDW^{-1} h^{-1}, which makes the distance of 5.3 mmol gDW^{-1} h^{-1} between the FBA-MBR prediction and experimental data relatively low (see Fig. 6.4a). The distance of 8.2 mmol gDW^{-1} h^{-1} between the FBA-MBR prediction and experimental data is slightly worse for pyruvate (see Fig. 6.4b), in which case the eccentricity of experimental observations is of 18.4 mmol gDW^{-1} h^{-1}. The disagreement between optimality predictions and experimental

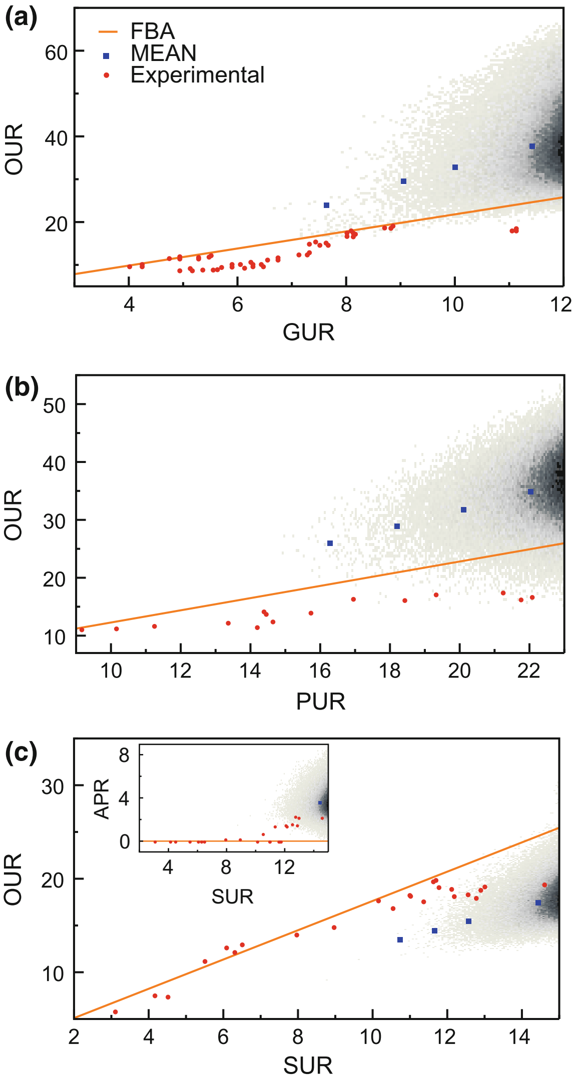


Fig. 6.4 Comparison of predicted phenotypes and experimental data. Sampled points in the FFP space with maximum carbon source upper bound are plotted in *shaded grey*, *darkness* is proportional to the number of points. Experimental data points are *red circles*. The *in silico*-defined line of optimality, representing FBA optimal growth solutions as a function of the upper bound uptake rate of the carbon source, is shown in *orange*. *Blue squares* correspond to FFP mean values for different carbon source upper bound uptake rates. **a** Oxygen versus glucose uptake rates, experimental data from [1]. The FFP space is sampled with glucose bounded to 12 mmol gDW^{-1} h^{-1}. **b** Oxygen versus pyruvate uptake rates, experimental data from Ref. [13]. The FFP space is sampled with pyruvate bounded to 23 mmol gDW^{-1} h^{-1}. **c** Oxygen versusu succinate uptake rates, experimental data from Ref. [14]. The FFP space is sampled with succinate bounded to 15 mmol gDW^{-1} h^{-1}. *Inset* Acetate production rate versus succinate uptake rate, experimental data from Ref. [14]. Extracted from Ref. [8] (color figure online)

data is much more significative in the case of succinate (see Fig. 6.4c), for which the eccentricity of experimental observations is only of 4.3 mmol gDW^{-1} h^{-1}, while the distance between the FBA-MBR prediction and experimental data is of 5.4 mmol gDW^{-1} h^{-1}, meaning that the FFP mean is indeed more adjusted to observations. The case of acetate production for this carbon source is even more conspicuous (see Fig. 6.4c *Inset*). While FBA-MBR is still reproducing well the experimental results of no acetate production in the low succinate uptake region, it cannot predict production of acetate at any succinate uptake rate due to the fact that FBA-MBR in minimal medium with unlimited oxygen does not capture the enzymatic oxygen limitation. The FBA-MBR solution diverts resources to the production of ATP entirely through the Oxidative Phosphorylation pathway. Thus, it fails to reproduce experimental observations of acetate production in the region of high succinate uptake rates [14, 16–18]. In contrast, most metabolic states in the FFP space are consistent with acetate production, so that in this case the FFP mean turns out as a good predictor of the experimentally observed metabolic behaviour.

In summary, while FBA-MBR predictions seem accurate for low carbon source uptake rate states in minimal medium as seen previously [14], the experimental points diverge from the FBA-MBR prediction state when increased values of carbon source uptake rates are considered. Note that, in general, it is not straightforward to quantify the significance of the divergence. Here, the FFP space is proposed as a benchmark. According to this calibration, one finds that FBA optimal growth predictions of oxygen needs versus glucose, pyruvate, or succinate uptake are worse the more downstream the position of the carbon source into catalytic metabolism. Using the core *E. coli* model, it has been checked that the ratio of the maximum ATP production rate to the maximum oxygen uptake (both calculated by FBA optimization of ATP production rate) for the three carbon sources glucose, pyruvate, and succinate are respectively 2.9, 2.6, and 2.4, so this ratio decreases as more downstream in the catalytic metabolism.

6.3 The High-Biomass Production Region of the FFP Space Displays Aerobic Fermentation in Minimal Medium with Unlimited Oxygen Uptake

The high-growth metabolic region of the core *E. coli* FFP space is resampled in glucose minimal medium with a glucose upper bound of 10 mmol gDW^{-1} h^{-1}. This region is defined by setting a minimal threshold for the biomass production of $\geq$ 0.4 mmol gDW^{-1} h^{-1} [19], and the new sample has a final size of 10^5 states. Note that phenotypes in this high-growth sample remain very close to the biomass yield threshold due to the exponential decrease of the number of feasible flux states with increased biomass production, as shown in the biomass flux profile in Fig. 6.1.

