Springer Theses
Recognizing Outstanding Ph.D. Research

Jens Karschau

Mathematical Modelling of Chromosome Replication and Replicative Stress

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

Jens Karschau

Mathematical Modelling of Chromosome Replication and Replicative Stress

Doctoral Thesis accepted by
the University of Aberdeen, UK

Springer

Author
Dr. Jens Karschau
Biological Physics Division
Max Planck Institute for the Physics
of Complex Systems
Dresden
Germany

Supervisor
Dr. Alessandro de Moura
Department of Physics
University of Aberdeen
Aberdeen
UK

ISSN 2190-5053 ISSN 2190-5061 (electronic)
ISBN 978-3-319-08860-0 ISBN 978-3-319-08861-7 (eBook)
DOI 10.1007/978-3-319-08861-7

Library of Congress Control Number: 2014943507

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Supervisor's Foreword

DNA replication is arguably the most crucial process in living cells. It is the mechanism by which organisms pass their genetic information from one generation to the next, and life on Earth would be unthinkable without it. Recent revolutionary advances in experimental techniques in molecular biology allow the dynamics of chromosome replication of whole genomes to be investigated with unprecedented accuracy, resolution and detail. This extraordinary wealth of data makes it possible to use quantitative and predictive mathematical models to investigate this crucial biological process. In fact, many of the recent advances in the field come from multidisciplinary approaches involving applications of modelling to address important biological questions.

This thesis makes important contributions to this line of research. In particular, it addresses two key questions in the area of DNA replication: what evolutionary forces drive the positioning of replication origins in the chromosome; and how the spatial organisation of replication factories observed in many organisms is achieved. These questions lie at the heart of much of cutting-edge research in the field, and the application of mathematical modelling described in the thesis yielded new insights as well as new predictions.

The first part of this work deals with the fundamental problem that locations on the DNA have to bind with proteins first to become an origin of replication. However this process is the result of a series of stochastic events, and hence the probability of a particular location to act as an origin of replication during a particular round of cell cycle varies; and so does the distance from one eventually active origin to its nearest neighbour. Some locations have a very high protein-binding affinity and activate in nearly every cell cycle; some others are less prone to do so. Many of the previous models on DNA replication however have taken origin loci and their activation probability as an input parameter, and neglected the question whether loci positions have been chosen in a manner that depends on the likelihood of their activation. A naive assumption would be to equally distribute origins along a chromosome to give cells minimum replication; then every pair of forks that emerges to either side from an origin travels the same distance until

it coalesces with another. Here, a mathematical model of replication timing that encompasses both parameters—origin position as well as origin activation probability—shows that however sparsely spread groups of origins achieve minimum replication time, if their activation probability is low in contrast to an equally spread out origin distribution. This monograph shows the importance of origin grouping in nature using a genetic algorithm (that mimics evolution by giving benefit to those individuals with least required time to replicate their genome, i.e. those with optimal origin locations). The result particularly relates to an example of yeast, where origin locations and activation probabilities are known; where Jens shows that low probability origins are predominantly found grouped together. He also adds two further examples to the discussion of origin grouping. One is with regard to early frog embryos, where there is seen variation in the actual time of origin activation, and the other concerns organisms with multiple origins on a circular chromosome, which show little grouping behaviour in nature.

The second part culminates in an intuitive explanation for the formation of replication factories: they result from random encounters of distal replication forks that are seen to localise together in an energetically preferred state. Conversely, there is no requirement for an active transport and controlling mechanism to bring them together, and random association can become achieved simply by means of diffusion. The herein developed mathematical model takes a set of experimentally measured association probabilities of neighbouring replication forks in yeast. These forks have a maximal possible separation from another, which is given by the length of the piece of DNA that connects them. A fit of the model to data shows that their probability of association decays as a function of their distance from another as well as their binding energy. This model then extrapolates well to the replication factory size distribution of an entire yeast cell, which experimental collaborators also have measured in vivo. Conclusively, this makes the process described in this thesis a classic example for developing a physical model of a biological process that not only produces a fit using known data, but is also able to correctly produce new predictions.

This work is the result of real cross-disciplinary collaborations between biologists and physicists, and as a result its findings represent advances in both physics/applied mathematics and molecular biology. This kind of intrinsically multidisciplinary research is becoming more and more necessary in the rapidly evolving field of molecular biology, and this work is a fine example of that.

Aberdeen, May 2014 Dr. Alessandro de Moura

Abstract

DNA replication is a common feature of life, and proper genome synthesis is crucial for error-free cell division to occur. Failure in this can be lethal for an entire generation of cells, or even give rise to cancer. Replication starting points (origins) play an important role for proper DNA synthesis. It is their distance from one to another that determines the replication fork travel time, and thus the time required until synthesis completion. Much of previous theoretical work on how DNA replication can be faithful neglected how these origins take their place and how replication time is affected when origins fail to activate. It is however crucial that origin loci are chosen so that too large gaps between them are avoided; otherwise the time until completion of chromosome replication becomes much longer than is allowed by the cell cycle.

We address this lack of knowledge here using mathematical modelling to describe swift progression through the cell cycle and efficient manners of copying the DNA. On one hand, the DNA synthesis rate is fixed, and thus the time for replicating DNA between origins should be too. On the other hand, origin activation is stochastic which might cause delays in replication completion times. It is therefore a balancing act to spread out origins in a certain manner to compensate for variations in activation. We show both analytically and through numerical simulations that there exist two regimes for origins, either positioned together in groups spaced far away from the next, or as equally scattered single origins depending on the uncertainty when activation occurs. We apply the model to known origin locations in yeast and show that grouping is a means of organisation driven by evolutionary pressure. The model is able to reproduce origin distributions of early frog embryos which are thought to be random, and shows contrarily that grouping must occur in order to swiftly complete replication. The model also holds when considering a circular DNA topology as for instance archaeal genomes have, as well as if applied to the whole replication profiling data of yeast.

We also introduce a model to account for the interaction of replication forks with each other which leads to their assembly into *replication factories*. For simplicity, cartoons often depict DNA replication on a straight one-dimensional line. In fact we deal with a polymer that is packed and modified on different levels yielding higher order structures of organisation. DNA replication also appears to be spatially organised within the cellular nucleus. Active replication forks are experimentally observed to organise in clusters of *replication factories*. We

initially investigate these by describing the process with a bead-on-a-string model. The initial model represents two active pairs of replication forks connected by DNA. We show using Boltzmann-statistics that their assembly into a factory is stochastic and matches experimental association probabilities. The model then extends to describe properties of experimental distributions such as fork numbers per cluster during the DNA synthesis phase for genome wide yeast replication data. Our in silico distribution of forks per factory matches in vivo data well; which suggests that active forks encounter each other randomly for an association into replication factories.

Acknowledgments

I would first like to thank my supervisor Alessandro de Moura. I feel very grateful to him for offering me his guidance, training and support to create and develop my own ideas. I would also like to thank my co-supervisor Julian Blow for his support, and thank him for his time, patience and constructive criticism during our discussions on how to model DNA replication. Most of all I am indebted to both of my supervisors and the Scottish Universities Life Science Alliance for giving me the chance to work on a fascinating project. It allowed me to establish further collaborations and I would like to thank all my collaborators for their experimental contribution to this thesis, particularly Peter Gillespie, Conrad Nieduszynski, Nazan Saner, Renata Retkute and Tomo Tanaka.

I would also like to acknowledge the rest of the Physics group at Aberdeen for their helpful and stimulating discussions, my office-mates Luca Ciandrini and Kelly Iarosz, my flat-mates Christopher Brackley and Nicolas Rubido-Obrer, my house-mates Tina and Aaron Schiavone, and not to forget all my other friends Michael Budnitzki, Thomas Burghagen, Stuart Campbell, Lucas Fernandes, Fiona Harden, Stefan Heldt, Martin Klauke, Thomas Pfau, Elahe Radmaneshfar, James Reid, Markus Rehberg, Julia Safier, Marcillio dosSantos, Ulli Seeger…

Finally I would like to thank my parents and my sister, as well as everyone I forgot to mention here and who have offered me support and encouragement during my Ph.D. studies.

Contents

List of Publications

I list here publications that have arisen from this work.

J. Karschau, J. J. Blow, A. P. S. de Moura Optimal placement of origins for DNA replication. *Physical Review Letters*, 108(5):058101 (2012).

N. Saner, J. Karschau, T. Natsume, M. Gierlinski, R. Retkute, M. Hawkins, C. A. Nieduszynski, J. J. Blow, A. P. S. de Moura, T. Tanaka Stochastic association of neighboring replicons creates replication factories in budding yeast. *Journal of Cell Biology*, 202(7):1001–1012 (2013).

A further publication, other than those described herein, is.

J. Karschau, C. de Almeida, M. C. Richard, S. Miller, I. R. Booth, C. Grebogi, A. P. S. de Moura A matter of life or death: modeling DNA damage and repair in bacteria. *Biophysical Journal*, 100(4):814–821 (2013).

Chapter 1
Introduction

To accomplish their numerous tasks cells must create and control an internal (spatial and temporal) order of processes by properly organising their resources. One of these processes—arguably the most crucial process of all—is DNA replication whose mechanisms appear to rely on random events. At first, this seems counter-intuitive as one would expect tremendous fluctuations in the time it takes cells to duplicate, but this is not the case: most cells have a well-timed cell cycle, and this is necessary if they are to have consistent growth rate and generation times. DNA acts as the blueprint of the entire protein machinery and cellular architecture, and its integrity when passed on from mother to daughter cells is therefore of particular importance. Diseases are a common consequence of replication failure, which can lead to embryonic death, cell apoptosis, or abnormal cell growth. This has the potential to imbalance tissue growth leading to malignant tumour growth—making replication failure one of the most common causes of cancer. In order to understand how such failure arises it is necessary to first comprehend how robust (precisely timed) DNA replication occurs under normal and healthy circumstances. Although the structure of the DNA has been known for over 50 years we still lack complete understanding of all facets of DNA replication. The aim of the work presented in this thesis is to address the lack of knowledge in this area using mathematical modelling, and to ultimately further our fight against diseases such as cancer.

1.1 The Cell Cycle

DNA replication occurs inside the nucleus of eukaryotic cells. Unlike bacteria, that do not have compartmental structures like the cellular nucleus and that can have concurrent rounds of DNA replication, DNA replication in eukaryotes is subject to a strict timing regime. This is the cell cycle, and it sets the time line of events in the life of a cell. Figure 1.1a depicts a typical cell cycle as we find it in most animal, human and yeast cells. It contains 4 phases, one of these is the mitosis phase (M-phase) where cells divide and produce offspring. Two further phases of the cell cycle are

J. Karschau, *Mathematical Modelling of Chromosome Replication and Replicative Stress*,
Springer Theses, DOI 10.1007/978-3-319-08861-7_1

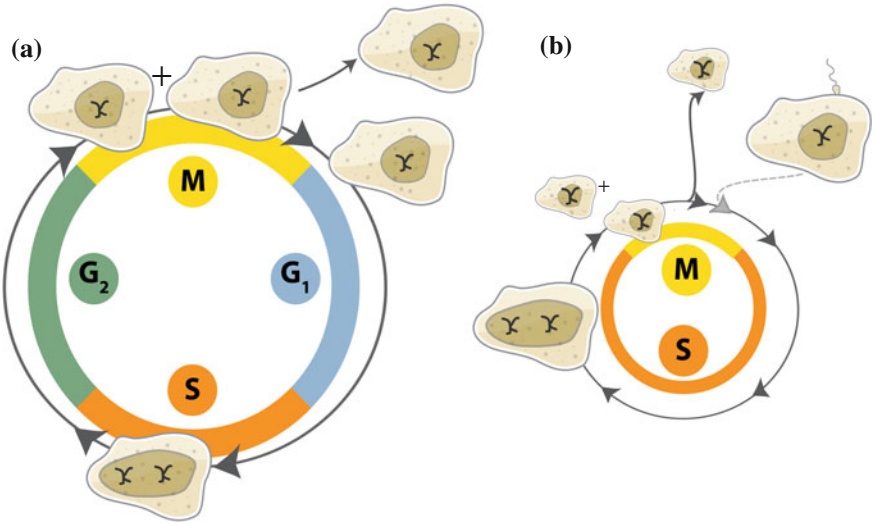

Fig. 1.1 The eukaryotic cell cycle. **a** The cell cycle of most eukaryotic organisms such as *Saccharomyces cerevisiae* contains four phases. Cell division occurs during M-phase, origin licensing during G_1-phase, DNA synthesis during S-phase, and cells grow and prepare for division in G_2-phase. **b** Early *Xenopus laevis* embryos have an abbreviated cell cycle which only consists of M-phase (during which the cells divide and origin licensing follows), and S-phase (during which the DNA is copied). Once fertilised, zygotes—eggs fused with spermii—begin replicating and dividing for 12 rounds each lasting 25 min. During this time cells only double their genome and then divide without going through G_1- and G_2-phase

G_1 and G_2 phase which are also often called gap phases as they sit in between the DNA synthesis phase (S-phase) and M-phase. Their role during the cell cycle is to either prepare the cell for DNA replication (G_1-phase) or to give the cell time to grow and prepare for division into two daughter cells (G_2-phase). The transition from one phase to another is regulated by biochemical agents called cyclines and cycline dependent kinases [1] which govern the timing of events in a cell. Their levels can be biochemically measured which can hint which cell cycle phase is currently active at a particular point during an observation of cells. There is an ever-growing body of also theoretical works that try to model cell cycle events using for example signalling networks or sets of ordinary differential equations. We invite the interested reader to explore the work by Radmaneshfar [2] for further information on modelling cell cycle as well as the consequences of stresses when exerted on a cell—particularly for the case of varying osmotic pressure.

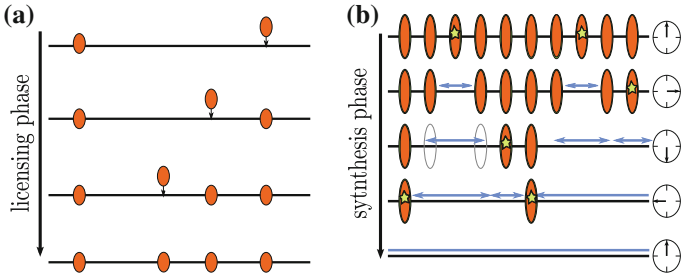

Fig. 1.2 DNA replication consists of two separate phases. **a** During the licensing phase origin-forming proteins bind to the DNA (*orange ovals*). **b** During the later synthesis phase, these origins become activated (indicated by *star*). From an activated origin replication forks emerge from either side of it (*blue arrows*) synthesising the DNA. Origins which have yet to become activated become unlicensed once DNA in this region has been replicated (*hollow ovals*). Their activation is then impossible

1.2 General Principles of the DNA Replication Process

Several mechanisms, the complexity of which is not yet fully understood, work to ensure that the DNA is properly copied—in its entirety—prior to cell division. A temporal separation of processes avoids multiple copies of DNA: the loading of inactive proteins onto potential replication initiation sites (*origins*) occurs only during a distinct phase of the cell cycle before actual origin activation [3, 4] (cf. Fig. 1.2a). The activation of licensed origins occurs in another phase when bidirectional forks emerge from origins (Fig. 1.2a). Depending on the organism the timing of these two phases can differ. For example, the yeast *Saccharomyces cerevisiae* has a cell cycle consisting of four phases (Fig. 1.1a), only two of which have relevance for DNA replication (G_1 and S-phase); it takes about 90 min to complete one round of the cell cycle. In contrast, early *Xenopus laevis* frog embryos have a shorter cell cycle consisting only of those DNA replication relevant phases (as shown in Fig. 1.1b), and completion of their cycle is within 25 min.

The first stage of DNA replication is often referred to as *licensing* (Fig. 1.2a); this is where various proteins bind to the DNA at the origin sites. Depending on the organism, licensing can be at random or sequence specific DNA positions [5], and although the licensing components are known to assemble into the *pre replication complex* (preRC), details of the interaction amongst the components is not yet fully understood. The current model suggests that proteins find their licensing binding site via diffusion, i.e. it is a stochastic process. Recently, it has also been shown that the choice of licensed origin sites in higher organisms varies from tissue to tissue or from one embryonal stage of development to another [6–8]. Another study suggests that changes to several limiting factors could lead to a prolonging of the replication time [9]; this also fits with another model where the cell cycle slows down during the mid-blastula transition in embryonal development [10]. Despite these advances, that work does not explain how the positions of the cohort of origins are chosen, and crucially how this results in a consistent timing for replication.

It is puzzling how this collection of stochastic processes in DNA replication can still yield reproducible timing, that is aligned with the cell cycle. Specifically, origins in *Xenopus laevis* appear to take random places and it is currently unknown how a random placement can give reliable replication completion times; this is the so called *random completion* or *random gap* problem [5, 11]. Besides the placement of origins, dynamic processes such as replication fork progression or stall (pausing due, e.g. to DNA damage) have also been shown to impact replication timing. For example, fork movement also plays some role in replication timing such as consequences of asymmetries in forks progressing either in a $3' \rightarrow 5'$ or $5' \rightarrow 3'$ direction [12], as well as does the DNA sequence appear to shape genomic positions which eventually leads to preferred origin locations, i.e. so termed timing domains [12–14].

Specifically, proper origin spacing is necessary because after licensing there is no further opportunity to lay out more origins once DNA synthesis begins. If origin sites are too far apart, replication fails and genomic information is lost because cells would divide before all DNA has been copied. Origin activation itself is also a stochastic process. So not only must the origins be sufficiently closely located, one needs them to be sufficiently closely loaded in the right ratio—active to inactive ones—without protein and energy resources being wasted. It has been suggested that particular higher order DNA modifications, for example DNA methylation (epigenetic factors) or histone modifications can hinder licensing and are sources of timing variation [15, 16].

1.3 Aims of This Thesis

The purpose of the work presented here is to resolve current questions surrounding the effects of noise in DNA replication. This will aid in solving problems such as the *random completion problem* and how replication forks interact inside the nucleus. This thesis elucidates three key elements centred around these problems using physical modelling. Specifically, these chief questions here are:

1. Why are origins located where they are? We consider *Saccharomyces cerevisiae*, where origin locations are encoded in the sequence and ask what are the optimal origin positions given noisy conditions during origin licensing and origin activation.
2. How should origins in *Xenopus laevis* be positioned to give minimum replication time?
3. Is there an interaction between replication forks? How do they interact with each other inside the cellular nucleus?

Answering these questions will further our understanding of the DNA replication process as a whole. Within the bigger picture this will help to identify particular targets within the DNA replication mechanisms which can be used to attack or avoid cancer.

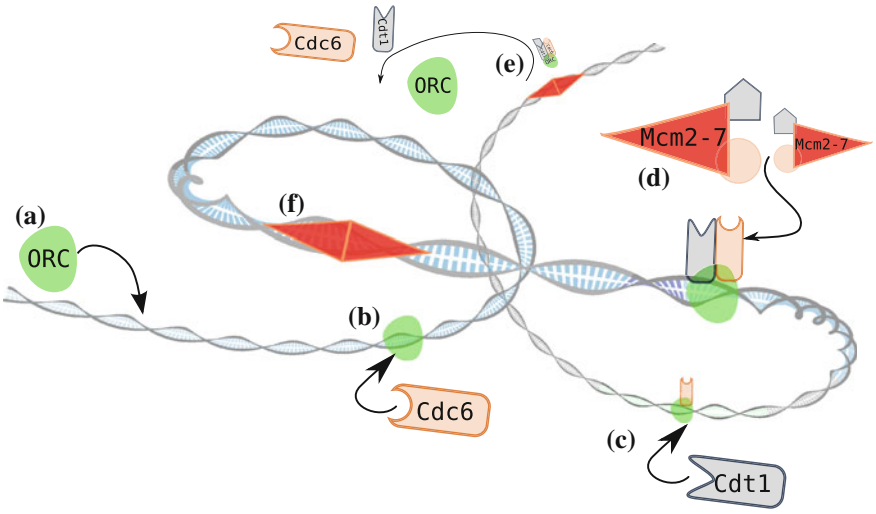

Fig. 1.3 Schematic representation of the steps involved in origin licensing. Not yet bound proteins are marked by their names, DNA-bound proteins are only marked with their symbols. **a** ORC binds to DNA which then recruits Cdc6 (**b**) and Cdt1 (**c**). These reactions are reversible. **d** In the final step, Mcm2-7 irreversibly binds as a double-hexamer. **e** This releases the previously bound components ORC, Cdc6 and Cdt1. Mcm2-7 remains irreversibly bound to DNA and forms the pre-replicative complex (preRC) in (**f**). This one can become active during S-phase as an origin of replication

1.4 Origin Licensing

As mentioned above, prior to the synthesis of DNA, replication starting points (*origins of replication*) are established at certain genomic locations (*origin loci*). This process is named origin licensing, and the fact that it is temporally separated from the DNA synthesis process means that DNA which has already been replicated does not become re-replicated.

Licensing consists of a sequential docking of proteins onto origin loci. The components have been found to be involved in a four-step mechanism as shown in Fig. 1.3. First a protein called the origin-recognition complex (ORC) binds to DNA (Fig. 1.3a). This is followed by the binding of two further scaffolding proteins called Cdc6 and Cdt1 that bind sequentially as depicted in Fig. 1.3b and c. Finally, the minichromosome maintenance complex (Mcm) consisting of several individual Mcm2-7 proteins is recruited to form the so-called *pre-replication complex* (preRC) (Fig. 1.3d), and triggers the release of ORC, Cdc6 and Cdt1 (Fig. 1.3e). However at this stage the preRC is still inactive and it shall remain inactive until the licensing phase is completed. Recent studies have shown that Mcm2-7 binds in the form of a pairwise hexamer (pMcm) [17, 18], i.e. there are always two Mcm2-7 complexes bound back-to-back with DNA [19] as shown in Fig. 1.3e and f. This suggests that Mcm2-7 is a driving force in processing the DNA during synthesis, and in fact it was shown that

Mcm2-7 acts as a helicase which unwinds the DNA [20–23], allowing the polymerase to access and copy it. So Mcm2-7 has an integral role in the replication machinery.

The way in which origin loci—those not yet licensed sites—are chosen differs depending on the organism or the tissue being studied. In this work we mainly focus on two systems, namely the yeast *Saccharomyces cerevisiae*, and the early embryos of the frog *Xenopus laevis*. It is well accepted that the yeast has signatures of specific origin loci encoded into its genomic sequence which ORC recognises. These have been termed *Autonomously Replicating Sequences* (ARS) [24, 25], and are distributed in distinctive 11 basepairs (bp) long DNA sequence specific motifs [26]. In contrast, in *Xenopus laevis* such clear and distinctive loci for ORC-binding do not exist, and ORC can bind anywhere on the chromosome [27, 28]. Both organisms, *Saccharomyces cerevisiae* and *Xenopus laevis* are good model systems for helping us to understand the organisation of licensing in man. Having these two model organisms allows us to study each mode of licensing—random locations and specific origin location—in isolation. Organisation of licensing in man does not seem as clear as in either *Saccharomyces cerevisiae* or *Xenopus laevis*. For example, a study finds that there are human genomic regions with sequence specific origin loci (as is the case for *Saccharomyces cerevisiae*) and there are also DNA segments in which there are no clear origin loci, and licensing appears to occur randomly (as is in the early *Xenopus laevis* embryo) [27]. Another investigation by Besnard et al. [29] showed that there exist a consensus guanine sequence (G-quadruplexes) which acts as a signature for origins of replication in man. G-quadruplexes consist of four guanine bases that become stacked in a way so that they form higher order structures amongst them. Yet more recently unpublished data by the Arneodo group suggests that epigenetic factors and histone-binding proteins such as H2A variants open up regions along the chromosome where origins can form (*personal communication with Alain Arneodo*). The inter-play of the accessibility of DNA with licensing has also been suggested previously in studies of DNA sequences which showed a jump in the ratio of its guanine and cytosine bases amongst leading and lagging strand DNA [30], which can act as DNA break point or as sites where origins of replication are established.

Origin licensing ends with a down-regulation of further assembly of proteins at origins [31]. The current model suggests that a degradation of the licensing factors in the cell nucleus can achieve this [32–34]; a further route to down-regulation is through blocking of the scaffolding proteins, i.e. the intermediate proteins which recruit the Mcm2-7 to form licensed origins. For example the protein Geminin blocks licensing by binding to Cdt1 preventing it from binding with an ORC-Cdc6 complex on the DNA, and ultimately inhibiting the recruitment of Mcm2-7 [35].

1.5 Origin Firing During DNA Synthesis

Once licensing completes, the cell cycle progresses to its next phase, where origin activation and the replication of DNA occurs. It is therefore called the synthesis or S-phase. The origins which were licensed during the previous phase lie *dormant*

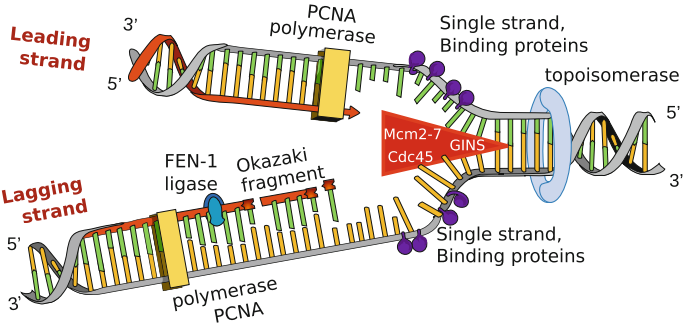

Fig. 1.4 The main components of a replication fork are a topoisomerase, a helicase (Mcm2-7), a polymerase (pol δ/ϵ), as well as further accessory proteins, e.g. Cdc45 and the GINS complex. Polymerase can only act in the $3' \rightarrow 5'$ direction. Thus it works continously in that direction on the leading strand, but forms *Okazaki* fragments on the lagging strand, which then require ligation and maturation through the ligase and FEN-1 proteins. This image is an adaption of [39]

(inactive) for some period of time until they become activated. Activation requires the binding of further proteins until the full replication machinery has assembled. In replication terminology one refers to this machinery as replication forks; they operate bi-directionally, meaning that forks emerge from either side of the origin with its machinery at their head, as shown in Fig. 1.4. The figure also shows that the polymerase (the DNA-copying element of the fork) only acts in one direction, namely in the $5'$ to $3'$ direction. So when replication forks progress, one strand, the *leading strand*, is copied continuously as the forks move in $3' \rightarrow 5'$ direction; the other is synthesised discontinuously, i.e. polymerases replicate a section of DNA in the $3' \rightarrow 5'$ direction, and then jumps back in the $5' \rightarrow 3'$ direction to replicate the next section. Such discontinuity creates intermediate fragments on the lagging strand called Okazaki fragments which require post-processing (*maturation*) to join them together (*ligate*) to form one continuous DNA strand.

The protein machinery of a replication fork consists of the following main components [36]:

- a helicase that unwinds and opens up the DNA, i.e. Mcm2-7,
- a polymerase which copies template DNA, e.g. a polymerase on the leading and another on the lagging strand.
- accessory proteins such as processivity factors that clamp polymerase tightly and prevents it from dissociating, i.e. PCNA,
- nucleases such as FEN-1 that cleave and prepares DNA for later Okazaki fragment ligation,
- further proteins for example DNA ligase I which joins Okazaki fragments or topoisomerase that unlinks parental strands,
- initiator proteins which are involved in the activation of the replication machinery and its progression [19, 37, 38]: Cdc45 and the GINS proteins.

