

DISSERTATION

submitted to the

Combined Faculty of
Natural Sciences and Mathematics

of Heidelberg University
Germany

for the degree of
Doctor of Natural Sciences

put forward by

M.Sc. Andreas Hofmann

born in: Bad Mergentheim

Oral examination: 05.02.2020

3D Organization of Eukaryotic and Prokaryotic Genomes

REFEREES:

PROF. DR. DIETER W. HEERMANN

PROF. DR. HEINZ HORNER

Abstract

There is a complex mutual interplay between three-dimensional (3D) genome organization and cellular activities in bacteria and eukaryotes. The aim of this thesis is to investigate such structure-function relationships.

A main part of this thesis deals with the study of the three-dimensional genome organization using novel techniques for detecting genome-wide contacts using next-generation sequencing. These so called chromatin conformation capture-based methods, such as 5C and Hi-C, give deep insights into the architecture of the genome inside the nucleus, even on a small scale. We shed light on the question how the vastly increasing Hi-C data can generate new insights about the way the genome is organized in 3D.

To this end, we first present the typical Hi-C data processing workflow to obtain Hi-C contact maps and show potential pitfalls in the interpretation of such contact maps using our own data pipeline and publicly available Hi-C data sets. Subsequently, we focus on approaches to modeling 3D genome organization based on contact maps. In this context, a computational tool was developed which interactively visualizes contact maps alongside complementary genomic data tracks. Inspired by machine learning with the help of probabilistic graphical models, we developed a tool that detects the compartmentalization structure within contact maps on multiple scales. In a further project, we propose and test one possible mechanism for the observed compartmentalization within contact maps of genomes across multiple species: Dynamic formation of loops within domains.

In the context of 3D organization of bacterial chromosomes, we present the first direct evidence for global restructuring by long-range interactions of a DNA binding protein. Using Hi-C and live cell imaging of DNA loci, we show that the DNA binding protein Rok forms insulator-like complexes looping the *B. subtilis* genome over large distances. This biological mechanism agrees with our model based on dynamic formation of loops affecting domain formation in eukaryotic genomes. We further investigate the spatial segregation of the *E. coli* chromosome during cell division. In particular, we are interested in the positioning of the chromosomal replication origin region based on its interaction with the protein complex MukBEF. We tackle the problem using a combined approach of stochastic and polymer simulations.

Last but not least, we develop a completely new methodology to analyze single molecule localization microscopy images based on topological data analysis. By using this new approach in the analysis of irradiated cells, we are able to show that the topology of repair foci can be categorized depending the distance to heterochromatin.

Zusammenfassung

Zwischen dreidimensionaler (3D) Genomorganisation und zellulären Aktivitäten in Bakterien und Eukaryoten gibt es ein komplexes Wechselspiel. Das Ziel dieser Arbeit ist es, solche Struktur-Funktions-Zusammenhänge zu untersuchen.

Ein Hauptteil dieser Arbeit befasst sich mit der Untersuchung der dreidimensionalen Genomorganisation unter Verwendung neuartiger Techniken zum Nachweis genomweiter Kontakte unter Verwendung von DNA-Sequenzierung der nächsten Generation. Diese auf "Chromosome Conformation Capture" basierenden Methoden, wie 5C und Hi-C, ermöglichen auch auf kleinen Skalen tiefe Einblicke in die Architektur des Genoms im Zellkern. Wir beleuchten die Frage, wie die immer weiter zunehmenden Hi-C-Daten neue Erkenntnisse über die räumliche Organisation des Genoms liefern können.

Zu diesem Zweck stellen wir zunächst den typischen Workflow der Hi-C-Datenverarbeitung vor, um Hi-C Kontaktmatrizen zu erhalten. Wir zeigen mögliche Fallstricken bei der Interpretation solcher Kontaktmatrizen unter Verwendung unserer eigenen Daten-Pipeline und öffentlich verfügbarer Hi-C-Datensätze auf. Anschließend konzentrieren wir uns auf Ansätze zur Modellierung der räumlichen Genomorganisation auf Basis von Kontaktmatrizen. In diesem Zusammenhang wurde ein Werkzeug entwickelt, das interaktiv Kontaktmatrizen und komplementäre Datenspuren visualisiert. Inspiriert durch maschinelles Lernen mit Hilfe probabilistischer graphischer Modelle haben wir ein Tool entwickelt, das die Kompartimentierungsstruktur in Kontaktmatrizen auf unterschiedlichen Skalen erkennt. In einem weiteren Projekt zeigen wir einen möglichen Mechanismus für die beobachtete Kompartimentierung innerhalb von Kontaktmatrizen unterschiedlicher Genome und testen ihn: Dynamische Schleifenbildung innerhalb von Domänen.

Im Kontext der räumlichen Organisation bakterieller Chromosomen präsentieren wir den ersten direkten Nachweis für eine globale Restrukturierung durch langreichweitige Wechselwirkungen eines DNA-bindenden Proteins. Unter Verwendung von Hi-C und Live-Cell Imaging von DNA-Loci zeigen wir, dass das DNA-bindende Protein Rok isolator-ähnliche Komplexe bildet, die das *B. subtilis* Genom über große Entfernungen schleifen. Dieser biologische Mechanismus bestätigt unser Modell, das auf der dynamischen Bildung von Schleifen basiert, die die Domänenbildung in eukaryotischen Genomen beeinflussen. Desweiteren untersuchen wir die räumliche Trennung des *E. coli* Chromosoms während der Zellteilung. Insbesondere interessieren wir uns für die Positionierung des *E. coli* Replikationsursprungs durch seine Wechselwirkung mit dem Proteinkomplex MukBEF. Wir verfolgen dabei einen kombinierten Ansatz aus stochastischen Simulationen und Polymer-simulationen an.

Im Rahmen einer weiteren Kollaboration wurde eine völlig neue Methode zur Analyse von Einzelmolekül-Lokalisationsmikroskopiebildern auf Basis topologischer Datenanalyse entwickelt. Mithilfe dieses neuen Ansatzes zur Analyse bestrahlter Zellen konnten wir zeigen, dass die Topologie von DNA-Reparaturzentren in Abhängigkeit ihrer Entfernung zu Heterochromatin kategorisiert werden kann.

Publications Related to this Thesis

Large parts of this thesis have already been published or are currently under peer-review. Papers in preparation are also listed. (Information as of January 06th, 2020)

- **A. Hofmann** and D.W. Heermann, The role of loops on the order of eukaryotes and prokaryotes. *FEBS Letters* (2015), 589, 2958-2965
DOI: <http://dx.doi.org/10.1016/j.febslet.2015.04.021>

- R.A. van der Valk, J. Vreede, L. Qin, G.F. Moolenaar, **A. Hofmann**, N. Goosen and R.T. Dame, Mechanism of environmentally driven conformational changes that modulate H-NS DNA-bridging activity. *eLife* (2017), 6, e27369.
DOI: <http://dx.doi.org/10.7554/eLife.27369>

- **A. Hofmann** and D.W. Heermann, Processing and Analysis of Hi-C Data on Bacteria. *Bacterial Chromatin. Methods in Molecular Biology* (2018), 1837, 19-31.
DOI: http://dx.doi.org/10.1007/978-1-4939-8675-0_2

- **A. Hofmann** and D.W. Heermann, Deciphering 3D organization of chromosomes using Hi-C data. *Bacterial Chromatin. Methods in Molecular Biology* (2018), 1837, 389-401.
DOI: http://dx.doi.org/10.1007/978-1-4939-8675-0_19

- **A. Hofmann***, M. Krufczik*, D.W. Heermann and M. Hausmann, Using Persistent Homology as a New Approach for Super-Resolution Localization Microscopy Data Analysis and Classification of γ H2AX Foci/Clusters. *International Journal of Molecular Sciences* (2018), 19, 2263.
DOI: <http://dx.doi.org/10.3390/ijms19082263>

- J. Jia, K. Li, **A. Hofmann** and D.W. Heermann, The Effect of Bending Rigidity on Polymers. *Macromolecular Theory and Simulations* (2019), 28, 1800071.
DOI: <http://dx.doi.org/10.1002/mats.201800071>

- **A. Hofmann***, J. Mäkelä*, D.J. Sherratt, D.W. Heermann and S.M. Murray, Self-organised segregation of bacterial chromosomal origins. *eLife* (2019), 8, e46564.
DOI: <http://dx.doi.org/10.7554/eLife.46564>

- **A. Hofmann**, J. Muggenburg, Frédéric Crémazy and D.W. Heermann, Bekvaem: Integrative Data Explorer for Hi-C Data. *Journal of Bioinformatics and Genomics* (2019), 2, 11.
DOI: <http://dx.doi.org/10.18454/jbg.2019.2.11.1>

*equal contribution

- **A. Hofmann**, F.Z. Rashid, F. Crémazy, R. T. Dame and D.W. Heermann, Domain Boundary Detection in Hi-C Maps: A Probabilistic Graphical Model Approach. (2020), in preparation, to be submitted to PLoS ONE.

- G. Dugar, **A. Hofmann**, D.W. Heermann and L.W. Hamoen, Robust interaction between insulator-like domains drives genome organization in bacteria. (2020), in preparation, to be submitted to Nature.

- S. Pandey, S. Sewitz, S. Schalbetter, J. Baxter, S.A. Juranek, V. Guryev, T. Schmitz, **A. Hofmann**, D.W. Heermann and K. Paeschke, Telomerase function and binding at endogenous G-rich regions. (2020), in preparation.

- E. Schwindt, S. Juranek, **A. Hofmann**, D.W. Heermann and K. Paeschke, Changes in the G4 landscape during vegetative growth and meiosis in *S. cerevisia*. (2020), in preparation.

Contents

1	Scope and Intentions	13
1.1	Introduction	13
1.2	Scope of this Thesis	14
1.3	Structure of this Thesis	15
2	Biological Background	17
2.1	Organization of Eukaryotic Chromosomes	17
2.1.1	The Three-Dimensional Genome	17
2.1.2	Domain Organization of the Genome	20
2.1.3	Enhancer-Promoter Interactions	23
2.2	Organization of the Bacterial Nucleoid	24
2.2.1	Bacterial Chromosome Packaging	24
2.2.2	Interplay between Gene Regulation and Genome Folding	25
2.3	Molecular Techniques for Unraveling the Three-Dimensional Genome	26
2.3.1	Fluorescence in Situ Hybridization (FISH)	26
2.3.2	Next-Generation Sequencing (NGS)	26
2.3.3	Chromosome Conformation Capture (3C) Methods	26
2.3.4	ChIP-seq and ChIA-PET	30
3	Theoretical Background	31
3.1	Simulation Methods	31
3.1.1	Polymer Physics	31
3.1.2	Monte Carlo Simulations	34
3.2	Statistical and Conformational Properties	37
3.2.1	End-to-End Distance and Gyration Radius	37
3.2.2	Mean Squared Displacement	38
3.2.3	Bending Rigidity and Persistence Length	39
3.2.4	Confinement	39
3.2.5	Contact Probability Measures	39
3.3	Computational Topology	42
3.3.1	Simplicial Complexes and Barcodes	42
3.3.2	Hausdorff Distance	43
4	Processing and Analysis of Hi-C Data on Bacteria	47
4.1	Introduction	48
4.2	Hi-C Data Processing	49
4.2.1	Mapping to the Reference Genome	49
4.2.2	Quality Control	50

4.2.3	Binning and Contact Matrix Generation	50
4.2.4	Balancing	51
4.2.5	Concluding Remarks	53
4.3	Hi-C Data Assessment	53
4.3.1	Hi-C Data Availability	53
4.3.2	How Many Reads?	53
4.3.3	What Kind of Reads?	54
4.3.4	Resolution of Hi-C Data and the Selection of Restriction Enzymes	56
4.4	Summary	56
5	Deciphering 3D Organization of Chromosomes using Hi-C Data	59
5.1	Introduction	60
5.2	Analyzing Hi-C Contact Maps	60
5.2.1	Comparison of Contact Maps	61
5.2.2	Feature Detection	62
5.2.3	Correlation-based Data Analysis	64
5.3	3D Modeling	65
5.3.1	3D Reconstruction	65
5.3.2	Polymer Simulations	67
5.4	Summary	68
6	Interactive Visualization of Hi-C Data	69
6.1	Introduction	70
6.2	Methods	71
6.3	Application	72
6.3.1	Comparison with Other Hi-C Browsers	72
6.3.2	Visualization Examples	76
6.4	Conclusion	76
7	Domain Boundary Detection in Hi-C Maps	81
7.1	Introduction	82
7.2	Approach	83
7.3	Methods	83
7.3.1	The Model	84
7.3.2	Algorithm	85
7.3.3	Validation of the Algorithm	86
7.4	Results	86
7.4.1	Domain Detection within Hi-C Contact Maps	86
7.4.2	Loop Detection within Hi-C Contact Maps	87
7.4.3	Multi-Scale Structure Identification within Hi-C Contact Maps	88
7.5	Discussion	88
8	The Role of Loops on the Order of Eukaryotes and Prokaryotes	91
8.1	Introduction	92
8.2	Current State of Modeling	93
8.3	Static Loop Domains	94
8.4	Dynamic Loop Interaction within Domains	96
8.5	Effect of Loops on the Nuclear Organization	99
8.6	Conclusion	100