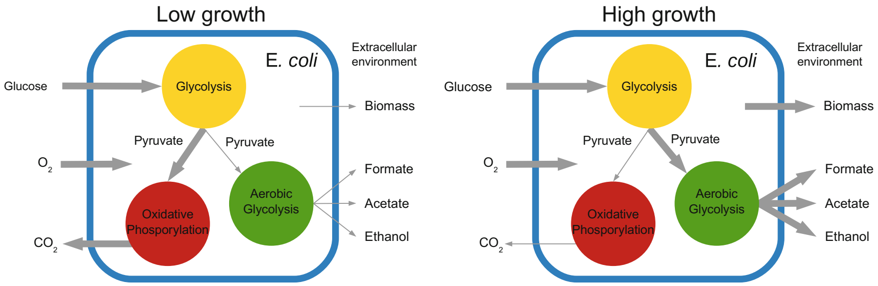


Fig. 6.5 Schematic of pathway utilization in high-growth versus low-growth conditions. Extracted from Ref. [8]

In this region, one can identify pathway utilization typical of proliferative microbial metabolism, even when considering a minimal medium and unlimited oxygen uptake. This metabolic behaviour is consistent with experimental data [1, 14, 20] but it is unreachable by FBA models based on optimality principles (unless optimization is accompanied by auxiliary constraints not assumed in standard FBA implementations, like the solvent capacity constraint [19], or by modeling beyond stoichiometric mass balance, for instance, thermodynamically feasible kinetics or enzyme synthesis [21, 22]). These by-products cannot be explained by FBA-MBR in minimal medium with unlimited oxygen supply since, in this optimization framework, metabolic fluxes are basically forced to ATP production through Oxidative Phosphorylation with excretion of CO_2 as waste. However, increasing the oxygen limitation in FBA-MBR results in secretion of formate, acetate, and ethanol -in that order-, with corresponding shifts in metabolic behaviour [15].

According to the FFP space of core *E. coli*, one can observe that the high-biomass production FFP subsample is characterized by the secretion of small organic molecules, even when the supply of oxygen is unlimited. This fact points to the simultaneous utilization of Glycolysis and Oxidative Phosphorylation to produce biomass and energy, as illustrated in the schematic shown in Fig. 6.5. Quantitative relationships between the production of small organic molecules and glucose and oxygen uptake rates are shown in the remaining panels of Fig. 6.6. Three-dimensional scatter plots for the production rates of formate, acetate, ethanol, and lactate are shown in Fig. 6.6a, c, e, g respectively, with projections into the three possible two-dimensional planes shown in Fig. 6.6b, d, f, h respectively. As the levels of glucose and oxygen uptakes are raised, metabolic phenotypes can achieve an increased production of formate, acetate, ethanol, and lactate even though the majority of feasible phenotypes remain at low production values. Due to the high-growth requirement, oxygen uptake is always high but its variability increases with glucose uptake increase around a value of approximately 41.2 mmol gDW^{-1} h^{-1}, which clusters the majority of high-growth metabolic phenotypes. Interestingly, this oxygen uptake rate value marks a region in the FFP space with maximum potential production rates of formate, acetate, ethanol,

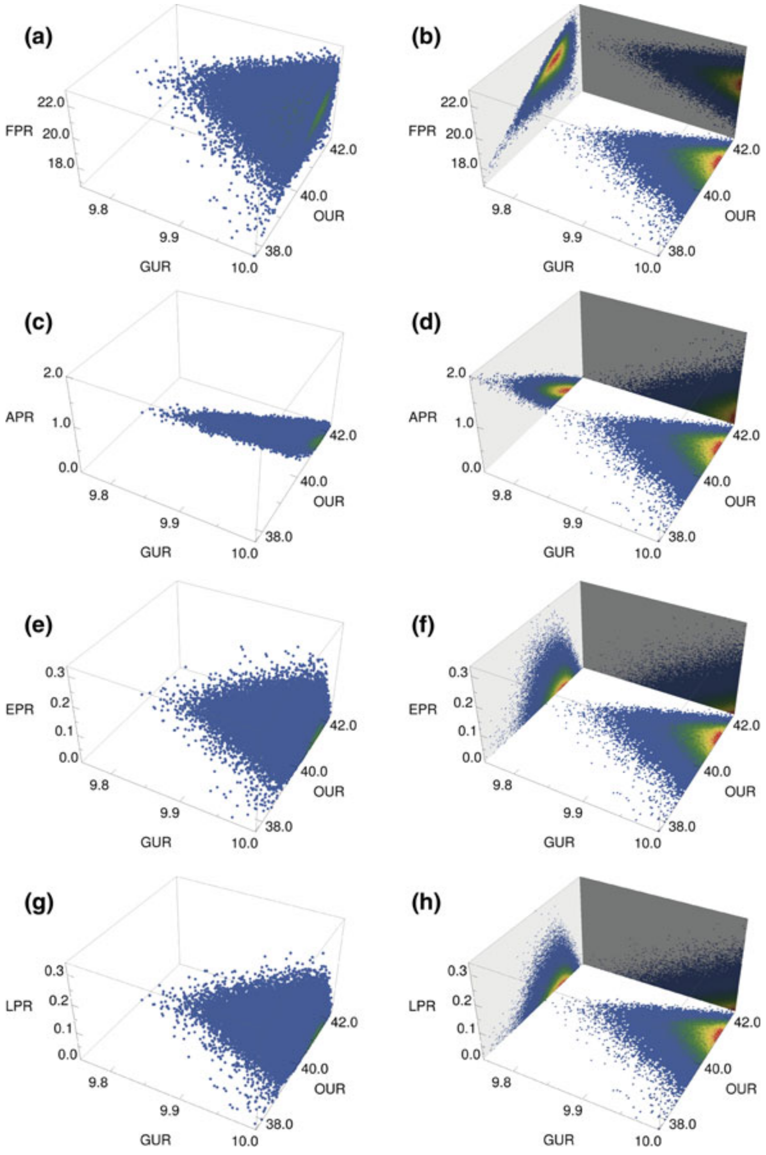


Fig. 6.6 High growth phenotypes of core *E. coli* on glucose minimal medium. **a** 3-Dimensional scatter plot of formate production rate versus glucose and oxygen uptake rates. **b** Density projections of **a** on each of the possible 2D planes, formate-glucose, formate-oxygen, and glucose-oxygen. **c** 3-Dimensional scatter plot of acetate production rate versus glucose and oxygen uptake rates. **d** Density projections of **c** on each of the possible 2D planes, acetate-glucose, acetate-oxygen, and glucose-oxygen. **e** 3-dimensional scatter plot of ethanol production rate versus glucose and oxygen uptake rates. **f** Density projections of **e** on each of the possible 2D planes, ethanol-glucose, ethanol-oxygen, and glucose-oxygen. **f** 3-Dimensional scatter plot of lactate production rate versus glucose and oxygen uptake rates. **g** Density projections of **f** on each of the possible 2D planes, lactate-glucose, lactate-oxygen, and glucose-oxygen. Extracted from Ref. [8]

and lactate. Above and below that value most states are concentrated in the range [39.0, 42.0] mmol gDW^{-1} h^{-1}.