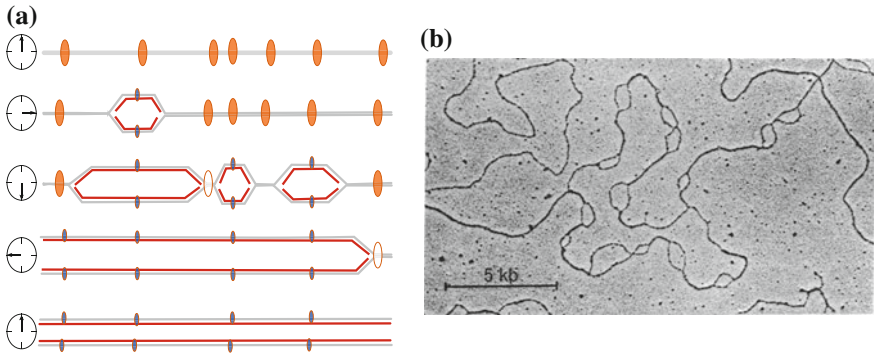

Fig. 1.5 Replication bubbles. **a** Replication starts from origins of replication shown as *oval shapes* (*top figure*). Once an origin activates replication bubbles form which carry newly synthesised DNA inside of them shown in *red*. The active origins sites are marked as *small blue circles* on either replicated DNA strand, not yet activated origins loose their ability to activate once a replication fork moves across them (*hollow ovals*). Replication finishes when forks reach the end of a chromosome or coalesce with another fork which originated elsewhere. **b** Electron micrograph image of replication bubbles in *Drosophila melanogaster*. This image was acquired from Fig. 2 from the original paper by Kriegstein and Hogness [41] who gave their kind permission to reprint it

As the protein machinery of a fork moves along the DNA, it leaves duplicated DNA in its wake, resulting in *replication bubbles*, as shown in Fig. 1.5. Replication forks flank the bubbles at either side, and a bubble carries newly synthesised DNA inside it. By comparing bubbles to linear DNA, i.e. not yet replicated DNA, one can infer the position of an origin of replication. Assuming furthermore at a constant speed this should be the midpoint of bubble. What can be gleaned from the study of replication bubbles is whether an origin was activated early or later during S-phase. For example, a large bubble suggests that forks have had longer to move further away from an origin at its centre, compared to a smaller replication bubble where replication might have only just started. However such information can also be ambiguous, because a large bubble could also be the result of the merging of two smaller bubbles. One way of visualising newly synthesised DNA uses two differently labelled nucleotides in the cellular growth medium. Their addition at different times to the medium during replication highlights some of the dynamics during replication [40]. For instance two differently labelled nucleotides can be added one at a time during replication. This will result in patterns on the DNA where those replicated from one labelled nucleotide will have one particular colour. This then aids in the identification of origin positions similar to studying the length of a replication bubble. The method also allows to distinguish between an early or late origin depending on the colour-coding of a replicated DNA stretch.

Localising origins of replication, as well as forks, is also possible using sequencing techniques (e.g. oligonucleotide microarrays [42], ChIP-chip sequencing [43], deep sequencing [44]) to examine DNA extracted from the cell during different times of the S-phase. These approaches use a set of known DNA snippets that hybridise

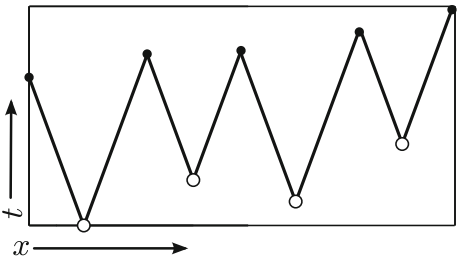

Fig. 1.6 A schematic representation of a replication timing profile is shown for the case of four origins (*hollow circles*). We assume this is how forks move if we observe a single cell. Origins are located at different positions x along the a chromosome. Origins activate at different times t and their forks progress from either side of the origin (*solid line*), until they encounter another fork or reach the end of the chromosome. Termination points are shown as *filled circles*

(are matched) with the sample DNA. Quantifying the extent of matches then relates back to the known DNA segment sequence and the marker at the time the sample was taken. For example, the study by Raghuraman et al. [42] labels replicated DNA with heavy isotopes. Doing this they can then sort replicated from unreplicated DNA and identify the corresponding known DNA sequence belonging to it. Sekedat et al. [43] make use of a different marker and they mark key proteins on the DNA during replication. They label a fork component (the GINS complex), and sort the DNA pieces that bound to it. They then determine the corresponding DNA sequence where the GINS complex sits and therefore location of replication forks over time. Samples taken at different times during DNA replication resolve fork movement spatially as well as temporally. The data produces *replication timing profiles* and these space-time plots reveal which origins activate when, when do forks merge, as well as which piece of DNA was replicated at which point in time during S-phase. A sketch of such a timing profile is shown in Fig. 1.6 for the case of four origins activating at different times when we observe this in a single cell.

When replication forks move along the DNA they might encounter a dormant origin, as is shown in the second last sequence of Fig. 1.5 (see also Fig. 1.2b). Since a dormant origin is inactive, it then becomes passively replicated [45] meaning that it loses its ability to act as a replication starting point. From there, replication then continues normally along the remaining stretch of DNA. Whenever a replication fork collides head-on with another fork, replication terminates at this point and the open ends of newly replicated DNA become ligated. Replication also terminates if a fork reaches the end of a chromosome. A replication fork might also stall, as a result of, for example, damaged DNA sites; in this case dormant origins play an important role. Their presence allows replication to restart from outside a stalled region, and they contribute not only to keep replication going, they also contribute in a manner to keep replication on-time [46]. Activating dormant origins helps to repartition not yet replicated DNA into smaller pieces which will reduce the overall replication completion time—the time the last segment of DNA takes to be fully replicated.

One particular drawback of most experimental procedures is that measurements usually observe a population average of cells (or origins). For example, experiments are normally carried out using synchronised cell cultures, so the onset of replication can be timed. However there is stochasticity in origin licensing and activation which then averages out some aspects of origins, e.g. probability of becoming licensed, the exact activation time. This makes it difficult to detect origins which have little probability to activate in a given round of cell cycle. For this single cell data is required and hard to come by, so one requires a theoretical approach as well as in silico models for a complete resolution of all facets of origins during licensing and during their activation in S-phase.

1.6 Replication Timing, Origin Positioning, and the Random Completion Problem

Replication timing and the timing of cellular division are strongly interlinked since cells would lose genomic information if they divided prior to the completion of replication. Previous research in the *Xenopus laevis* early embryo system showed that after fertilisation egg cells double at a constant speed for the first twelve rounds [10]. The time between divisions is ∼25 min which is comparably short to a 24 hour-long division cycle of most differentiated, somatic cells. However the genomic content of both cell types is the same, i.e. an information content of $6.2 \cdot 10^9$ bp (about the same magnitude as for humans).

There are several points that form a conundrum of how such rapid doubling within 25 min can be achieved in the early frog embryo system. The speed at which replication forks process the DNA in eukaryotes is apparently fixed at about 1 kbp/min with its exact value depending on the specific organism, e.g. 1.5 kbp/min in *Saccharomyces cerevisiae* [42, 43] and 0.5–0.6 kbp/min in *Xenopus laevis* [47]. As a consequence of *Xenopus laevis* early embryos undergoing cellular division within 25 min, the distance from one origin to the next must be no larger than ∼20 kbp to achieve this. However DNA saturates at approximately one Mcm2-7 per 1.5 to 3 kbp [19, 48] which means that there are about 200,000 origins from which replication starts in every cell cycle. Keeping in mind that in *Xenopus laevis* embryos origins can assemble at any given DNA sequence, then a random distribution of origin sites of this extent juxtaposed with the possibility to lay them out along $6.2 \cdot 10^9$ bp results in a high probability of having at least one gap >20 kbp [5, 11][1]; which would prolong S-phase and retard the cell cycle. However experiments show that this is not the case, moreover origins appear to have a bias towards a regular spacing with one

[1] The maximal gap allowed to complete replication on time is given by $20\,\text{min}/(2 \cdot 0.5\,\text{kb/min}) = 20\,\text{kbp}$. This is if we consider that all origins activate at the same time under the conditions of two forks replicating each at 0.5 kb/min for a period of 20 min.

origin every 5–15 kbp [11]. This conundrum has therefore been termed the *random completion problem* and also *random gap problem* which to-date has not been fully resolved [49, 50].

1.7 Mathematical Modelling of DNA Replication

Within the school of mathematical modelling of DNA replication there are two views on the importance of licensing-defined origin sites [50]. Whether or not to include licensing in a modelling approach is of special importance to resolve the aforementioned *random completion problem*.

In the first view, origins can literally form anywhere, and modelling of the replication process depends only on an initiation function. It governs the amount of origins that activate at a particular time in not yet replicated DNA regions [51, 52], and it thus acts as a rate of origin activation over time. The essence of this approach is a one-dimensional nucleation and growth model which holds a long-history and myriad applications [53] in the statistical physics realms known as the *Kolmogorov-Johnson-Mehl-Avrami* (KJMA) model [54]. The KJMA model considers random nucleation events along a one-dimensional line whose rate of nucleation depends on the particulars of the initiation function. From each nucleation point the line undergoes a transition (e.g. from unreplicated to replicated) at some speed to either side from it. The elegance of the KJMA model is its direct provision for analogy to the replication mechanism: nucleation points are origins of replication, transformation of DNA acts as changing DNA from an unreplicated to a replicated state [54] (cf. also Fig. 1.2b). However this approach falls short of biological details; it mainly addresses the question of how cells are able to replicate their genome in a short time lacking the relationship between the stochasticity of the location of origin *and* their activation times [13, 51, 54, 55]. The KJMA model of DNA replication works only properly (works out the *random completion problem*) if the initiation rate increases towards the end of S-phase [56–58]. However an increasing initiation rate requires an arbitrary amount of origins to be licensed which lacks the known biological model of *Xenopus laevis* embryos. In other words the KJMA model always requires the potential presence of a licensed origin at every arbitrary position to allow for random initiation. However there is only a finite amount of origins loaded during licensing [59], and once licensing completes the origin positions are fixed. The KJMA model is therefore incomplete and lacks a full explanation for a solution to the *random completion problem* from a licensing point of view.

Despite this shortcoming of explaining the origin positions, applying the pioneering KJMA model to *Saccharomyces cerevisiae* replication profiles yields insight into the likelihood of their known location to act as an origin [60]. One can hence use this model to extract the probability of an origin locus to contribute in a particular round of the cell cycle, i.e. its *efficiency* [13]. There exist also several other modelling approaches to extract information by also taking into account several cell cycle progression regulators [61, 62].

The second view of modelling DNA replication, considers two ingredients for a possible model. It divides the entire process into two separate phases as are in biology—licensing and activation—because an origin can only activate if it was previously licensed. For example Moura et al. developed a model in such a way, and as a result they require extra parameters [63] compared to the KJMA model. In their scenario, an origin locus has a probability to actually become licensed or not. It is therefore considered to have certain *competence* to activate later for DNA synthesis. During S-phase the licensed origin has a further probability attached to it to become activated over time. The authors achieve this by introducing a time-dependent activation distribution [64, 65] which is defined by mean activation time and standard deviation per origin. Such model can then be fitted to *Saccharomyces cerevisiae* experimental replication timing profile data to extract these relevant parameters as well as describing an origins' overall efficiency as was in the aforementioned KJMA modelling approach. There have also been further approaches [66, 67] which consider such separation of phases for DNA replication as well as linking them to players of the cell cycle for a holistic model of a population of yeast cells. As modelling separated into licensing and origin activation results in a more realistic approach, we will investigate in Chap. 2 what the best origin positions are; whether those give minimum replication time according to their parameters. We then reverse engineer the optimal origin positions by simulating an evolutionary process and find that origins take positions so that replication completes quickly.

1.8 The Spatio-Temporal Organisation of Replication Forks

A typical eukaryotic chromosome has a length of $2{\cdot}10^8$ bp and it is contained within the nucleus of a typical eukaryotic cell. It would be about 6 cm [68] long, if it were fully stretched out. So this is much longer than the actually size of a cell or even its nucleus where chromosomes are located. DNA must thus be packaged and compacted to fit into the cellular nucleus, and it is known that this occurs at different scales [1], for example DNA is wound around histones forming chromatin, chromatin condenses further to form chromosomes. Chromatin as well as chromatin organisation occur in three spatial dimensions which has been of long-standing interest by experimental and theoretical groups alike and is also of particular interest in polymer science (as reviewed in [69]). It is hypothesised that chromatin organisation also plays a role in the activation or down-regulation of a particular gene by compacting the DNA sequence of a protein and shielding it from transcription factors that try to bind with it [1]. DNA is thus much more active than the usually projected picture which only sees it as a sole means to encode for information.

When DNA condenses particular genomic regions come into contact with another which, if they were stretched out linearly in 1D would be otherwise far away from each other. For example, Duan et al. [70] suggested that packaging DNA inside the nucleus creates higher order structures that organise into regions inside the nucleus. They further suggest that localisation organises the DNA into structures which aid

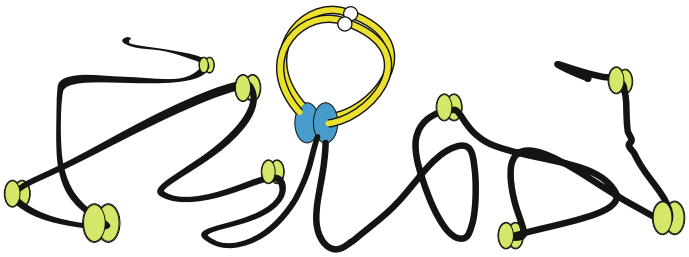

Fig. 1.7 A replisome pair. Active replication forks (*blue ovals*) synthesise unreplicated DNA (*black line*) by pulling it through them from one side and ejecting the newly synthesised DNA (*yellow*) on the other. In doing this sister replisomes, i.e. the active replication forks, move away from their origin of replication (*white circle*), but always stay attached with each other. Not yet activated origins which are shown as two *green ovals* to represent DNA-bound Mcm2-7 double hexamers waiting to become activated

in accessing a particular gene or help during replication. Boulos et al. also suggested such functional dependency of intra- and inter-chromosomal interaction during their network analysis of human chromatin regions which have potential to play a role in DNA replication. Previous to these theoretical works there has also been convincing experimental evidence for a spatial organisation of DNA during replication (and mRNA transcription) into compartment-like structures that have no physical boundaries; these observations are termed *replication factories* [71]. It is currently unresolved how these structures form and stay together despite no clear compartment wall. Work by Cisse et al. [72] suggests their formation to be more dynamic, and assembly and disassembly transiently occur with an average life-time of 5 s of equivalent structure called *transcription factories*, i.e. where DNA is copied into messenger RNA. Data we use for our analysis of replication factories in Chap. 3 also displays dynamic rates of close localisation and dislocation of unreplicated DNA regions with each other, however once DNA becomes replicated movement of these regions becomes constrained (see also [73]) that we interpret as a strong interaction between nearby replicating DNA regions. A possible difference in the organisation of replication and transcription is that transcription only acts on one DNA strand whereas during DNA replication both strands become duplicated simultaneously which might require a stronger force to keep the replication machinery associated.

The models of replication kinetics which are discussed in the previous Sect. 1.7 considered DNA replication to happen spatially in only one dimension: replication forks move away to either side from an origin. Experiments however suggest that in a three dimensional space of a nucleus replication forks nucleated from one origin stay bound together as *sister replisomes* [74–76]. Sister replisomes are sketched in Fig. 1.7. They spool unreplicated DNA from one side and then spool out replicated DNA to the other. The replisome pairs stay associated for the course of replication of a DNA segment, but more importantly they were also shown to group together with other replisomes eventually forming *replication factories* [75, 77, 78]—regions with a high fork content. A proposed function for factories is that there is a high pool of

essential proteins required for DNA synthesis localised around them [79]. This can aid to, for example, activate dormant origins under replicative stress such as DNA damage [80–82], and when there is a large amount of origins required to activate quickly. To-date, it is not clear what keeps these factories together, however previous modelling shows that an energy barrier must be overcome to bear the entropic cost from forming DNA loops to have factory structures [83–85].

We address this lack of knowledge in Chap. 3. Using Boltzmann statistics and numerical simulations we show that replisomes randomly associate with each other on a chromosome. We establish a model for two sister replisomes to pair and then extend our model to account for replisomes associating on a genome-wide scale. This model is firmly grounded at the core biological question and supplements experimental data of the probability of observing associations in vivo in *Saccharomyces cerevisiae* cells.

References

1. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994)
2. E. Radmaneshfar, *Mathematical Modelling of the Cell Cycle Stress Response* (Springer International Publishing, Switzerland, 2014). doi:10.1007/978-3-319-00744-1
3. H. Nishitani, Z. Lygerou, Control of DNA replication licensing in a cell cycle. Genes Cells **7**(6), 523–534 (2002)
4. A.C. Porter, Preventing DNA over-replication: a Cdk perspective. Cell Div. **3**, 3 (2008). doi:10.1186/1747-1028-3-3
5. O. Hyrien, K. Marheineke, A. Goldar, Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. Bioessays **25**(2), 116–125 (2003). doi:10.1002/bies.10208
6. J. Nordman, T.L. Orr-Weaver, Regulation of DNA replication during development. Development **139**(3), 455–464 (2012). doi:10.1242/dev.061838
7. M. Méchali, K. Yoshida, P. Coulombe, P. Pasero, Genetic and epigenetic determinants of DNA replication origins, position and activation. Curr. Opin. Genet. Dev. **23**(2), 124–131 (2013). doi:10.1016/j.gde.2013.02.010
8. O. Hyrien, C. Maric, M. Méchali, Transition in specification of embryonic metazoan DNA replication origins. Science **270**(5238), 994–997 (1995)
9. C. Collart, G.E. Allen, C.R. Bradshaw, J.C. Smith, P. Zegerman, Titration of four replication factors is essential for the xenopus laevis midblastula transition. Science **341**(6148), 893–896 (2013). doi:10.1126/science.1241530
10. M. Kirschner, J. Newport, J. Gerhart, The timing of early developmental events in Xenopus. Trends Genet. **1**, 41–47 (1985). doi:10.1016/0168-9525(85)90021-6
11. J.J. Blow, Control of chromosomal DNA replication in the early Xenopus embryo. EMBO J. **20**(13), 3293–3297 (2001). doi:10.1093/emboj/20.13.3293
12. A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. D'Aubenton-Carafa, C. Thermes, Multi-scale coding of genomic information: from DNA sequence to genome structure and function. Phys. Rep. **498**(2–3), 45–188 (2011). doi:10.1016/j.physrep.2010.10.001
13. A. Baker, B. Audit, S.C.-H. Yang, J. Bechhoefer, A. Arneodo, Inferring where and when replication initiates from genome-wide replication timing data. Phys. Rev. Lett. **108**(26), 268101 (2012). doi:10.1103/PhysRevLett.108.268101

14. T.J. Newman, M.A. Mamun, C.A. Nieduszynski, J.J. Blow, Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts. Nucleic Acids Res. **41**(21), 9705–9718 (2013). doi:10.1093/nar/gkt728
15. S. Jun, J. Herrick, A. Bensimon, J. Bechhoefer, Persistence length of chromatin determines origin spacing in Xenopus early-embryo DNA replication: quantitative comparisons between theory and experiment. Cell Cycle **3**(2), 211–217 (2004). doi:10.4161/cc.3.2.655
16. A.G. Evertts, H.A. Coller, Back to the origin: reconsidering replication, transcription, epigenetics, and cell cycle control. Genes Cancer **3**(11–12), 678–696 (2012). doi:10.1177/1947601912474891
17. D. Remus, F. Beuron, G. Tolun, J.D. Griffith, E.P. Morris, J.F.X. Diffley, Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. Cell **139**(4), 719–730 (2009). doi:10.1016/j.cell.2009.10.015
18. C. Evrin, P. Clarke, J. Zech, R. Lurz, J. Sun, S. Uhle, H. Li, B. Stillman, C. Speck, A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. Proc. Natl. Acad. Sci. U.S.A. **106**(48), 20240–20245 (2009). doi:10.1073/pnas.0911500106
19. A. Gambus, G.A. Khoudoli, R.C. Jones, J.J. Blow, MCM2-7 form double hexamers at licensed origins in Xenopus egg extract. J. Biol. Chem. **286**(13), 11855–11864 (2011). doi:10.1074/jbc.M110.199521
20. A. Bielinsky, S. Gerbi, Where it all starts: eukaryotic origins of DNA replication. J. Cell Sci. **114**(4), 643–651 (2001)
21. M.L. Bochman, A. Schwacha, The Mcm2-7 complex has in vitro helicase activity. Mol. Cell **31**(2), 287–293 (2008). doi:10.1016/j.molcel.2008.05.020
22. M.L. Bochman, A. Schwacha, The Mcm complex: unwinding the mechanism of a replicative helicase. Microbiol. Mol. Biol. Rev. **73**(4), 652–683 (2009). doi:10.1128/MMBR.00019-09
23. T.J. Takara, S.P. Bell, Putting two heads together to unwind DNA. Cell **139**(4), 652–654 (2009). doi:10.1016/j.cell.2009.10.037
24. C.S. Newlon, J.F. Theis, The structure and function of yeast ARS elements. Curr. Opin. Genet. Dev. **3**(5), 752–758 (1993). doi:10.1016/S0959-437X(05)80094-2
25. K. Shirahige, T. Iwasaki, M.B. Rashid, N. Ogasawara, H. Yoshikawa, Location and characterization of autonomously replicating sequences from chromosome VI of Saccharomyces cerevisiae. Mol. Cell. Biol. **13**(8), 5043–5056 (1993). doi:10.1128/MCB.13.8.5043
26. S.P. Bell, A. Dutta, DNA replication in eukaryotic cells. Annu. Rev. Biochem. **71**, 333–374 (2002). doi:10.1146/annurev.biochem.71.110601.135425
27. C. Cvetic, J.C. Walter, Eukaryotic origins of DNA replication: could you please be more specific? Semin. Cell Dev. Biol. **16**(3), 343–353 (2005). doi:10.1016/j.semcdb.2005.02.009
28. M. Méchali, Eukaryotic DNA replication origins: many choices for appropriate answers. Nat. Rev. Mol. Cell Biol. **11**(10), 728–738 (2010). doi:10.1038/nrm2976
29. E. Besnard, A. Babled, L. Lapasset, O. Milhavet, H. Parrinello, C. Dantec, J.-M. Marin, J.-M. Lemaitre, Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. **19**(8), 837–844 (2012). doi:10.1038/nsmb.2339
30. C.-L. Chen et al., Replication-associated mutational asymmetry in the human genome. Mol. Biol. Evol. **28**(8), 2327–2337 (2011). doi:10.1093/molbev/msr056
31. J.J. Blow, A. Dutta, Preventing re-replication of chromosomal DNA. Nat. Rev. Mol. Cell Biol. **6**(6), 476–486 (2005). doi:10.1038/nrm1663
32. T. Kondo et al., Rapid degradation of Cdt1 upon UV-induced DNA damage is mediated by SCF-Skp2 complex. J. Biol. Chem. **279**(26), 27315–27319 (2004). doi:10.1074/jbc.M314023200
33. T. Senga, U. Sivaprasad, W. Zhu, J.H. Park, E.E. Arias, J.C. Walter, A. Dutta, PCNA is a cofactor for Cdt1 degradation by CUL4/DDB1-mediated N-terminal ubiquitination. J. Biol. Chem. **281**(10), 6246–6252 (2006). doi:10.1074/jbc.M512705200
34. J. Hu, Y. Xiong, An evolutionarily conserved function of proliferating cell nuclear antigen for Cdt1 degradation by the Cul4-Ddb1 ubiquitin ligase in response to DNA damage. J. Biol. Chem. **281**(7), 3753–3756 (2006). doi:10.1074/jbc.C500464200

35. M.L. DePamphilis, J.J. Blow, S. Ghosh, T. Saha, K. Noguchi, A. Vassilev, Regulating the licensing of DNA replication origins in metazoa. Curr. Opin. Cell Biol. **18**(3), 231–239 (2006). doi:10.1016/j.ceb.2006.04.001
36. T.K. GS Brush, DNA replication mechanisms. DNA Replication Eukaryot. Cells. **71**, 333–374 (1996)
37. A. Zembutsu, S. Waga, De novo assembly of genuine replication forks on an immobilized circular plasmid in Xenopus egg extracts. Nucleic Acids Res. **34**(13), e91 (2006). doi:10.1093/nar/gkl512
38. A. Gambus, R.C. Jones, A. Sanchez-Diaz, M. Kanemaki, F. van Deursen, R.D. Edmondson, K. Labib, GINS maintains association of Cdc45 with MCM in replisome progression complexes at eukaryotic DNA replication forks. Nat. Cell Biol. **8**(4), 358–366 (2006). doi:10.1038/ncb1382
39. M. Ruiz. DNA replication, http://en.wikipedia.org/wiki/File:DNA_replication_en.svg. Accessed 28 October 2013
40. J. Herrick, P. Stanislawski, O. Hyrien, A. Bensimon, Replication fork density increases during DNA synthesis in X. laevis egg extracts. J. Mol. Biol. **300**(5), 1133–1142 (2000)
41. H.J. Kriegstein, D.S. Hogness, Mechanism of DNA replication in Drosophila chromosomes: structure of replication forks and evidence for bidirectionality. Proc. Natl. Acad. Sci. U.S.A. **71**, 135–139 (1974). doi:10.1073/pnas.71.1.135
42. M.K. Raghuraman et al., Replication dynamics of the yeast genome. Sci. **294**(5540), 115–121 (2001). doi:10.1126/science.294.5540.115
43. M.D. Sekedat, D. Fenyö, R.S. Rogers, A.J. Tackett, J.D. Aitchison, B.T. Chait, GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. Mol. Syst. Biol. **6**, 353 (2010). doi:10.1038/msb.2010.8
44. C. A. Müller et al. The dynamics of genome replication using deep sequencing. Nucleic Acids Res. (2013), gkt878. doi:10.1093/nar/gkt878
45. A.M. Woodward, T. Göhler, M.G. Luciani, M. Oehlmann, X. Ge, A. Gartner, D.A. Jackson, J.J. Blow, Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. J. Cell Biol. **173**(5), 673–683 (2006). doi:10.1083/jcb.200602108
46. D. McIntosh, J.J. Blow. Dormant origins, the licensing checkpoint, and the response to replicative stresses. Cold Spring Harb. Perspect. Biol. **4**(10) (2012). doi:10.1101/cshperspect.a012955
47. H.M. Mahbubani, T. Paull, J.K. Elder, J.J. Blow, DNA replication initiates at multiple sites on plasmid DNA in Xenopus egg extracts. Nucleic Acids Res. **20**(7), 1457–1462 (1992)
48. H.M. Mahbubani, Cell cycle regulation of the replication licensing system: involvement of a cdk-dependent inhibitor. J. Cell Biol. **136**(1), 125–135 (1997). doi:10.1083/jcb.136.1.125
49. N. Rhind, DNA replication timing: random thoughts about origin firing. Nat. Cell Biol. **8**(12), 1313–1316 (2006). doi:10.1038/ncb1206-1313
50. J. Bechhoefer, N. Rhind, Replication timing and its emergence from stochastic processes. Trends Genet. **28**(8), 374–381 (2012). doi:10.1016/j.tig.2012.03.011
51. J. Herrick, S. Jun, J. Bechhoefer, A. Bensimon, Kinetic model of DNA replication in eukaryotic organisms. J. Mol. Biol. **320**(4), 741–750 (2002)
52. S. Jun, J. Bechhoefer, Nucleation and growth in one dimension. II. application to DNA replication kinetics. Phys. Rev. E **71**(1), 011909 (2005). doi:10.1103/PhysRevE.71.011909
53. M. Fanfoni, M. Tomellini, The Johnson-Mehl- Avrami-Kohnogorov model: A brief review. Nuovo Cim. D **20**(7-8), 1171–1182 (1998). doi:10.1007/BF03185527
54. S. Jun, H. Zhang, J. Bechhoefer, Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model. Phys. Rev. E **71**(1), 011908 (2005). doi:10.1103/PhysRevE.71.011908
55. A. Goldar, M.-C. Marsolier-Kergoat, O. Hyrien, Universal temporal profile of replication origin activation in eukaryotes. PLoS One **4**(6), e5899 (2009). doi:10.1371/journal.pone.0005899
56. H. Zhang, J. Bechhoefer, Reconstructing DNA replication kinetics from small DNA fragments. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **73**(5 Pt 1), 051903 (2006)
57. A. Goldar et al., A dynamic stochastic model for DNA replication initiation in early embryos. PLoS One **3**(8), e2919 (2008). doi:10.1371/journal.pone.0002919