8.7	Methods	100
9	Insulator-like Domains Drive Genome Organization in Bacteria	103
9.1	Introduction	104
9.2	Discussion	111
9.3	Methods	112
9.3.1	Cell Growth	112
9.3.2	Strain Construction	112
9.3.3	SICO-seq	112
9.3.4	RNA Isolation and RNA-seq	113
9.3.5	Mapping and Visualization of SICO-seq, RNA-seq and ChIP Data	113
9.3.6	Motif Density Analysis	113
9.3.7	Chromosome Capture by Hi-C	113
9.3.8	Hi-C Data Mapping and Contact Matrix	114
9.3.9	Hi-C Data Visualization	114
10	Self-organized Segregation of Bacterial Chromosomal Origins	117
10.1	Introduction	118
10.2	Results	119
10.2.1	ori is attracted towards MukBEF foci	119
10.2.2	Model of ori positioning by self-organised MukBEF reproduces mid-cell positioning	122
10.2.3	Preferential loading leads to stable and accurate partitioning	126
10.2.4	Accurate partitioning during growth	128
10.2.5	Directed movement of ori can arise from spatially-dependent looping interactions	130
10.3	Discussion	132
10.4	Predictions	135
10.5	Outlook	136
10.6	Materials and Methods	137
10.6.1	Review of the Model	137
10.6.2	Stochastic Simulations	138
10.6.3	Apparent ori Diffusion Constant and Drift Rate	140
10.6.4	ori Drift and Diffusion Parameters	142
10.6.5	Entropic Repulsion of ori	142
10.6.6	Polymer Simulations	142
10.6.7	Experiments	143
11	Topological Data Analysis for Super-Resolution Localization Microscopy	145
11.1	Introduction	146
11.2	Results	149
11.2.1	SMLM Data Processing	149
11.2.2	γ H2AX Cluster Recognition and Cluster Classification	150
11.2.3	Topological Analysis of the Clusters	153
11.3	Discussion	159
11.4	Materials and Methods	161
11.4.1	Sample Preparation	161
11.4.2	Single Molecule Localization Microscopy (SMLM)	161
11.4.3	Sample Irradiation	162

11.5 Conclusions	162
12 Conclusion and Outlook	163
12.1 Short Summary of the Results	163
12.2 Outlook	165
Acknowledgments	169
Conference/Workshop Participation	170
References	171

Chapter 1

Scope and Intentions

A Short Overview over Topics and Aims

1.1 Introduction

The study of the concerted interplay of structure and function of DNA and chromatin goes back several centuries. How DNA is organized in three dimensions inside the cell nucleus and how this impacts on gene expression is indeed one of the most important questions in cell biology. The human genome project [1] completed in 2003 deciphered the sequence of the over 3×10^{12} base pairs long human DNA. Stretched out and placed end to end that amounts to about 2 meters.

How can this amount of DNA fit inside a micron-sized cell nucleus? And how does the solution to this organizational challenge then further impacts on gene expression or, in

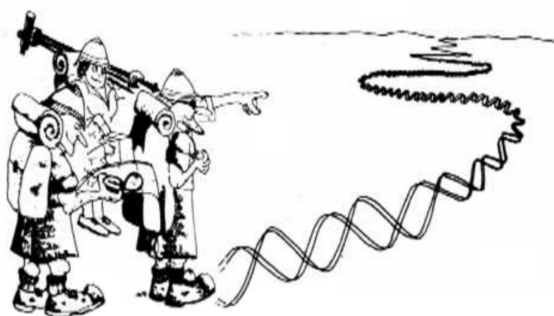


Figure 1.1: Illustration of the enormous size of the genome. Image adapted from the National Human Genome Research Institute (<https://www.genome.gov>).

other words, how can the genetic material can be sufficiently compactified while providing dynamic access to the genetic information relevant for nuclear processes like transcription, replication, DNA repair and recombination? The knowledge of the sequence of the genome alone cannot answer these fundamental questions in cell biology. However, the combination of the knowledge of both sequence and three-dimensional (3D) organization of the genome does provide insights into these questions.

Bacterial genomes face similar questions as eukaryotic genomes: The *Escherichia coli* (*E. coli*) chromosome is approximately 1000 times longer than the confining bacterial nucleoid [2] and replication, segregation and transcription of the genetic material must be in harmony with the needed level of compaction.

What is the influence of the 3D organization of bacterial and eukaryotic genomes on essential cell functions and, vice versa, how do nuclear processes impact on genome folding?

The answers to these questions require research on genome folding and the underlying physical principles. For this purpose, computational modeling coupled to “chromosome conformation capture” (3C) techniques as well as to high-resolution microscopy are being used. Genome-wide 3C methods, such as Hi-C, allow for probing the 3D structure of the genome by determining the number of contacts between any pair of genomic loci, thus generating a genome-wide snapshot of genomic self-interactions [3]. While 3C-based techniques measure contact frequencies between DNA segments for a population of millions of cells, high-resolution (live) cell imaging can provide information on the single-cell level.

1.2 Scope of this Thesis

In this thesis, we want to address important questions for a better understanding of 3D genome organization of both eukaryotes and bacteria by means of computational modeling approaches coupled to experimental evidence gathered from data provided by both our different collaborators as well as by publicly accessible databases. We first target the aspect of analyzing, interpreting and building models based on Hi-C data. Our intention is to work out a clear guidelines and to this end provide two in-house developed tools. Furthermore, we present a physical model that can account for the experimentally observed Hi-C contact maps of both eukaryotic and bacterial genomes. Next, we focus on the spatial segregation of the *E. coli* chromosome during cell division and present our important new findings. Last but not least, we present a fundamentally new approach to analyze high-resolution microscopy images in terms of topology.

Hi-C allows for probing the 3D structure of the genome in great detail like never before. Thanks to massive improvements in throughput, ever-increasing amounts of data are being produced and deposited in publicly accessible databases like Gene Expression Omnibus (GEO) [4]. Although there is a variety of data analysis tools for processing raw Hi-C data [5], the illustration of the final contact maps is not standardized. Most importantly, one should differentiate contact frequency maps and contact probability maps. While contact frequency maps are composed of the individual total number of contacts between any pair of genetic loci, contact probability maps, on the other hand, are doubly stochastic matrices that specify the probability of any pair of genetic loci to be in contact. Consequently, the term “contact maps” is imprecise and should be initially defined. We are abbreviating contact probability maps as contact maps throughout this thesis. Conventionally, contact maps are illustrated as heat maps, what allows setting different colormaps as well as different colorbar ranges. And regardless of how unimpressive this may seem, it has a major impact on the visual inspection as will be discussed later. More importantly, this issue highlights the necessity for a flexible and interactive visualization. Further important points include the assessment of the quality of Hi-C experiments, the comparison of contact maps and the pattern recognition within contact maps.

It is emerging that chromosome folding requires several levels of compaction, from chromatin loops connecting genes and enhancers on the small scale to chromosomal domains and nuclear compartments on the large scale [6]. Recently, high-resolution Hi-C studies of several eukaryotic genomes provide evidence for the existence of intrachromosomal domains, so called “topological domains” [7, 8]. These clusters appearing in the contact maps of such studies are characterized by pronounced long-range interactions between loci within the same domain. In contrast, the crosslinking probability of loci located in adjacent domains is found to be lowered. As a consequence, chromosomes are composed of a string of domains being topologically separated from each other. These findings lead to the question of how domains and loops are interconnected.

The faithful and timely segregation of genetic material is essential for all cellular life. In eukaryotes the mechanism behind chromosome segregation is the well-understood mitotic spindle. In contrast, the mechanisms underlying bacterial chromosome segregation are considerably less understood, but are just as critical for cellular proliferation [9]. The starting point for chromosomal replication, the origin (*ori*), plays a crucial role in chromosome organization as well as in chromosome segregation since its dynamic spatial position defines the position of other chromosomal regions in the nucleoid [10, 11]. The mechanisms underlying *ori* positioning remain unclear [9], especially for the case of *E. coli* [12]. It is known that MukBEF, a functional homolog of ubiquitous Structural Maintenance

of Chromosomes (SMC) complexes, plays a role in both *E. coli* chromosome organization and segregation [13, 14]. However, the interesting question of whether and how MukBEF could position chromosomal origins is not yet known.

Microscopy images are conventionally analyzed as follows: First, images are segmented to recognize multiple objects, then quantitative measurements are carried out and finally, the morphology is characterized by using specialized algorithms [15–17]. However, this task is difficult when the concept of an object is not well-defined, as in the case of microscopy images of the cell nucleus. This particular case requires a new approach to extract the relevant conformational information by topological means which are maintained under different perspectives and different deformations. Such an approach is automatically not restricted to a certain scale, parameter-free and can deal with high-resolution microscopy images.

1.3 Structure of this Thesis

The motivation of the work presented in this thesis is to investigate a wide range of complex biological systems. Even though these systems vary in their dynamics, length scales, structural complexity and functional purpose, all of them can be described by physical models. The aim of this thesis is to analyze complex and big data from state-of-the-art experiments and to develop physical models in order to expand the current knowledge on structure-function relationships in the context of eukaryotic and bacterial cells.

- Since the work presented in this thesis is highly interdisciplinary, the next two chapters introduce selected topics in biology and physics. **Chapter 2** describes the basic biology of eukaryotic and prokaryotic cells and highlights the hierarchy of 3D genome organization. A special focus is put on the two most important experimental methods for the investigation of large-scale DNA organization: 3C-based methods and DNA techniques.
- **Chapter 3** provides an introduction to physical and mathematical methods that are important for the modeling and analysis in the different projects. First, basic concepts on polymer physics and the essential simulation technique, Monte Carlo simulations, are presented. Second, measures for the description of statistical and conformational properties of polymers are introduced. The third major part of this chapter introduces the field of computational topology.
- In **chapter 4** we present the data analysis workflow of Hi-C experiments. This comprises the processing of the Hi-C sequencing data to generate a final contact probability map as well as strategies to assess the quality of Hi-C data sets. We describe the different steps using our own data pipeline and publicly available Hi-C data sets.
- After having introduced the biological and bioinformatic background of Hi-C experiments, we focus on their potential to decipher 3D genome organization in **chapter 5**. To this end, we first describe computational methods for pattern recognition within the contact map and, second, discuss approaches to modeling 3D genome organization.

- In **chapter 6** we present our developed tool to visualize Hi-C contact maps alongside complementary genomic data tracks interactively in any web browser. Besides the technical implementation we show some case studies and point out the advantages of our tool over others.
- We have also developed a probabilistic graphical model for pattern recognition within Hi-C contact maps that is presented in **chapter 7**. Its unique feature is an inherent multi-scale detection of the compartmentalization structure within Hi-C contact maps and its suitability for loop detection.
- In **chapter 8** we focus on the role of loops on the spatial organization of both eukaryotic and prokaryotic genomes. We show that a model based on the dynamic formation of loops within domains can account for the experimentally observed Hi-C contact maps not only of certain mammalian genomes but also of bacterial genomes.
- **Chapter 9** employs the role of loops especially on the 3D organization of bacterial genomes. Using Hi-C and live cell imaging of DNA loci, we show that the DNA binding protein Rok forms insulator-like complexes looping the *B. subtilis* genome over large distances. This biological mechanism agrees with our model developed in chapter 8 and shows that the dynamic formation of loops affects domain formation both in eukaryotic and prokaryotic genomes.
- **Chapter 10** deals with the spatial segregation of the *E. coli* chromosome during cell division. We provide an explanation for the positioning of the chromosomal replication origin region (*ori*) based on the self-organization of the protein complex MukBEF.
- In the final project presented in **chapter 11**, we introduce a new methodology based on topological data analysis to analyze super-resolution localization images. Using cells that were exposed to ionizing radiation, we first show that repair foci can be classified into two major groups dependent on their location in the genome. Subsequently, using our developed method, we analyze how these two groups of repair foci differ in their morphology.
- Finally, in **chapter 12** we provide a concise summary of all research projects and our results. We close with an outline of the future challenges in our field of research.