Taken together, these results indicate that, contrarily to standard FBA predictions, a high level of glucose uptake combined with enough oxygen can maintain the requirements of proliferative metabolism for biomass formation through aerobic fermentation even if the rest of nutrients are scarce and restricted to the minimum. At the same time, additional oxygen uptake diverts glucose back towards more efficient ATP production through Oxidative Phosphorylation. Hence, oxygen has the potential of regulating the glucose metabolic switch in which glucose uptake rates larger than a critical threshold around 5.0 mmol gDW^{-1} h^{-1} [19] lead to a linearly increasing maximum organic by-products production by a gradual activation of aerobic fermentation and a slight decrease of Oxidative Phosphorylation.

6.4 Conclusions

The information content of the full FFP space of metabolic states in a certain environment provides with an entire map to explore and evaluate metabolic behaviour and capabilities. While optimality goals need to be tailored to conditions and produce singular optimal solutions that may not be consistent with experimental observations, we have nowadays sufficient computational and methodological capacity to produce and analyse full FFP maps. The latter offer a reference framework to put into perspective the likelihood of particular phenotypic states that, as shown, enables to uncover metabolic behaviours that are unreachable using standard models based on optimality principles. In fact, the location of metabolic flux distributions into precise optimal states has been challenged recently by the proposal that metabolic flux evolve under the trade-off between two forces, optimality under one given condition and minimal adjustment between conditions [7]. In this way, resilience to changing environments necessarily forces flux states to near-optimal but suboptimal regions of feasible flux states in order to maintain adaptability.

In the FFP map of core *E. coli* in aerobic minimal medium, optimal growth states appear as eccentric and far from the bulk of more probable phenotypes represented by the FFP mean, which offers an ergodic perspective of the FFP space in which all states can be explored at random with equal probability. One of the uses of the method is precisely to evidence the effects of evolutionary pressure on organisms, which may actually result in eccentric flux states. On the other hand, the FFP space gives a standard to calibrate the deviation of optimal phenotypes from experimental observations. Oxygen consumption is a particularly interesting target for analysis since it has been identified as a trigger of metabolic shifts [15, 23]. Interestingly, according to the FFP map as a reference standard, it is found that, in high-growth conditions, FBA-MBR predictions of experimental observations for unlimited oxygen needs versus glucose, pyruvate, or succinate uptakes are worse the more downstream the uptake of the carbon source into the catalytic metabolic stream. This is consistent with the fact that the FBA-MBR solution diverts resources to the production of ATP entirely through the Oxidative Phosphorylation pathway, so that the more is the effective potential of the carbon source to recombine with oxygen to produce energy the more convergent will be the *in silico* prediction and the observed states.

In order to correct FBA in high-growth conditions, some investigations restricted the solution space beyond mass balance and uptake bounds through additional thermodynamic, kinetic or physiological constraints, like the solvent capacity constraint quantifying the maximum amount of macromolecules that can occupy the intracellular space [19]. Alternatively, the objective function implemented in FBA has been modified to non-linear maximization of the ATP or biomass yield per flux unit [6], or modelling beyond stoichiometric mass balance, like thermodynamically feasible kinetics or enzyme synthesis, has been considered [21, 22]. While these FBA modifications enhance some predictions, their effectiveness depends on the estimation of kinetic coefficients using empirical or experimental data. In contrast, the FFP map naturally displays all high-growth feasible states which show characteristic metabolic behaviours, like aerobic fermentation with unlimited oxygen uptake even in minimal medium, without the need to determine additional constants. This aerobic fermentation, apparently inefficient in terms of energy yield as compared to Oxidative Phosphorylation, has been demonstrated to be a favourable catabolic state for all rapidly proliferating cells with high glucose uptake capacity [19], and from this analysis it turns out as a probable metabolic phenotype even in minimal medium.

Beyond theoretical implications, FFP maps of microbial organisms can be of particular interest as tools for biotechnological applications, for instance in the engineering of *E. coli* fermentative metabolism as a fundamental cellular capacity for valuable industrial biocatalysis [24]. In biomedicine, the investigation of FBA optimal phenotypes in the framework of the FFP map can help to contextualize disease phenotypes in comparison to normal states. For instance, FBA proved suitable for modelling complex diseases like cancer as it assumes that cancer cells maximize growth searching for metabolic flux distributions that produce essential biomass precursors at high rates [25, 26]. The analysis of the entire region of high-growth phenotypes will allow to reach and study a variety of suboptimal feasible flux states close to optimality but which cannot be reproduced by optimality principles, and so it opens new avenues for the understanding of general and fundamental mechanisms that characterize this disease across subtypes.

6.5 Summary

- FFP maps offer a reference framework to put into perspective the likelihood of particular phenotypic states. It enables to uncover metabolic behaviours that are unreachable using standard models based on optimality principles [8].
- Optimal FBA growth states are eccentric and appear far from the bulk of more probable phenotypes represented by the FFP mean [8].
- The FFP space gives a standard to calibrate the deviation of extreme phenotypes from experimental observations [8].

- The FFP map naturally displays all high-growth feasible states which show characteristic metabolic behaviours like aerobic fermentation with unlimited oxygen uptake even in minimal medium without the need to force additional constants [8].

References

1. Ibarra RU, Edwards JS, Palsson BØ (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420:186–189
2. Blank LM, Kuepfer L, Sauer U (2005) Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6(6):R49
3. Frick O, Whittmann C (2005) Characterization of the metabolic shift between oxidative and fermentative growth in *Saccharomyces cerevisiae* by comparative 13C flux analysis. 4:30
4. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the warburg effect. Science 324:1029–1033
5. Menendez J, Joven J, Cufí S, Corominas-Faja B, Oliveras-Ferraros C, Cuyàs E, Martin-Castillo B, Lopez-Bonet E, Alarcón T, Vazquez-Martin A (2013) The Warburg effect version 2.0. Cell Cycle 12(8):1166–1179
6. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol Syst Biol 3:119
7. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. Science 336:601–604
8. Güell O, Massucci FA, Font-Clos F, Sagués F, Serrano MÁ (2015) Mapping high-growth phenotypes in the flux space of microbial metabolism. J R Soc Interface 12:20150543
9. Lovász L (1999) Hit-and-run mixes fast. Math Progr. 86(3):443–461
10. Orth JD et al (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. Mol Syst Biol 7:535
11. Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Philos Mag Ser 6 2(11):559–572
12. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
13. Fong S, Marciniak JY, Palsson BØ (2003) Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. J. Bacteriol. 185(21):6400–6408
14. Edwards JS, Palsson BØ (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19:125–130
15. Varma A, Boesch BW, Palsson BØ (1993) Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. Appl Environ Microb 59:2465–2473
16. Reiling HE, Laurila H, Fiechter A (1985) Mass culture of *Escherichia coli*: medium development for low and high density cultivation of *Escherichia coli* B/r in minimal and complex media. J. Biotechnol. 2:191–206
17. El-Mansi EM, Holms WH (1989) Control of carbon flux to acetate excretion during growth of *Escherichia coli* in batch and continuous cultures. J Gen Microbiol 135:2875–2883
18. Wolfe AJ (2005) The acetate switch. Microbiol Mol Biol R 69:12–50
19. Vázquez A, Beg QK, deMenezes MA, Ernst J, Bar-Joseph Z, Barabási AL, Boros LG, Oltvai ZN (2008) Impact of the solvent capacity constraint on *E. coli* metabolism. BMC Syst Biol 2:7
20. Varma A, Palsson BØ (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl Environ Microb 60(10):3724–3731
21. Molenaar D, van Berlo R, de Ridder D, Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. Mol Syst Biol 5(1):323
22. Wortel MT, Peters H, Hulshof J, Teusink B, Bruggeman FJ (2014) Metabolic states with maximal specific rate carry flux through an elementary flux mode. FEBS J 281(6):1547–1555