58. M.G. Gauthier, P. Norio, J. Bechhoefer, Modeling inhomogeneous DNA replication kinetics. PLoS One **7**(3), e32053 (2012). doi:10.1371/journal.pone.0032053
59. M. Oehlmann, A.J. Score, J.J. Blow, The role of Cdc6 in ensuring complete genome licensing and S phase checkpoint activation. J. Cell Biol. **165**(2), 181–190 (2004). doi:10.1083/jcb.200311044
60. S.C.-H. Yang, N. Rhind, J. Bechhoefer, Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Mol. Syst. Biol. **6**, 404 (2010). doi:10.1038/msb.2010.61
61. T.W. Spiesser, E. Klipp, M. Barberis, A model for the spatiotemporal organization of DNA replication in Saccharomyces cerevisiae. Mol. Genet. Genomics **282**(1), 25–35 (2009). doi:10.1007/s00438-009-0443-9
62. M. Barberis, T.W. Spiesser, E. Klipp, Replication origins and timing of temporal replication in budding yeast: how to solve the conundrum? Curr. Genomics **11**(3), 199–211 (2010). doi:10.2174/138920210791110942
63. A.P.S. de Moura, R. Retkute, M. Hawkins, C.A. Nieduszynski, Mathematical modelling of whole chromosome replication. Nucleic Acids Res. **38**(17), 5623–5633 (2010). doi:10.1093/nar/gkq343
64. R. Retkute, C. Nieduszynski, A. de Moura, Dynamics of DNA replication in yeast. Phys. Rev. Lett. **107**(6), 068103 (2011). doi:10.1103/PhysRevLett.107.068103
65. R. Retkute, C.A. Nieduszynski, A. de Moura, Mathematical modeling of genome replication. Phys. Rev. E **86**(3), 031916 (2012). doi:10.1103/PhysRevE.86.031916
66. A. Brümmer, C. Salazar, V. Zinzalla, L. Alberghina, T. Höfer, Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. PLoS Comput. Biol. **6**(5), e1000783 (2010). doi:10.1371/journal.pcbi.1000783
67. K. Koutroumpas, J. Lygeros, Modeling and analysis of DNA replication. Automatica **47**(6), 1156–1164 (2011). doi:10.1016/j.automatica.2011.02.007
68. N. A. Campbell, J. B. Reece, and L. G. Mitchell. Biology. Benjamin/Cummins, 1999.
69. B.I. Vaidyanathan Shantha, M. Kenward, G. Arya, Hierarchies in Eukaryotic Genome Organization: Insights from Polymer Theory and Simulations. BMC Biophys. **4**(1), 8 (2011). doi:10.1186/2046-1682-4-8
70. Z. Duan et al., A three-dimensional model of the yeast genome. Nature **465**(7296), 363–367 (2010). doi:10.1038/nature08973
71. P.R. Cook, The nucleoskeleton and the topology of replication. Cell **66**(4), 627–635 (1991)
72. I.I. Cisse et al., Real-time dynamics of RNA polymerase II clustering in live human cells. Sci. **341**(6146), 664–667 (2013). doi:10.1126/science.1239053
73. N. Saner et al., Stochastic association of neighboring replicons creates replication factories in budding yeast. J. Cell Biol. **202**(7), 1001–1012 (2013). doi:10.1083/jcb.201306143
74. A. Falaschi, Eukaryotic DNA replication: a model for a fixed double replisome. Trends Genet. **16**(2), 88–92 (2000). doi:10.1016/S0168-9525(99)01917-4
75. E. Kitamura, J.J. Blow, T.U. Tanaka, Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. Cell **125**(7), 1297–1308 (2006). doi:10.1016/j.cell.2006.04.041
76. A. Ligasová, I. Raska, K. Koberna, Organization of human replicon: singles or zipping couples? J. Struct. Biol. **165**(3), 204–213 (2009). doi:10.1016/j.jsb.2008.11.004
77. R. Berezney, D.D. Dubey, J.A. Huberman, Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma **108**(8), 471–484 (2000)
78. P.J. Gillespie, J.J. Blow, Clusters, factories and domains: The complex structure of S phase comes into focus. Cell cycle **9**(16), 3218–3226 (2010). doi:10.4161/cc.9.16.12644
79. T. Natsume, T.U. Tanaka, Spatial regulation and organization of DNA replication within the nucleus. Chromosome Res. **18**(1), 7–17 (2010). doi:10.1007/s10577-009-9088-0
80. D.S. Dimitrova, D.M. Gilbert, Temporally coordinated assembly and disassembly of replication factories in the absence of DNA synthesis. Nat. Cell Biol. **2**(10), 686–694 (2000). doi:10.1038/35036309

81. X.Q. Ge, J.J. Blow, Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories. J. Cell Biol. **191**(7), 1285–1297 (2010). doi:10.1083/jcb.201007074
82. A.M. Thomson, P.J. Gillespie, J.J. Blow, Replication factory activation can be decoupled from the replication timing program by modulating cdk levels. J. Cell Biol. **188**(2), 209–221 (2010). doi:10.1083/jcb.200911037
83. D. Marenduzzo, C. Micheletti, P.R. Cook, Entropy-driven genome organization. Biophys. J. **90**(10), 3712–3721 (2006). doi:10.1529/biophysj.105.077685
84. D. Marenduzzo, K. Finan, P.R. Cook, The depletion attraction: an underappreciated force driving cellular organization. J. Cell Biol. **175**(5), 681–686 (2006). doi:10.1083/jcb.200609066
85. D. Marenduzzo, I. Faro-Trindade, P.R. Cook, What are the molecular ties that maintain genomic loops? Trends Genet. **23**(3), 126–133 (2007). doi:10.1016/j.tig.2007.01.007

Chapter 2
Optimal Origin Placement for Minimal Replication Time

Eukaryotic genomes vary in their size and are much larger than their bacterial counterpart, e.g. that of *Saccharomyces cerevisiae* is $\sim 10^7$ bp long, those of *Xenopus laevis* or humans are $\sim 10^9$ bp in length, whereas *Escherichia coli* is $\sim 10^6$ bp. Bacteria also only have one single origin locus of replication from which they start replication, with each fork propagating at ~ 4 kbp/min [1, 2]; this allows for replication completion in under 40 min. Eukaryotic replication forks however exhibit a much slower characteristic speed, and experimental data shows that the speed of synthesis is ~ 1.5 kbp/(min·fork) [3, 4] in *Saccharomyces cerevisiae* and at ~ 0.6 kbp/(min·fork) in early *Xenopus laevis* frog embryos [5]. Let us then consider the time required for *Saccharomyces cerevisiae* DNA replication here if there were only one single origin of replication: it would take *Saccharomyces cerevisiae* almost three days to complete its genome replication.[1] In a laboratory environment yeast completes replication of its entire genome in less than about 30 min [6, 7], more than 100 times faster than what we calculated—it is clearly not the case that there is only one of replication.

The time until replication completion is accelerated by partitioning the chromosome into smaller replication domains; each of these requires an origin of replication that has formed at an origin locus. Origin loci therefore need to be placed in a manner such that replication time is minimal, i.e. replication completes by the end of S-phase. An initial guess is to space origin loci at regular intervals across a chromosome (Fig. 2.1a), if we assume origins always become licensed and activate at the same time. Such a scenario is the optimal case to result in quickest replication as compared to having the same number of origins sparsely spaced but instead groups (Fig. 2.1). Within a group only one origin is able to become active which then means that replication forks must travel farther prolonging the overall replication process (Fig. 2.1b). Therefore grouping seems to be a waste of origin resources. However we do show in this chapter that grouping is necessary to achieve minimum replication time. This is if there is uncertainty for a locus to become licensed. To compensate for not activated origins, replication forks need to travel farther than in the ideal

[1] The replication time of the yeast genome for the case of replication starting from one single origin of replication is $1.2 \cdot 10^7$ bp/$(2 \cdot 1.5 \cdot 10^3$ bp/min$) = 2.9$ days.

J. Karschau, *Mathematical Modelling of Chromosome Replication and Replicative Stress*,
Springer Theses, DOI 10.1007/978-3-319-08861-7_2

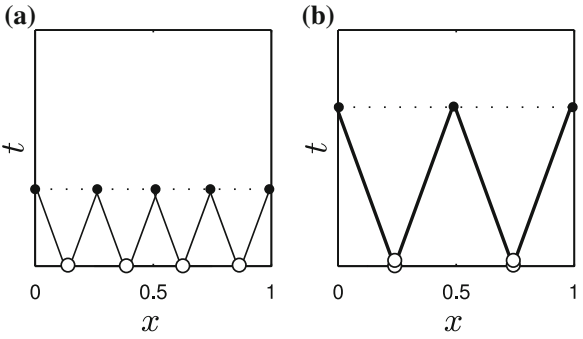

Fig. 2.1 Space–time diagram of four origins of replication (*hollow circles*). Schematic representation of origin loci distributed along the x-coordinate of a unit-sized chromosome. They have all been licensed and activate at the same time $t = 0$ min. The replication fork movement (*black line*) along the x coordinate at a given point in time is shown, and forks terminate when they coalesce or arrive at an end of the chromosome (*black filled circles*). **a** The resulting replication time is minimal if all four origins are regularly spaced. **b** If origins are grouped (shown on top of each other) only one origin of a group is able to activate. The forks must travel a longer distance at the same speeds as in (**a**); the replication time is hence longer

scenario (Fig. 2.1a). It becomes a balancing act of either spreading out origins but risking failure and longer fork travelling times, or grouping origins to compensate for the likelihood of failure and initially have longer gaps between groups. Using mathematical modelling we show that there exist certain regimes between grouped and separated origin loci positions depending on the likelihood of activation.

We relate our modelling to budding yeast *Saccharomyces cerevisiae*, which has origin loci at specific genomic positions on a chromosome—some origins in groups and some separated. For the *Saccharomyces cerevisiae* origin distribution we investigate through our model what the optimal origin distribution must be, and find that grouping of origin loci is present within *Saccharomyces cerevisiae* origin distribution to minimise replication time. This is done through an evolutionary model which searches for loci positions to give minimum replication time, and our simulations results of optimal origin positions compare well to the experimental origin distribution. We also extend our model of specific genomic positions to apply it to the case of a circular chromosome. Finally, we also introduce uncertainty in origin activation time. An origin might never have the chance to activate if it has a high chance of activating later than other origins so that it always becomes replicated by forks that originated elsewhere. We show that in such a scenario origin grouping is also a means to minimise replication time. We use the example of *Xenopus laevis* where origins appear to take random positions. In experiments, groups of origins however appear to be regularly spaced [8] which we show gives indeed minimum replication time in our model.

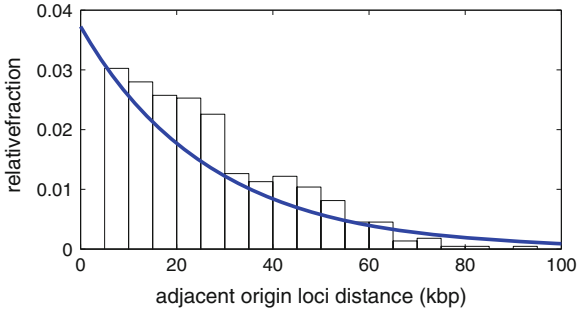

Fig. 2.2 Histogram of *Saccharomyces cerevisiae* inter-origin distances. The separation from one origin to its nearest neighbour is determined and then binned at intervals of 5 kbp (*black bars*). The origin position data was kindly provided by Hawkins et al. [9]. The mean of this data is 26 kbp, which is used to plot an exponential distribution with the same mean value (*blue solid line*)

2.1 Properties of Origins of Replication in *Saccharomyces cerevisiae*

If the origin loci in *Saccharomyces cerevisiae* would take their position within the genome randomly, then their nearest neighbour distances should be exponentially distributed. A histogram of inter-origin loci distances *Saccharomyces cerevisiae* however shows that this is not case. We show this in Fig. 2.2 where we plot a histogram of a recent study by Hawkins et al. [9]. The mean distance of the experimental data is 26 kbp which does not fit an exponential distribution with the same mean value. Also the inspection of a map for loci on the *Saccharomyces cerevisiae* genome reveals that there are groups of two or three very closely spaced origin loci which are prominent in most chromosomes [10]. We show such a map of *Saccharomyces cerevisiae* origin loci in Fig. 2.3 from the origin location data that was used in the study by Hawkins et al. [9]. Furthermore a similar map of origin loci of the fission yeast *Schizosaccharomyces pombe* gives a similar predominant grouping behaviour of origin loci (Fig. 2.4). It is to note that *Schizosaccharomyces pombe* has fewer but longer chromosomes than *Saccharomyces cerevisiae* which still require a large cohort of possible origin sites that have to be spaced with minimal gaps between to allow replication within the time allowed by the cell cycle. The data was taken from the oriDB database [10], and origin loci are shown for those classified as 'confirmed' or as 'likely'.

Previous theoretical works on *Saccharomyces cerevisiae* have used the experimentally determined loci as given parameters, without attempting to understand why the origins are located where they are [11–14]. Here, we will first show an analysis of *Saccharomyces cerevisiae* origin data addressing this, and then use mathematical modelling to explain the origin loci distribution for a specific chromosome.

As discussed in Sect. 1.4, DNA replication is divided into two distinct phases; the licensing phase and the synthesis phase (S-phase). Origins in budding yeast carry a

Fig. 2.3 Map of *Saccharomyces cerevisiae* origin positions. The location of origins (*red bars*) is shown along each individual chromosome as numbered (*blue horizontal line*). For reference, the length of the smallest chromosome, chromosome 1, is 230 kbp. The origin position data was kindly provided by Hawkins et al. [9]

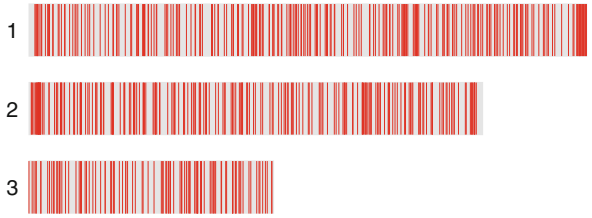

Fig. 2.4 Map of *Schizosaccharomyces pombe* origin positions. The location of origins (*red bars*) is shown along each individual chromosome as numbered (*blue horizontal line*). For reference, the length of the smallest chromosome, chromosome 3, is 2450 kbp. The origin position data was taken from the oriDB data base [10], and only those classified as either 'confirmed' or 'likely' have been considered here

certain sequence motif which allows ORC to specifically bind to a target location during licensing. This means that *Saccharomyces cerevisiae* proteins take fixed origin positions along chromosomes, and we term these positions *origin loci* to distinguish them later from licensed positions to which we refer to as *origins*. Although origin loci are at specific sites on the *Saccharomyces cerevisiae* genome this does not mean that every origin locus is going to become an active origin during each and every round of the cell cycle; i.e. not all origin loci become licensed every time. This is

because there are stochastic factors involved that hinder ORC from finding its DNA binding motif; also ever once licensed an origin might not become activated during S-phase.

For the analysis, we assign to every origin locus certain (simplifying) properties. The first one is what we term *competence*, and it describes the likelihood of an origin locus to actually become licensed which will give it the ability to become activated. It is a value between zero and one and for example, a 50 % competent origin becomes on average licensed in every other round of the cell cycle, a 25 % competent one is licensed in every fourth—the larger the value, the higher the likelihood of licensing. The second property defines the time when a licensed origin activates; it is the origin activation time distribution assuming that the origin is not passively replicated. We characterise this probability density to activate in S-phase by a distribution, which in case of a Gaussian distribution has mean time of activation μ and standard deviation σ. Previous analysis of the origin activation time distribution suggests a bell-shape-like function [15], and thus a Gaussian distribution is a good first approximation. In a previous mathematical model of DNA replication which incorporates these origin properties Hawkins et al. [9] determined parameters of the entire origin population in budding yeast using a model developed by Retkute et al. [16]. They fitted their model to experimental replication timing curves of *Saccharomyces cerevisiae* to determine the competence, mean and spread of an origin activation time distribution. For their study, Hawkins et al. and Retkute et al. chose a Hill-type function to represent their origin activation time distribution which depends on two parameters t_{12} and t_w which are similar to mean and standard deviation of a Gaussian distribution. Their choice of a variant function manifests in the possibility of having origin activation prior to the begin of S-phase, which is biologically unphysical. A Hill-type function however gives origin activation times well defined between zero and later times although any other choice of function can display replication time data equally well (*personal communication with Renata Retkute*). Hawkins et al. study uncovers valuable information on the spatial distribution of origins along chromosomes, and the parameters of origin loci.

We here analyse their data which we will discuss for the remainder of this section. Of particular interest is whether specific genomic regions for origin loci are random or whether their spacing depends on the competence value of their neighbours. We calculate the sum of the competence values for adjacent origin pairs, and look at a plot of this against their genomic separation. Figure 2.5 shows that this separates groups with a low value from those with a high value. We expect that most points would be roughly near the diagonal, and the two off-diagonal corners to be empty.

Plotting the distribution of origin data shows a somewhat linear trend between the competence of neighbouring loci pairs and their separation (Fig. 2.6a). We emphasise on the left-hand tail of the distribution which shows that low competent origins *per se* are closely located for a certain parameter regime up to about 2/3. Highly competent pairs tend to be further separated from their nearest neighbour whereas low competent pairs have a tendency to be very close to each other; although there are also close nearby pairs for the case of highly competent origins. This tendency is also reflected in the correlation coefficient of 0.331 (p-value $\sim 10^{-13}$) for this data. As we show

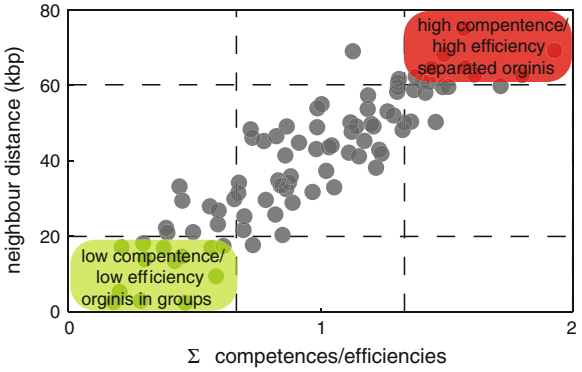

Fig. 2.5 Scheme for plotting origin neighbour distances. We plot the distance of adjacent origins of certain group size versus the sum of this competence or efficiency value of such a group. We expect that group with low competent/efficiency values have low distance to their neighbours and will be found in the *bottom left corner* (*green region*). As for highly competent/efficient origins, we expect these to be far away from their nearest neighbour and will be shown in the *top right hand corner* (*red region*)

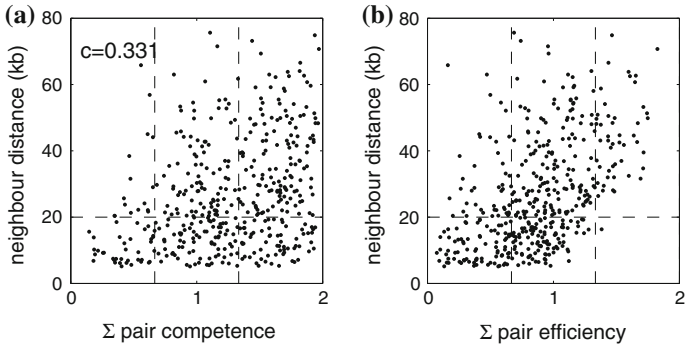

Fig. 2.6 Competence, efficiency and pair-wise neighbour distance in *Saccharomyces cerevisiae*. **a** Pairwise origin nearest-neighbour distance is plotted against their pairwise sum $\left(\sum\right)$ of competences are. Highly competent pairs are found on the *right-hand side* of the *vertical line* at 4/3, and low competent ones at the *left-hand side* of the *vertical line* at 2/3. **b** Pairwise origin nearest-neighbour distances plotted here against the sum of their efficiency, i.e. the probability to become activated per round of the cell cycle

in Fig. 2.6b, a stronger trend for separation of highly competent origins holds for our analysis of efficiency—the probability of an origin being competent and also becoming activated in a particular round of the cell cycle. We emphasise that for the case of efficiency that there are no close and highly efficient origin pairs (bottom right corner) Fig. 2.6b.

This trend also persists if one considers sets of three nearest neighbouring origins. In Fig. 2.7a, b we compare the sum of competences with the maximal or minimal distance between direct origin neighbours out of a group of three adjacent origins.

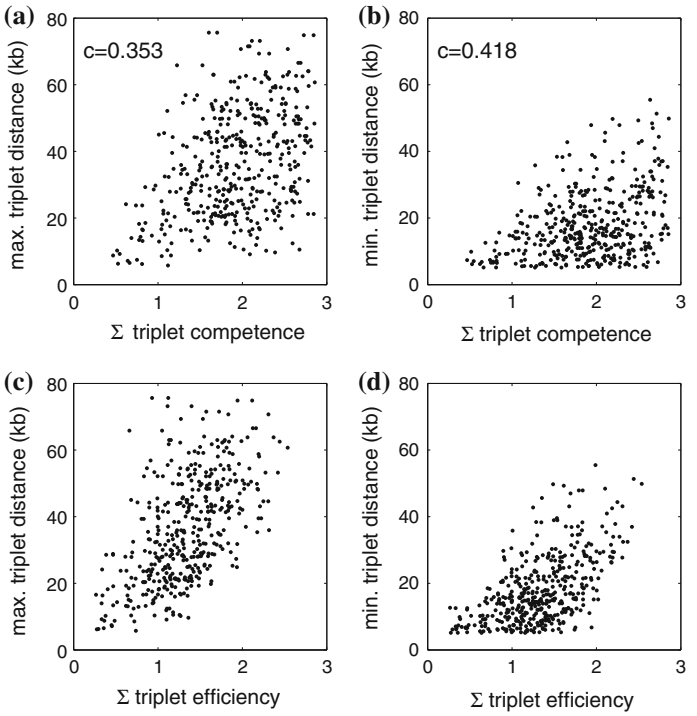

Fig. 2.7 Sets of three adjacent origins (*triplets*) are taken and either their maximum (**a**, **c**) or minimum (**b**, **d**) distance from one another within their triplet are plotted against either the $(\sum)$ sum of their competences (**a**, **b**) or the sum of their efficiencies (**c**, **d**)

There is a striking difference in maximal, or minimal separation when considering low competent and highly competent groups of three, consistent with data for a group of two. The linear correlation between origin separation and their ability to eventually activate is even clearer when we also consider efficiency (Fig. 2.7c, d). Figure 2.7c also shows that as the efficiency of a group of three origins increases at least one origin becomes further and further separated from the other two origins. This also applies to the minimum distance of a group (Fig. 2.7d). The data gives reason to speculate that origin positions have thus been chosen preferably to compensate for origins that have little likelihood to activate by others in their surroundings.

So this data in Figs. 2.6 and 2.7 show that the proximity of origin loci correlates with their competence. These properties are therefore not independent. The remaining question is however under what conditions do origins group and whether the positions of origin loci have been favourably selected to minimise the average replication time.

2.2 A Mathematical Model for Optimal Origin Positions

The data showed that the separation of origin loci correlates with origin competence and efficiency of their neighbour(s). Yet it is unclear whether those position found in experiments are actually optimal loci positions—i.e. those giving the minimum replication time for an average of a cell population. To re-phrase the question, we can ask whether evolution has driven origin loci to their positions on the chromosome where they are found today.

2.2.1 A Simplified Two Origin Model

In a first attempt to establish a many origin model we consider the case of having only two origin loci that are positioned on a stretch of DNA. We also simplify further that origins only have a probability to activate (or fail). In other words, we only consider competence p_i for the ith origin locus. The DNA is modelled as a one-dimensional line of unit length, and we denote competences of two loci p_1 and p_2. We initially make the assumption that origins activate at a well-defined time, $t = 0$. All replication forks travel at the same unit speed across the DNA. Specifically, we consider the geometry depicted in Fig. 2.8a where d_1 (d_2) is the distance from the left (right) end of the chromosome to the left (right) most locus. If both loci fail to be licensed we postulate that replication will eventually take place anyway, with a replication time T_0—for example, we can imagine that this stretch of DNA will be replicated by forks originating from origins outside of the region we are considering. It will be clear shortly that our results do not depend on T_0; this is just a mathematical device to prevent us dealing with infinite replication times.