Chapter 2

Biological Background

In this chapter, the fundamental biological principles of 3D folding of eukaryotic and prokaryotic genomes is discussed. Additionally, we present the most important experimental techniques for analyzing the 3D structure of the genome and also go into detail on the bioinformatics methods related to chromosome conformation capture methods that play an important role throughout this thesis.

2.1 Organization of Eukaryotic Chromosomes

It was the German biologist Ernst Haeckel who first proposed the idea that the nucleus takes care of inheritance in eukaryotic cells. However, the understanding that the nucleus contains deoxyribonucleic acids (DNA), or “nuclein” as it was initially called, was first proposed in 1871 by the Swiss physician and biologist Friedrich Miescher. In 1879 the German biologist Walther Flemming discovered tiny thread-like structures within the nucleus which he named “chromatin” and which were later known as chromosomes. Today, chromatin refers to the combination of DNA and proteins that make up eukaryotic chromosomes and it has been shown that chromosomes play a key role in inheritance: Before cell division, the chromosomes are first replicated, then divided into two sets and finally segregate to the two daughter cells.

Nowadays it is basic textbook knowledge that the entirety of the genetic material of an organism is the genome which includes both: genes, which encode the relevant instructions for producing proteins, as well as non-coding sequences. The process that leads from the genetic information, the DNA, to the required protein is called gene expression and consists basically of two steps: Initially, the DNA segment, that encodes for the needed protein, is read and a complementary RNA copy (mRNA) is built. This initial step, the so-called transcription, is followed by the second step, called translation, in which the mRNA is converted into the required protein.

Furthermore, it is for sure, that the units in the cell nucleus carrying genes are long and highly dynamic molecules that consist of DNA and are called chromosomes. The human genome consists of several ten thousands of genes [18] that are distributed amongst 23 different chromosomes with the first 22 simply numbered from 1 to 22 and the last chromosome called the sex chromosome as it determines the sex of a person. Most cells are diploid, i.e. they possess two homologous copies of each chromosome. As a consequence the human genome is composed of 46 chromosomes.

2.1.1 The Three-Dimensional Genome

Scientists attached great importance to the three-dimensional positioning of genomic elements from early on [19]. With the advent of the FISH technology (see subsection 2.3.1)

which allowed for visualization of the positioning of specific sites within the nucleus [20] the importance of the three-dimensional positioning of genomic elements was realized. This process accelerated once again with the invention of the so-called chromosome conformation capture (3C) technology [21] that can quantify interaction frequencies between selected regions in the genome (see subsection 2.3.3 for an in-depth discussion).

Cell Cycle Dependent Chromosome Packaging

Most eukaryotic cells have a life cycle that involves sexual reproduction that leads to two genetically identical daughter cells. The cell cycle consists of two main parts: interphase and mitosis. During interphase, in which cells spend most of their life, the chromosomes are duplicated in preparation for mitosis. In mitosis, cell division occurs and two daughter cells are formed, each containing a complete set of chromosomes. Hence it is the part of the cell cycle in which the asexual transfer of genetic information happens. Interphase chromosomes are far less tightly packed compared to the extremely condensed mitotic chromosomes.

Hierarchy of Chromosome Packaging

Each of our cells contains up to two meters of DNA. However, the chromosomes are folded into the cell nucleus on length scales of a few microns. This high degree of compaction is achieved by a hierarchical packaging of the chromosomes. The first level of compaction of the double-helical DNA is the nucleosome, a complex consisting of a segment of DNA wound around eight histone proteins. These complexes are interconnected by linker DNA forming a rather loose structure that is called "beads-on-a-string" chromatin fiber. The existence of a further filament with 30 nm in diameter [22] as observed in electron microscopy experiments is still under debate for living cells [23]. Further levels of chromatin architecture as illustrated in Fig. 2.1 are dictated by compartmentalization, such as the formation of distinct chromosome territories [24] on the level of chromosomes and domains within single chromosomes (see the next subsection 2.1.2).

Transcription Factories

An important aspect of genome organization is the idea that genes and regulatory elements can cluster themselves spatially within the nucleus in order to be transcribed in a concerted fashion [27]. This concept of so-called transcription factories, which is illustrated in Fig. 2.2, has been speculated to be formed as a way for genes to reposition themselves in spatially confined regions with high concentration of polymerases [28,29]. Importantly, transcription factories are dynamic. Albeit the existence of transcription factories today is not disputed, the question whether transcription factories are a cause or a consequence of gene expression is still unanswered [30]. Current experimental evidence indicates that the number of transcription factories per nucleus varies from a few hundreds to several thousands [31]. The factories are thought to form during cell-differentiation upon activation, and remain even after the genes in the factory are no longer active [32]. Evidence of specialization of individual transcription factories, such as clustering of genes that belong to the same pathway, is still lacking [30]. However, spatial clustering of active globin genes in mouse and human cells has been reported [29].

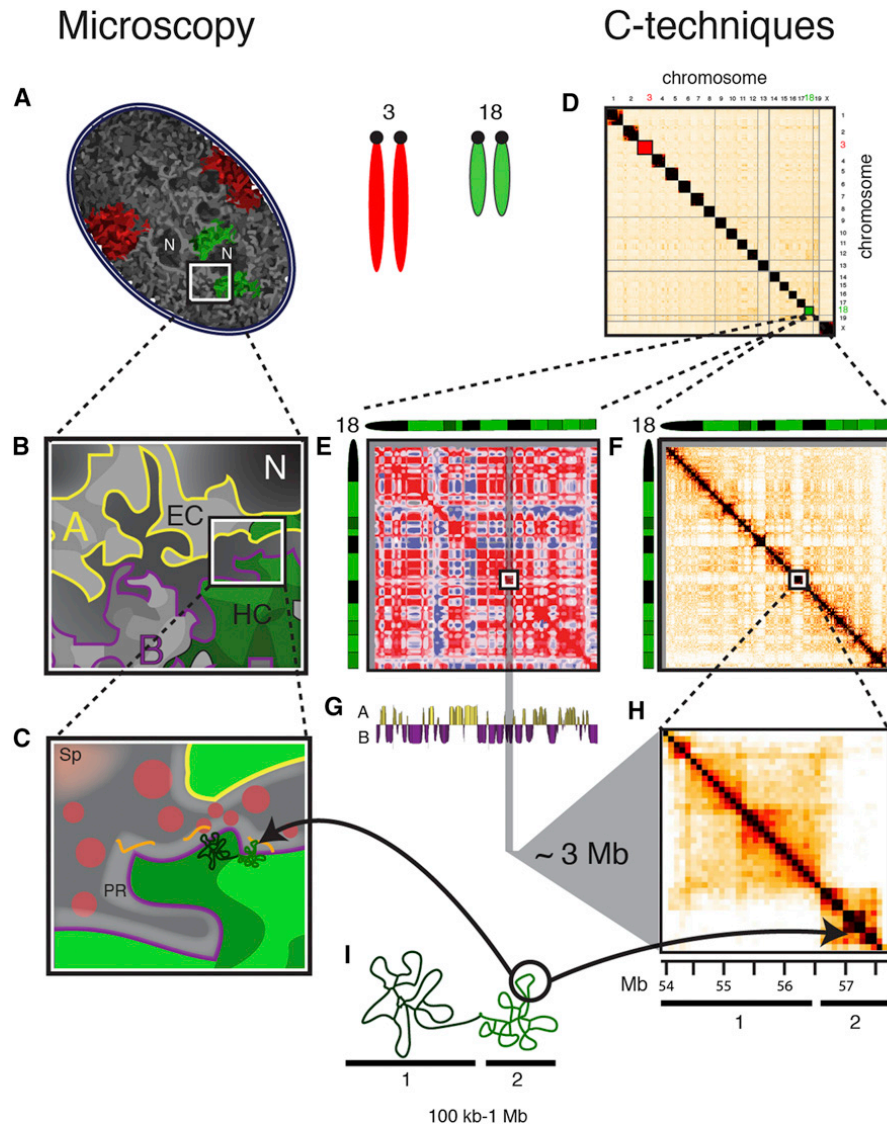


Figure 2.1: Large-Scale Nuclear Organization in Mammals. **A.** The nucleus is composed of chromosome territories (for example, mouse chromosome 3 and 18 are depicted in red and green, respectively). DNA is organized in more or less condensed regions, as can be shown by staining (gray intensities). **B.** The inset shows a more detailed architecture of the nucleus with (so called A and B) compartments, heterochromatin (HC) and euchromatin (EC). **C.** Zoomed-in view of hypothetical chromosome domains. Foci of factors interacting with looping chromatin in the perichromatin region (PR) are depicted as pink circles, RNA as orange lines. **D.** All-by-all chromosome matrix showing the interactions within and between chromosomes. **E.** Red and blue “plaid” pattern of chromosome 18 emphasized through Pearson correlation shows the separation into two chromosomal domains (colored red and blue). **F.** The cis-interaction matrix for chromosome 18. The inset indicates a ~ 3 Mb large B compartment. **G.** The clustering into compartments A and B. **H.** Detailed version of the 3 Mb large B compartment from (F), revealing the organization of topological domains (1 and 2). **I.** Representation of looping of chromatin as can be found at the PR (see C) or in deeper structures within topological domains (see H). Image and caption adapted from [25].

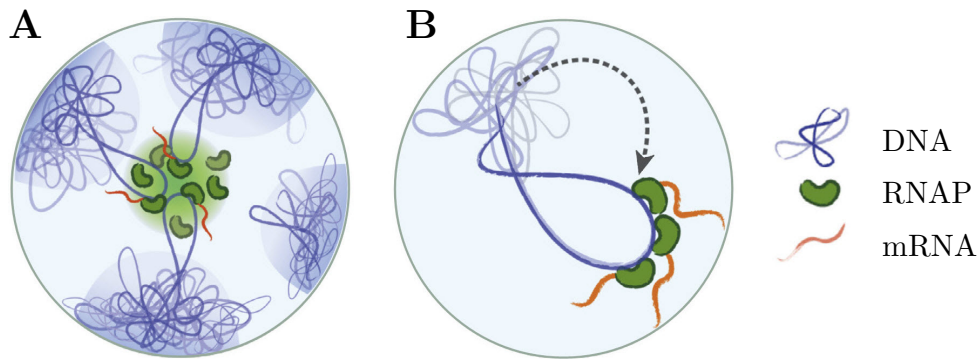


Figure 2.2: Two possible ways for the spatial organization of genes. **A.** Multiple genes spatially cluster together into transcription factories to be coregulated. **B.** A gene may move out of its ordinary domain to be activated. Image adapted from [26].

2.1.2 Domain Organization of the Genome

The idea that chromosomes are organized in a domain-like architecture goes back to the microscopy observations of euchromatin and heterochromatin and the banding patterns of mitotic chromosomes upon staining with particular dyes.

Compartments

In 2009, Lieberman et al. published their results on a study of the entire human genome using a method which they called Hi-C [3] (discussed in subsection 2.3.3). By coupling chromosome conformation capture with high-throughput sequencing, this method detects contact frequencies between loci covering the entire genome. An important finding of this study is the organization of the human genome into two separate compartments that the authors called A (or open) and B (closed) compartments (see Fig. 2.1) and that have a characteristic size of ~ 5 megabases each. Genomic interactions were found to be formed mostly within compartments, and much less frequently between compartments. Additionally, A compartments were associated with euchromatic, transcriptionally active and gene-rich regions, while B compartments were associated with gene-poor, inactive regions.