23. Losen M, Frolich B, Pohl M, Buchs J (2004) Effect of oxygen limitation and medium composition on *Escherichia coli* fermentation in shake-flask cultures. Biotechnol Progr 20:1062–1068
24. Orencio-Trejo M, Utrilla J, Fernández-Sandoval MT, Huerta-Beristain G, Gosset G, Martinez A (2010) Engineering the *Escherichia coli* fermentative metabolism. Adv Biochem Eng Biot 121:71–107
25. Price ND, Shmulevich I (2007) Biochemical and statistical network models for systems biology. Curr Opin Biotech 18(4):365–370
26. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7:501

Chapter 7
Conclusions

This thesis presents a study of cell metabolism from a systems-level approach trying to unveil new mechanisms and responses impossible to reach by traditional reductionist procedures. Different methods and analysis techniques have been used, and each one has allowed to extract new insights about the properties of cell metabolism. Tools that belong to the complex network science and Systems Biology have been used. On what follows, the conclusions of this thesis are given, answering to the objectives stated in Chapter 1.

The thesis starts by considering the study of the topology, *i.e.*, the connectivity pattern, of metabolic networks. From this point of view, it is possible to check whether the structure of metabolic networks has evolved towards increasing its robustness against external perturbations. It is important to notice that, at this stage, reaction fluxes are not considered.

From the obtained results of the first chapter of this thesis, one can conclude that the structure of the metabolic networks of *Escherichia coli*, *Staphylococcus aureus*, and *Mycoplasma pneumoniae* has evolved towards robustness against individual and multiple reaction failures, which produce a reduced damaged compared to failures in degree-preserving randomized counterparts. *M. pneumoniae* is an exception in relation to individual reaction failures. This feature can be explained in terms of its simpler structure. Moreover, it is found that failures provoked by pairs of reactions generate an amplification effect which arises due to the non-linear interactions between the two damaging cascades propagating in the networks. In addition, a predictor of damage propagation for single cascades computed locally accounts for damage spreading. Also at the local level, a series of structural motifs can explain amplified failure patterns in double reaction cascades.

When the study is extended to gene failures, one finds that the method to compute cascades captures most of the scenarios of experimentally determined lethality in *M. pneumoniae*. Furthermore, when referring to multiple failures, the proposed analysis allows to find that (1) for failure cascade spreading, the distribution of cluster sizes is more important than the actual composition of the clusters, and (2) the regulation of high-damage genes tends to appear isolated from that of other genes, a kind of functional switch in metabolic networks that at the same time acts as a kind of

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_7

genetic firewall. In any case, it is important to notice that a cascade may not only be interpreted as the harmful spreading of failures, but also as the ability to efficiently regulate metabolism. Large cascades may point at the evolutionary requirement of regulating large parts of metabolism through the regulation of small sets of enzyme-coding genes. Therefore, evolutionary pressure seems to favour the ability of efficient metabolic regulation at the expense of robustness to reaction knockouts.

This study can be complemented taking into account the fluxes flowing through the biochemical reactions with the aim to describe more appropriately real features of metabolic operation. These investigations permit to know how reactions adapt to different situations, extending for instance the previous study of gene knockouts, or additionally, looking at responses to changes of the composition of the external environment. Flux Balance Analysis is used to compute fluxes of biochemical reactions. This method is based on different suppositions, principally that (1) metabolic networks work at steady state and that (2) the biological target of organisms is to grow as much as possible. In this way, FBA can be used to go beyond the mere analysis of the structure of metabolic networks and to identify metabolic fluxes that cannot be resolved using only a topological analysis. When FBA is applied to single reaction knockouts in *E. coli*, the main conclusion is that there exists a set of reactions which must be always active in order to ensure viability. However, non-essential reactions deserve special attention, either considering their role as growth enhancers or their potential participation in synthetic lethal pairs.

The study of synthetic lethal pairs allows to understand new protection mechanisms that metabolism has developed to survive. Synthetic lethal reaction pairs can be classified into two classes, plastic and redundant, depending on whether one reaction is active for maximum growth in the medium under consideration and the second inactive (plasticity) or, conversely, both reactions have simultaneously non-zero fluxes (redundancy). This particular study is made in both *E. coli* and *M. pneumoniae*. On the one hand, plasticity is a sophisticated mechanism that is able to reorganize metabolic fluxes turning on inactive reactions when coessential counterparts are removed so as to maintain viability, working as a backup mechanism. On the other hand, redundancy corresponds to a simultaneous use of different flux channels, ensuring in this way viability and increasing the growth rate of the organism. Furthermore, plasticity requires a higher degree of functional organization, using at the same time less resources for maximum growth. It takes place more often in *E. coli* than in *M. pneumoniae*.

The previous study is completed by analysing how plasticity and redundancy depend on the external environment for *E. coli*. One finds that plasticity and redundancy are conserved independently of the composition of the medium which acts as environmental condition for growth. Moreover, this conservation takes place also when this environment is enriched with non-essential compounds or overconstrained to decrease the maximum growth rate.

One can further exploit FBA, assuming conditions of growth optimality, in order to assess evolution or adaptation characteristics of metabolic networks. A filtering method called disparity filter allows to reduce the density of links of metabolic networks while preserving their main features. The metabolic networks of *E. coli* and *M. pneumoniae* are filtered to extract their backbones. First of all, it is checked that the disparity filter is, indeed, very efficient in order to decrease the link density of the studied metabolic networks using FBA fluxes as the weights of the links.

The analysis of the connected components of the metabolic backbones of both *E. coli* and *M. pneumoniae* in a glucose minimal medium allows to identify that these components mainly contain reactions that belong to ancient pathways, *i.e.*, pathways showing long-term evolution. Moreover, for both organisms, the presence of pathways related to energy metabolism -like Glycolysis, Citric Acid Cycle, and Oxidative Phosphorylation for *E. coli*, or Glycolysis and Pyruvate Metabolism for *M. pneumoniae*- could mean that these pathways have an important role in maximizing the growth and have evolved towards maximum efficiency to obtain chemical energy, something very important in case of nutrient scarcity and hence energy deficiency.