If only one of the loci fails to become licensed, the replication time depends on the time it takes for the fork to reach the furthest end of the segment, so $T_{d_1} = 1 - d_1$ for locus 1 and $T_{d_2} = 1 - d_2$ for locus 2. If both loci have been licensed the replication time $T_{d_1,d_2} = \max\{d_1, d_2, (1 - d_1 - d_2)/2\}$ is defined by the longest time for a fork to reach the end of the segment or for two forks to collide. Figure 2.8b illustrates that the replication time of an asymmetric placement of loci is never less than a corresponding symmetric configuration (that is, with $d_1 = d_2$). Therefore we consider only symmetrical locus placements, and use $d_1 = d_2 = d$ with $0 \leq d \leq 1/2$. The average replication time is then given by

$$T_{\text{rep}}(d) = (1 - p_1)(1 - p_2)T_0 + (p_1 + p_2 - 2p_1p_2)(1 - d) + p_1p_2 \max\{d, (1 - 2d)/2\}. \tag{2.1}$$

This is a piecewise-linear function with discontinuity in its first derivative at $d = 1/4$, and with domain $[0\ 1/2]$. Hence, T_{rep} can only have a minimum at $d = 0$, $d = 1/2$, or at $1/4$. Placing loci at the end of a segment ($d = 0$) is obviously not a minimum of T_{rep}. Placing both loci in the middle ($d = 1/2$) we assume that

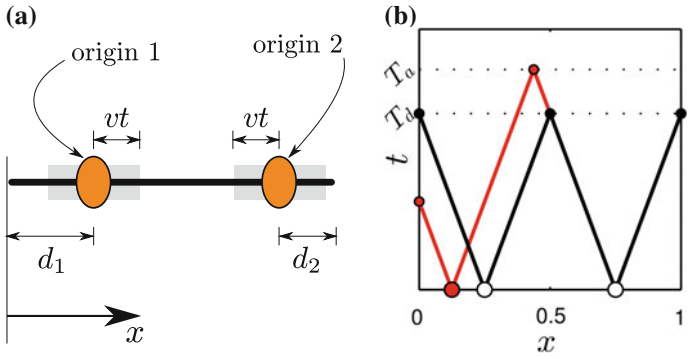

Fig. 2.8 Two origin model of DNA replication. **a** Coordinate system for origin loci with d_1, d_2 being the distance from the *left-* or *right-end* of the chromosome, respectively. x is the position coordinate along the chromosome. Replication forks travel at a speed v away from the origins. The *grey regions* show the replicated DNA at time t. **b** Space–time diagram of replication fork movement for the case of both origins starting replication at the same time $t = 0$ min. Forks move from each origin position and replication is completed once a fork reached the end of the chromosome or the last pair of forks coalesced. A symmetric placement of origins gives minimal replication time whereas an asymmetric one requires more time, i.e. $T_d < T_a$

both can activate at the same time, however the replication time is then 1/2 for the last term in Eq. (2.1) as well as for the second term when only one activates. The replication times for $d = 1/4$ and $1/2$ are

$$T_{\text{rep}}(d = 1/2) = (1 - p_1)(1 - p_2)T_0 + (p_1 + p_2 - p_1 p_2)/2$$

and

$$T_{\text{rep}}(d = 1/4) = (1 - p_1)(1 - p_2)T_0 + (3p_1 + 3p_2 - 5p_1 p_2)/4.$$

We conclude that the two loci group together ($d = 1/2$) to achieve minimum replication time if $T_{\text{rep}}(d = 1/2) < T_{\text{rep}}(d = 1/4)$, which leads to the condition

$$p_2 < \frac{p_1}{3p_1 - 1}. \tag{2.2}$$

Notice here that T_0 drops out. The inequality Eq. (2.2) defines two regions on the p_1–p_2 plane, corresponding to grouped or isolated loci being optimum. This is shown in Fig. 2.9a, where this analytical result is confirmed by stochastic simulations. These simulations are done employing a minimisation algorithm (using genetic algorithms [17]) which searches for the minimal replication time. The principal ingredients to the algorithms are as follows. First, origin loci are selected. Each origin locus is checked whether it will activate given its competence value, i.e. checking a random number against this probability. Finally the replication time is calculated, and this procedure repeats for several times to establish the average replication time. The positions of the origin loci are then changed, and the average replication time is

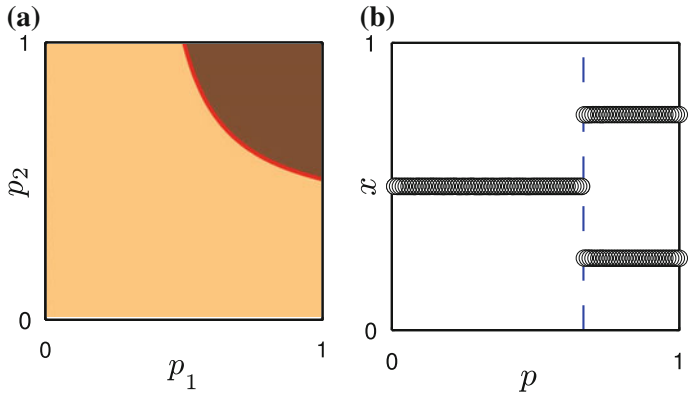

Fig. 2.9 Optimal locations in a two origin loci model. **a** Simulation results, showing optimal loci to achieve minimal T_{rep} for 2 loci with different competences, are shown for p_1–p_2 combinations on a lattice grid. Colour indicates $d_1, d_2 = 1/2$ (*beige*) or $d_1, d_2 = 1/4$ (*brown*). The two regimes are separated by a coexistence line matched by the condition Eq. (2.2) in *red*. **b** Optimal position of 2 identical loci with respect to their competence p to minimize the replication time T_{rep} (*circles*) and $p = 2/3$ (*dashed line*)

calculated for this new configuration. It is then compared to other randomly selected loci positions to whether or not it results in minimum replication time. In Fig. 2.9a, the region above the curve corresponds to competences for which T_{rep} is minimized by loci being apart ($d = 1/4$) and below the curve for organising these in a group ($d = 1/2$). In general, if one of the loci has low competence grouping gives the minimum replication time. In fact, it can be shown that if one of the loci has a competence lower than 50 %, grouping is the optimal situation regardless of the competence of the other—even if the other is close to 100 % competent. This becomes clear with if one imagines that once a replication fork from an origin has to cover a distance more than 1/2, such a grouped configuration becomes favourable. Figure 2.10 shows how the individual replication time (T_{rep}) terms change depending on how many origins become activated.

For the case of equal competences, $p_1 = p_2 = p$, the grouped configuration is optimal if $p < 2/3$. We ran a numerical optimization algorithm again to find the loci corresponding to the least replication time for a range of p; these results are shown in Fig. 2.9b. The same transition also takes place for non-identical values of p_1 and p_2—whenever one crosses from the dark to the beige region of Fig. 2.9a.

The above results may seem at first quite counter-intuitive; one might expect that the configuration with the least replication time would correspond to isolated loci ($d = 1/4$). However, if the origins have a significant chance of failing to activate, this configuration would mean that often one side of the chromosome would have to wait for a fork which originated at the origin on the other site to replicate it, therefore increasing T_{rep}. So in the case of low competences, it becomes advantageous to have

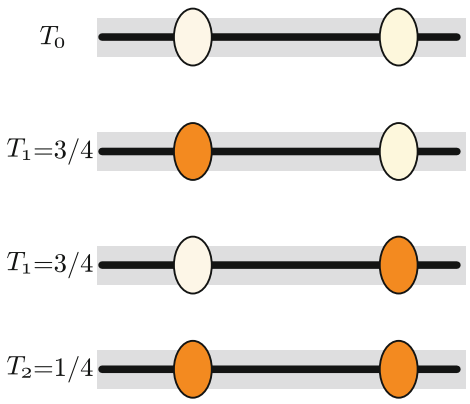

Fig. 2.10 The time it takes to replicate a given piece of DNA T_{rep} depends on the number of origins that activate (*orange filled ovals*) or not activating. This contributes to the different terms as for instance in Eq. (2.1)

both loci centered, which is near any point in the chromosome. This explains the condition for grouping if $p < 2/3$.

2.2.2 Many Origin Loci

In reality eukaryotic chromosomes have more than two loci [18], so next we investigate the case of a chromosome on which there are many loci and examine the conditions under which it becomes favourable to have isolated origin loci compared to groups. In this analysis we will assume for simplicity that the loci all have identical competence.

We consider a group of loci as one single locus with an effective competence p_{eff}. For a group consisting of m loci p_{eff} is the competence that at least one locus will be licensed there, and is given by

$$p_{\text{eff}} = 1 - (1 - p)^m. \tag{2.3}$$

We assume that one large group of n identical loci breaks up into two groups of equal size, each consisting of $n/2$ loci. A locus organized with others in a group of size $m = n/2$ rather than with n loci will give minimum T_{rep}, as long as the locus' competence is larger than its critical probability p_c, given by $p_{\text{eff}} = 2/3$, which yields

$$p_c = 1 - 1/\sqrt[n]{9}. \tag{2.4}$$

Figure 2.11 confirms our analytical result showing the value of p_c for increasing group sizes in our simulations. These results clearly show that large groups of many highly competent loci are unfavorable, but that groups tend to form for

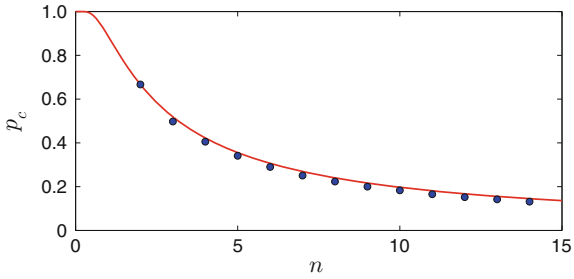

Fig. 2.11 Many origins with variable competence. **a** Probability at which groups separate p_c versus loci/group n. Shown are simulations (*circles*) and analytical prediction for $p_c = 1 - 1/\sqrt[n]{9}$ (*line*)

low-competence loci. Our formula is also a good approximation to predict the probability at which a transition occurs for an odd number of origins in a group.

So we would expect for example a group of four origins, to break up at $p_c \approx 0.42$. Our simulations do show this to be the case (Fig. 2.12a). However the groups of four origins does not break up symmetrically into two groups of two origins, but rather into three groups. As p_{eff} increases through p, we first see two origins move out to positions $x = 1/4$ and $x = 3/4$ leaving two at $x = 1/2$. Only at a slightly larger p do we get two clusters of two. This is due to the fact that we assumed the simple case of two origins can be directly applied to the more complicated case of more origins. Figure 2.12b shows this as well where we plot the replication time for the individual configurations. This also illustrates that at first only two loci break out of the four origin group which is the crossover of the black with the blue line in Fig. 2.12b; before the blue line crosses with the red one.

2.2.3 Evolutionary Pressure Drives Yeast Origin Loci to Optimal Positions

Our hypothesis from this modelling is that selective pressure has influenced the position of origin loci through the minimization of the replication time. The theoretical result—low competence loci group, high competence loci are spread out—is also in line with our data analysis presented in Sect. 2.1. The competence data used for the analysis there however resulted in silico by model fitting to experimental data. So it required a proxy that could be potentially biased, and as a further example we now use a *Saccharomyces cerevisiae* chromosome for which origin positions and competence values are experimentally known. We then apply a search algorithm for it to find the optimal loci positions to achieve minimal replication time. This will show that in silico optimisation matches a known set of locations.

We show in Fig. 2.14 locus competence and location data for *Saccharomyces cerevisiae* chromosome VI, which has been studied extensively [12, 19]. Competences

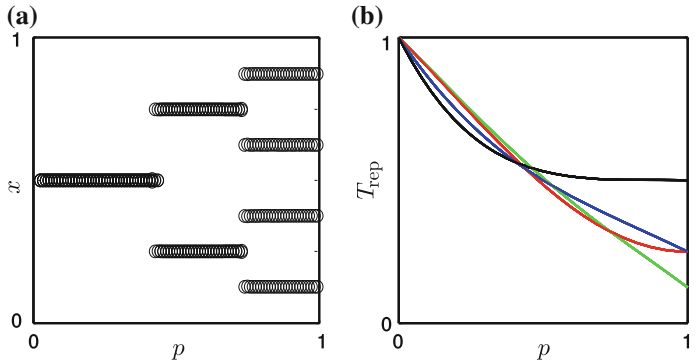

Fig. 2.12 Four origins with variable competence. **a** Simulation results for positions x of four identical origin loci with probability p. As p increases spreading loci along the chromosome of unit length results in minimal replication time. **b** The average replication time T_{rep} of arranging four origin loci positions with the same probability [corresponding to configuration shown in (**a**)]. The different colours of the curves correspond to: all 4 loci clustered at the middle position (*black*); 2 loci at $x = 1/2$, and 2 loci at either $x = 1/4$ or $x = 3/4$ (*blue*); groups of 2 individual loci at either $x = 1/4$ or $x = 3/4$ (*red*); individual loci $x = 0.2, 0.4, 0.6, 0.8$ (*green*)

cannot be measured for all loci (in white), because either they are too close to the end of the chromosome or to an adjacent locus. We performed a search for the optimal position for the loci in the region with known competences using a genetic algorithm [17]. The algorithm mimicks an evolutionary process by first selecting sets of random origin locations for a parent generation of 50 individuals. The parent generation is then tested for its individual set location to give minimum replication time. The most optimal of the minimal sets are selected for the next round of iteration. They then become reshuffled amongst each other to yield a new collection of origin loci positions on this chromosome. The sets of locations are in tournament. A pair of randomly selected individuals is set to tournament, meaning the one with lower replication time succeeds. Ten new sets of location are drawn randomly and replace the ten worst (maximal replication time) location sets out of the tournament. The remaining sets produce children. They result from crossing over 85 % of the parents which are selected randomly, i.e. 15 % of the best part of a population remains unchanged to the next generation. The selection of new locations from parents results from crossover of the two parental sets of locations, i.e. either picking location 1 from parent 1 or parent 2 and so forth. They produce two children sets so that each child inherits a particular location from a particular parent to 50 %; termed crossover. Note that the number of origins always stays fixed. We then determine the replication time for the individual position sets just as before. The genetic algorithm was run with a population of 100 chromosomes of the parent generation and optimised over 2,000 iterations meaning the genomes evolve over 2,000 generations. The procedure repeats for 18,000 times with different seeds of the random number generator. Figure 2.13 summarises the algorithm detailed above. The details of parameters here lead to a local minimum set of origin location in a reasonable amount of computation time.

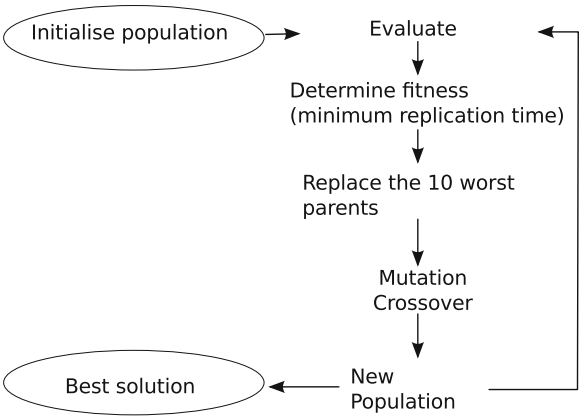

Fig. 2.13 Genetic algorithm. Summary of the steps of the genetic algorithm to determine the set of origin location to give minimum replication time

Although an appropriate choice of biological evolutionary-like parameters can be used to mimic evolution to occur over millions of years however this requires a substantial amount of extra computation time as most of the runs end in a local minimum similar to the one we find below (Figs. 2.14 and 2.15).

We remark that in our example of chromosome VI is an identifiability problem as all strong loci have $p \sim 90\,\%$, and we therefore constrained the ordering during the optimisation. Although in this result we do not consider inter-origin variations in the origin activation time, the predicted locus distribution from these simulations bears a good resemblance to the actual spacing with a score of $F = 0.11$[2]; in particular we recover the group in the middle, in which an origin locus with 58 % competence is placed next to one with 88 % competence. Even multiple repeats of the optimisation algorithm produce minimum replication time solutions which have on average $F = 0.12$ (Fig. 2.15). This indicates that evolution has generated a near optimal solution for the proper placement of origin loci over many generations. Our study here shows a possible means to minimise replication time by choosing optimal origin loci positions. Mutations such as the translocation of genetic sequences occur frequently in unicellular, eukaryotic organisms such as yeast [20]. The rearrangement of genetic sequences—origin loci in our model—over many generations is therefore also a legitimate device in an evolutionary context to achieve minimal replication timing.

[2] $F = \frac{1}{9}\sum_{i=1}^{n=9} d_i^o/d_i^r$ is a measure of the difference between the gap distribution of the optimised and random cases. A gap is defined as the separation between the i^{th} experimental locus position p_i^e and that of the optimization p_i^o: $d_i^o = |p_i^e - p_i^o|$. d_i^r is akin; the average separation that arises from placing a locus uniformly randomly and p_i^e. $F = 0$ means that the optimization fits the experimental loci positions perfectly; $F \sim 1$ indicates no difference to that of a random placement.

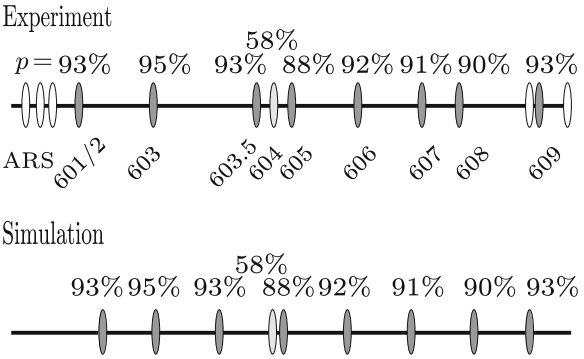

Fig. 2.14 Distribution of origin loci on yeast chromosome VI with known (*grey*) and unknown competences [12, 19]. The distribution results from our simulation in search for minimum T_{rep} (only *grey* origins considered). The group in the middle of the chromosome with a low and highly competent locus was recovered

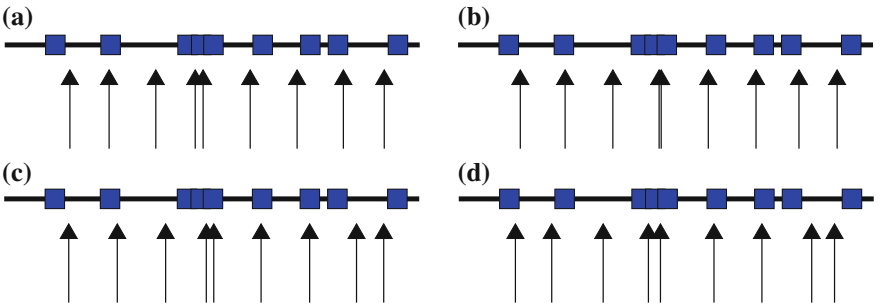

Fig. 2.15 Optimisation results for finding the minimum T_{rep} by varying the origin loci positions given their competences. The *blue boxes* are the experimental origin loci positions, the *arrows* show the positions found in simulations of individual runs. **a** Origin loci distribution that has the overall minimum replication time corresponds to Fig. 2.14. **b–d** Some distributions that give minimum replication time close to the overall minimum solution

2.2.4 Loci Competence and Circular Chromosomes

Most prokaryotes, for example the bacterium *Escherichia coli*, carry their genomic information on a single, circular chromosome. They have no compartmentalisation, meaning DNA is contained within the cytoplasm and not within a nucleus. Therefore there is no separation of licensing and origin activation as is in eukaryotes, and prokaryotes can start replication as soon as their origin locus becomes replicated. So here we can have re-replication since there is no separation of licensing from synthesis. This way they can produce concurrent copies of their DNA during exponential, unlimited growth conditions.

Their organisation of DNA replication on circular chromosome also has the advantage of only one replication fork being able to replicate its entire genome. For instance,

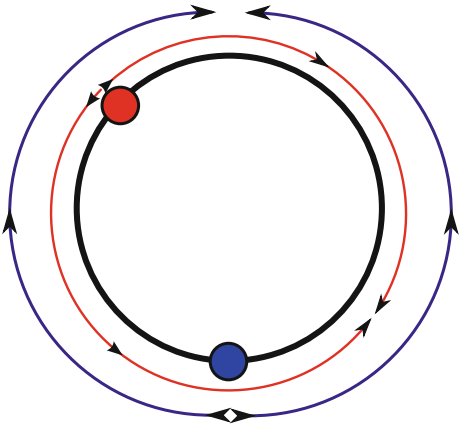

Fig. 2.16 One origin model of circular chromosome. The replication time for one origin (*blue or red*) is independent of its location due to the symmetry of a *circle*. Replication forks will always meet after travelling half circumference

the fork can start from any position on a chromosome, from which it takes the same time until to complete synthesis; it completes a full circle. This is different from the previous case of having a linear chromosome. There fork movement is more constrained because a fork cannot go around and one requires at least two forks travelling from either direction of an origin to complete DNA or have a fork starting from an edge of the chromosome. This edge effect was shown in Sect. 2.2.1 to result in preferred origin locations; only two locations that are symmetrically around the centre of a DNA segment result in the minimal replication time.

A circular chromosome also has advantage over failing origins or stalling replication forks to be easily recovered by a fork travelling towards them from elsewhere on the circle as illustrated in Fig. 2.16. So we note that all positions on a circle with a circumference we set to unit length result in the same replication time of $T_{\text{rep1}} = 1/2$ (2 forks, each replicating half of the circle); and therefore any position serves equally well to act as an origin locus. In principle, we will always observe the same T_{rep1} for a population of cells no matter where each individual cell starts its replicating from. The remaining question is whether there also exist similar origin placement conditions as we observed previously—grouped or separated; and if, so how many origins are required along with their competence value to achieve minimal replication time. We consider growth to be limiting, so that there are at maximum two copies of a chromosome and not multiple ones as during exponential growth, and again ask the question which loci positions give minimum replication time.

2.2.4.1 Two Origins

The case of two origins, shown in Fig. 2.17, results in a shorter replication time of $T_{\text{rep2}} = 1/4$, if both origins origins are maximally apart as is the case for a symmetric placement in Fig. 2.17a. An asymmetric placement however results in a replication time less optimal, depending on the maximum distance between the two origins it

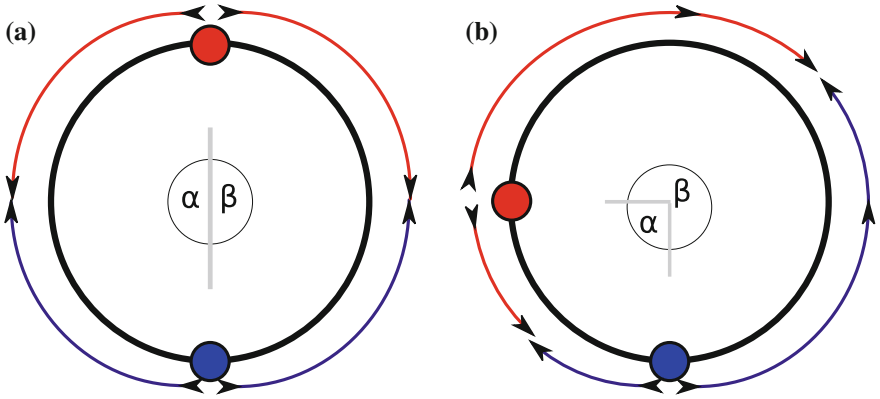

Fig. 2.17 Two-origin loci model with the angle α and β between them indicated by the *grey bar*. **a** The minimum replication time of $T_{\text{rep2}} = 1/4$ is achieved by placing the origin loci furthest apart from each other with distances clockwise and anticlockwise to the other origin being equal. **b** An asymmetric placement results in a longer replication time, because it takes longer for two forks to coalesce

will be $1/4 \leq T_{\text{rep2}} \leq 1/2$ (Fig. 2.17b). The other extreme is placing both origins on top of each other, for which we recover the same result as in the one-origin case. There exists only one optimal configuration, which is placing origins furthest apart which we show analytically. We define the angle between adjacent origins α and β. We note that the time of the replicated piece of the chromosome by two forks is defined by $T = \alpha/(2 \cdot 360)^\circ$, which then gives the mean replication time

$$\begin{aligned} T_{\text{rep2}} = &(1-p_1)(1-p_2)T_0 + p_1(1-p_2)\frac{1}{2} + p_2(1-p_1)\frac{1}{2} \\ &+ p_1 p_2 \max\left\{\frac{\alpha}{2\cdot 360^\circ}, \frac{\beta}{2\cdot 360^\circ}\right\}. \end{aligned} \tag{2.5}$$

The first term accounts for neither of the origins activating, the second and third terms account for only either origin to activate and the last term if both do. The angles are constrained by one full round around the circle $360^\circ = \alpha + \beta$ which gives $\beta = 360^\circ - \alpha$. We can only find the minimum of Eq. (2.5) at either $\alpha = 0^\circ$, $\alpha = 360^\circ$ or the discontinuity of the maximum function $\alpha = 360^\circ - \alpha$ which is for $\alpha = 180^\circ$. $\alpha = 0^\circ$ and $\alpha = 360^\circ$ mean that both origins would sit on top of each other; $\max\{\alpha, \beta\} = 360^\circ$. This only leaves the configuration shown in Fig. 2.17a with both origins maximally apart to give minimum replication time.

We now write Eq. (2.5) in terms of different competence values p_1 and p_2 and include our knowledge that the minimum replication time can only be found for either $\alpha = 180^\circ$ or $\alpha = 0^\circ$, i.e. if both origins activate $T = 1/4$ or $T = 1/2$, respectively (cf. Fig. 2.17). We set $T_0 = 1$. The average replication time of both cases is then given by

$$T^b_{\text{rep2}}(p_1, p_2) = \frac{1}{4}p_1p_2 - \frac{1}{2}p_1 - \frac{1}{2}p_2 + 1, \text{ and} \tag{2.6}$$

$$T^b_{\text{rep2}}(p_1, p_2) = \frac{1}{2}p_1p_2 - \frac{1}{2}p_1 - \frac{1}{2}p_2 + 1. \tag{2.7}$$

The minimum is found using the configuration for $\alpha = 180°$ [Eq. (2.6)], because $T^a_{\text{rep2}}(p_1, p_2) < T^b_{\text{rep2}}(p_1, p_2)$ for $p_1, p_2 \in (0, 1]$. So even for origins with different competence it is always best to be farmost apart from each other. This result differs from our analysis of a linear chromosomes in Sect. 2.2.1. We showed that there exists a sharp transition from finding origins together or apart depending on the parameter p_1 and p_2 for a linear chromosome.

2.2.4.2 Three Origin Loci Break Circular Symmetry: And Group Together

We now examine an odd number of origin loci and continue our analysis in terms of the time a fork travels. We take the example of three origins and place them as depicted in Fig. 2.18. The case which results in minimum $T_{\text{rep3}} = 1/6$ is again placing all origins maximally apart from each other (Fig. 2.18a). Maximum replication time is achieved by placing all three origin loci on top of each other, which is obviously not the preferred configuration to achieve an optimal replication time. This leaves two possible scenarios to arrange the origins. We place two of them maximally apart and the third one on top of any of the two (Fig. 2.18b), or the third origin is placed somewhere in the remaining halves (Fig. 2.18c). We note that: if all origin loci are always competent to activate ($p = 100\,\%$) then the resulting $T_{\text{rep2}} = 1/2$ which is independent of the arrangement of the third origin locus. In a more general approach, we write an expression for the average replication time

$$T_{\text{rep3}}(p) = (1-p)^3T_0 + 3p(1-p)^2T_1 + 3p^2(1-p)T_2 + p^3T_3, \tag{2.8}$$

with which we show analytically that placing origin loci maximally apart is the only optimal configuration. The four different terms in Eq. (2.8) account for the possible number of origins activating during a round of the cell cylcle. T_0 is the time resulting of all origins failing, but we note that it can be chosen arbitrarily as it will not influence our analysis. We choose $T_0 = 1$, as this is the longest time it takes for one single fork to complete replication. T_1 accounts for the time, if only one of the three origins activates is always independent of the placement of the failing origins. We know from the case of one origin locus that $T_1 = T_{\text{rep1}} = 1/2$. T_2 and T_3 both depend on the chosen configuration for the origin loci, and are defined by when the last coalescence event happens, so by the maximum distance a fork must travel.