TADs

Experimental methods, such as Hi-C and 5C, which map interacting loci in a specific genomic region or in the complete genome, can identify structural features of chromosomes. Recently, such studies in fly [33] and mammalian [7,8] cells suggested that chromosomes are subdivided into discrete topologically (associated) domains (TADs). These TADs have a size of hundreds of kilobases and therefore differ from the larger A and B compartments which typically span a few megabases. Furthermore, TADs can be both active or inactive. Visual inspection of high-resolution 5C interaction maps, such as that of the 4.5 Mb long region depicted in Fig. 2.3 A, of undifferentiated (as well as differentiated) mouse embryonic stem cells reveals a series of large structural domains. Loci within these domains have a higher chance of interacting with each other than with loci located outside. The authors found that both the human and mouse genomes consist of more than 2000 TADs, covering

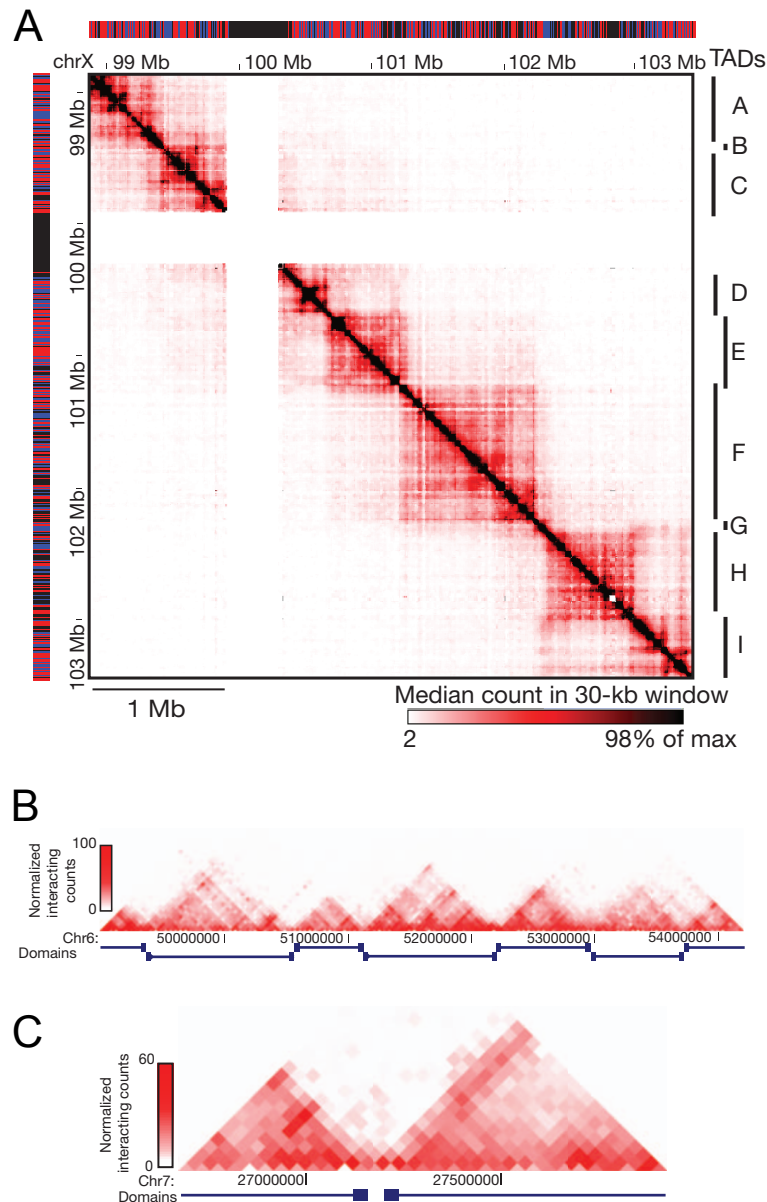


Figure 2.3: Chromosome partitioning into topological associated domains (TADs). **A.** 5C data from XY undifferentiated mouse embryonic stem cells (ESCs) as a heat map. Chromosomal contacts are organized into discrete genomic blocks (TADs A-I). **B, C.** Hi-C interaction frequencies both from (**B**) ESCs and (**C**) human IMR90 cells. Note the two different representations of actually the same kind of interaction data: The left figure depicts the data by means of a two-dimensional heat map, whereas the right ones only show the upper triangular part of the heat map rotated by 90 degrees counter-clockwise. Images adapted from [8] and [7].

over 90% of the genome, suggesting an evolutionary conserved and important function of this organization.

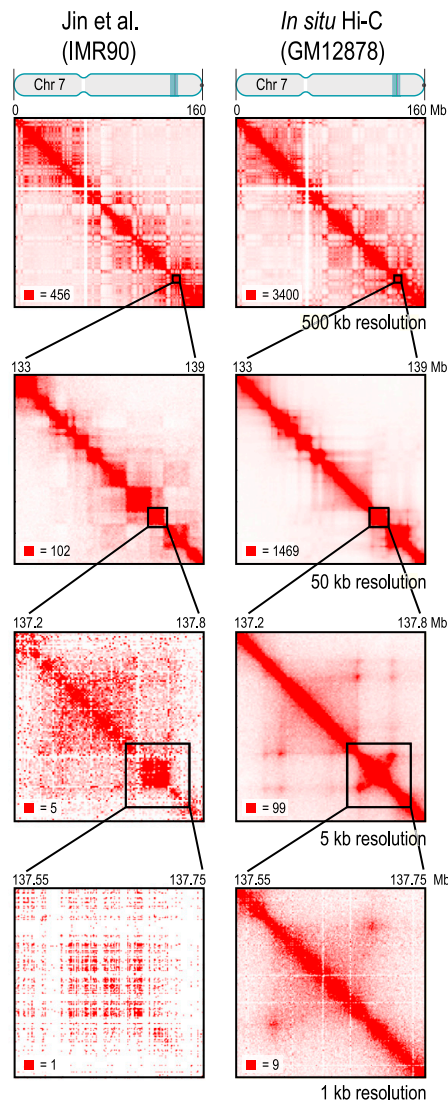
It is reported that the domain borders are demarcated by CTCF and cohesion [7, 8, 34], but that at the same time both proteins also frequently bind sites located within TADs. Moreover the spatial clustering into TADs remains largely intact after the depletion of CTCF and cohesion. As a consequence, it is still unclear which mechanisms establish

TAD boundaries. However, gene-expression profiles were found to be highly correlated within TADs, as opposed to between TADs, suggesting that regulation of genes via cis-regulatory elements may happen in a concerted fashion. We will come back to this idea in the following subsection 2.1.3.

Loop Domains

Just recently, Rao et al. published an in-situ Hi-C study of the human genome [36] that provides a huge amount of probed genomic interactions. This study outperforms the just presented ones with respect to the possible maximum resolution of contact maps. What we actually mean by “resolution” will become clear in chapter 4. For the moment, it suffices to understand that constructing a contact matrix with a “1 Mbp resolution” only means that the linear genome is split up into 1 Mbp bins and the contact frequency between each pair of bins is recorded. By constructing up to 5 and even 1 kbp resolution contact maps (compare Fig. 2.4), Rao et al. show that zooming into individual topological domains reveals that most of them are actually composed of even smaller domains. Furthermore, they also report the identification of ~ 10000 loops that frequently link promoters and

Figure 2.4: Comparison of Hi-C maps of chromosome 7 from two recent Hi-C studies. The left column shows Hi-C maps of primary human fibroblasts cells (IMR90) from the study of Jin et al. [35] at different resolutions. Although the authors claim having reached a resolution of 10 kbp, it is reasonable to visualize the contact maps at a resolution of 50 kbp (2nd row) instead. The right column depicts Hi-C maps of human GM12878 B-lymphoblastoid cells containing 4.9 billion pairwise contacts from the study of Rao et al. [36]. This Hi-C data set allows for a resolution of 1 kbp (4th row). The uppermost row shows the intrachromosomal contact map of chromosome 7 spanning about 159 Mbp (thus representing between 5 and 5.5 % of the total DNA in cells) at a resolution of 500 kbp. The following rows successively zoom into specific regions of chromosome 7 while simultaneously increasing the resolution. The visual examination of the Hi-C data of Rao et al. at high resolution, i.e. 1 or 5 kbp, already reveals distinct peaks in the contact map indicating the presence of loops. Indeed, a large fraction of peaks coincide with the corners of a contact domain, i.e. the peak loci are located at domain boundaries. Moreover the loops are often preserved across cell types and from human to mouse. Image adapted from [36].



enhancers. Interestingly, loop anchors typically occur at domain boundaries creating so called “loop domains” (compare the 5 and 1 kbp resolution contact maps in the second column of Fig. 2.4).

2.1.3 Enhancer-Promoter Interactions

One of the most widely studied classes of genomic three-dimensional interactions are interactions between regulatory elements, such as enhancers and promoters. This causes an activation of transcription in tissues where interaction occurs. The exact mechanisms for the concerted action of promoters and enhancers to activate transcription was debated at the end of the 20th century [38]. However, with the advent of the 3C-based techniques, the so called “looping model” [39] has become the accepted for the explanation of such interactions. According to this model, proteins, including transcription factors with affinity to motifs on the DNA, bind to the enhancer region forming a complex that has an affinity to proteins bound at the promoter. This causes them to form a loop between themselves via binding of the two protein complexes. This looping mechanism is also used by repressor proteins that bind to silencer regions to silence or downregulate genes. The action of enhancers can occur at large distances as well as either upstream or downstream relative to the promoter, and can even be positioned within the transcription unit itself. Enhancers are also able to activate multiple promoters, and can combine with other enhancers to activate a single promoter. Recently, by applying 5C experiments to 1% of the human genome, it has been shown that only a small amount of looping interactions occur with the nearest gene [40]. Additionally, the authors found evidence for several complex networks of interacting promoters and enhancer elements with functional effects on gene expression.

The existence of topological domains suggests that looping interactions between genes

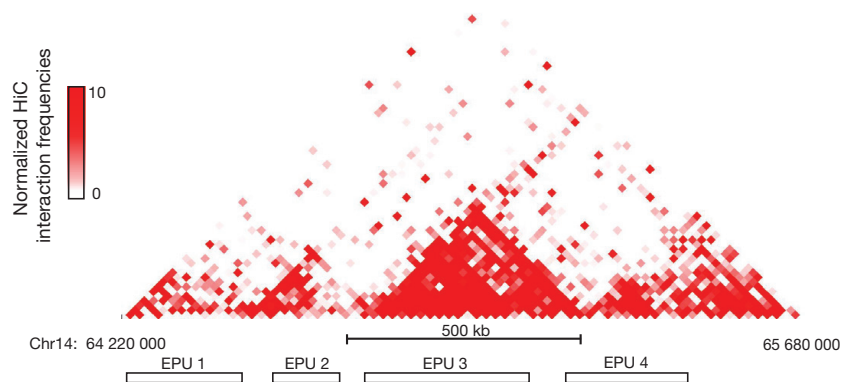


Figure 2.5: Comparing EPU blocks and physical partitioning of the mouse cortex. The upper part of the figure illustrates the normalized Hi-C interaction frequencies in mouse cortex as two-dimensional heat map where the special representation of the rotated upper triangular part of the contact matrix is chosen for demonstration purposes, since this allows the visual comparison to the identified EPUs and ChIP-seq data. The Hi-C interaction data shows that the physical partitioning of the genome is highly correlated with the EPUs that encompass gene clusters on chromosome 14. Image adapted from [37].

and distal regulatory elements are spatially constrained. Indeed, just recently, a ChIP-seq study of the mouse genome [37] showed that there are significantly more enhancer-promoter interactions within topological domains as illustrated in Fig. 2.5. As a result, these findings hint at the importance of genomic domain architecture in shaping the local regulatory landscape of genomes.

2.2 Organization of the Bacterial Nucleoid

While eukaryotic cells are usually part of multicellular organisms, prokaryotic cells are typically single-celled organisms and are lacking a defined cell nucleus. Instead, prokaryotic cells have a nucleoid region which contains a single circular, double-stranded DNA molecule. In general, prokaryotes can be classified into two domains, archaea and bacteria. Within this thesis we mostly focus on bacteria.

2.2.1 Bacterial Chromosome Packaging

Bacterial genomes are circular double-stranded DNA molecules that are typically several million base pairs in size. By adopting a highly compact but orderly structure the DNA can adapt to the spatial conditions of a bacterial cell while at the same time enabling replication and transcription.