In addition, the study of the dependence of *E. coli* backbones on different environments allows to identify environment specific pathways displaying short-term adaptation. First, the analysis of the metabolic backbone obtained in a rich medium allows to demonstrate that the nutritionally-rich medium induces a large increase in the growth rate of *E. coli* due to nutrient abundance. The instantaneous response of *E. coli* to environment is to synthesize as much as membrane lipids as possible, since fast-growing cells must synthesize membrane components more rapidly to satisfy the high lipid demand to generate new cells. Second, with the study of the different backbones obtained from different minimal media, one finds that the distribution of the fluxes is little dependent on the nutrients present in the environment. In addition, it is also possible to extract that the pathway Alternate Carbon Metabolism is, for *E. coli*, the pathway with more capabilities to respond to external stimuli.

It is worth remarking that FBA makes the supposition that the biological target of organisms is to grow as much as possible. This may be plausible in some situations but there exist other in which the biological target of an organism is not to maximize growth. Hence, a study of the entire space of possible flux solutions can help to assess whether the FBA solution is representative of the whole space or not. The whole space encompassing the entire set of flux solutions, referred to as the full feasible flux phenotypes (FFP) space, is computed for *E. coli*. The information contents of the FFP space of metabolic states in a certain environment provides with an entire map to explore and evaluate metabolic behaviour and capabilities. In fact, FFP maps can answer the question of whether FBA gives a representative solution of the flux space. The main conclusion is that optimal growth states obtained via FBA computations appear as eccentric and far from the bulk of more probable phenotypes.

In addition to the eccentricity of the FBA solution, the FFP space also gives a standard to calibrate the deviation of phenotypes obtained using FBA from experimental observations. Thus, it serves to compare FBA predictions with experimental results. For instance, the analysis of oxygen needs versus glucose, pyruvate, or succinate uptakes show that FBA results are worse the more downstream the uptake of the carbon source into the catalytic metabolic stream. This is explained due to the fact that the FBA solution diverts resources to the production of ATP entirely through Oxidative Phosphorylation. In this way, the more the effective potential of the carbon source to recombine with oxygen to produce energy using Oxidative Phosphorylation, the more convergent will be the FBA prediction with respect to experimental results.

On the other hand, the FFP space naturally displays all high-growth feasible states which show characteristic metabolic behaviours, like aerobic fermentation with unlimited oxygen uptake even in minimal medium. This is an important feature, since these metabolic behaviours cannot be obtained under FBA maximum growth computations without using additional constraints. This reinforces the idea that the FFP map contains valuable information about metabolic states.

It is important to point out that the used methodology in this thesis is not restricted to bacteria, and that it could also be applied to metabolic networks of other species. In particular, the results of the study of structural stress may have potential implications in areas like metabolic engineering or disease treatment. The study of complex systems under structural stress poses a number of formidable challenges critical to understand their behaviour as well as towards proposing successful strategies for prediction and control. In this framework, the study of structural stress in human pathogens may help to develop more sophisticated forms of identifying new and more efficient drug targets.

Plasticity and redundancy are very important concepts for biological complex systems in general. Whether they are adaptive in cell metabolism or, as it has been argued for metabolism in changing environments [1, 2], they are rather a by-product of the evolution of biological networks toward survival, these regulatory mechanisms are key to understand how complex biological systems protect themselves against malfunction. Among the many different applications of synthetic lethality, one of them is to determine the accuracy of gene essentiality of new genome-scale reconstructions of metabolic networks [3].

Since the application of the disparity filter in metabolic networks can be used to recognize pathways and reactions which (1) are more sensitive to environmental changes, and (2) which are involved in the maximization of the growth rate of an organism due to evolutionary pressure, its use could be appropriate in the field of biotechnology. For example, it could be useful for the targeting of the most important pathways present in cancer cells which are in charge of their high growth rate. Therefore, this could help to understand the biochemical mechanisms that cancer cells use to proliferate. In this way, it will be possible to find a way to decrease the high performance achieved by cancer cells in terms of growth efficiency.

Finally, FFP maps of microbial organisms can be of particular interest as tools for biotechnological applications, for instance in the engineering of *E. coli* fermentative metabolism as a fundamental cellular capacity for valuable industrial biocatalysis [4]. In biomedicine, the investigation of FBA phenotypes in the framework of the FFP map can help to contextualize disease phenotypes in comparison to normal states. For instance, FBA proved suitable for modelling complex diseases like cancer as it assumes that cancer cells maximize growth searching for metabolic flux distributions that produce essential biomass precursors at high rates [5, 6]. The analysis of the entire region of high-growth phenotypes will allow to reach and study a variety of suboptimal feasible flux states close to optimality but which cannot be reproduced by optimality principles, and so it opens new avenues for the understanding of general and fundamental mechanisms that characterize this disease across subtypes.

References

1. Harrison R, Papp B, Pál C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci USA 104:2307–2312
2. Wang Z, Zhang J (2009) Abundant indispensable redundancies in cellular metabolic networks. Genome Biol Evol 1:23–33
3. Larocque M, Chénard T, Najmanovich R (2014) A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. BMC Syst Biol 8(1):117
4. Orencio-Trejo M, Utrilla J, Fernández-Sandoval MT, Huerta-Beristain G, Gosset G, Martinez A (2010) Engineering the Escherichia coli fermentative metabolism. Adv Biochem Eng Biotechnol 121:71–107
5. Price ND, Shmulevich I (2007) Biochemical and statistical network models for systems biology. Curr Opin Biotechnol 18(4):365–370
6. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7:501

Appendix A
Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov test [1] is a test used in statistics which compares the probability distribution obtained from a sample with a reference probability (one-sample K-S test), or which compares two samples (two sample K-S test). It basically quantifies a distance between the cumulative distribution function of the sample and the cumulative distribution function of the reference distribution, or between the cumulative distribution functions of two samples. The null hypothesis of this test assumes that the samples are obtained from the same distribution (two sample K-S test) or that the sample is drawn from the reference distribution (one sample K-S test). The two-sample KS test is one of the most useful methods for comparing two samples, which is the variant that has been used in this thesis.

To compare two samples, first of all one has to compute the maximum distance $K - S$ between the two cumulative distribution functions

$$K - S = \max |F_{1,n}(x) - F_{2,n'}(x)| \tag{A.1}$$

where $F_{1,n}(x)$ and $F_{2,n'}(x)$ are the cumulative distribution functions of the first and second sample, and n and n' are the sizes of each sample respectively (see Fig. A.1). To compute the associated significance of the value of $K - S$, one has to calculate the p-value applying the following expression:

$$p = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2\, j\, l^2) \tag{A.2}$$

where $l = K - S \cdot (\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}})$, and $N = \frac{n\, n'}{n+n'}$. Then, one compares this p-value with the chosen reference, usually $\alpha = 0.05$. If $p < \alpha$, one can consider that both distributions are drawn from the same distribution, otherwise they are considered significantly different.