The average replication times T^a_{rep3}, T^b_{rep3} and T^c_{rep3} for a circle of circumference $c = 1$ and origin loci at the positions shown in Fig. 2.18a–c are given by

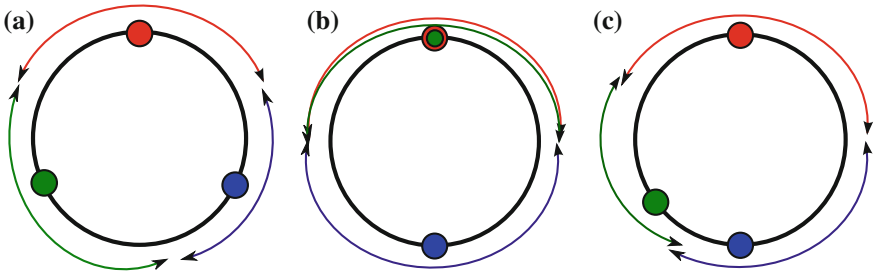

Fig. 2.18 Three origin loci on a circular chromosome. Three origin loci model with origins in either *green, blue* or *red*, and their corresponding forks shown as *lines*; origin loci 1, 2 and 3 respectively. Forks coalesce at the positions where two *arrowheads* meets. There exist three possible configurations to achieve minimum replication time. **a** All origins are spaced maximally apart. **b** Two origins are at either side along the diameter of the *circle* and the third origin at the same location of one of the two other. **c** The third origin can be placed in either half of the chromosome. However it does not contribute to the minimum replication time since it will always take longer to replicate the *right-hand side* of the *circle*

$$T^{a}_{\text{rep3}} = -1/3p^3 + p^2 - \frac{3}{2} + 1, \tag{2.9}$$

$$T^{b}_{\text{rep3}} = -1/4p^3 + p^2 - \frac{3}{2} + 1, \tag{2.10}$$

$$T^{c}_{\text{rep3}} = -1/4p^3 + p^2 - \frac{3}{2} + 1. \tag{2.11}$$

We note that $T^{a}_{\text{rep3}} < T^{b}_{\text{rep3}} = T^{c}_{\text{rep3}}$ as well as $T^{b}_{\text{rep3}} = T^{c}_{\text{rep3}}$ is for all origin sites with the same p; a group of two origin loci can either be situated at the top half or the bottom of the circle [configurations (2) and (3) in Fig. 2.19a]. We conclude that for all identical origin loci T^{a}_{rep} is the only optimal configuration, i.e. three origin loci are best placed maximally apart from each other. The cases for T^{b}_{rep3} and T^{c}_{rep3} both result in the same average replication time; the open boundary allows replication forks to travel around the circle. Those cases however are relevant for origin loci that differ in their competence as we show below.

We fix two loci with competence equal to 1, say $p_3 = p_2 = 1$ (red and blue loci respectively). Using a general expression for the average replication time [Eq. (2.8) for individual p_i values] one can show that the positioning of the third origin locus with variable competence has no contribution to the average replication. This is for as long as its competence value is below 0.5. We give the analytic expression of the average replication time for the configuration shown in Fig. 2.18a, c ($p_2 = p_3 = 1$), which we call $T^{a^*}_{\text{rep3}}$ and $T^{c^*}_{\text{rep3}}$ respectively:

$$T^{a^*}_{\text{rep3}} = 1/3 - 1/6p_1, \tag{2.12}$$

$$T^{c^*}_{\text{rep3}} = 1/4. \tag{2.13}$$

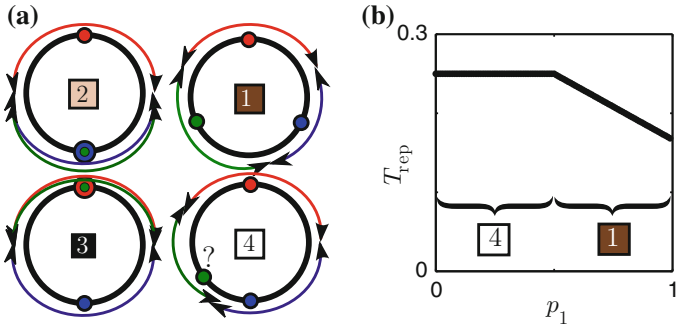

Fig. 2.19 Distribution of three loci on a circular chromosome. Origin 1, 2 and 3 have colours *green, blue* and *red*, respectively. **a** Three loci can be distributed in four different ways on a circular chromosome. In configuration (*1*) all origins are equally spread out, in configurations (*2*) and (*3*) one locus pairs with another one, and in configuration (*4*) the third locus can positioned anywhere. **b** Average replication time T_{rep} of a 3 origin system with two loci of competence 100 % and one origin having varying competence p_1. T_{rep} is independent of p_1 for $p_1 < 0.5$ [configuration (*4*) in (**a**)] and for values $p_1 > 0.5$ it contributes [configuration (*1*) in (**a**)]

We see that Eq. (2.13) is independent of p_1, the green origin, which is confirmed through stochastic simulations shown in Fig. 2.19a. There are only two possible configurations for this setting which are depicted as configuration (1) and (4) in Fig. 2.19b. Origin loci are either best placed far apart from each other, or only two origins contribute to the replication time. Minimum average replication time is achieved for the condition $T^{a^*}_{\mathrm{rep3}} < T^{c^*}_{\mathrm{rep3}}$ for $p_1 > 0.5$. Therefore a less competent origin will not influence the average replication time if combined with two highly competent origins.

Now we vary the competence of two origins, say the red origin that has $p_3 = 1$ here. We will see that there are four different configurations for this case. These are shown in Fig. 2.19a. Again using the general expression Eq. (2.8), we find that the other two green and blue origins cluster together; the red origin, origin 3, stays isolated as in Fig. 2.19a configuration (2). This is if the following condition is justified

$$p_1 < \frac{p_2}{3p_2 - 1}, \tag{2.14}$$

which corresponds to the beige region in Fig. 2.20a. The relative position of origin 1 and 2 (green and blue loci) to origin 3 (red locus of Fig. 2.19) is plotted in this figure; beige indicates locus 1 and 2 group together [configuration (2) in Fig. 2.19a], black they are 1/2 apart from each other [configuration (3) in Fig. 2.19a], brown all loci are maximally apart from each other [configuration (1) in Fig. 2.19a].

We now lower p_3 as in for example Fig. 2.20b–d where $p_3 = 0.75,\ 0.50,\ 0.25$, respectively. This makes the above mentioned four regions more visible; each corresponds to an optimal configuration. If two origin loci have same competence, the location of the weaker third origin locus can be chosen freely as it will not affect

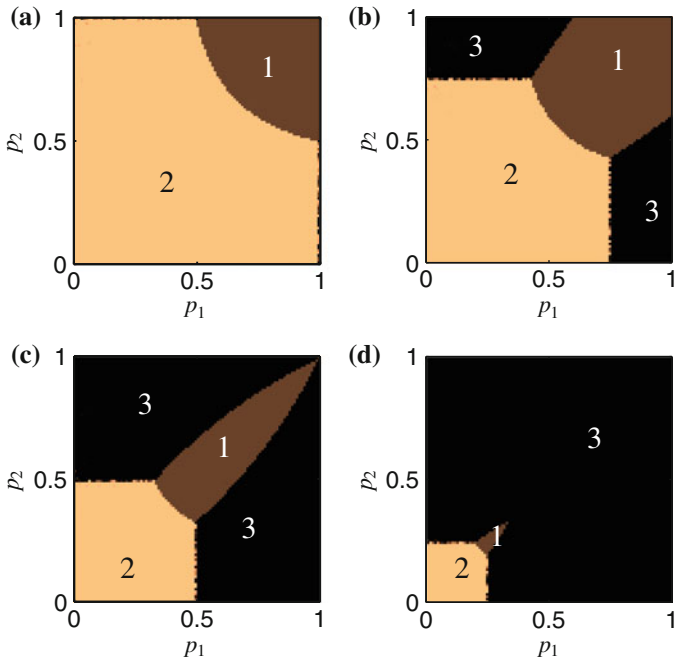

Fig. 2.20 The configurations giving minimal replication are shown for the case of three origin loci. The competence of locus 3 is fixed to either the value of $p_3 = 1.00$ in (**a**), $p_3 = 0.75$ in (**b**), $p_3 = 0.50$ in (**c**), or $p_3 = 0.25$ (**d**). Competences p_1 and p_2 of loci 1 and 2 are varied. The colour code and numbering correspond to the configurations as shown in Fig. 2.19a. The *brown* (*1*) colour indicates complete separation of all loci. *Beige* (*2*) and *black* (*3*) colours correspond to grouping of two loci, i.e. configurations (*2*) and (*3*) of Fig. 2.19a. There is a fourth regime along $p_1 = p_2 = 1.00$ in (**a**), $p_1 = p_2 = 0.75$ in (**b**), $p_1 = p_2 = 0.50$ in (**c**), and $p_1 = p_2 = 0.25$ in (**d**) where the position of the fourth locus positions can be chosen arbitrarily. As the competence p_3 decreases the number of possible configurations of (*1*), where we find maximal separation, decrease; as do those for configuration (*2*)

the result of the average replication time, [cf. Fig. 2.19a configuration (4)]. This is the case for the randomly coloured shades as one crosses from the beige to the black region at $p_2 = p_3 = 0.50$; the configurations change from (2)→(4)→(3) (Fig. 2.19a) in Fig. 2.20b. As p_3 decreases even further the region of configuration (1) shrinks even further. Once p_3 drops below 0.50, i.e. going from Fig. 2.20c, d, the regime of configuration (3) increases. This is in agreement with Fig. 2.19b where we showed that grouping or not requires a minimum value to contribute to the average replication time.

This analysis shows that origin loci grouping is also a means of minimising replication time for a circular chromosome. If all origins are sufficiently competent they will be furthest apart from each other. A transition from where it becomes best to group two origins if they are weaker compared to a third. Then the individual loci and the group of two take a configuration similar to a two origin model; they are

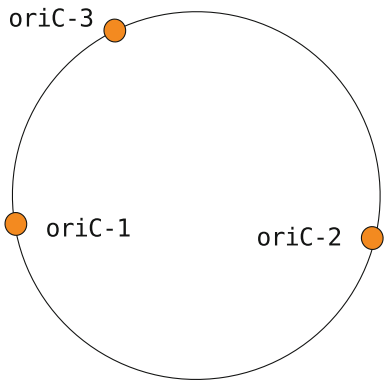

Fig. 2.21 An archaeal chromosome with 3 origin loci. Schematic representation of the arrangement of the origin loci (oriC-1, oriC-2, oriC-3) of the archaea *Sulfolobus solfataricus*

1/2 apart from each other. There are only a few examples in nature where there are three origins on a circular chromosome. Most of the organisms with circular chromosomes and multiple origins are part of the kingdom of archaea [21, 22]. In Fig. 2.21, we show an example of a *Sulfolobus solfataricus* chromosome with three origin loci [23]. The arrangement of its origin loci bears resemblance with what we have shown here to be the optimal positions for loci with high competence (see also Figs. 2.19a and 2.18), and there are also several further examples as for instance in *Haloferax volcanii* [24] or *Sulfolobus islandicus* [25] with similar loci arrangements.

2.3 Optimal Origin Loci and Stochasticity in Origin Activation Time

The above discussions focused on the case of pre-defined loci in yeast and archaea, and ignored additional noise such as the variation in origin activation time. Stochastic origin activation is also well accepted by biologists and we now examine the case of stochastic activation time for *Xenopus laevis* embryos as a model organism. We remind that unlike loci in *Saccharomyces cerevisiae*, any DNA locus in a *Xenopus laevis* embryo is capable of binding with pMcm to become an origin. Surprisingly, biologists find roughly equally-spaced groups of 5–10 pMcms separated by approximately 10 kbp [26–28]. However do these give minimal replication time for biological relevant parameters with such an activation time distribution?

We first turn to the case where origin loci have been licensed, and there is a delay during their activation given by some activation time distribution. For simplicity we assume that the pMcms at an origin can activate with uniform probability at any time within a window which has a lower boundary at $t_0 = 0$ min and an upper at t_b, which is at maximum the length of an S-phase (20 min). The probability for an origin to activate at some time t is distributed according to

$$f_T(t) = \begin{cases} 1/t_b & 0 \le t \le t_b \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

This distribution has mean $\mu = t_b/2$ and standard deviation $\sigma = t_b/\sqrt{12}$ and represents, for example the grey area in Fig. 2.22a. As an origin activates later than its neighbour the overall replication is delayed as well. In this scenario replication completes when all forks have either coalesced or reached the end of the DNA segment. If an origin does not activate by the time its locus is replicated from a replication fork which originated elsewhere, it then cannot become activated anymore. The replicating fork then has to continue synthesis until it reaches the end of the chromosome; which prolongs overall replication time.

We will use the same approach as for a linear chromosome in Sect. 2.1 now incorporating the uniform activation time distribution. In this case, an "origin" is defined as a locus where at least one pMcm has bound to it, and so it corresponds to 100 % competent locus in the notation we have used so far. In addition, pMcms are assumed to be all identical with the same activation probability distribution (standard deviation $\sigma = t_b/\sqrt{12}$). We apply this probability distribution to the two-origin model depicted in Fig. 2.22a, and we also use the genetic search algorithm [17] to find the positions resulting in minimum replication time as σ increases. The expectation is that we will again see a transition of the optimal configuration from isolated pMcms to groups as σ increases; this is akin to varying competence in our previous scenario. If for most cases an origin activates too late it becomes replicated and cannot activate anymore. The active replication fork then has to travel a much farther to complete replication at a much later time as if all origins had activated. We test this prediction using the two-origin model with one pMcm bound to one origin; we find numerically the optimal (minimum average replication time) positions for the origins as a function of σ which are shown in Fig. 2.22b. These results show that origin grouping is also preserved in the two-origin model with stochastic variation in origin activation time. Grouping is important for swift replication under conditions of low competence and large noise which we will explain in the remainder of this chapter. We again use a segment of unit length and forks progress at unit speed of $v = 1\,\text{kbp/min}$. We observe a sharp transition at $\sigma \approx 0.25\,\text{min}$, above which it is best to place both origins in the middle of the segment, as observed in the case with varying competence. This is consistent with Fig. 2.22c which shows the average replication time. A minor difference between this case and the previous one in Sect. 2.1 is that for $\sigma < 0.25$, the optimal location of the origins is not constant. Origins move by a small amount further towards the edges of the chromosome. Using a Gaussian activation time distribution, as suggested by for example Goldar et al. [15] or Herrick et al. [29], also gives the transition from separated to grouped origin for a similar σ-value of around 0.25 if we fix the mean at zero (cf. also Fig. 2.23b). A uniform distribution is thus a good approximation and further has the advantage that replication can only occur after a set time $t = 0$. A Gaussian distribution however has the complication that by its definition an activation prior to the begin of S-phase is possible. The transition for the uniform distribution we use here is also

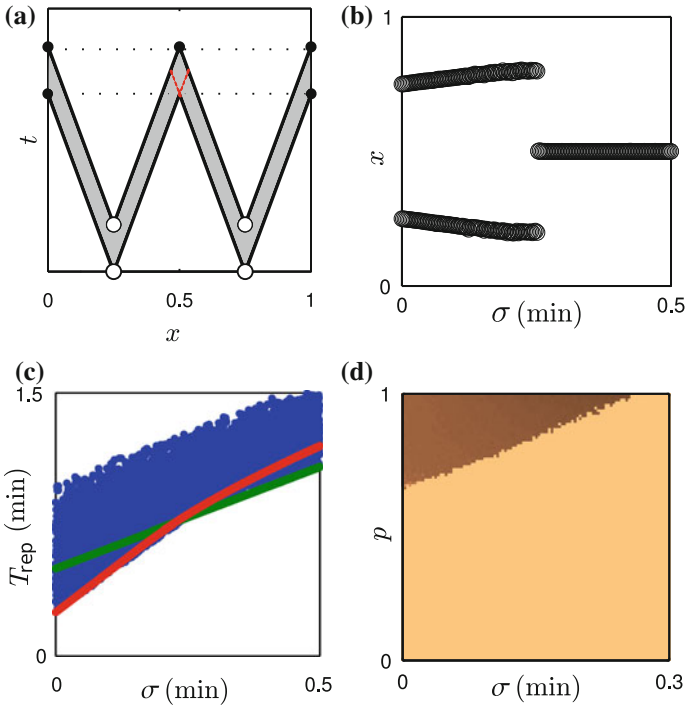

Fig. 2.22 **a** Schematic representation of space–time diagram for a two-origin system. Origins (*hollow circles*) can activate randomly within a time window (*grey area*). This will change the replication completion time (*dotted lines*). Forks arrive later at an edge (*filled circle*), and also forks from an early origin have to travel further until they coalesce with those of a late origin (*red dotted line*). **b** Origin position x so that the average replication T_{rep} for 2 pMcms is minimal on a segment of unit length, when the standard deviation σ for their activation time increases. **c** T_{rep} curves for two-origin systems of (**b**) at fixed positions (*green* $x = 1/2$; *red*: $x = 1/4$ and $x = 3/4$); or both at random sites (*blue*). **d** Phase diagram of the two-origin model to minimize replication time with changing competence and increasing the σ. Colour indicates origin position relative to chromosome ends $d_1, d_2 = 1/2$ (*beige*) or $d_1, d_2 = 1/4$ (*brown*)

reflected in Fig. 2.22c where the fixed positions at $x = 1/4$ and $x = 3/4$ (red solid line) result in a slightly higher replication time than compared to a random sampling of all possible configurations (blue area). Intuitively speaking, the origins group if the fork travelling towards the middle position needs to travel beyond the position of the other, i.e. it travels a distance longer than 0.5 and then has to continue until it reaches the end (see also 2.22a). Figure 2.22d shows that the transition between the group and ungrouped regimes also holds if we vary competence as well as varying σ of the uniform activation time distribution.

We also remark that our result is independent from the particulars of origin activation time distribution. Figure 2.23 depicts examples for the case of two origins and using either a distribution where an origin can activate with probability 1/2 at

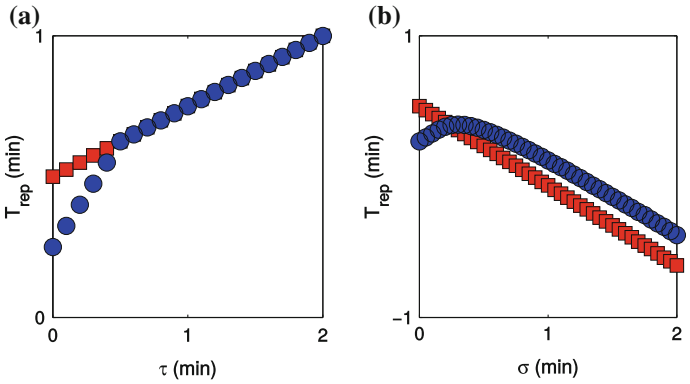

Fig. 2.23 Spread of replication times using different activation time distributions and varying the time between activations. Two origins are placed on a line of unit length either at positions $x = 1/4$ and $x = 3/4$ (*blue circles*), or both at position $x = 1/2$ (*red squares*). In (**a**) an origin can either activate at time $t = 0$ min or later with equal probability, thus τ denotes the difference between those times. In (**b**) the origin activation time is given by a Gaussian activation time distribution with zero mean. We increase its standard deviation σ. This allows for activation at times earlier than zero, hence the decreasing (and 'negative') average replication times

t=0 min or with equal probability at some later time. The difference between these times is shown as τ in Fig. 2.23a. It is clear that once the difference in origin activation time is larger than 1/2 the configuration of having origins positioned at 1/4 and 3/4 does not display any advantage compared to the case of having both origins at the middle position. Similarly as the standard deviation σ of a Gaussian activation time distribution passes over a threshold value the grouped configuration gives minimum replication time (Fig. 2.23b). We note here that once the Gaussian activation time distribution becomes very wide we achieve minimum replication times as the mean is fixed at zero, however left hand tail of the distribution stretches towards negative value allowing (at least one of the) origins to start at some 'negative' time.

We now apply this model for more origins and pMcms, using realistic parameters so that we can relate the results to what is experimentally known about pMcm distribution of *Xenopus laevis*. We model a stretch of DNA of size 100 kbp and $v = 1$ kbp/min [3]. To determine whether the minimum-replication-time configuration requires pMcm grouping, we distributed 64 pMcms in total, i.e. that there is on average $1/1.5$ pMcm/kbp as found in nature [26]. The pMcms are then placed in $64/n$ groups of $n \in \{1, 2, 4, 8, 16, 32, 64\}$ origins, so that the origins are uniformly distributed through the 100 kbp chromosome, or completely random. As the group size decreases the spacing between origins becomes closer as for instance shown in Fig. 2.24. Other authors have identified σ to be 6–10 min and $\mu \sim 15$ min (Gaussian-like) in *X. laevis* [29, 30] as well as in *S. cerevisiae* [3, 4, 12, 14]. As works by Herrick et al. and Goldar et al. [29, 30] have identified the activation distribution at a fixed mean in *Xenopus laevis*, using a uniform distribution and varying σ is a good approximation for our analysis here. Our results (Fig. 2.25a) indicate

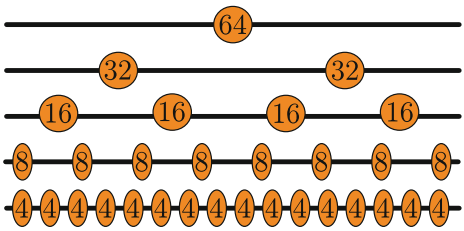

Fig. 2.24 Cartoon illustration of distributing a total of 64 pMcms at origins in groups of varying sizes to simulate the pMcm distribution in *Xenopus laevis*. As the groups of pMcms at an origin decrease the separation between individual origins decreases as well

that grouping with an equal spacing of up to 12.5 kbp achieves precise and fast DNA synthesis before the end of S-phase (20 min) for σ within these limits. We also find that 8 groups of 8 pMcms gives the advantage of a 1.1 min quicker T_{rep} than using random loci; even when the number of pMcms at these 8 groups varies, a quicker T_{rep} is achieved (data not shown). Grouping pMcms also protects the overall replication process against fluctuations from one round of the cell cycle to another; a similar problem is discussed in [31]. This is because one initiation event at an origin is sufficient to activate replication forks and result in a shorter mean time for an activation event at an origin, as we show below.

The probability of the ith pMcm activating by the time t^* given our uniform activation time distribution is

$$P(X_i = t^*) = \frac{t^*}{t_b}, \tag{2.16}$$

and the probability that a pMcm activates later than t^* is

$$\begin{aligned} P(X_i > t^*) &= \int_{t^*}^{t_b} \frac{t'}{t_b} \mathrm{d}t' \\ &= \frac{t_b - t^*}{t_b}, \end{aligned} \tag{2.17}$$

We consider there to be a group of n identical pMcms at an origin. The probability of at least one of those activating by t^* then follows as

$$\begin{aligned} P(\min(X_i) = t^*) &= \sum_{i=1}^{n} \left\{ P(X_i = t^*) \prod_{j=1, j\neq i}^{n} P(X_j > t^*) \right\}, \\ &= n P(X_i = t^*) P(X_j > t^*)^{n-1}, \qquad (2.18) \\ &= n \frac{1}{t_b} \left(\frac{t_b - t^*}{t_b} \right)^{n-1} \qquad (2.19) \end{aligned}$$

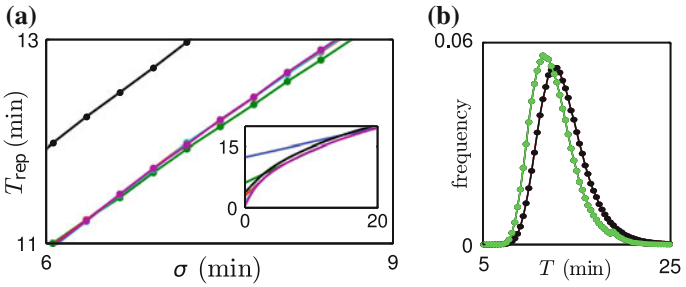

Fig. 2.25 Replication timing in *Xenopus laevis*. **a** Inset: T_{rep} as a function of σ for realistic parameters as given in the text. Origins are distributed in 4 equally-spaced groups of 16 pMcms (*blue*); 8 groups of 8 pMcms (*green*); 16 groups of 4 pMcms (*red*); 32 groups of 2 pMcms (*cyan*); 64 single pMcms (*magenta*); 64 pMcms placed randomly (*black*). Main: zoom around realistic $\sigma \sim 8$ min. For $6 < \sigma < 20$ min minimal T_{rep} is achieved for groups of 8 pMcms. **b** Distribution of replication times T for a 100 kbp chromosome under the condition that $\sigma = 8$ min. Shown is the distribution of 64 pMcms in 8 equally-spaced groups of 8 pMcms (*green*) and placed randomly (*black*)

According to this origin activation time distribution the mean activation time is $t_b/(n+1)$. This shows that activation is earlier for a certain group of pMcm compared to an individual that has mean activation time $t_b/2$. So as the average activation increases through t_b it becomes a balancing act to be able to activate before a replication fork has moved across from another origin elsewhere. The origin must also not be too sparsely placed to leave small enough gaps between groups to replicate on time. Grouping is therefore a useful tactic to achieve this by lowering the overall activation time of a group of pMcm.

In a natural environment, one might expect that there would not be strict equal spacing of groups as we show it here. We now relax our previous assumption by taking evenly-spaced groups and perturb the location of each group by a small random amount drawn from a Gaussian distribution. The introduction of such variation allows us to compare our simulation with available experimental data of replicated genomic regions, which were captured as centre-centre distances at around 5 min after the onset of replication (for instance in Blow et al. [27]). Figure 2.26 shows that our result is in agreement with the current understanding of the biological community, i.e. groups of 5–10 pMcms about every 10 kbp. This may be achieved by a regulation of pMcm—loading proteins, whose affinity to bind decreases around existing origins [32, 33]. Although a random placement represents the data similarly well, T_{rep} remains smaller in this case where the origin groups are not equally-spaced as seen before (cf. Fig. 2.25b).

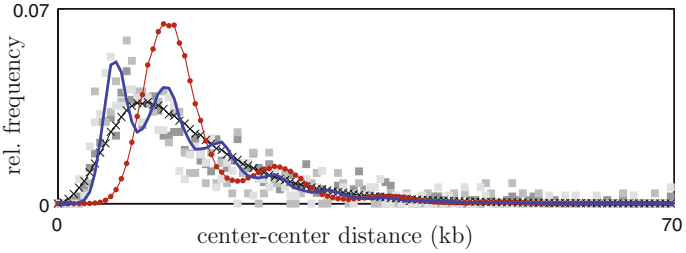

Fig. 2.26 Centre–centre distribution from three experiments [27] (*squares*) and from simulations 5 min after replication started. The simulation is positioning groups of 4 pMcms every 6.3 kbp (*solid line*), groups of 8 pMcms every 12.5 kbp (*circles*), or all randomly (*crosses*). A small random amount was added to the group location of fixed distances which was picked from a Gaussian distribution with $\sigma \sim 16\,\%$ of group distances. The pMcm/length ratio was fixed with a total of 64 origins distributed per 100 kbp of DNA (cf. Fig. 2.25a)

2.4 Summary

Grouping is a means by which replication time is minimised and it strongly depends on the parameters of an origin. Some of the previous models of DNA replication neglected this fundamental question of where origins should be placed to minimise replication time.