The well-studied bacterial model organisms *Escherichia coli* (*E. coli*) and *Caulobacter crescentus* (*C. crescentus*) are good examples showing that bacterial chromosomes are organized on multiple levels as depicted in Fig. 2.6. This hierarchical fashion is a common feature of bacterial and eukaryotic chromosomes. On the lowest level, the negatively supercoiled *E. coli* chromosome forms plectonemic loops, which are actively maintained by

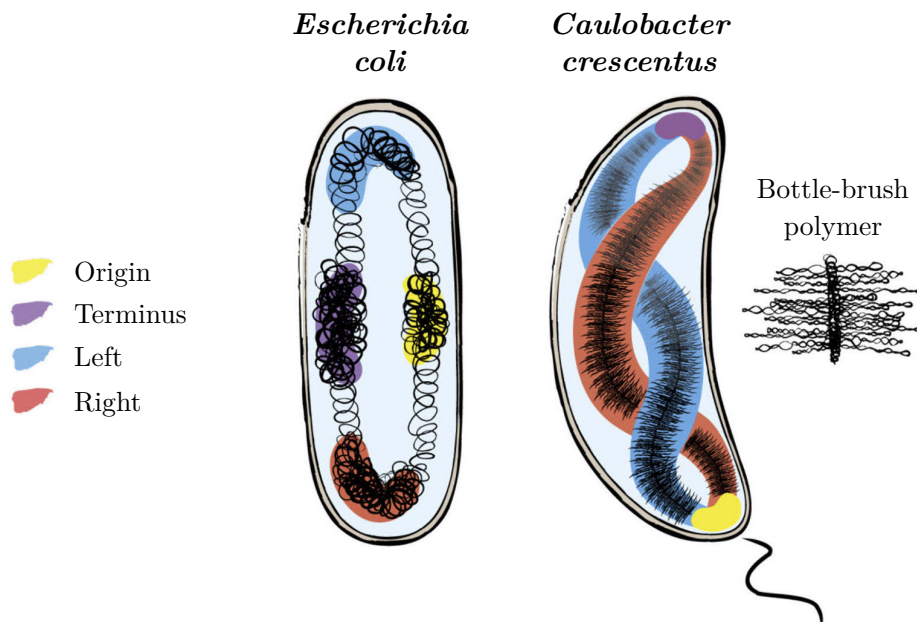


Figure 2.6: Models for nucleoid organization in *Escherichia coli* and *Caulobacter crescentus*. Chromosomal domains are colored as indicated by the legend. Image and caption adapted from [26].

the interplay of the enzymes DNA gyrase and DNA topoisomerase I as well as the so-called nucleoid-associated proteins (NAPs). These NAPs, such as H-NS (histone-like nucleoid structuring protein), HU (histone-like protein), Fis (factor for inversion stimulation), and IHF (integration host factor), have an influence on DNA structure both locally by bending and wrapping DNA segments [42] and globally by DNA looping [43]. Additionally, NAPs establish boundaries for “microdomains” [44] with a typical size of 10-100 kilobases that are located stochastically on the chromosome [45]. On a higher level, the organization of the *E. coli* genome is characterized by mainly four “macrodomains” with a size of approximately 1 megabase (origin, terminus, left and right) [46]. However, a recent genome-wide 3C-based study highlights an important role for the terminus region in the organization of the *E. coli* genome [47]. The features of 3D genome organization of *E. coli* are similar to those of other bacterial model organisms, such as *Caulobacter crescentus* [48].

2.2.2 Interplay between Gene Regulation and Genome Folding

The interplay of 3D genome organization and transcriptional activity indicates that, irrespective of their genetic location on the chromosome, genes are spatially positioned with respect to their transcriptional activities. In principle, there are two general ways for organizing genes spatially as depicted in Fig. 2.2.

A recent approach investigating the spatial distribution of H-NS in *E. coli* suggests

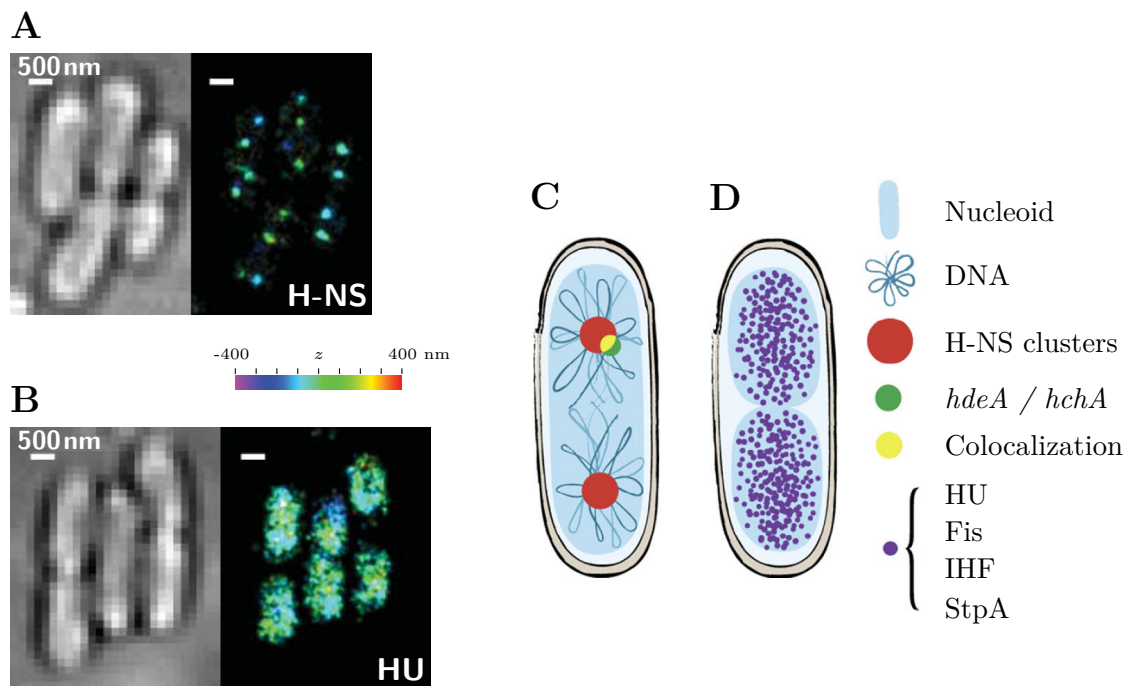


Figure 2.7: Different spatial distributions of NAPs in *E. coli* cells identified by using STORM. **A.** Compact H-NS clusters in the nucleoid. *E. coli* cells imaged using bright-field microscopy are depicted on the left and NAPs in living *E. coli* cells imaged using localization-based super-resolution microscopy (STORM) on the right. **B.** Scattered distribution of H-NS in the nucleoid. (Left) Bright-field image; (right) 3D STORM image. **C.** Schematic drawing of the H-NS clusters colocalizing with H-NS regulated genes *hdeA* and *hchA*. **D.** Schematic drawing of the homogeneous distribution of HU, Fis, IHF and StpA within the nucleoid. Image adapted from [41] and [26].

clustering of genes [41]. Using single-molecule based super-resolution imaging, to be precise, stochastic optical reconstruction microscopy (STORM), H-NS was shown to form on average two clusters per chromosome in living *E. coli* cells (see Fig. 2.7 A,C). Two-color colocalization imaging revealed that these clusters spatially overlap with genes regulated by H-NS.

2.3 Molecular Techniques for Unraveling the Three-Dimensional Genome

In the last decades, several novel techniques for analyzing the chromatin structure have been developed. Particularly, the combination of next-generation sequencing and traditional chromosome conformation capture techniques depicts a milestone in the field for the exploration of the three-dimensional shape of the genome.

In this section, we introduce the most important methods for studying the three-dimensional architecture of the genome. The focus is set on 3C-based technologies that are most relevant for this thesis.

2.3.1 Fluorescence in Situ Hybridization (FISH)

FISH is a combined molecular and cytological approach where fluorescently labeled DNA probes are hybridized to complementary sequences on chromosomal preparations fixed on slides. The probes are then visualized using microscopy. Even though FISH was invented 30 years ago, it is still widely used both in research and diagnostics [49]. The wide usage of FISH is attributed to the fact that it provides spatial information at intermediate degree resolutions at single cell level. Techniques utilizing FISH are still being refined and diversified into more specialized versions. The three most widely used FISH variants in three-dimensional genome analysis are Cryo-FISH, 3D-FISH and Immuno-FISH.

2.3.2 Next-Generation Sequencing (NGS)

With the advent of “next-generation sequencing”, traditional Sanger-based sequencing techniques have virtually been entirely replaced. As its synonym, “massively parallel sequencing”, implies, the approach allows for the simultaneous interrogation of millions of sequences based on clonal amplification of DNA fragments. To this end, the sequences are often spatially separated on plates or slides, and interrogated using a high-resolution camera. There are several next-generation sequencing technologies available, such as 454 (Roche Diagnostics), SOLiD (Applied Biosystems), and Solexa (Illumina). Most of them enable paired-end sequencing, where the two ends of the same DNA molecules are sequenced from both sides.

2.3.3 Chromosome Conformation Capture (3C) Methods

In what follows, the technologies based on chromosome conformation capture (3C) are discussed. The underlying concept of all these methods is the same and consists in the quantification of ligation junctions by digestion and re-ligation of fixed chromatin in cells and in the fact that the quantified DNA contact frequencies reflect spatial proximity within the nucleus [50].

Chromosome Conformation Capture (3C)

3C was invented in 2002 by Dekker et al. [21]. Originally, contact frequencies were quantified using quantitative polymerase chain reaction (qPCR), whereas today paired-end sequencing is used to a much larger extent. 3C-based methods, contrary to microscopy-based techniques, allow for more systematic and quantitative characterization of genome topology and a higher resolution at the same time. On the other hand, the essential drawback is the fact that the conventionally ensemble 3C-based methods are mostly performed on large populations of cells, whereby the information at the single-cell level is lost.

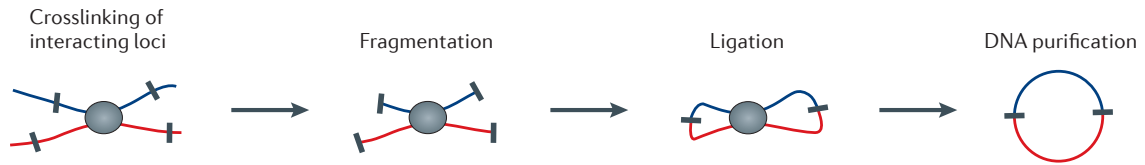
All 3C-based techniques start with the same steps, which aim to isolate DNA fragments in spatial proximity. The first step is cross-linking (fixation) of chromatin by addition of formaldehyde. This causes interacting chromatin segments, i.e. those in close spatial proximity, to be covalently linked together. The fixed chromatin is then cut with a restriction enzyme, such as HindIII, DpnII, EcoRI and many others, chosen such that the frequency of cuts provides the desired resolution for the given analysis. The sticky ends of the fragmented cross-linked DNA are then re-ligated under diluted conditions to favor intramolecular ligation of the cross-linked fragments. The re-ligated DNA molecules thereby form a hybrid consisting of two DNA fragments from the two segments that were cross-linked. After DNA purification, qPCR or sequencing is used to quantify the number of such hybrid DNA-molecules.

In 3C, primers are designed near the ends of the restriction fragments of interest, enabling quantification of selected ligation junctions. Ligation frequencies, as measured by the amount of ligation product between the selected primer combinations, are then used to infer which fragments were spatially proximal [50]. 3C therefore allows for focused quantification of contact frequencies at selected regions in a *one-versus-one* fashion as depicted in Fig. 2.8. It is important to recognize that there are two major types of ligation junctions that are over-represented. One is the junction forming between neighboring restriction fragments due to incomplete digestion. The other one is the junction that forms when one end of the fragment ligates with the other end of the same fragment. Therefore, the quantification step also involves determining whether DNA segments contact each other more than expected simply due to the (linear) genomic proximity between them. A further observation is that over large genomic distances ligation products become very infrequent and quantification using qPCR therefore becomes infeasible. Furthermore, the differences in primer efficiencies need to be controlled for, by making a PCR control template with all ligation products in equal amounts [51].

Chromosome Conformation Capture-on-Chip (4C)

As the name already indicates, chromosome conformation capture-on-chip (4C) combines 3C technology with microarrays in order to quantify the contact frequency between one locus (or “viewpoint”) and all other genomic loci represented on the array [53]. The definition of this single viewpoint and the subsequent screening of the genome for sequences that contact this selected site makes 4C a *one-versus-all* type of analysis, in contrast to the one-versus-one nature of 3C.