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

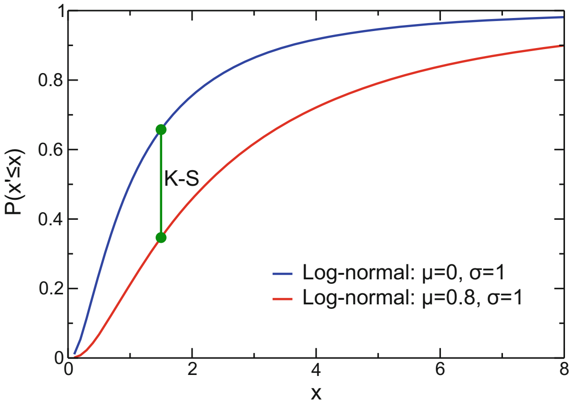


Fig. A.1 Visualization of the value $K - S$ used in the K-S test computed using two Log-normal distributions [2] with different means and the same standard deviation. After computing the maximum difference, this value is transformed into a p-value

References

1. Smirnov NV (1948) Tables for estimating the goodness of fit of empirical distributions. Ann Math Stat 19:279
2. Aitchison J, Brown JAC (1957) The lognormal distribution with special reference to its uses in economics

Appendix B
Spearman's Rank Correlation Coefficient

The Spearman's rank correlation coefficient [1], often denoted by the Greek letter ρ, is a nonparametric measure used in statistics which measures statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. Spearman's coefficient can be used both for continuous and discrete variables, including ordinal variables.

The Spearman's coefficient is basically the Pearson correlation coefficient between the ranked variables. The ranks of both samples are compared and the value of ρ_S is computed with the following expression:

$$\rho_S = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{B.1}$$

where x_i and y_i are the ranks of the values of the sample X_i and Y_i.

To assess the significance of the measure, a permutation test is done in this thesis, where the values of X_i and Y_i are reshuffled and then, for each realization, ρ_S is calculated. After doing this for all realizations, one keeps the maximum and minimum value of the obtained ρ_S, which gives the interval that belongs to the null model. Thus, if the value of ρ_S of the original sample lies within this interval, it implies that there is no correlation between the ranks of both samples. Otherwise, if the value of ρ_S of the original sample lies outside the range of the null model, one can consider that there exists a correlation between both samples.

Reference

1. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

Appendix C
Point-Biserial Correlation Coefficient

The point biserial correlation coefficient (r_{pb}) [1] is a correlation coefficient which is used when one variable is continuous and the other is dichotomous. This dichotomous variable can either be a truly dichotomous variable, like male/female, or an artificially dichotomized variable, obtained by using a threshold on a continuous variable. However, in most situations it is not advisable to dichotomize variables artificially and thus it is more appropriate to use specific statistical tests for continuous variables.

The point-biserial correlation is equivalent to the Pearson correlation. To calculate the point-biserial correlation coefficient, one assumes that the dichotomous variable can have the values 0 and 1. Therefore, one can divide the data between two groups, the first group which corresponds to the value 1 on the dichotomous variable, and the second group which corresponds the value 0 on the dichotomous variable. Thus, the point-biserial correlation coefficient is calculated as follows:

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{C.1}$$

where M_1 is the mean value of the continuous variable for all data points in the first group, and M_0 is the mean value of the continuous variable for all data points in the second group. Further, n_1 is the number of data points in the first group, n_0 is the number of data points in the second group, and n is the total sample size. s_n is the standard deviation computed as follows:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \nu)^2} \tag{C.2}$$

where x_i is continuous variable and ν is its average value. It is possible to compute a t-value (associated to a Student's t-distribution) from this correlation coefficient:

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

$$t = r_{pb}\sqrt{\frac{n_1 + n_0 - 2}{1 - r_{pb}^2}} \tag{C.3}$$

where r_{pb} is the the point-biserial correlation coefficient. From this value of t, a p-value of the significance can be obtained by computing the area of the Student's t-test from $-\infty$ to the computed value of t with $(n_1 + n_0 - 2)$ degrees of freedom. If the p-value is lower than a chosen critical value of the significance (usually 0.05), one can consider that there is a significant correlation between the continuous and the dichotomous variable. Otherwise, one must conclude that there is not a significant correlation between both variables.

Reference

1. Kornbrot D (2005) Point biserial correlation. Wiley statsRef: statistics reference online

Appendix D
Disparity Filter

The disparity filter [1] takes advantage of the local fluctuations present in the weights of the links between nodes. It is useful to define the strength s_i of a node i as the sum of the weights (ν_{ij}) of the links associated to this node, $s_i = \sum_j \nu_{ij}$. The filtering method starts by normalizing the weight of the nodes $p_{ij} = \frac{\nu_{ij}}{s_i}$, where ν_{ij} is the weight of a link j of the node i, since one needs a measure of the fluctuations of the weights attached to a node at the local level. The key point is that a few links have a large value of p_{ij} being thus more significant than the others, as computed by the disparity measure defined as $\Upsilon_i(k) \equiv k \sum_j p_{ij}^2$, where k is the degree of the node and p_{ij} is the normalized weight of the link between node i and node j.

In the application of this method to metabolic networks, $\Upsilon_i(k)$ characterizes the level of local heterogeneity of a metabolite i, and so p_{ij} stands for the normalized weight of the link between metabolite i and reaction j, with ν_{ij} the flux of reaction j. Under perfect homogeneity, when all the links share the same amount of the strength of the node, $\Upsilon_i(k)$ equals 1 independently of k, whereas for perfect heterogeneity, when one of the links carries the whole strength of the node, $\Upsilon_i(k)$ equals k. Usually, an intermediate behavior is observed in real systems.

To assess the deviations of the weights of the links, a null model is used which provides the expectation of the disparity measure of a node in a random case. The null hypothesis consists on the fact that the normalized weights that correspond to a certain node are produced by a random assignment coming from a uniform distribution. Notice that, since in this chapter directed metabolic networks are used, one has three kinds of links. Bidirectional links are decoupled into incoming and outgoing links, leading to a network where nodes have incoming and outgoing links. Each kind of links are treated independently, each one having its own probability density function. The filter then proceeds by identifying which links must be preserved. To do this, one computes the probability α_{ij} that a weight p_{ij} is non-compatible with the null model. This probability is compared to a significance level α, and thus links that carry weights with a probability $\alpha_{ij} < \alpha$ can be considered non-consistent with the null model and they are considered significant for the metabolite. The probability

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

α_{ij} is computed with the expression $\alpha_{ij}^{in/out} = (1 - p_{ij}^{in/out})^{k^{in/out}-1}$. Note that, for nodes with only one incoming or outgoing connection, one uses the prescription to preserve those links.

Reference

1. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106:6483–6488

Appendix E
Hit-And-Run Algorithm

The feasible flux phenotypes (FFP) space of different metabolic models in specific environments has been explored using different sampling techniques [1–4]. Here, the Hit-And-Run (HR) algorithm is used, tailoring it to enhance its sampling rate and to minimize its mixing time [4]. On what follows, the key points and ideas behind the HR algorithm are stated.