We have shown that random fluctuations in the formation of origins, and the subsequent activation of proteins lead to variations in the replication time. We analysed these stochastic properties of DNA and derive the positions of origins corresponding to the minimum replication time. This was done calculating the relation between the competence of the origin to activate and the replication time; low-competence replication origins tend to group in order to minimise replication, and so do origins with long delay in their activation time. This delay is independent of the shape of the activation time distribution of the origins. Moreover we intuitively showed that origin grouping occurs to compensate for the origin failure. It thus only depends on whether or not an origin had become activated before it becomes passively replicated by a replication fork that originated elsewhere.

We have related this to experimental data in a number of species. All of those organisms show that origin grouping on linear as well as circular chromosomes is a means for minimising replication time. We finally showed arguments to prove our hypothesis that evolution has driven origins to the locations where they are found today. For this we used *Saccharomyces cerevisiae* as an example, however we propose that our results also applies to other yeast species such as *Schizosaccharomyces pombe*.

References

1. R. Reyes-Lamothe, C. Possoz, O. Danilova, D.J. Sherratt, Independent positioning and action of Escherichia coli replisomes in live cells. Cell **133**(1), 90–102 (2008). doi:10.1016/j.cell.2008.01.044
2. T.M. Pham, K.W. Tan, Y. Sakumura, K. Okumura, H. Maki, M.T. Akiyama, A single-molecule approach to DNA replication in Escherichia coli cells demonstrated that DNA polymerase III is a major determinant of fork speed. Mol. Microbiol. (2013). doi:10.1111/mmi.12386
3. M.K. Raghuraman et al., Replication dynamics of the yeast genome. Science **294**(5540), 115–21 (2001). doi:10.1126/science.294.5540.115
4. M.D. Sekedat, D. Fenyö, R.S. Rogers, A.J. Tackett, J.D. Aitchison, B.T. Chait, GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. Mol. Syst. Biol. **6**, 353 (2010). doi:10.1038/msb.2010.8
5. H.M. Mahbubani, T. Paull, J.K. Elder, J.J. Blow, DNA replication initiates at multiple sites on plasmid DNA in Xenopus egg extracts. Nucleic Acids Res. **20**(7), 1457–1462 (1992)
6. A. Lengronne, P. Pasero, A. Bensimon, E. Schwob, Monitoring S phase progression globally and locally using BrdU incorporation in TK+ yeast strains. Nucleic Acids Res. **29**(7), 1433–1442 (2001)
7. C.A. Müller et al., The dynamics of genome replication using deep sequencing. Nucleic Acids Res. **42**(1), e3 (2013). doi:10.1093/nar/gkt878
8. J.J. Blow, Control of chromosomal DNA replication in the early Xenopus embryo. EMBO J **20**(13), 3293–3297 (2001). doi:10.1093/emboj/20.13.3293
9. M. Hawkins, R. Retkute, C.A. Müller, N. Saner, T.U. Tanaka, A.P. de Moura, C.A. Nieduszynski, High-Resolution Replication Profiles Define the Stochastic Nature of Genome Replication Initiation and Termination. Cell Rep. **5**(4), 1132–1141 (2013). doi:10.1016/j.celrep.2013.10.014
10. C.A. Nieduszynski et al., OriDB: a DNA replication origin database. Nucl. Acids Res. **35**, 40–46 (2007). doi:10.1093/nar/gkl758
11. T.W. Spiesser, E. Klipp, M. Barberis., A model for the spatiotemporal organization of DNA replication in Saccharomyces cerevisiae. Mol. Genet. Genomics **282**(1), 25–35 (2009). doi:10.1007/s00438-009-0443-9
12. A.P.S. de Moura, R. Retkute, M. Hawkins, C.A. Nieduszynski, Mathematical modelling of whole chromosome replication. Nucleic Acids Res. **38**(17), 5623–5633 (2010). doi:10.1093/nar/gkq343
13. S.C.-H. Yang, N. Rhind, J. Bechhoefer, Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Mol. Syst. Biol. **6**, 404 (2010). doi:10.1038/msb.2010.61
14. A. Brümmer, C. Salazar, V. Zinzalla, L. Alberghina, T. Höfer, Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. PLoS Comput. Biol. **6**(5), e1000783 (2010). doi:10.1371/journal.pcbi.1000783
15. A. Goldar, M.-C. Marsolier-Kergoat, O. Hyrien, Universal temporal profile of replication origin activation in eukaryotes. PLoS One **4**(6), e5899 (2009). doi:10.1371/journal.pone.0005899
16. R. Retkute, C.A. Nieduszynski, A. de Moura, Mathematical modeling of genome replication. Phys. Rev. E **86**(3), 031916 (2012). doi:10.1103/PhysRevE.86.031916
17. D. Levine, Users guide to the PGAPack parallel genetic algorithm library. (1996), http://ftp.mcs.anl.gov/pub/pgapack/, doi: 10.2172/366458
18. O. Hyrien, A. Goldar, Mathematical modelling of eukaryotic DNA replication. Chromosome Res. **18**(1), 147–161 (2010). doi:10.1007/s10577-009-9092-4
19. K. Shirahige, T. Iwasaki, M.B. Rashid, N. Ogasawara, H. Yoshikawa, Location and characterization of autonomously replicating sequences from chromosome VI of Saccharomyces cerevisiae. Mol. Cell. Biol. **13**(8), 5043–5056 (1993). doi:10.1128/aANMCB.13.8.5043
20. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994)

21. L.M. Kelman, Z. Kelman, Multiple origins of replication in archaea. Trends Microbiol. **12**(9), 399–401 (2004). doi:10.1016/j.tim.2004.07.001
22. O. Hyrien et al., From simple bacterial and archaeal replicons to replication N/U-domains. J. Mol. Biol. **425**(23), 4673–4689 (2013). doi:10.1016/j.jmb.2013.09.021
23. I.G. Duggin, N. Dubarry, S.D. Bell, Replication termination and chromosome dimer resolution in the archaeon Sulfolobus solfataricus. EMBO J. **30**(1), 145–153 (2011). doi:10.1038/emboj.2010.301
24. C. Norais, M. Hawkins, A.L. Hartman, J.A. Eisen, H. Myllykallio, T. Allers, Genetic and physical mapping of DNA replication origins in Haloferax volcanii. PLoS Genet. **3**(5), e77 (2007). doi:10.1371/journal.pgen.0030077
25. R.Y. Samson et al., Specificity and function of archaeal DNA replication initiator proteins. Cell Rep. **3**(2), 485–96 (2013). doi:10.1016/j.celrep.2013.01.002
26. H.M. Mahbubani, Cell Cycle Regulation of the Replication Licensing System: Involvement of a Cdk-dependent Inhibitor. J. Cell Biol. **136**(1), 125–135 (1997). doi:10.1083/jcb.136.1.125
27. J.J. Blow, P.J. Gillespie, D. Francis, D.A. Jackson, Replication origins in Xenopus egg extract Are 5–15 kilobases apart and are activated in clusters that fire at different times. J. Cell Biol. **152**(1), 15–25 (2001)
28. M.C. Edwards, A.V. Tutter, C. Cvetic, C.H. Gilbert, T.A. Prokhorova, J.C. Walter, MCM2-7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in Xenopus egg extracts. J. Biol. Chem. **277**(36), 33049–33057 (2002). doi:10.1074/jbc.M204438200
29. J. Herrick, S. Jun, J. Bechhoefer, A. Bensimon, Kinetic Model of DNA Replication in Eukaryotic Organisms. J. Mol. Biol. **320**(4), 741–750 (2002). doi:10.1016/S0022-2836(02)00522-3
30. A. Goldar et al., A dynamic stochastic model for DNA replication initiation in early embryos. PLoS One **3**(8), e2919 (2008). doi:10.1371/journal.pone.0002919
31. S.C.-H. Yang, J. Bechhoefer, How Xenopus laevis embryos replicate reliably: investigating the random-completion problem. Phys. Rev. E: Stat. Nonlin. Soft Matter Phys. **78**(4), 41917 (2008)
32. A. Rowles, S. Tada, J.J. Blow, Changes in association of the Xenopus origin recognition complex with chromatin on licensing of replication origins. J. Cell Sci. **112**, 2011–2018 (1999)
33. M. Oehlmann, A.J. Score, J.J. Blow, The role of Cdc6 in ensuring complete genome licensing and S phase checkpoint activation. J. Cell Biol. **165**(2), 181–90 (2004). doi:10.1083/jcb.200311044

Chapter 3
Actively Replicating Domains Randomly Associate into Replication Factories

In the previous chapter DNA was treated as a stiff, one-dimensional line. However within a cellular environment DNA diffuses and organises into structures on different scales as for instance being wrapped around nucleosomes or forming chromatin. This brings otherwise far-away genomic regions into physical contact with each other. Yet it is unclear what leads to such an organisation of structures in three dimensions—especially during DNA replication. Recent advances using chromosome conformation capture data, e.g. HiC, 3C, 4C techniques, shed some light on the way chromosomal regions (domains) interact with each other for example during protein expression or genome duplication (see review and commentary [1, 2]). The technology captures the organisation of chromosomes in form of contact maps that can help to understand the organisation of chromosomes in a given cell subject to particular (growth) conditions. There is also an increasing body of work modelling chromosome interaction data using techniques from graph theory [3] or polymer theory [4, 5]. One disadvantage of these technologies is that they also captures random collisions of chromosomal region with another. The experimental result thus also contains information of unspecific interaction that occurred among genomic loci and the signal must be sufficiently from those regions that specifically come into contact. A further confounding factor is that the majority of experiments, that use the chromosome capture techniques, average over population measurements when establishing interaction and chromosome contact data. They therefore pursue a top–down approach by inferring single-cell operations from population studies. It is difficult to draw conclusions from these studies about mechanism occurring on a single-cell level.

We present here—on basis of single-cell experiments—the mechanism that causes replication forks (replisomes) in *Saccharomyces cerevisiae* to establish chromosomal interaction which affects further organisation into *replication factories*. We show that these factories stem from random interaction events of adjacent replication forks. Our theoretical model establishes from experimental grounds. Data is derived from a technique initially established by Kitamura et al. [6] for single *Saccharomyces cerevisiae* cells to visualise replication on DNA. Their data shows that during DNA synthesis genomic regions (domains) undergoing replication—so called *replicons*—

J. Karschau, *Mathematical Modelling of Chromosome Replication and Replicative Stress*,
Springer Theses, DOI 10.1007/978-3-319-08861-7_3

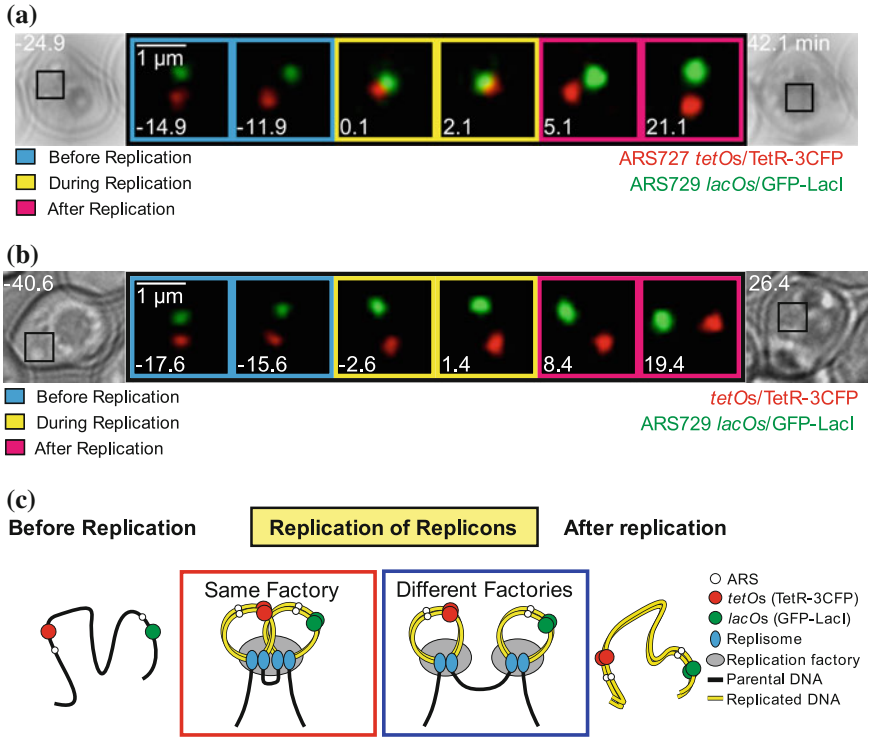

Fig. 3.1 Observations of two replicating dots. Two genomic loci have been labelled either in *green* or in *red*. If these loci undergo replication the observed intensity at these loci doubles. With this information one can establish whether or not both loci are seen in close contact (**a**), or whether they are localised far apart (**b**) during DNA replication. Panel (**c**) shows a graphical interpretation of close dot localisation, which can lead to replication in the same factory, or not as in (**a**) and (**b**). Details for the experimental are provided in Saner et al. [8]

become associated with each other. They label DNA near origin loci sites which allows them to observe when a replicon becomes replicated; the fluorescent intensity at a replicated region doubles. This is illustrated in Fig. 3.1. Moreover labelling two regions also shows that those sometimes become associated (Fig. 3.1a) forming *replication factories* of about 90 nm in diameter [7]. In some other cases, labelled regions do not associate when a region undergoes replication as for example in Fig. 3.1b. It is currently an open question how association occurs and an extension of the labelling technique by our experimental collaborators using multiple labelled sites produces new experimental data of replicon associations (see also Saner et al. [8]).

This data alone is incomplete and requires a physics approach for a holistic comprehension to whether or not association is a deterministic or stochastic process. We complement their data using a mathematical model to allow further insight. We test our model numerically first using a Metropolis-Monte-Carlo algorithm. This then

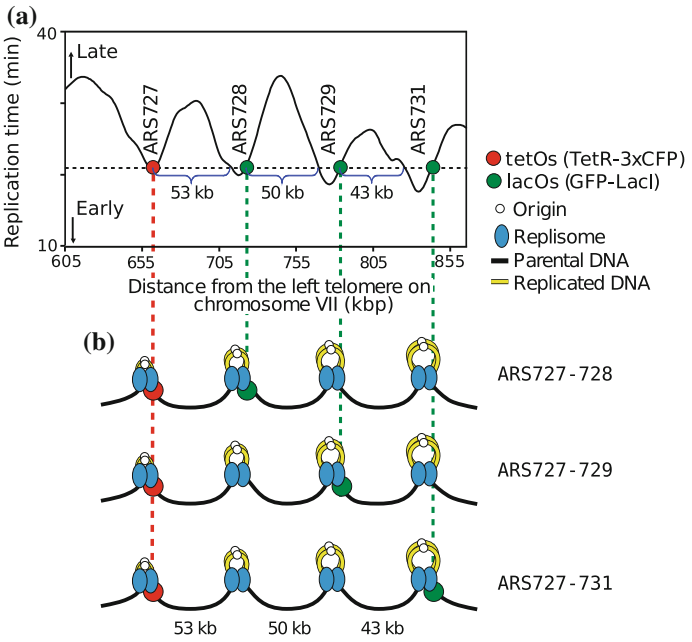

Fig. 3.2 Tagged replisome pair distances on chromosome VII. **a** Replication profile of the relevant genomic region on chromosome VII obtained from [9]. Origin loci (ARS727–731) are the values of the timing profile and tetO (*red*) or lacO (*green*) where integrated in three different strains as a combination of ARS727–728, ARS727–729, and ARS727–731. **b** Diagram of the length scale of sites which were labelled near replicating origins. Each of three different strain had ARS727 labelled as a reference (*red dot*) and in *green* in that strain either ARS728, ARS729, or ARS731 was labelled (*green dot*). Length scales between neighbouring points is given in kbp. The chromosomal distance d between relevant replisome pairs upon replication of fluorescent dots (i.e., tetO and lacO arrays) is estimated assuming that upon replication sister replisomes stay together. Thus, to obtain the chromosomal distance d between replisome pairs, only the integration sites of tetOs and lacOs **a** need to be considered, but not the length of these arrays, as shown by the distances

allows us to extend it from a four origin in vivo experiment to a whole-genome in silico one. Without further need for any parameter we relate back to observations of entire cells and the size distribution of their replication factories.

3.1 Summary of Experimental Procedure

To investigate the organisation of replication factories Saner et al. [8] analysed the behaviour of replicons using live-cell microscopy in *Saccharomyces cerevisiae*. They chose a region on chromosome VII with four adjacent replication origin loci (Fig. 3.2) and selected one locus on each replicon such that all four loci show the same average replication timing [9]. They integrated two DNA sequences into a chromosome.

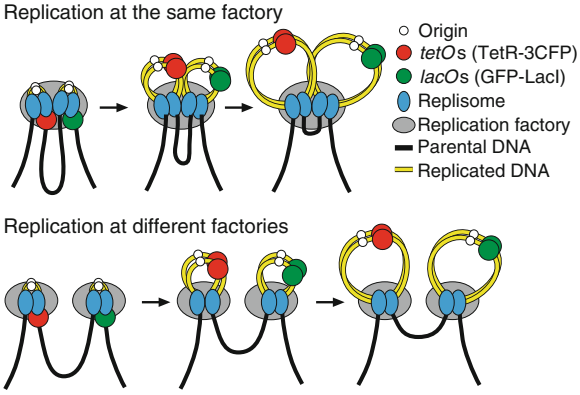

Fig. 3.3 Concept of replication factories. Models for replication of two replicons at the same factory and at different factories. If two replicons are processed for replication in the same factory, CFP and GFP fluorescent dots should come into close proximity during replication i.e. when the intensity of the dots increases. In contrast, if replicons are processed in different factories, they do not come closer during replication

Those sequences are normally repressed by the binding of a fluorescent protein. Those inserted genes are tetO and lacO arrays producing a separate strain for each of the three different red–green combinations of Fig. 3.2. These arrays bound TetR and LacI proteins, fused with cyan and green fluorescent proteins (CFP, GFP), respectively, and were thus visualised as small fluorescent dots. The fluorescent dots increased in intensity upon DNA replication as the number of arrays was doubled, thus defining their replication timing by microscopy [6].

To analyse how replicons are gathered into factories, only cells whose two marked loci replicated with similar timing, i.e. their difference in activation was <3 min, were taken into account. When both loci replicated and they were observed in close spatial proximity of less than 350 nm apart for more than 2 min, they were considered to be replicating in the same factory (Fig. 3.3). In contrast, if replicons are processed in different factories, the fluorescent dots do not come close during replication.

Using this protocol, it was found that the two marked loci in the first strain (ARS727–728: 53 kb apart) replicated in the same factory in 43 % of cells (10 out of 23) and in different factories in 57 % of cells (13 out of 23). This suggests that grouping of replicons within factories can vary from cell to cell. In contrast, in the other two strains ARS727–729 and ARS727–731, the two marked loci replicated in the same factory less frequently: 11 % (2/19) and 5 % (1/19), respectively. Thus, replicons that are close along a chromosome were often processed for replication in the same factory, but replicons that are farther apart replicated more frequently in different factories. The manner by which this occurs is not directly inferable from the data alone and requires additional modelling to explain the association of replisomes and the amount replisomes per factories. We therefore introduce a

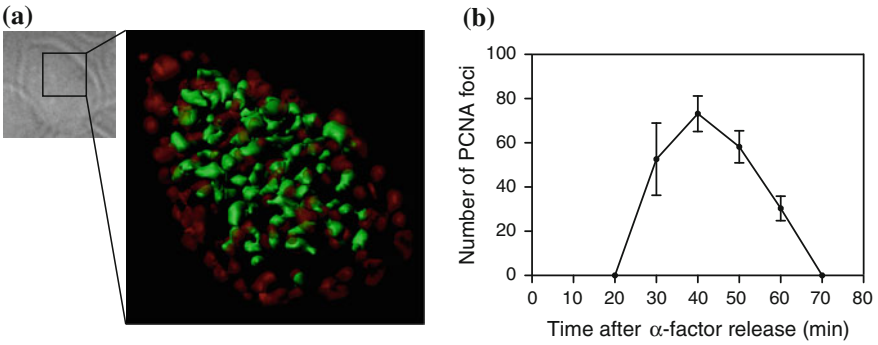

Fig. 3.4 Replication factories observed by super-resolution microscopy. Cells (T8375) with GFP-POL30 (PCNA), SPC42-mCherry (a component of spindle pole body, SPB), and NIC96-mCherry (a component of the nuclear pore complex, NPC) were released from α-factor treatment (defined as 0 min). **a** A bright-field image, a fluorescence image (GFP, *green*; mCherry, *red*), and the 3D rendering of a fluorescence image (GFP, *green*; mCherry, *red*) in a representative cell, which was fixed at 40 min. **b** The number of PCNA foci (mean $\pm$ SD) within the nucleus along the time course. For further experimental procedures refer to Saner et al. [8]

mathematical model that will help us to understand the meaning of the above mentioned association probabilities.

A further set of data from this study contains single-cell images of fluorescent PCNA; a unique replication fork component which shows where replication forks localise (cf. Sect. 1.5, page 10ff). Images as shown in Fig. 3.4a contain size distributions and the count of PCNA foci at a particular time point (Fig. 3.4b). We will use our mathematical model and apply it to genome-wide yeast replication simulation to establish the number of forks per factory in silico as well as determine this number independently in vivo.

3.2 The Diffusion Time Scale of Two Replicating Dots

In the above experiment for replicon grouping probabilities, two dots appear associated if they are in close proximity. This can occur via two routes (i) replication forks have reached the location of an adjacent dot or (ii) two dots meet randomly due to diffusion. In experiments, the diffusion coefficient of one dot was obtained and found to be $D_1 = 0.2\,\mu\text{m}^2/\text{min}$. The typical diffusion time scale for one dot to travel some distance L is calculated using [10]:

$$t_D = \frac{L^2}{2D}. \tag{3.1}$$

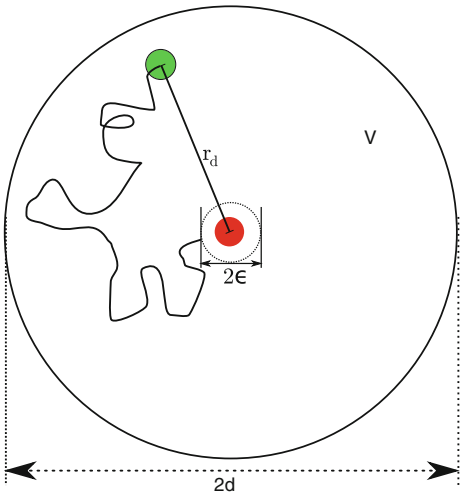

Fig. 3.5 Target finding time. To establish the target finding time we fix one replisome pair at the centre of spherical volume V which has radius r. The moving pair (*green*) performs diffusion with twice the diffusion coefficient of one particle until it hits the centre particle (*red*). This target region has a diameter of 2ϵ, i.e. twice of that of one particle. illustration

Since we have two particles diffusing, this is equivalent to one particle diffusing with $D = 2D_1$. Hence, we double the diffusion coefficient in the Eq. (3.1). The DNA is not as fully stretched as is shown in Fig. 3.2b for instance. We therefore apply a chromatin packaging ratio of 10 nm/kbp, based on a ratio of measured spatial distances over the chromosomal distances between two fluorescently labelled chromosomal loci. This matches a reported value [11]. For the maximum distance between the origins in the second strain with ARS727–729, $L = 1.3\,\mu$m corresponding to chromosomal distance of 129 kbp between the fluorescent dots [12]. The diffusion time-scale for this length is $\sim$2 min. In this strain it takes 4–6 min after replication initiation at ARS727 and ARS729 until tetO/lacO dots are replicated (i.e. until replication forks reach the middle of tetO/lacO arrays, which are 10–11 kb in length). Replication from one dot to another, i.e. replicating a distance of 50 kbp (cf. Fig. 3.2), takes about 15 min. So diffusion can in principle account for two dots to meet and to be seen associated under the microscope. However the diffusion time scale alone does not mean that the two dots will actually meet within this time.

We therefore establish a further property for a system of two diffusing dots which is the *mean target finding time*. This measure allows to estimate whether the dots not only have time to explore a certain distance, but also whether they actually meet with another, i.e. the time scale until one dot will encounter another. We derive the target finding time in analogy to Sneppen and Zocchi [10] by rephrasing our problem as depicted in Fig. 3.5. Instead of describing two sister replisome pairs diffusing in a spherical volume, it is equivalent to fix one sister replisome pair (red) at the center—it becomes the target—and the second pair (green) diffuses around this target (shown in Fig. 3.5). Hence, the diffusion constant of the second pair D is the sum of its own diffusion constant and that of the target, which we place at a fixed position. The spherical volume V, in which the mobile pair diffuses to find its target, has radius r_d. This radius $r_d = d$ is defined by the distance of the pairs to be at a distance d

maximally apart from each other (cf. Fig. 3.2). The radius of the spherical volume (dotted circle) that two sister replisome pairs occupy once the mobile pair meet the target, is twice the diameter of one and therefore 2ϵ.

We describe this process using the standard diffusion equation

$$\frac{\partial p(\vec{r},t)}{\partial t} = D\nabla^2 p(\vec{r},t), \tag{3.2}$$

with position probability density $p(\vec{r},t)$ for finding the mobile pair at a particular position $\vec{r}$ at a time t. We are interested in the equilibrium state, that is for once the systems has reached steady-state. We therefore set the left hand-side of Eq. (3.2) to zero as the system will be independent from time

$$0 = D\nabla^2 p(\vec{r},t). \tag{3.3}$$

We simplify the equation by assuming that $p(\vec{r},t)$ is independent from any particular direction (isotropic), and only depends on the radial distance r_d to the fixed pair. Using spherical coordinates we show that the probability distribution satisfies

$$\int_{r=0}^{d}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} p(r,\theta,\phi)\, r^2 \,\sin\theta \,\mathrm{d}r\mathrm{d}\theta\,\mathrm{d}\phi = 1. \tag{3.4}$$

It then follows from the position probability density distribution $p(r,\theta,\phi) = \text{const}$ and

$$\int_{r=0}^{d}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} r^2 \,\sin\theta \,\mathrm{d}r\mathrm{d}\theta\,\mathrm{d}\phi = V \tag{3.5}$$

that

$$p(r,\theta,\phi) = 1/V. \tag{3.6}$$

In order to solve the second-order differential equation, Eq. (3.3), we choose two boundary conditions:

1. When the mobile pair is away from the fixed one ($r_d > \epsilon$), there is a flux I towards the fixed origin
$$I = -4\pi D r^2 \frac{dp(r_d)}{dr_d}, \tag{3.7}$$
2. When the mobile pair is far away from the fixed one, the probability of finding the particle is $p(\infty) = 1/V$.