Briefly, 4C technology, includes a second ligation step, in which self-circularized DNA fragments are created. Subsequently, inverse PCR is used to amplify and identify DNA sequences contacting these DNA circles. 4C-seq [54] is a variant of the 4C method that uses NGS instead of microarray analysis for the identification step, thereby saving costs, offering a higher resolution and quantifying DNA contact frequencies more accurately. A

A 3C: converting chromatin interactions into ligation products**B Ligation product detection methods**

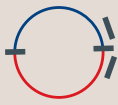
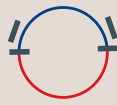
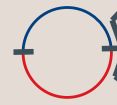
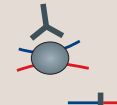

3C	4C	5C	ChIA-PET	Hi-C
One-by-one All-by-all	One-by-all	Many-by-many	Many-by-many	All-by-all
			<ul style="list-style-type: none"> • DNA shearing • Immunoprecipitation 	<ul style="list-style-type: none"> • Biotin labelling of ends • DNA shearing 
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing	Sequencing

Figure 2.8: **A.** Initial steps in the chromosome conformation capture (3C) procedure. **B.** Schematic illustration of the key concepts for the various 3C-based technologies. Image adapted from [52].

point worth mentioning is that in the resulting data consisting of a genome-wide profile of ligation events with peaks corresponding to significant interactions an unspecific background signal needs to be filtered out. This is usually done using time series methods, such as moving average or median.

Chromosome Conformation Capture Carbon Copy (5C)

5C captures all interactions between a set of selected regions [55] and is hence referred to as a *many-versus-many* technology. 5C requires specifically designed oligonucleotides, i.e. small bits of nucleic acids that can be manufactured with any user-specified sequence, also referred to as primers (\curvearrowright “5C primers”), that anneal to the ends of the restriction fragments. A ligation between two interacting fragments in the 3C library therefore results in two 5C primers annealing adjacent to each other on each side of the ligated restriction sites. Because all 5C primers have a universal sequence at one of their ends, it is possible to amplify all ligation products simultaneously. Finally, ligation junctions are identified using either microarray analysis or NGS. The resulting data is a matrix of interaction frequencies between the two sets of restriction fragments selected prior to the analysis. Again, just as for 3C data in general, fragments that are close in terms of genomic distance have an increased probability of forming contacts, expressed through a pronounced main diagonal in the resulting contact matrix.

The resolution of 5C is determined by the distance between neighboring primers on the chromosome template. 5C has relatively low coverage since 5C primers cannot be designed for every unique end of a restriction fragment. For this reason, and because there is also the need to use many primers simultaneously, it is not possible to reach the resolution of 4C or Hi-C with a 5C assay. On the other hand, and contrary to 4C, 5C provides a contact

matrix of interaction frequencies for many pairs of sites. This makes it especially suited for focused studies, such as reconstructing the (average) three-dimensional conformation of selected, but possibly, larger genomic regions [8, 40, 56].

Hi-C

Hi-C is the first truly genome-wide 3C based technology. It was introduced by Lieberman-Aiden et al. in 2009 [3]. Realizing a genome-wide quantification of interactions in an *all-versus-all* fashion, it constitutes a major breakthrough in the study of chromatin architecture. The Hi-C protocol (see Fig. 2.9) differs from the standard 3C protocol therein that there is an extra step needed before ligation. It consists in filling in the sticky ends after restriction enzyme cutting with biotin-labeled nucleotides. After purification and shearing of the DNA, the biotin marks are pulled down in order to ensure that only ligation junctions are selected for further analysis. By mapping the reads back to the reference genome, a genome-wide contact matrix is constructed. Due to the quadratic nature of “all-versus-all” data, an extremely high throughput is needed. As an illustration, the resolution of 950 bp, referred to as “kilobase resolution”, of a very recent in situ Hi-C study of the human genome [36] is based on 4.9 billion valid paired-end reads.

A technique similar to Hi-C, called genome conformation capture (GCC), has recently been applied for the mapping of yeast chromosome interactions [57] as well as for studying the spatial organization of the *Escherichia coli* nucleoid [58].

Various biases affect the read counts in genome-wide contact maps like those obtained by Hi-C. Thus, correcting and normalizing these genome-wide interaction data is not only necessary, but essential for an adequate interpretation of such data (see chapter 4).

Lately, a single-cell version of Hi-C was published [59]. Nagano et al. modified the conventional or “ensemble” Hi-C protocol to create a method to determine the contacts in an individual nucleus. However, at this stage this approach allows the capture of only 2.5% of all interactions in a cell.

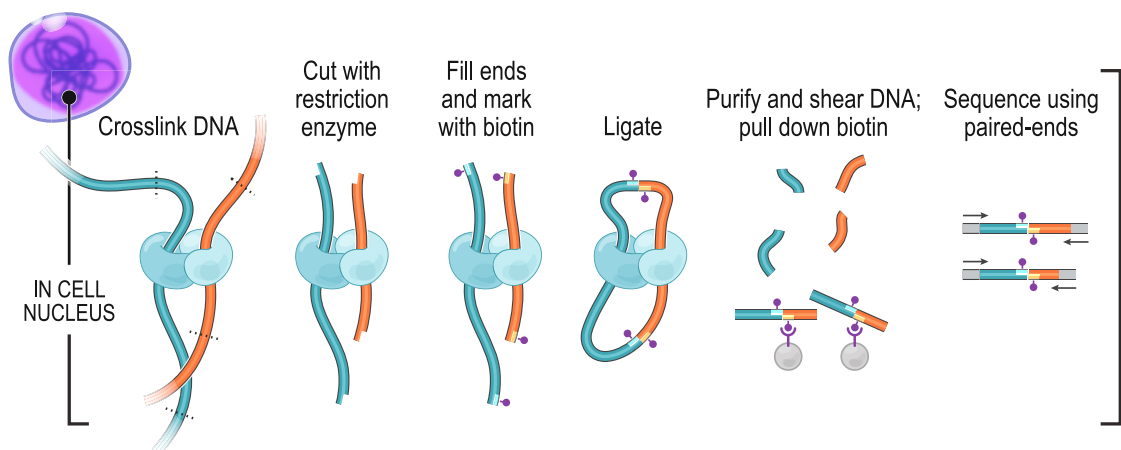


Figure 2.9: Schematic illustration of the (in situ) Hi-C process. The DNA-DNA proximity ligation is performed in intact nuclei. Image adapted from [36].

2.3.4 ChIP-seq and ChIA-PET

Chromatin Immunoprecipitation (ChIP) is an experimental method used to analyze protein interactions with DNA in living cells. The objective is to establish whether certain proteins are bound to specific DNA segments. Hence, for example, if transcription factors are bound to promoters or enhancers. ChIP-sequencing (ChIP-seq) combines ChIP and NGS, thereby enabling the genome-wide identification of binding sites of DNA-associated proteins.

Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) combines ChIP with 3C-type analysis and enables the identification of all chromatin interactions between regions bound by a given protein [60]. As such, it is a *many-versus-many* type of analysis (see Fig. 2.8). So far, ChIA-PET has been applied to DNA sites bound by the estrogen receptor α (ER α) [60] and CTCF [61].

It is important to note that ChIA-PET, unlike other 3C-based methods, exclusively identifies interactions between regions that are bound on both sides by the same protein of interest. As a consequence, it is not compatible with protein knock-down or knock-out and is not able to unravel a possible connection between identified loops and the selected protein. A further disadvantage is that the datasets so far produced by ChIA-PET show a rather low signal-to-noise ratio [50].

Chapter 3

Theoretical Background

In this chapter, we first present an overview of basic approaches to model macromolecules using polymer models. We then provide a basic introduction into Monte Carlo simulations and introduce the Bond Fluctuation Model that is later used to study the architecture of different genomes.

Furthermore, we discuss the statistical and conformational properties reflecting the characteristics of the polymer systems we are simulating. We present results for quantities, such as the end-to-end distance distribution and the mean squared displacement. We also introduce the concept of spatial confinement. Our focus rests on contact probability measures. Next, we introduce the basic concepts of computational topology. This rapidly emerging field of research at the interface between mathematics and computer science is dedicated to the investigation of efficient algorithms for topological problems. Since this is a vast and complex area, we provide an overview of the most important concepts of topological data analysis rather than introducing the topic in a mathematically rigorous manner. For a more comprehensive presentation, we refer to the work of R. Christ [62] and the textbook by Afra J. Zomorodian [63].

3.1 Simulation Methods

3.1.1 Polymer Physics

Biological macromolecules can be considered as polymers composed of smaller subunits called monomers. Polymer physics is a branch of statistical physics which studies polymers, their mechanical properties, as well as their conformational motion. The areas of application are diverse and range from materials science, condensed matter physics to biophysics and molecular biology. A prominent example is DNA where the nucleotides are the monomers that make up the DNA polymer. Using coarse-grained models, even complex systems like lipid membranes can be considered as polymers.

In this section we present the main polymer models and properties to describe them.

Simple Polymer Models

The Freely-Jointed Chain Model

A freely-jointed chain or ideal chain is the simplest model to describe polymers, such as DNA and proteins. The polymer is represented by a chain of N bond vectors \mathbf{b}_i of fixed length b . It assumes a polymer as a random walk (RW) of $N + 1$ steps or monomers, thus all directions of the vectors \mathbf{b}_i are independent from each other.

The **end-to-end distance** vector is an important measure for the spatial size of a polymer. It is simply given by the vector from the first to the last monomer:

$$\mathbf{R}_e = \mathbf{r}_{N+1} - \mathbf{r}_1 = \sum_{j=1}^N \mathbf{b}_j ,$$

where \mathbf{r}_i denotes the position vector of the i^{th} monomer.

In polymer physics we are mostly interested in mean values over ensembles. We denote the ensemble average by $\langle \cdot \rangle$ in the following. The mean-square end-to-end distance of a freely-jointed chain is given by

$$\langle \mathbf{R}_e^2 \rangle = \sum_{i=1}^N \sum_{j=1}^N \langle \mathbf{b}_i \cdot \mathbf{b}_j \rangle .$$

Since all the vectors are independent, it is $\langle \mathbf{b}_i \cdot \mathbf{b}_j \rangle = b^2 \delta_{ij}$ and thus get

$$\langle \mathbf{R}_e^2 \rangle = Nb^2 ,$$

which is often referred to via the scaling law

$$\langle \mathbf{R}_e^2 \rangle \propto N^{2\nu} ,$$

with the scaling exponent ν . For the ideal chain we have $\nu = 0.5$.

The **probability distribution function** (*pdf*) of the end-to-end distance $p(\mathbf{R}_e)$ for ideal chains is determined using the central limit theorem because of the independence of the bond vectors \mathbf{b}_i . Hence, the end-to-end distance is Gaussian distributed in the limit $N \rightarrow \infty$.

The **radius of gyration** of a polymer is defined as the average distance of monomers from its center of mass. As such, it measures the effective size of a polymer chain. The mean-square radius of gyration is

$$\langle R_g^2 \rangle = \frac{1}{N} \sum_{i=1}^N (\mathbf{r}_i - \mathbf{r}_{\text{cm}})^2 ,$$

where \mathbf{r}_{cm} denotes the center of mass of the polymer. The mean-square radius of gyration and the mean-square end-to-end distance of ideal chains show the same scaling behavior.

“Real” biopolymers can be considered as ideal chains on large length scales in the case of highly dense systems with an uniform spatial distribution of interacting polymers. Otherwise, the scope of the ideal chain model is limited due to missing excluded-volume interactions.

The Gaussian Chain Model

Contrary to the fixed length of bond vectors \mathbf{b}_i in the ideal chain model, the Gaussian chain model models the bond length as fluctuating. This accounts for the fact that chemical bonds possess a certain intrinsic flexibility. According to the name of the model the bond vectors are Gaussian distributed

$$G(\mathbf{b}_i) = \left(\frac{3}{2\pi b^2} \right)^{3/2} \exp \left(-\frac{3\mathbf{b}_i^2}{2b^2} \right) ,$$

where $\langle \mathbf{b}_i^2 \rangle = b^2$.