One must start by noticing that all points in the FFP space must simultaneously satisfy mass balance conditions and uptake limits for internal and exchanged metabolites. The former requirement defines a set of homogeneous linear equalities, whose solution space is K, while the latter defines a set of linear inequalities, whose solutions lie in a convex compact set V. From a geometrical point of view, the FFP space is thus given by the intersection $S = K \cap V$. A key step of the HR approach used here consists on realizing that one can directly work in S by sampling V in terms of a basis spanning K. This allows to retrieve all FFPs that satisfy mass balance in the medium conditions under consideration, without rejection. Additionally, sampling in S allows to perform a drastic dimensional reduction and to decrease considerably the computation time. Indeed, assuming to have N reactions, I internal metabolites, and E exchanged metabolites ($N > I + E$), one has that $S \subset \mathbb{R}^{N-I}$, which is typically a space with greatly reduced dimensionality with respect to $V \subset \mathbb{R}^{N}$.

Once a basis for K is found, the main idea behind HR is fairly simple. Given a feasible solution $\boldsymbol{\nu}_{\mathrm{o}} \in S$, a new, different feasible solution $\boldsymbol{\nu}_{\mathrm{n}} \in S$ can be obtained as follows:

1. Choose a random direction $\boldsymbol{u}$ in $\mathbb{R}^I$
2. Draw a line ℓ through $\boldsymbol{\nu_o}$ along direction $\boldsymbol{u}$:

$$\ell : \boldsymbol{\nu_o} + \lambda \boldsymbol{u}, \qquad \lambda \in \mathbb{R}$$

3. Compute the two intersection points of ℓ with the boundary of S, parametrized by $\lambda = \lambda_-, \lambda_+$:

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

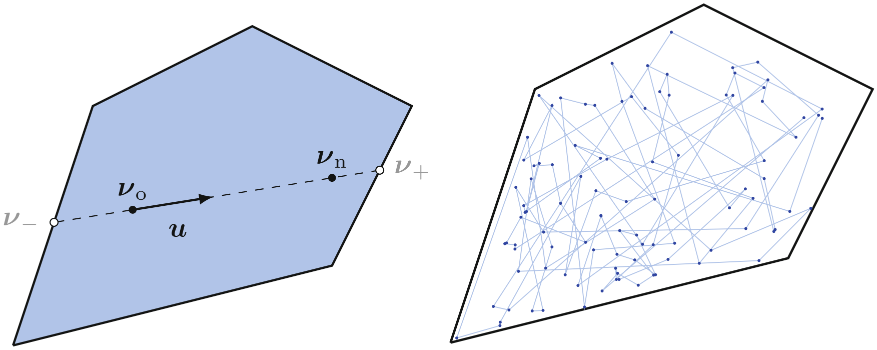


Fig. E.1 Illustrative representation of the HR fundamental step, which generates a new feasible state ν_n $\boldsymbol{\nu_n}$ from a given one $\boldsymbol{\nu_o}$. Extracted from Reference [6]

$$\boldsymbol{\nu_-} = \boldsymbol{\nu_o} + (\lambda_-)\boldsymbol{u}$$
$$\boldsymbol{\nu_+} = \boldsymbol{\nu_o} + (\lambda_+)\boldsymbol{u}$$

4. Choose a new point $\boldsymbol{\nu}_\mathrm{n}$ from ℓ, uniformly at random between $\boldsymbol{\nu}_-$ and $\boldsymbol{\nu}_+$. In practice, this implies choosing a value λ_n in the range (λ_-, λ_+) uniformly at random, and then

$$\boldsymbol{\nu}_\mathrm{n} \equiv \boldsymbol{\nu}_\mathrm{o} + \lambda_\mathrm{n}\boldsymbol{u}$$

This procedure is repeated iteratively so that, given an initial condition, the algorithm can produce an arbitrary number of feasible solutions (see Fig. E.1 for an illustrative representation of the algorithm). The initial condition, which must be a feasible metabolic flux state itself (i.e., it must belong to S), is obtained by other methods. In this work, the algorithm called MinOver is used, see References [4, 5], but any other technique is valid. In particular, in cases where small samples of the FFP space have been already obtained by other sampling techniques, such points can be used to feed the HR algorithm and produce a new, larger sample.

It was proven [7] that, by iterating steps (1–4), the samples obtained are asymptotically unbiased, in the sense that the whole FFP space is explored with the same likelihood in the limit of very large samples. In practice, one must always work with a finite sample, and hence the following additional measures are taken so as to ensure that the used samples were truly representative of the whole FFP space. In particular:

1. Only one every 10^3 points generated by HR is included in the final sample. This effectively decreases the "mixing time" of the algorithm, since the correlation among the points that are actually retained decays fast.
2. Different initial conditions are used. Results show no dependence on the initial condition, as expected for large samples. Even so, the first 30% of points are discarded, in order to rule out any subtler effect of the initial condition on the final results.

3. Results are recalculated using subsamples of size 10% of the original sample. Qualitative differences between the two sets are not found.

Since the HR algorithm is very efficient itself and due to the dimensionality reduction that this implementation adds, very large samples can be generated in reasonable time. For each model, samples of size 10^9 are initially created, giving rise to a final set of 10^6 feasible solutions uniformly distributed along the whole FFP space.

References

1. Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886–897
2. Wiback SJ, Famili I, Greenberg HJ, Palsson BØ (2004) Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. J Theor Biol 228:437–447
3. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427(6977):839–843
4. Massucci FA, Font-Clos F, De Martino A, Pérez Castillo I (2013) A novel methodology to estimate metabolic flux distributions in constraint-based models. Metabolites 3:838–852
5. Krauth W, Mezard M (1987) Learning algorithms with optimal stability in neural networks. J Phys A 20(11):L745–L752
6. Güell O, Massucci FA, Font-Clos F, Sagués F, Serrano MÁ (2015) Mapping high-growth phenotypes in the flux space of microbial metabolism. J R Soc Interface 12:20150543
7. Lovász L (1999) Hit-and-Run mixes fast. Math Program 86(3):443–461

Appendix F
Principal Component Analysis

The computation of reaction pairs correlations may be exploited to detect how global flux variability emerges in the system through Principal Component Analysis (PCA) [1, 2] and to quantify, in turn, the closeness of optimal phenotypes to the bulk of the feasible flux phenotypes (FFP) space. On what follows, PCA is briefly described, while an illustrative example is also provided (see Fig. F.1).