We then obtain,

$$p(r_d) = \frac{I}{D4\pi r_d} + p(\infty) \tag{3.8}$$

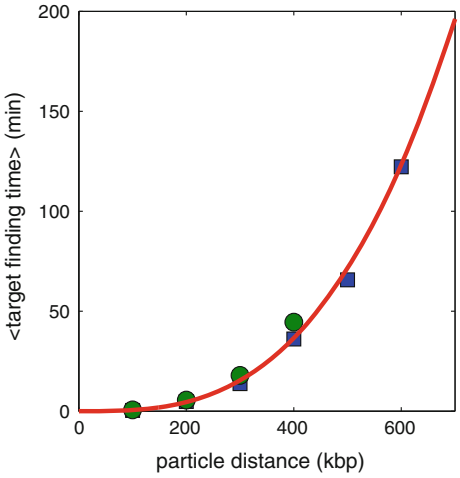

Fig. 3.6 Simulation of average target finding times. Two particles move freely in a given volume (*green squares*), or one dot moves with the sum of the diffusion constant of both particles for the average time it takes for both particle to be found associated if they diffuse in the given volume, i.e. given by the maximum separation allowed. The solid red line shows the analytical result for the average target finding time τ_{on} given by Eq. (3.9) for $D = 0.2\,\mu\text{m}^2/\text{min}$ and $\epsilon = 62.5$ as chosen in these simulations

as the solution for Eq. (3.3). The probability of finding the mobile origin at the target is $P(\epsilon) = 0$. We now solve Eq. (3.8) for the flux I which is the inverse of the time it takes for the mobile origin to find the target

$$\tau_{on} = \frac{1}{|I|} = \frac{V}{4\pi D\epsilon} = \frac{r_d^3}{3D\epsilon}. \tag{3.9}$$

We check our analytical result of Eq. (3.9) numerically by simulating the diffusion of both origins given some maximum separation distance. We also check this against the consideration of one being fixed at the origin and the other moving with a diffusion coefficient that is the sum of both. Figure 3.6 shows the results for this and it also confirms that our analytical derivation of τ_{on} appropriately describes the average target finding time. We remark that the simulation result here is for illustration purposes only as we have $D = 0.2\,\mu\text{m}^2/\text{min}$ and $\epsilon = 62.5$ chosen in these simulation; their actual values become refined in the discussion further below.

The above derivation assumed that replisomes find each other from initial maximally-extended DNA length, i.e. the distance is always as shown in Fig. 3.2. In an experimental setting this clearly not the case, cells are selected at random, and replisomes within each cell can take their initial position during replication anywhere within a volume that has a maximal diameter as the length of the DNA connecting them. We therefore calculate the mean radial distance $\langle r_{\text{act}} \rangle$ of those two. The probability density distribution to be anywhere in that spherical volume of fixed V

(thus $d = \text{const}$) is the same throughout (uniform). Hence the probability density $p(d_{\text{act}})$ of replisome pairs being at an actual distance d_{act} away, i.e. the moving replisome lies on a spherical segment $4\pi d_{\text{act}}^2$, is given by

$$p(d_{\text{act}}) = \frac{4\pi d_{\text{act}}^2}{V}, \tag{3.10}$$

which satisfies

$$\int_{d_{\text{act}}=0}^{d} p(d_{act})\, \mathrm{d}d_{\text{act}} = 1.$$

In a collection of individual cells we therefore find the average distance between replisomes as

$$\langle d_{act} \rangle = \int_0^d d_{\text{act}} p(d_{act})\, \mathrm{d}d_{\text{act}},$$
$$\langle d_{act} \rangle = \frac{3}{4} d. \tag{3.11}$$

This means that we rescale the known chromosomal distance d by 3/4 which yields for the average target finding time of a collection of cells

$$\tau_{\text{on}}^{\text{pop}} = \frac{1}{|I|} = \frac{(3/4d)^3}{3D\epsilon} = \frac{9d^3}{64D\epsilon}. \tag{3.12}$$

We now determine the duration for two sister replisome pairs to meet each other based on Eq. (3.12). This is the time required for the system to approach equilibrium. As a result of taking both replication and diffusion processes into account, the time needed for two sister replisome pairs to find each other is approximately in the the range of 1–5 min in strain ARS727–728 and ARS727–729, respectively. The time of diffusion suffices so that nearby replisomes can associate by the replication of the tetO/LacI arrays. It further allows us to draw an adiabatic assumption—the association probability relaxes to its local equilibrium on a time-scale which is smaller than replication time-scale. This is important to establish an equilibrium model of random associations of replisomes in the following section.

In a wider context, the organisation of DNA inside the cellular nucleus can influence diffusion. The data used to determine the diffusion in the experiment by Saner et al. [8] displays a mean squared displacement which shows saturation at longer times (Figs. 3.7, 3.8 and 3.9). This is characteristic for anomalous diffusion as for example occurs in crowded environments such as the cellular nucleus. The saturation is particularly visible in Figs. 3.8 and 3.9. Previous work, as for example the one by Heun et al. [13], also showed that when cells undergo replication diffusion

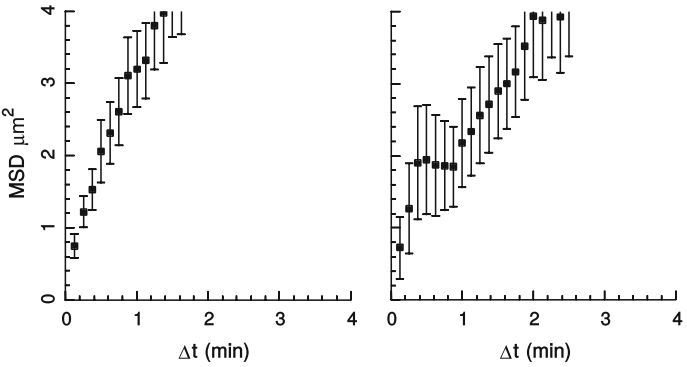

Fig. 3.7 Diffusion of a DNA locus. Mean squared displacement (MSD) of a tagged DNA locus is plotted as a function of the time interval of observation Δt. This was done for a stage of S-phase in the experiment by Saner et al. [8] observing the locus at 25–33 min after α-factor release. Shown here are examples for two cells

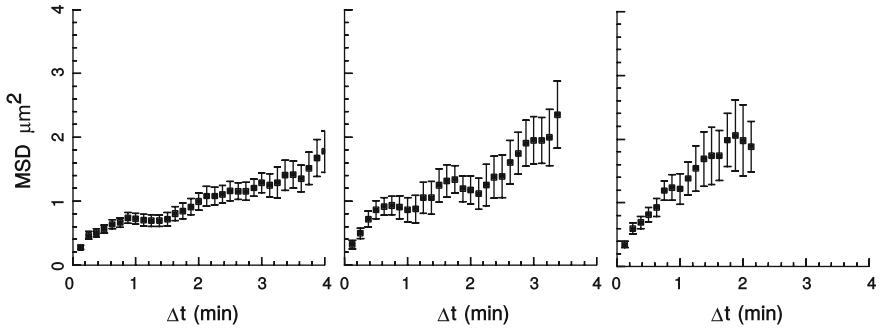

Fig. 3.8 Diffusion of a DNA locus. Mean squared displacement (MSD) of a tagged DNA locus is plotted as a function of the time interval of observation Δt. This was done for a stage of S-phase in the experiment by Saner et al. [8] observing the locus at 35–43 min after α-factor release. Shown here are examples for three cells

slows down, which authors claim to depend on the openness of the chromatin. Of relevance here for our estimate of the target finding is diffusion at short time scales ($\sim$2 min) which allows us to draw a rough estimate of the target finding time as we have done above. In particular the plots shown in Fig. 3.8 which correspond to the time point relevant of replication of the origin loci tagged in the experiment by Saner et al. [8]. However we remark a model also accounting for anomalous diffusion will give a more accurate approximation.

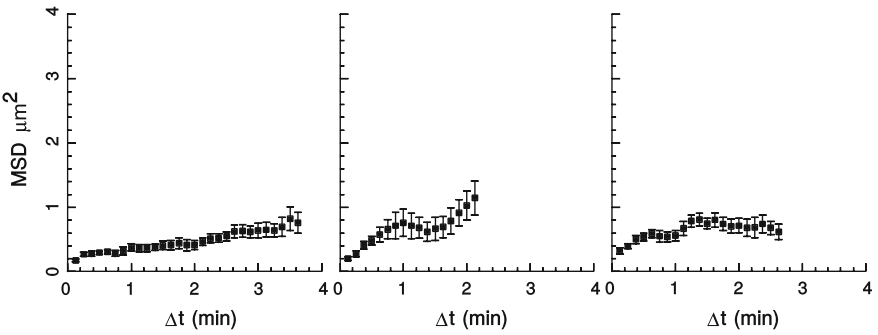

Fig. 3.9 Diffusion of a DNA locus. Mean squared displacement (MSD) of a tagged DNA locus is plotted as a function of the time interval of observation Δt. This was done for a stage of S-phase in the experiment by Saner et al. [8] observing the locus at 45–53 min after α-factor release. Shown here are examples for three cells

3.3 Binding Energy

Using the above assumption that we can treat the system to be in thermodynamic equilibrium allows us to use an analogy for the different strains. Each strain has two replisomes connected by DNA of varying length.

We therefore rephrase our problem as two particles tethered by a string of length d. This describes two sister replisome pairs that are apart from each other at a chromosomal distance d (Fig. 3.10). Each sister replisome pair is thereby a particle fixed to the end of the string. Each particle is considered to be a sphere with diameter ϵ. We assume the string has no stiffness, given that the persistence length of yeast chromatin is short (2.5 kbp [12]) relative to the distance between replication origins and between marked chromosome loci analysed here. The particles perform a random walk within a sphere of radius $d/2$ in three dimensions (illustrated in two dimensions, for simplicity, in Fig. 3.10). If both particles come close within interaction radius, i.e. the distance between their centres is ϵ or less, they associate. Note that we here place the centre of reference at the middle of the string (the centre of mass), compared to the derivation in the previous section were the coordinate system was fixed at a particle.

The two particle system can be in two conditions. First, when particles are separated; the energy of the system is then E = 0 (Fig. 3.10a, b). Second, when particles are in close proximity and become associated; the energy of the system is then E = J, where J is a binding energy (Fig. 3.10c). Therefore, J is negative, meaning that the particles' interaction is attractive.

Our aim is to estimate the probability of finding the system in any of these conditions—particles separated or particles associated—depending on the string length between them. The probability that the two particles meet and associate with each other when the system is in thermodynamic equilibrium is

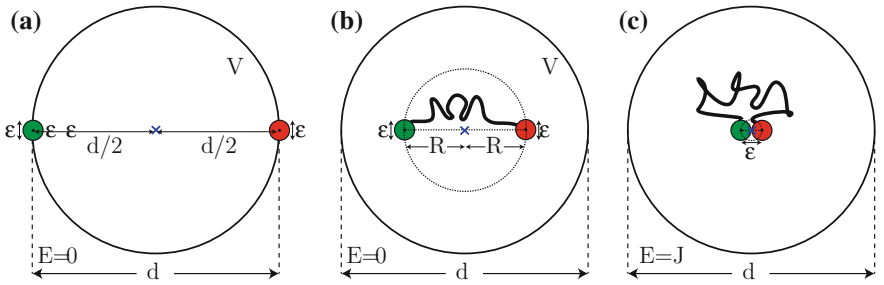

Fig. 3.10 Schematic diagram of the *particles on a string* model. Two particles are connected with a string of length d. The two particles and the string represent two pairs of sister replisomes and the chromosome region between them, respectively. **a**, **b** The particles move by a random walk and the energy remains $E = 0$ as long as they are separated (*left*, *central panels*). **c** If the two particles are associated with each other, i.e. the distance between the centres of the particles is, the energy is reduced ($E = J$, J is the binding energy with a negative value)

$$P_a = \frac{n_a B_a}{n_a B_a + n_s B_s}, \tag{3.13}$$

where n_a and n_s are the normalised numbers of states in which particles are associated and separated, respectively. B_a and B_s represent corresponding Boltzmann factors (weighing factors). Each Boltzmann factor $B = e^{-E/(k_B T)}$ depends on temperature, T, and energy, E, of the system. k_B is the Boltzmann constant.

The normalized number of states is derived as follows. As the reference frame is centered at the midpoint, states corresponding to the particles separated by a distance R lie on a spherical shell of radius $R/2$. Particles are considered to associate once the distance between their centres is less than ϵ, i.e., they are within a sphere of radius $\epsilon/2$ around the origin. The volume of this sphere is

$$V_a = \frac{4}{3}\pi \left(\frac{\epsilon}{2}\right)^3. \tag{3.14}$$

We normalise the number of states to the total volume

$$V = \frac{4}{3}\pi \left(\frac{d}{2}\right)^3. \tag{3.15}$$

The normalised number of states in which two particles associate is then given by

$$n_a = \frac{V_a}{V} = \left(\frac{\epsilon}{d}\right)^3. \tag{3.16}$$

The energy of the system at this association state is minimised, thus $E_a = J$, with the corresponding Boltzmann factor

$$B_a = e^{-J/(k_B T)}. \tag{3.17}$$

The normalised number states in which particles are not associated is

$$n_S = \frac{V - V_a}{V} \approx 1, \tag{3.18}$$

for a small interaction radius ($\epsilon << d$). The energy of the system when the particles are apart is $E_S = 0$, and the Boltzmann factor is $B_s = 1$.

Therefore the association probability from Eq. (3.13) is then

$$P_a = \frac{\left(\frac{\epsilon}{d}\right)^3 B_a}{1 + \left(\frac{\epsilon}{d}\right)^3 B_a} = \frac{1}{Ad^3 + 1}, \tag{3.19}$$

where $A = e^{J/(k_B T)}/(\epsilon^3)$ is a constant with constant temperature. Equation (3.19) describes the probability of two particles, separated by a distance $\leq d$, being associated in an equilibrium system.

3.4 Test of the Analytical Result Versus Computer Simulations

In the previous Sect. 3.3, we derived an analytical expression for the probability of observing two dots in a joined configuration. Our assumption is that we can treat the system to be at equilibrium and that the state of having both particles in a joined configuration minimises the energy of the system. In this section, we test our theory with help of computer simulations, and we use an algorithm known as Metropolis Monte Carlo simulations [14].

It allows for testing the different configurations of the system and the sampling of transitions from one state to another depending on the binding energy J, e.g. from associated to non-associated particle configurations. For the simulations here we simplify to problem by placing the particles on a cubic grid. For example a particle can be at position (0, 5, 0), i.e. particle positions are direct integer coordinates in our simulations. If both particles have the same coordinates that means they occupy the same grid point and we recognise them to be in an associated configurations and the energy of the system is E, or they are not associated and the energy of the system is 0 otherwise. Note that the interaction radius is thus of the order of a unit cell. When a new, random configuration for the particles is chosen the energy is calculated and whether or not the new configuration is accepted depends on the energy from the old to the new configuration. Specifically, the algorithm is outlined as follows:

1. Generate a random configuration Z_0 for both particles to be at some coordinates that satisfies the maximum separation constraint, i.e. their distance must be $\leq d$.
2. Determine the corresponding energy J_0 of the system, i.e. test whether particles are associated or not.

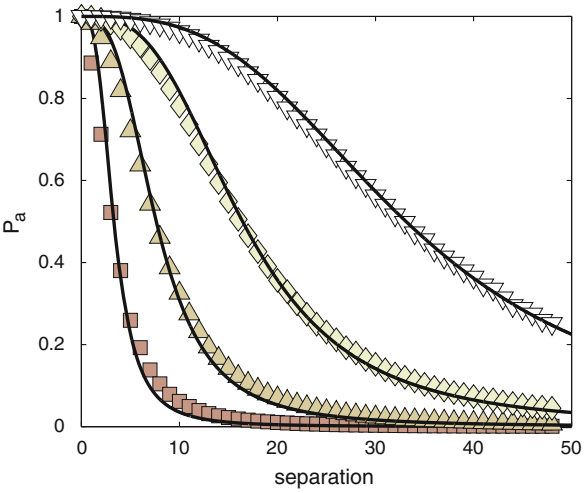

Fig. 3.11 Metropolis-Monte Carlo simulations. Fitting of the model for the association probability $P_a(d) = 1/(Ad^3+1)$ [Eq. (3.19)] of individual replisome pairs to simulation results. For simplicity we set $k_BT = 1$, so that $A = \exp(-J)$. Particles move on a three dimensional grid with grid points along a dimension indicated by the maximum separation. The association energy J increases for the different simulation results: $J = -4$ (*squares*), $J = -6$ (*upward triangles*), $J = -8$ (*diamonds*), $J = -10$ (*downward triangles*). The solid lines indicate the best fit of the simulation data with the association model Eq. (3.19); their values in ascending order are: -3.6, -6.1, -8.4, -10.5

3. Choose a new, random configuration for both particles again under the constraint that both particles can only have some maximum separation d.
4. Determine the energy J_1 for this new configuration.
5. Calculate the energy difference of both configuration $\Delta J = J_0 - J_1$ and the corresponding Boltzmann factor: $F_B = \exp\{\frac{\Delta J}{k_BT}\}$.
6. If $F_B > 1$, the new configuration is accepted.
7. If $F_B < 1$, the new configuration is a priori not immediately accepted.
8. Test a uniformly chosen random number $w \in [0, 1)$ against F_B:
   (a) If $w < F_B$ accept the new configuration.
   (b) If $w > F_B$ reject the new configuration, and keep the old one.
9. Chosen configuration becomes Z_0. Continue the simulation and return to step 3.

We simulate the process in accordance with the rules above as well as under the constraints of maximum separation and association energy. For simplicity we set $k_BT = 1$ so that F_B becomes $\exp(\Delta J)$. We simulate for a series of parameters ranging over $J = -4, -6, -8, -10$ and plot the result for the probability of particle association in Fig. 3.11. We find that if we fit our numerical results to the equation of particle association Eq. (3.19), we can recover the energies we had used in these simulations. Although the fit is does not match to 100 % with the assigned binding energies, we remark that we have simplified the problem from a spherical, continuous

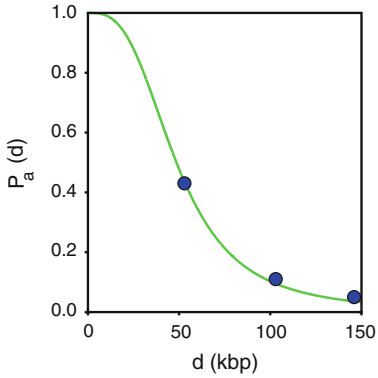

Fig. 3.12 Fitting of the model for the association probability $P_a(d) = 1/(Ad^3 + 1)$ [Eq. (3.19)] of individual replisome pairs with another. In experiments (*blue circles*) origin loci were tagged and their association frequency was determined depending on their chromosomal distances d from each other. The fitting for the single model parameter results in $A = 8.7 \cdot 10^6\,\mathrm{kbp}^{-3}$

one to a cubic, discrete one. This leads to slight differences in our fit to Eq. (3.19) and the actual parameters used in our simulations. Overall our simulation confirms the trend that we expect for the probability of particle association given by Eq. (3.19), so that we can now use our formula for biological relevant experimental data for sister replisome pair associations, and the investigation of their dependence on maximum allowed separation (cf. Saner et al. [8]).

3.5 Fit to Experimental Data of Replisome Association

We fit Eq. (3.19) to the biological data of probability of replicons grouping in the three yeast strains in order to determine the single parameter A. The function $P_a(d)$ fits the data well (Fig. 3.12), with the best fitting for $A = 8.7 \cdot 10^{-6}\,\mathrm{kbp}^{-3}$ ($\mathrm{R}^2 = 0.99$, Fig. 3.12). The binding energy of two sister replisome pairs is $J = k_B T \ln(A^3) = -5.1\,k_B T = -12.5\,\mathrm{kJ/mol}$, for the best-fitting A, $\epsilon = 90\,\mathrm{nm}$ and $T = 298.2\,\mathrm{K}$. Here we estimate the diameter of a single sister replisome pair from the minimum size of a replication factory of $\epsilon = 90\,\mathrm{nm}$ [7]. We again apply the chromatin packaging ratio of 10 nm/kb to account for the actual distance between replicating dots [12]. The calculated binding energy of sister replisome pairs ($-12.5\,\mathrm{kJ/mol}$) is in the range of a typical weak protein–protein interaction [15, 16]. It is also in agreement with the estimated energy for the association of DNA polymerases bound on two replication origins [17]. So there is a relatively strong force which keeps replisomes together in factories once they meet. The experimental data of Saner et al. [8] show that dots randomly co-localise and stay closely together for periods of about 2 min. Their data also shows that when loci undergo replication in the same factory their movement is more restrained than compared to those loci that do not co-localise

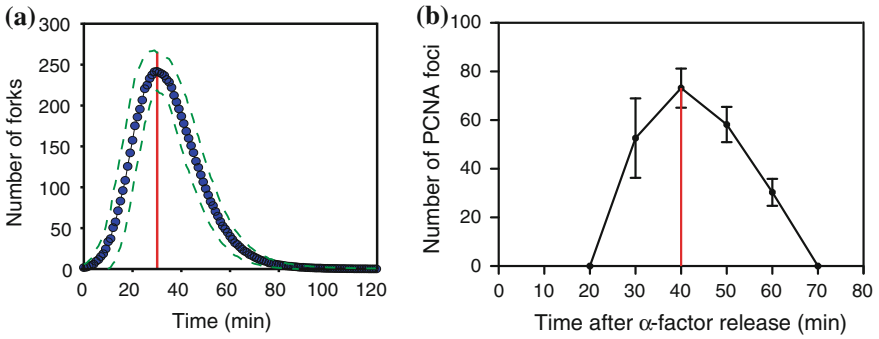

Fig. 3.13 **a**, **b** In silico and in vivo (respectively) distributions of the number of active replication forks. The experimental distribution **b** shows the number PCNA foci (*dots*) which is a unique component of a replication fork. This number is however incomplete and requires the in silico data so that it can be related to the number forks present in vivo. We chose to relate this number at the peak of S-phase (*red line*), when most replication forks are present in the cell

during replication [8]. Between intervals of co-localisations, the loci are seen to move rapidly before and after replication independent from their localisation during replication (see for instance Fig. S2 in Saner et al. [8] and also Kitamura et al. [6]). This then further suggests a strong interaction of replisomes as we calculate here.

3.6 Genome-Wide Replication Data and the Number of Forks Per Factory

Next we establish the size distribution of replication forks that correspond to microscope images of the distribution of PCNA foci (see Fig. 3.4 on page 53). In experiments, cells were observed at the peak of S-phase, i.e. when most forks are active corresponding to 40 min after α-factor release.

This data showed that the number of replication factories increased to a peak value of 73 ± 8 (mean $\pm$ standard deviation) in mid-S-phase (Fig. 3.13b). We next evaluated the number of replisome pairs present in each replication factory. Published replication profiles showing the replication timing of the whole genome [9, 18] are an average from a large number of cells and do not accurately represent replication in individual cells. To estimate the total number of forks, we used data, which Hawkins et al. kindly provided to us [19–21]. This fitting determined origin parameters such as competence, mean activation time, and standard deviation of activation timing for origins in *Saccharomyces cerevisiae* (unpublished data); excluding ribosomal DNA. We applied Hawkins et al.'s data to our dynamical model which simulates stochastic origin activation, according to these parameters, as well as fork progression at a speed of 1.5 kbp/min [18]. In simulations, origins are first selected to become licensed or not; which is done by testing a uniform random number against the competence value

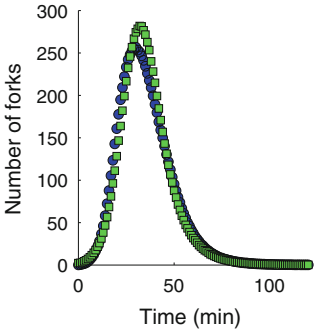

Fig. 3.14 The number of replication forks during S-phase using a Hill- (*blue circles*) or Gaussian-type (*green squares*) activation time distribution. We simulate the data without using a variability of ±4 min in onset of S-phase as we do in Fig. 3.13a. The data is in good qualitative agreement with Fig. 3.13b independent of the activation type used

of each origin. Second, origins are assigned their activation times which are drawn randomly from their activation time distribution with mean and standard deviation. Finally, a point in time during S-phase is selected and fork progression up to then is recorded, e.g. within an array representing a chromosome '0' marks a places with unreplicated DNA, '1' represents replicated DNA. The simulation runs per chromosome and the number of active forks results as the number of edges of regions with '1'. Repeating the simulations several times produces the statistics of fork numbers at a particular time which is plotted in Fig. 3.13a. This allows to correlate the size distribution of replication foci which had been measured earlier (Fig. 3.4a). Using our simulation, we estimated that 242 ± 24 (mean ± standard deviation) replication forks were present at the peak of DNA replication (Fig. 3.13). We also remark that using a Hill-type function for origin activation over time or some other type such as a Gaussian origin activation time function will not change our result. Either type shown in Fig. 3.14 is in good qualitative agreement with the experimental distribution in Fig. 3.13b.

Using published data [22, 23], we estimated that 60 replication forks were present, on average, in the ribosomal DNA region. This brings the estimate to 302 forks (242 + 60) to be found in the whole nucleus at the peak of DNA replication. In the cell imaging data these 302 replication forks are assigned to each of the replication factories, assuming that the integrated GFP-PCNA signal in each factory is proportional to the number of forks it contains (Fig. 3.16). We also show the distribution of the number of forks that are found in a particle cell at 25, 30, and 35 min in Fig. 3.15. It becomes apparent that at the peak of S-phase (30 min), the distribution is thinner in comparison to some earlier or later times. In this way, we are able to estimate the in vivo number of replisome pairs (at sister replication forks) present in each replication factory—in particular at the experimental relevant time at the peak of S-phase.