Since the model still assumes the bond vectors to be independent, the mean squared end-to-end distance is given by $\langle \mathbf{R}_e^2 \rangle = Nb^2$. The *pdf* of the end-to-end distance in the case of the Gaussian chain model is also a Gaussian distribution.

The Worm-like Chain Model

A fundamental characteristic of biomacromolecules is chain flexibility or rather chain rigidity. Important biopolymers such as DNA and proteins are semi-flexible and can be described by the worm-like chain or Kratky-Porod model. Within this model, semi-flexible polymers with contour length L are parameterized by the path $\mathbf{r}(s)$ along the chain of length L . The unit tangent vector to the chain at point $s \in [0, L]$ is defined as $\mathbf{u}(s) \equiv \frac{\partial \mathbf{r}(s)}{\partial s}$. The energy associated with the bending of the polymer is given by the Hamiltonian

$$\mathcal{H} = \frac{\kappa}{2} \int_0^L ds \left(\frac{\partial^2 \mathbf{r}(s)}{\partial s^2} \right)^2,$$

with the bending rigidity $\kappa = l_p kT$ which is proportional to the polymer's characteristic persistence length l_p .

We can analyze the worm-like chain using the correlation of two unit tangent vectors $\mathbf{u}(i)$ and $\mathbf{u}(j)$ which are separated by a distance $|i - j| \in [0, L]$ along the polymer chain as

$$C(|i - j|) = \langle \mathbf{u}(i) \mathbf{u}(j) \rangle,$$

where $\langle \cdot \rangle$ denotes the ensemble average of all polymer conformations. For polymers that are not fulfilling self-avoidance the orientational correlation function decays exponentially

$$\langle \mathbf{u}(i) \mathbf{u}(0) \rangle = \exp(-|i - j|/l_p).$$

The orientational correlation function for chain segments that are separated far enough vanishes, i.e. $C(|i - j| \rightarrow \infty) \rightarrow 0$.

Self-avoiding Polymers

As mentioned above, the ideal chain is the simplest model for polymers. More “realistic” polymer models consider the excluded volume of the chain segments. Since monomers constitute physical subunits of the polymer, comparable to atoms within a molecule, they can not occupy space that is already occupied by another part of the polymer. Excluded-volume interactions are extremely strong, short-range repulsive interactions leading to swelling of polymer chains as opposed to ideal chains.

The chemist Paul J. Flory was the first to introduce and theoretically describe self-avoiding walks (SAWs). According to his theory the mean squared end-to-end distance vector of SAW polymers shows the scaling law

$$\langle \mathbf{R}_e^2 \rangle \propto N^{2\nu},$$

with $\nu \approx 3/5$.

3.1.2 Monte Carlo Simulations

Molecular dynamics (MD) and Monte Carlo (MC) simulations are the two most important modeling techniques for the study of biological macromolecules. In MD simulations, the time dependent behavior of atoms and molecules is determined by numerically solving Newton's laws of motion for a system of interacting particles. Forces between the particles and potential energy are defined by potentials or molecular mechanics force fields. Monte Carlo algorithms, on the other hand, are based on repeated random sampling to get numerical results. In order to obtain the distribution of an unknown probabilistic entity simulations are typically repeated many times.

Metropolis Monte Carlo

We use a Metropolis Monte Carlo algorithm for our simulations. This algorithm leads to thermal equilibrium after a certain number of steps (see the next subsection 3.1.2 for further details) and works as follows. We start from an initial configuration of the particles in a system. The algorithm proceeds by randomly attempting to change the configuration of the particles, i.e. a Monte Carlo move. The move is either accepted or rejected based on the Metropolis acceptance criterion guaranteeing that the sampled configurations are drawn from the Boltzmann distribution with the correct Boltzmann weight. After having either accepted or rejected a move, we compute the quantity in question. The algorithm proceeds by randomly attempting to move about the sample space and eventually, after many moves have been made, it yields a reliable average value of the quantity in question.

Autocorrelation Time

Markov Chain Monte Carlo (MCMC) methods generate a new state based on its previous state. Thus obtained samples by MCMC algorithms are statistically dependent on each other or correlated. The autocorrelation time helps in obtaining statistically independent or uncorrelated conformations in simulations. It is determined using the integrated autocorrelation time τ_{int} which is computed using the autocorrelation function $C(t)$ and the normalized autocorrelation function $\rho(t)$. This scheme allows the calculation of the correlation between polymer conformations separated by t Monte Carlo steps and described by a particular measure or observable. The autocorrelation function of an observable $A(t)$ is defined as

$$C(t) = \langle A(s+t) \cdot A(s) \rangle_s - \langle A(s) \rangle_s^2$$

and the normalized autocorrelation function is given by $\rho(t) = \frac{C(t)}{C(0)}$, where $\langle \cdot \rangle_s$ is defined as the mean of the ensemble at time s .

We use the windowing method by Sokal [64] in order to estimate the integrated autocorrelation time as

$$\tau_{\text{int}} = \frac{1}{2} \sum_{t=1}^M \rho(t).$$

The integer M is chosen such that $M > c \cdot \tau_{\text{int}}$. The value c can vary between four for exponential decaying $\rho(t)$ to ten for slower decay [64]. Two subsequently obtained conformations are considered to be uncorrelated when they are separated by more than $5\tau_{\text{int}}$ steps. In each of our simulations, $10\tau_{\text{int}}$ steps are prepended in order to equilibrate our artificially generated starting configuration.

We use this windowing scheme of Sokal in favor of simply fitting an exponential model to the autocorrelation function because we are also simulating both large and quite stiff

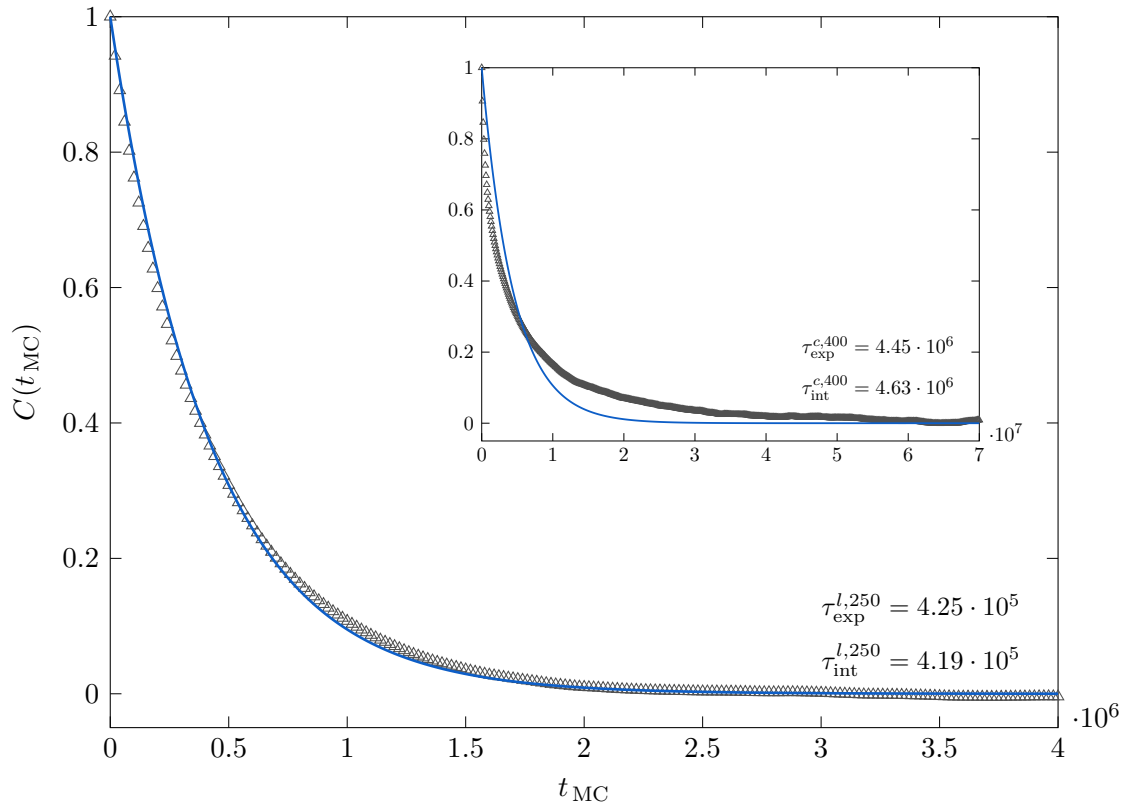


Figure 3.1: The main figure shows a rapidly sloping autocorrelation function (in gray) of the mean squared radius of gyration of a linear polymer consisting of $N = 250$ monomers. The blue curve illustrates an exponential fit and aligns perfectly with the computed values of the autocorrelation function. However, the autocorrelation function of the mean squared radius of gyration of a circular polymer with $N = 400$ monomers (see inset graph) is not well described by an one-parameter exponential model obeying the functional form $C(t_{MC}) = \exp(-\tau_{\text{exp}}^{-1} \cdot t_{MC})$. The resulting exponential autocorrelation time τ_{exp} is illustrated in the graphs and also compared to the computed integrated autocorrelation time τ_{int} .

polymers leading to a slowly decreasing autocorrelation function of the mean squared radius of gyration (see the inset graph in Fig. 3.1).

The Bond Fluctuation Model

The Bond Fluctuation model (BFM) is a well established lattice model for polymers. The BFM includes excluded-volume interactions and preserves the topological state of the polymers by preventing bond crossings. It is a Monte Carlo method characterized by especially high acceptance rates making it a good choice for dense polymer systems. A detailed description of the BFM can be found in [65, 66].

A long polymer on a three-dimensional cubic lattice consists of N monomers, numbered from one to N . Each monomer occupies one box (i.e. eight lattice sites) on the lattice and thus the polymer can be described by the set of bond vectors of its comprising monomers $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{N-1}\}$. Volume interactions are integrated into the model by forbidding one box to be occupied by two or more monomers. As one monomer occupies eight lattice sites, there always has to be at least one empty box between two monomers. The maximum

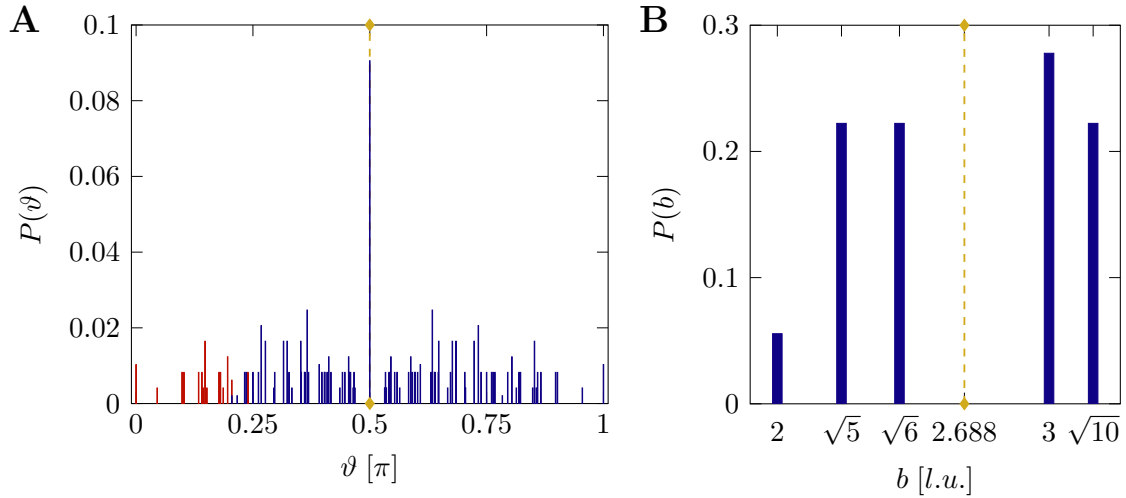


Figure 3.2: The a priori probability distribution of **A.** the bond angles ϑ within the BFM as well as **B.** those of the bond lengths b . In contrast to the bond lengths, not all possible bond angles are allowed within the BFM fulfilling self-avoidance. The forbidden bond angles violating excluded volume are colored in red whereas allowed ones are depicted in blue.

bond length is restricted to ten, limiting the distance between neighboring monomers and thus preventing the chain from developing gaps. There are further constraints on the bond vectors in order to avoid bond crossings and for ensuring the preservation of the topology of the polymer. On a three-dimensional lattice 108 different bond vectors can be realized. The a priori probability distributions of both the bond angles and the bond lengths are depicted in Fig. 3.2 (forbidden bond angles are colored red). The possibility for fluctuating bonds is a key ingredient of the BFM since this leads to an increased probability for local moves of the monomers resulting in quicker relaxation towards equilibrium.