One starts by writing down the matrix C_{ij} of correlations between all reaction pairs i, j. In doing this, one measures how much the variability of a reaction flux ν_i affects the flux ν_j (and viceversa). In mathematical terms, for each pair of reactions i, j, one has:

$$C_{ij} = \frac{\langle \nu_i \nu_j \rangle - \langle \nu_i \rangle \langle \nu_j \rangle}{\sqrt{\left(\langle \nu_i^2 \rangle - \langle \nu_i \rangle^2\right)\left(\langle \nu_j^2 \rangle - \langle \nu_j \rangle^2\right)}}, \tag{F.1}$$

where $\langle \ldots \rangle$ denotes an average over the sampled set and the denominator of the fraction is simply the product of the standard deviations of ν_i and ν_j. This matrix is shown in Fig. 6.4e in Chap. 6.

Matrix $\boldsymbol{C}$ is real and symmetric by definition and, thus, diagonalizable. This means that, for every eigenvector ρ_κ, one has $\boldsymbol{C}\rho_\kappa = \lambda_\kappa \rho_\kappa$. Note that matrix $\boldsymbol{C}$ describes paired flux fluctuations in a reference frame centered on the mean flux vector. The eigenvectors ρ_κ of $\boldsymbol{C}$ express, in turn, the directions along which such fluctuations are taking place. In particular, the eigenvectors ρ_1, ρ_2 associated with the first two largest (in modulo) eigenvalues dictate the two directions in space where the sampled FFP displays the greatest variability (see Fig. F.1). This implies that sampled phenotypes lie closer to the plane spanned by ρ_1 and ρ_2 than the ones produced by any other linear combination of $\boldsymbol{C}$ eigenvectors. Projecting all sampled FFP onto this plane allows thus to perform a drastic dimensional reduction yet retaining much of the original variability and allows to have a direct graphical insight on where phenotypes lie, on where the bulk of the FFP is located, and on how the Flux Balance Analysis (FBA) solution compares to them. In such plot, each phenotype $\jmath$ is described by two coordinates that may be parametrized via a radius $r_\jmath$ and an angle $\theta_\jmath$. Since the projection is normalized, it follows that $r_\jmath \leq 1$. Furthermore, the closer $r_\jmath$ to one, the

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6

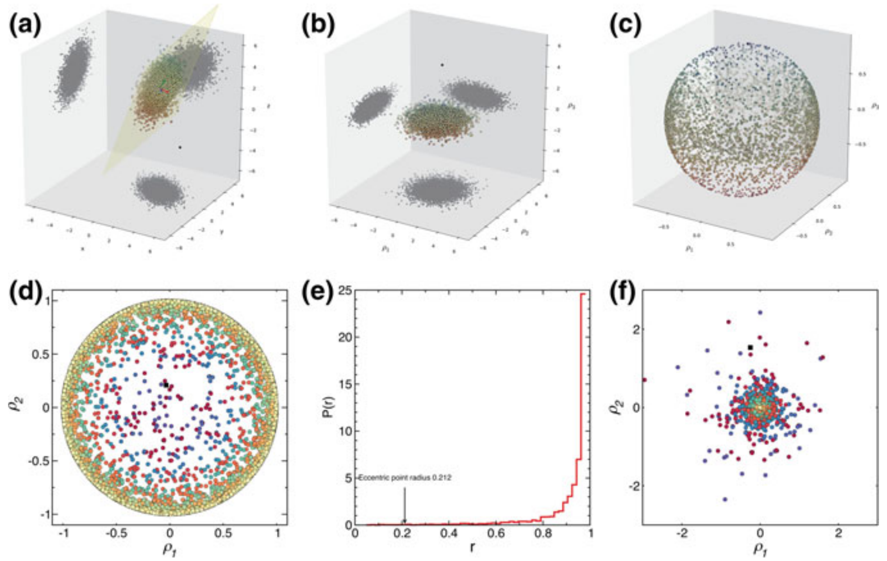


Fig. F.1 An example to describe PCA analysis. **a** FFP sampling produces a cloud of points in a multidimensional space that, when projected along the (x, z), (y, z) and (x, y) planes, is seen to span a wide range of values. Finding the eigenvectors of the correlation matrix, one can see that such points are actually clustered around a plane (plotted as a *yellow grid*). By diagonalizing the (3×3) correlation matrix, one finds the three vectors (plotted in *blue*, *red* and *green*, respectively) identifying the direction in space where the points show most variation, in a decreasing manner. A *black square* is also plotted as a reference eccentric point. **b** By projecting the sampled FFP along vectors ρ_1, ρ_2, and ρ_3, all points are squeezed in a thin region close to the (ρ_1, ρ_2) plane. This shows that the greatest variability of the sampled points actually occurs in the ρ_1, ρ_2 directions. In this representation, the eccentric *black square point* is seen to lie far from the plane with a large ρ_3 coordinate. **c** Normalizing the projection in (**b**) over the modulus of the vector identifying the point coordinate allows to quantify the closeness to the (ρ_1, ρ_2) plane. In such way all points are projected over the unit radius sphere, with the majority of points scattered near the equator, i.e., the (ρ_1, ρ_2) plane. Therefore, in this representation, eccentric points like the *black square* are close to the pole. **d** Points on the unit sphere may in turn be projected on the (ρ_1, ρ_2) plane only. In this way all points are constrained within the unitary radius *circle*, with points close to the equator in plot (**c**) now close to the *circle* and the ones close to the pole in (**c**) near the origin. In this representation, typical points, i.e., those originally closer to the *yellow plane* in (**a**), have larger radius (close to one, but smaller than that) and eccentric points have a smaller radius, like the *black square*. **e** Plotting the distribution of the points radius, as in Fig. 6.3, one sees that $P(r)$ has indeed a peak in one, with very low probability of finding a point with a radius close to zero. Similarly to Fig. 6.3 the radius of the eccentric point is indicated, highlighting how low r, eccentric points are indeed unlikely. **f** Similarly to Fig. 6.2e in Chap. 6, the points on the (ρ_1, ρ_2) plane are re-projected, but with a negative log radius. Here all points plotted in panel (**d**) appear with the same angular coordinate they have in (**d**) but with a radius $r' = -\log(r)$. In this way, typical points that in (**d**) have almost unitary radius now coalesce towards the origin and atypical points, that in (**d**) lie close to zero, are now pushed away from the origin, like the *black square*. A similar pattern is observed in Fig. 6.2e in Chap. 6, where the majority of points converge towards the origin and FBA is seen to be a rather eccentric outlier. Extracted from Reference [3] (color figure online)

better the phenotype J is described by only looking at variability along ρ_1, ρ_2. As r_J is one at the most and since one has so many phenotypes clustered together, it is possible to choose to plot the PCA projection by using an effective radius $r'_J = -\log r_J$, as in Fig. 6.4e. In this way one could better discriminate among different phenotypes and got a 'closest to the origin, closest to the ρ_1, ρ_2–plane' setup.

References

1. Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Philos Mag Ser 6 2(11):559–572
2. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
3. Güell O, Massucci FA, Font-Clos F, Sagués F, Serrano MÁ (2015) Mapping high-growth phenotypes in the flux space of microbial metabolism. J R Soc Interface 12:20150543