Next, we relate our simulated fork profiles, that contain information of fork position along the DNA at the peak of S-phase, to the in vivo replication factory dis-

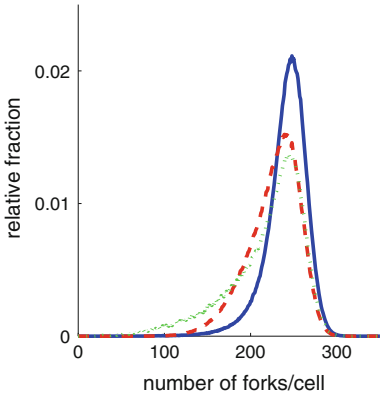

Fig. 3.15 In silico distributions of the number of active replication forks per cell at 25 min (*green dotted line*), 30 min (*blue solid line*), and 35 min (*red dashed line*). The data indicates that at the peak of S-phase, i.e. at 30 min, the distribution centres at around 242 forks per cell

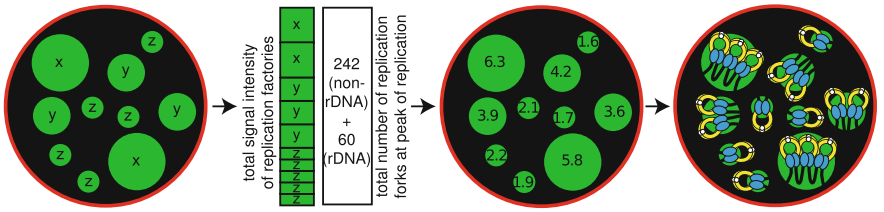

Fig. 3.16 Schematic diagrams explaining how the number of replisomes (at replication forks) was estimated at each factory. The intensity of each replication factory from the first sequence was correlated with simulation data for the number of replication forks present at the peak of S-phase (242 forks + 60 forks in ribosomal DNA regions) in the second sequence. This binned number of the third sequence was binned to even integers as sister replisomes associate [6]. This represents the number of replication forks per factory in the fourth sequence

tribution. We derived the dependence of grouping probability on fork distances in the previous section [Eq. (3.19)] which we now apply to our in silico fork distributions. This establishes an in silico distribution of sister replisome pairs per replication factory in the manner depicted in Fig. 3.17. Specifically, we ran the simulation in one million cells and took snapshots of replisome positions on chromosomes at the peak of replication after cells had entered S-phase with 4 min variation to allow for noise in cell synchronisation during experiments. Based on these snapshots, we determined whether adjacent sister replisome pairs were grouped in the same factory or not, depending on the chromosomal distance between them and corresponding probability of grouping [Fig. 3.12 and Eq. (3.19)].

Let us designate each sister replisome pair along a chromosome as A, B, C, D, …etc. in order from left to right. To determine whether A and B are grouped to the same factory, we drew a uniformly distributed random number from (0, 1], which is tested against the distance-dependent value $P_a(d)$ of those two pairs. If the random

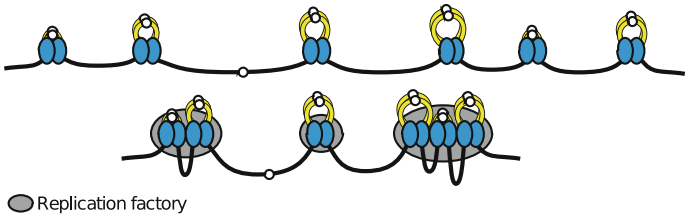

Fig. 3.17 Pairing of replication forks into replication factories. The top panel shows the genomic distances between active replicating units (replication forks in *blue*). Using Monte-Carlo simulations along with Eq. (3.19) one can then establish the number of active replication forks that associate within a replication factory

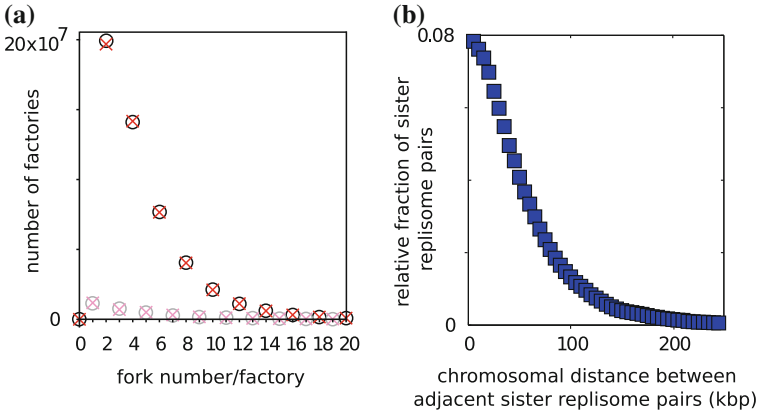

Fig. 3.18 In silico fork and fork distance distributions. **a** The computer algorithm used for pairing neighbouring replication forks into factories is directionally independent, i.e. it does not depend whether association starts from the left (*circles*) or the right (*crosses*) end of a chromosome. It produces the same fork per factory distribution. **b** Distribution of the distance (replicated DNA is not counted) between neighbouring sister replisome pairs along a chromosome, obtained from the simulation. Relative fractions of the pairs at the indicated distance (each 5 kbp window) are obtained from one million simulations, at the peak of DNA replication (at 30 min in Fig 3.13)

number was below $P_a(d)$, the pairs are assumed to be part of the same factory. Next, we examine the association of pairs B and C in the same way. If A and B are in the same factory and if B and C were in the same factory, then we conclude that A, B, and C are grouped together in the same factory; if not then C has the chance to form a factory with D etc. We performed this pairwise clustering of adjacent sister replisome pairs into factories in the rightward direction along each chromosome. Nonetheless, we also confirmed that clustering in the left direction gave a very similar result (Fig. 3.18).

In this study, we assume that sister replisomes are always associated with each other during replication of a relevant replicon. This assumption was based on Kitamura et al. [6] previous results that sister replisomes were associated in vivo

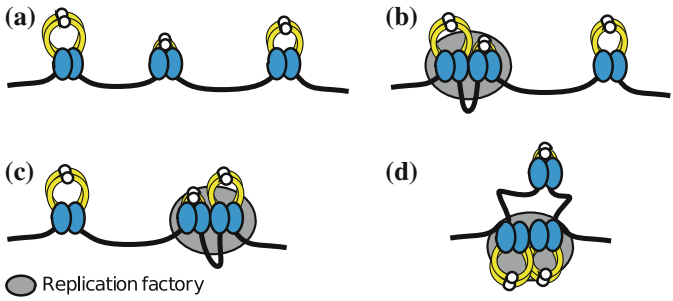

Fig. 3.19 Factory grouping for an example of three neighbouring replisome pairs. We consider three directly adjacent replisome pairs. They are initially isolated (**a**). Some other possible configurations are the immediate neighbours association (**b**) and (**c**), or the association of the farthest neighbours (**d**). We show in the main text that this configuration has a small likelihood

in most of the cells. We also assume that, when two replisome pairs encounter one another (head-on-head fork collision and coalescence), one sister replisome in each pair disappears, leaving the remaining two replisomes associated. This mechanism allows the new pair to undergo further replication. This assumption is consistent with a low energy state of associated replisome pairs; i.e., once two pairs become associated, we can expect that they stay associated for some time (see also diffusion time scales in Sect. 3.2, page 53). Nonetheless, in the above mathematical simulation, we observe also a low number of cases where one replisome is present without its sister, producing an odd number of forks per factory (Fig. 3.18a). This happens when one replisome has completed replication at the end of a chromosome (which is linear) while its sister is still engaged in replication. This led to generation of a small number of replication factories containing odd numbers of replisomes (Fig. 3.18a). However, for a direct comparison of the distribution of sister replisome pairs in factories obtained from in vivo and in silico data, we partitioned factories with odd numbers of forks (replisomes) proportionally to the nearby categories with even numbers of forks. For example factories with three forks were recategorised to those with two and four forks proportionally to their factory numbers.

In the above mathematical modeling, we assume that replisome pairs A and C only associate when both A/B and B/C associate. In other words, we consider association between immediate neighbours but not between others. It is actually difficult to consider direct association between A and C because we would need to consider all possible permutations: the presence and absence of A/C (Fig. 3.19d) association separately depending on whether A/B (Fig. 3.19b) and B/C (Fig. 3.19c) association is present or not. Doing this for genome-wide simulations is impractical. Nonetheless, our assumption—direct pairwise association only—is justified only when A/C (second neighbour) association is relatively low compared with A/B and B/C (direct neighbour) association. We test this in a simplified case of equal chromosomal distances d between A and B and between B and C. The ratio of the A/C association probability to the A/B and B/C association probability is $\eta = P_a(2d)/(2P_a(2d))$.

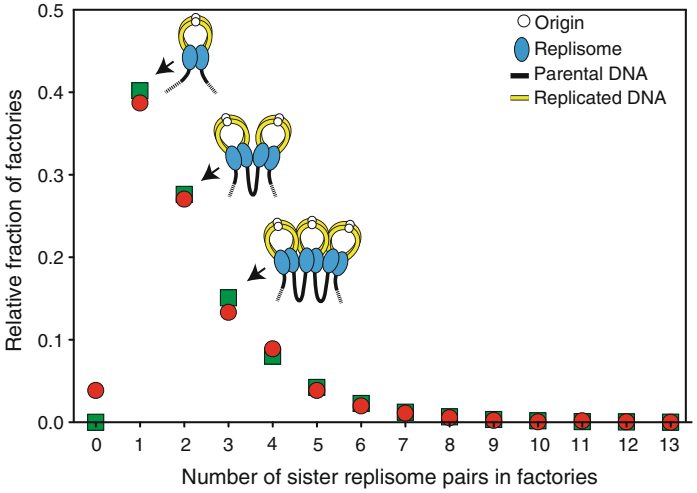

Fig. 3.20 Genome-wide association probabilities of forks per factories. Simulations are shown as *green squares* and in vivo observation are shown as *red circles*. There is an offset at the origin in the experimental data. This is an artifact of noise in the experimental technique. Depicted are also the number of sister replisome pairs (*ovals*), replicated (*yellow*) as well as unreplicated DNA (*black*) and the origin of replication (*small white circles*)

Considering the median chromosomal distances between two neighbouring replisome pairs (Fig. 3.18b) $d = 36\,\text{kbp}$ and $\text{A} = 8.7 \cdot 10^{-6}\,\text{kbp}^{-3}$ equates to $\eta = 0.17$. From first principles, $\eta = 0.17$ is not negligible however taken into account the accuracy of the experimental procedure we estimate this to be of similar magnitude, and also an association of A/B and B/C is about six times more likely than A/C. In the initial experimental setting of observing two dots it is for instance not possible to distinguish when A/C are seen associated whether this is in a configuration of A/C, A/B/C. So there is also an intrinsic error introduced in the measurement when reporting A/C association.

In this way, we are able to compare in vivo and in silico estimates of the number of replisome pairs in each replication factory as shown in Fig. 3.20. Simulations and microscopy observation are very similar. Thus, from the frequency that adjacent sister replisome pairs associate with one another we are able to accurately recapitulate the genome-wide distribution in replication factories by assuming stochastic assembly of replicons. Our result shows that it is mainly neighbouring replicons on a chromosome that are brought together in factories; the association of replicons is random. The result further suggests that factory organisation is intra-chromosomal (replicons on the same chromosome), albeit other factors constituting to inter-chromosomal factory formation for a minor part of the population of factories.

In Fig. 3.21, we also compare the experimentally obtained distribution with in silico experiments at some time before (25 min) and after (35 min) the peak of S-phase (30 min). The tail of the distribution for large factories still matches the experimental well, however diverges for the data point when two sister replisome pairs associate,

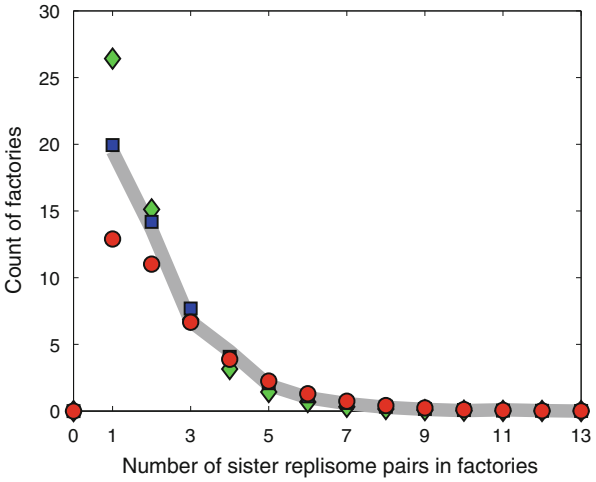

Fig. 3.21 Genome-wide association of sister replisome pairs per factories at different times in silico. Simulations are shown as an average of 10,000 in silico cells. This is done by taking snapshots at different times for each of the 10,000 individual simulations: t = 25 min (*green diamonds*), t = 30 min (*blue squares*), t = 35 min (*red circles*). The *grey line* shows the trend of the experimental from the count of two sister replisome pairs per cell (as seen in Fig. 3.20). For an actual comparison of the different settings we here show the actual count of factories rather than the relative fraction as is done in Fig. 3.20

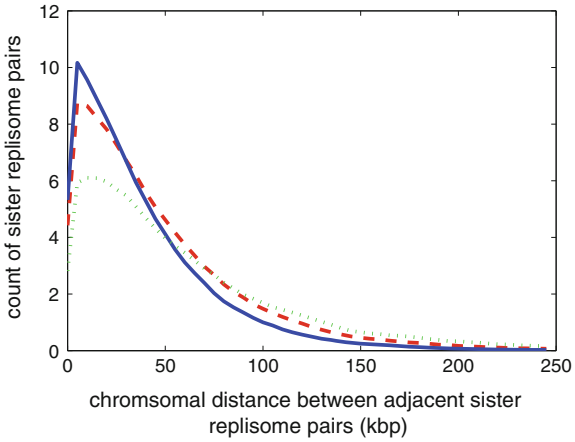

Fig. 3.22 In silico fork distribution of the distance (replicated DNA is not counted) between neighbouring sister replisome pairs along chromosomes. The count of the pairs at the indicated distance (each 5 kbp window) are obtained at t =25 min (*green dotted line*), t = 30 min (*blue solid line*) at the peak of DNA replication (cf. Fig 3.13), and at t =35 min (*red broken line*). This was done for 10,000 simulation runs for each time point

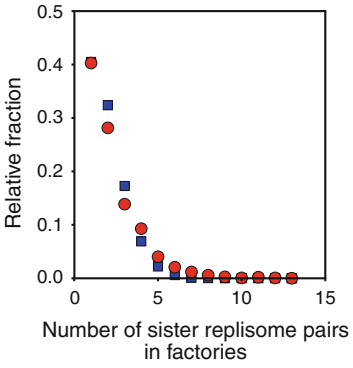

Fig. 3.23 The distribution of in vivo sister replisomes per replication factory (*red circles*) approximately fits a Poisson distribution with $\lambda = 1.60$ $R^2 = 0.98$ (*blue squares*)

i.e. one pair of those. Along with Fig. 3.15 as well as with Fig. 3.22 the data indicates that the immediate neighbour interactions lead to their increase or decrease over the course of replication. Its consequence is then felt mostly in the number of two sister replisome pairs because their formation is driven by forks are either being further or closer away from a neighbouring one. This is in agreement with recent data published by Cisse et al. [24] who showed that replication factories constantly become assembled and disassembled over time course of S-phase rather than displaying a completely fixed entity within the cellular nucleus. We propose that it is a natural consequence of fork coalescence events (decreasing the number of two sister replisome pairs), or for example the nucleation nearby some other fork which then leads to the formation of factories of size two. A further point of consideration is that in a factory containing many sister replisome the energy required to break it up into to single individual sister replisomes has to be rather large; although this has to be tested in a different kind of simulation that studies their dynamics over the entire course of S-phase at an individual cell level instead as we did here where we took an equilibrium assumption for our model. Constant assembly and disassembly also occurs for other structures, e.g. the growth of filamentous structures inside cells, the formation of centromeres, and thus is of particular interest to further understanding of physical mechanisms that lead and control cellular functions.

In summary, our result is in line with the current biological model [11, 25] and it is consistent with observed clustering of active replicons on DNA fibre [26] as well as a high rate of association of neighbouring DNA sequences observed in chromosome conformation capture assays [27]. We also supplement this random association hypothesis showing a fit of *zero-truncated* Poisson distribution, i.e. a Poisson distribution under the assumption that there is no observation for a count of $k = 0$ forks per factory. The probability distribution is then given by

$$P(Z = k|X > 0) = \frac{P(X = k)}{P(X > 0)} = \frac{\lambda^k e^{-\lambda}}{k!(1 - e^{-\lambda})}, \tag{3.20}$$

with $\lambda = 1.60$ as a result of a least-squares fit with the in vivo distribution and Eq. 3.20 (Fig. 3.23).

3.7 Summary

We have shown that replisomes associate randomly with each other using an adiabatic assumption for our model, and we verify our analytical results numerically using Metropolis-Monte-Carlo stimulations. A stochastic assembly mechanism may provide robustness to factory organisation. It is relatively easy to establish—all that is required is that some replisome components have an affinity for another replisome component. In a deterministic assembly scheme, failure to incorporate one component might cause failure of the entire factory network, whereas in a stochastic scheme, each individual interaction is independent of the status of the others. This has particular importance for example in human and animal cells responding to replication stress when a replication factory defines the boundary, inside of which dormant origins can initiate and complete replication for the region between two stalled replication forks [28, 29].

In addition to organising DNA replication, replication factories (foci) are likely to represent a fundamental feature of chromosome organisation [11, 30]. Using *Saccharomyces cerevisiae* as a model organism for our mathematical modelling, we find that individual replication factories creating replicons are highly variable from cell to cell. Their group size also depends on a particular of the cell size as distances from one replisome to another constantly changes and fork movement bringing replisome pairs closer into contact, hence promoting the formation of replication factories. Our results show adjacent replicons assemble stochastically and stay associated together to maintain replication factories in a stable manner. Their formation is also essential to then build up replication factories containing a larger number of sister replisome pairs. Our study elucidates the importance of not only organisation of DNA replication within the nucleus, but also to general mechanisms by which chromosomes organise sub-nuclear structures such as transcription factories and repair foci [31, 32]. Their further investigation is required, especially in light of new experimental data which shows that transcription factories are very dynamic structures which constantly assemble and disassemble inside a cluster with a typical life time of 5 s [24]; rather than sticking together for long times (>2 min) as is the case of replication factories here.

References

1. M. Tark-Dame, R. van Driel, D.W. Heermann, Chromatin folding-from biology to polymer models and back. J. Cell Sci. **124**(6), 839–845 (2011). doi:10.1242/jcs.077628
2. V. Dion, S.M. Gasser, Chromatin movement in the maintenance of genome stability. Cell **152**(6), 1355–1364 (2013). doi:10.1016/j.cell.2013.02.010
3. R.E. Boulos, A. Arneodo, P. Jensen, B. Audit, Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. Phys. Rev. Lett. **111**(11), 118102 (2013). doi:10.1103/PhysRevLett.111.118102
4. M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, M. Nicodemi, Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl. Acad. Sci. U. S. A. **109**(40), 16173–16178 (2012). doi:10.1073/pnas.1204799109

5. S. Hahn, D. Kim, Physical origin of the contact frequency in chromosome conformation capture data. Biophys. J. **105**(8), 1786–1795 (2013). doi:10.1016/j.bpj.2013.08.043
6. E. Kitamura, J.J. Blow, T.U. Tanaka, Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. Cell **125**(7), 1297–1308 (2006). doi:10.1016/j.cell.2006.04.041
7. D. Baddeley et al., Measurement of replication structures at the nanometer scale using super-resolution light microscopy. Nucleic Acids Res. **38**(2), e8 (2010). doi:10.1093/nar/gkp901
8. N. Saner et al., Stochastic association of neighboring replicons creates replication factories in budding yeast. J. Cell Biol. **202**(7), 1001–1012 (2013). doi:10.1083/jcb.201306143
9. N. Yabuki, H. Terashima, K. Kitada, Mapping of early firing origins on a replication profile of budding yeast. Genes Cells **7**(8), 781–789 (2002)
10. K. Sneppen and G. Zocchi, Timescales for target location in a cell. Phys. Mol. Biol. 178–182 (2005)
11. R. Berezney, D. D. Dubey, and J. A. Huberman, Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma **108**(8), 471–484 (2000)
12. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation. Science **295**(5558), 1306–1311 (2002). doi:10.1126/science.1067799
13. P Heun, T Laroche, K Shimada, P Furrer, and S. M. Gasser, Chromosome dynamics in the yeast interphase nucleus. Science 294(5549), 2181–2186 (2001). doi:10.1126/science.1065366
14. D.P. Landau, K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2005). doi:10.1017/CBO9780511614460
15. C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, M.D. Eldridge, Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins **33**(3), 367–382 (1998). doi:9829696
16. K. Rippe, Dynamic organization of the cell nucleus. Curr. Opin. Genet. Dev. **17**(5), 373–380 (2007). doi:10.1016/j.gde.2007.08.007
17. D. Marenduzzo, C. Micheletti, P.R. Cook, Entropy-driven genome organization. Biophys. J. **90**(10), 3712–3721 (2006). doi:10.1529/biophysj.105.077685
18. M.K. Raghuraman et al., Replication dynamics of the yeast genome. Science **294**(5540), 115–121 (2001). doi:10.1126/science.294.5540.115
19. A.P.S. de Moura, R. Retkute, M. Hawkins, C.A. Nieduszynski, Mathematical modelling of whole chromosome replication. Nucleic Acids Res. **38**(17), 5623–5633 (2010). doi:10.1093/nar/gkq343
20. R. Retkute, C.A. Nieduszynski, A. de Moura, Mathematical modeling of genome replication. Phys. Rev. E **86**(3), 031916 (2012). doi:10.1103/PhysRevE.86.031916
21. M. Hawkins, R. Retkute, C.A. Müller, N. Saner, T.U. Tanaka, A.P. de Moura, C.A. Nieduszynski, High-resolution replication profiles define the stochastic nature of genome replication initiation and termination. Cell Rep. **5**(4), 1132–1141 (2013). doi:10.1016/j.celrep.2013.10.014
22. M.H. Linskens, J.A. Huberman, Organization of replication of ribosomal DNA in Saccharomyces cerevisiae. Mol. Cell. Biol. **8**(11), 4927–4935 (1988)
23. P. Pasero, A. Bensimon, E. Schwob, Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. Genes Dev. **16**(19), 2479–2484 (2002). doi:10.1101/gad.232902
24. I.I. Cisse et al., Real-time dynamics of RNA polymerase II clustering in live human cells. Science **341**(6146), 664–667 (2013). doi:10.1126/science.1239053
25. P.J. Gillespie and J.J. Blow, Clusters, factories and domains: The complex structure of S phase comes into focus. Cell Cycle **9**(16) (2010). doi:10.4161/cc.9.16.12644
26. S. Tuduri, H. Tourrière, P. Pasero, Defining replication origin efficiency using DNA fiber assays. Chromosome Res. **18**(1), 91–102 (2010). doi:10.1007/s10577-009-9098-y
27. Z. Duan et al., A three-dimensional model of the yeast genome. Nature **465**(7296), 363–367 (2010). doi:10.1038/nature08973
28. X.Q. Ge, J.J. Blow, Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories. J. Cell Biol. **191**(7), 1285–1297 (2010). doi:10.1083/jcb.201007074

29. A. M. Thomson, P. J. Gillespie, and J. J. Blow, Replication factory activation can be decoupled from the replication timing program by modulating Cdk levels. J. Cell Biol. 188(2), pp. 209–221 (2010). doi:10.1083/jcb11037
30. D.A. Jackson, Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J. Cell Biol. **140**(6), 1285–1295 (1998). doi:10.1083/jcb.140.6.1285
31. M. Lisby and R. Rothstein, DNA damage checkpoint and repair centers. Curr. Opin. Cell Biol. 16(3), pp. 328–334 (2004). doi:10.1016/j.ceb.03.011
32. H. Sutherland, W.A. Bickmore, Transcription factories: gene expression in unions? Nat. Rev. Genet. **10**(7), 457–466 (2009). doi:10.1038/nrg2592

Chapter 4
Summary and Conclusions

Life is only possible because the genomic information in a cell's chromosomes is copied from one generation to the next, by means of DNA replication. Components of the complex biological machinery responsible for replication in eukaryotes act in concert to ensure that replication takes place rapidly and accurately, at the right time within the cell cycle, and crucially that every piece of the chromosome is replicated once and only once per round of the cell cycle. The two key elements of this process—establishing the starting points for replication (*origins*) and the activation of replication forks at those origins—are stochastic events which occur during two distinct phases of the cell cycle. From each origin, replication forks propagate from either side, synthesising DNA at an apparently fixed speed. Hence, the time required to replicate the DNA content of a cell is dictated by the distances between origins.

The work presented here examined theoretically, and in close collaboration with experimentalists, how replication can be brief and on time as is the case in nature, how robust timing is possible under fluctuating and noisy conditions, and how replication forks organise spatially within the cell.

In brief, we developed a timing–model of DNA replication which has identified optimal origin positions induced depending on failure probabilities and evolutionary pressure. The emergence of such optimal positions is still to be investigated, however previous experimental results in *Xenopus laevis* suggest that this could be achieved as an effect of limitting space on DNA due to pMcm binding. We also introduced a general theory for active forks in this work which shows that their random assembly leads to higher–order structures or *replication factories*, and for the first time theory accurately predicts factory size distributions from genome–wide yeast data in silico which agree with quantitative in vivo experiments.

In detail, the work here described those aspects as follows.

1. The balancing act to spread out origins in a certain manner to compensate for variations in activation timing and lack of proteins stochastically binding to sequences. We showed both analytically and through numerical simulations (similar to a 1D nucleation and growth process) that there exists two regimes for origins, either positioned together in groups spaced far away from the next, or

J. Karschau, *Mathematical Modelling of Chromosome Replication and Replicative Stress*,
Springer Theses, DOI 10.1007/978-3-319-08861-7_4

as equally–scattered single origins depending on the uncertainty when activation occurs. We applied the model to known origin locations in yeast and showed that grouping is a means of organisation driven by evolutionary pressure. The model is able to reproduce origin distributions of Xenopus which are thought to be random, and showed contrarily that grouping must occur in order to swiftly complete replication. The model also holds when considering a circular DNA topology as for instance archaeal genomes have, as well as if applied to the whole replication profiling data of yeast.

2. The second topic aimed at the organisation of replication forks within the cellular nucleus. For simplicity, cartoons often depict DNA replication on a straight one–dimensional line. In fact we deal with a polymer that is packed and modified on different levels yielding higher order structures of organisation. A project in experimental collaboration focussed on the aspect of spatial organisation of active replication forks during the time–course of replication. These forks are observed to organise in clusters of *replication factories* which we investigated by describing the process with a particles on a string model. We calculated analytically the probability for forks to meet using Boltzmann–statistics. The model was then used to describe properties of measured experimental distributions such as fork numbers per cluster during the DNA synthesis phase. Analysis was extended to the whole yeast–genome which yielded a near–perfect match with the data suggesting that actively replicating units of DNA randomly associate with each other to form *replication factories*.

Particular emphasis on the stochastic processes that work here at different scales allowed us to describe key aspects of replication. The models developed here allow for further extensions to describe DNA replication at various scales of organisation; e.g. at the DNA sequence level, an investigation of origin positioning under perturbing conditions, i.e. epigenetic factors that forbid binding to a chromosomal region, and the effect on replication completion times. At larger scales, formation of these origins is linked to fluctuating levels of proteins, e.g. during embryogenesis or under starvation conditions. More importantly at tissue–level, a treatment of noise in these processes combined with a modelling of replication fork movement could then provide a comprehensive model for tissue homeostasis.

Ultimately, we expect the outcome of such studies to also provide understanding of stochasticity in protein synthesis and usage of cellular resources. The theoretical investigation and new modelling tools can be integrated with experiments. In the long–term view, extensions to the work presented here has potential to address challenges in cancer therapy. For example, a model could be used alongside data to predict critical shifts in the replication pattern that trigger cancer or how to achieve optimal cellular growth conditions. This must be tested using a holistic model incorporating the physical process involved in replication licensing, origin activation. Such a model is then testable against experimental replication timing profiles and yields insight into organisation of DNA replication at a single cell level.