Dynamic Looping Mechanism

In our simulations we make use of the dynamic loop (DL) model developed by Bohn and Heermann [67]. The DL model is based on the BFM and incorporates the ability of non-adjacent monomers to become linked by a bond vector. Whenever two monomers come close to each other by diffusion, there is a looping probability p_{loop} for them to form an additional bond. When this happens, a crosslink of the fiber is created with a lifetime, drawn from a Poisson distribution with mean value τ . Thus, loops can form and dissolve dynamically. The size of the loops is restricted, monomers must have at least a genomic distance of three to be able to form loops. The maximum allowed size of the loops as well as the number of bonds starting from one monomer can be restricted.

The dynamic and probabilistic crosslinking mimics the effect of surrounding proteins which mediate the process of loop formation. It causes a coiling and local collapsing of the chromatin fiber, which is anticipated to have implications on the shape and the mechanical properties of the polymer.

3.2 Statistical and Conformational Properties

3.2.1 End-to-End Distance and Gyration Radius

The end-to-end distance and the gyration radius (see the previous section 3.1) provide measures of the space that a polymer coil occupies. Hence, they give information about the size of the simulated chromosomes.

The comparison between the *pdf* of the end-to-end distance for polymers with varying degree of dynamic looping as well as self-avoiding walk (SAW) and random walk (RW) polymers as depicted in Fig. 3.3 shows that dynamic loop formation can induce compact polymers. The mean squared end-to-end distance for a SAW polymer amounts to $\langle \mathbf{R}_e^2 \rangle \sim 70$ lattice units [l.u.], for a RW polymer to $\langle \mathbf{R}_e^2 \rangle \sim 40$ l.u. whereas it is in the range of

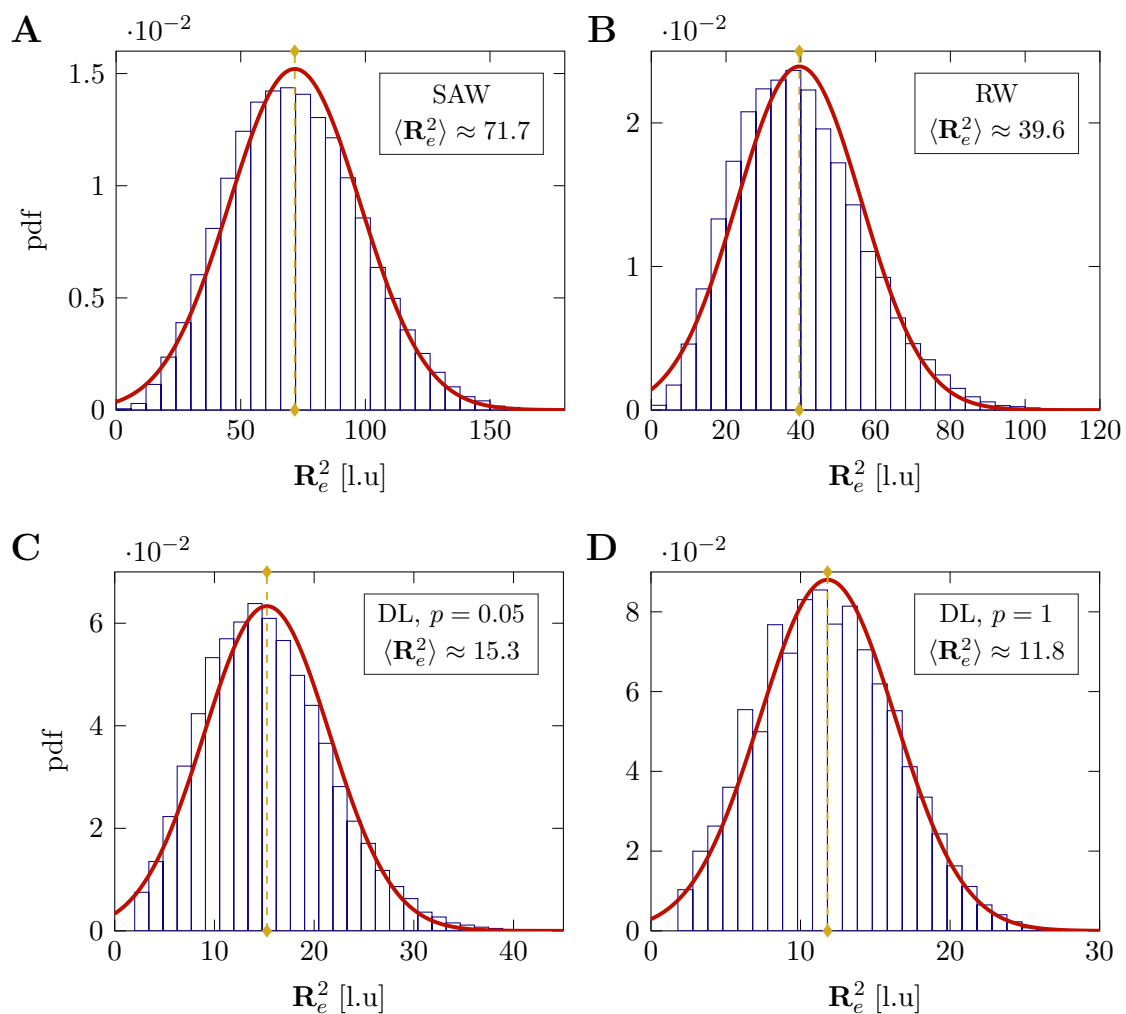


Figure 3.3: The probability density function of the squared end-to-end distance $\langle \mathbf{R}_e^2 \rangle$ for a linear polymer configuration with $N = 250$ monomers as a function of the underlying polymer model. The distributions for **A.** self-avoiding, **B.** random walk polymers and **C, D.** polymers with dynamic looping ($p_{\text{loop}} = \{0.05, 1\}$) are similar, they only fluctuate around different expectation values. As the probability and therefore the number of loops increases, the randomly looped configurations become denser and together with this, fluctuations decrease.

10 – 15 l.u. and can be adjusted to one's needs for polymers with dynamic looping.

Therefore, it is justified to view the dynamic loop formation as an adjustable compaction mechanism. The Gaussian shape of the distributions of the end-to-end distance of polymers with dynamic looping indicates that they still have the character of self-avoiding walks, although they can be much more compact.

3.2.2 Mean Squared Displacement

The mean squared displacement (MSD) is given by

$$\langle R_n^2 \rangle = \frac{1}{(N-n)N_C} \sum_{j=0}^{N_C-1} \sum_{i=0}^{N-n-1} \|r_{j,i} - r_{j,i+n}\|_2^2, \quad (3.1)$$

where N is the total number monomers, N_C is the number of polymer conformations over which the ensemble average is computed and n is the contour length, i.e. the genomic distance between monomers i and j .

In Fig. 3.4, a typical example for the MSD or intrachain distance statistics is given. Loop formation has a strong influence on the mean squared displacement of the monomers, resulting in a leveling off for distances when loops can be formed (the smallest loop size is three monomers). The fact that the mean square displacement does not increase with increasing contour length, means that monomers stay in closer spatial proximity to each other.

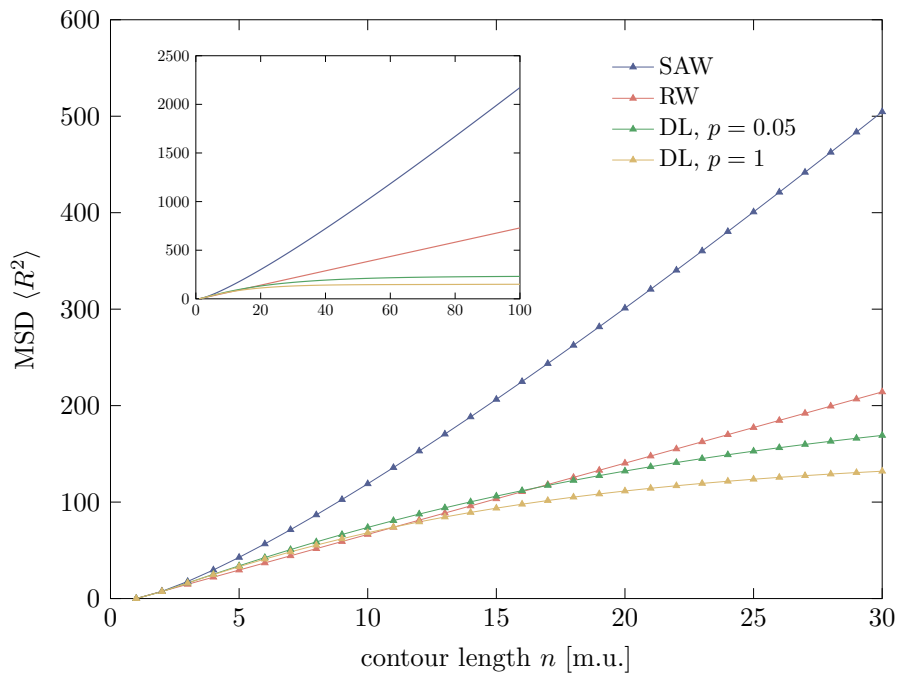


Figure 3.4: Results for the mean-square distance $\langle R_n^2 \rangle$ for various simulated polymer chains (with systems size $N = 250$), namely the simple random walk (RW) model, the self-avoiding walk (SAW) model and furthermore the model incorporating dynamic looping (DL) for two different looping probabilities. In comparison to the MSD resulting from the RW and the SAW model, that of a loopy polymer displays a leveling-off and adopts a plateau level at contour lengths above about 30 monomeric units (see inset).

3.2.3 Bending Rigidity and Persistence Length

The persistence length l_p is a basic mechanical property of semiflexible polymers that quantifies the bending stiffness and is defined as the contour length over which directional correlations due to thermal fluctuations are lost within the chain backbone. The parts of the polymer that are shorter than the persistence length can be described as rigid whereas parts of the polymer which are larger than the persistence length can be considered as a random walk. However, just recently a study investigating conformational properties of bottle-brush polymers showed that standard definitions of persistence length do not describe the local “intrinsic” stiffness of real polymer chains [68]. Rather, the decay of the orientational correlation function can be interpreted as an effective persistence length describing conformational properties on a global scale.

We can integrate bending rigidity U_{bond} into the BFM by introducing a weight factor for the bond angles given by

$$U_{\text{bond}} = \kappa_{\theta}(1 - \cos(\theta)) .$$

The distribution of the a priori probability distribution of the bond angles within the BFM, which is illustrated in Fig. 3.2, is especially important for arguing that bending rigidity can be actually introduced within the framework of the BFM.

3.2.4 Confinement

The conformation of biological macromolecules is highly dynamic, rapidly adapting to physiological and environmental conditions, such as confined spaces. The bacterial nucleoid is a striking example of an environment that strongly influences the packaging of the genetic material of prokaryotes.

Confining polymers affects their conformation by excluding volume interactions and thus entropy. For the simulation of ring polymers within the scope of the study of the bacterial nucleoid it is important to introduce the concept of spatial confinement. The interplay of confinement, bending rigidity as well as excluded volume interactions has been already studied in detail [69]. Nevertheless we reproduce important findings of this theoretical study (see Fig. 3.5), namely the bond angle correlation function for circular polymers within cubic and rectangular confining geometries.

3.2.5 Contact Probability Measures

3C-based technologies, such as Hi-C, are experimental methods that can quantify the contact frequency between different sites of the DNA molecule as outlined in chapter 2. Fortunately, in our simulations the contact frequency can be measured comparatively simple since we know the exact configuration of our polymer, i.e. the position of each single monomer in the three-dimensional space, at each point in time. In order to be able to quantify the contact frequency, we initially have to define a criterion for two monomers being in contact with each other. We recall that within the framework of the BFM we restrict the bond length between a monomer k and its neighboring monomer j such that it meets the condition

$$\|\mathbf{r}_k - \mathbf{r}_j\|_2 \leq \sqrt{10} .$$

Analogously, we choose the same upper limit of $\sqrt{10}$ l.u. for two monomers k and j being in contact with each other. For making the step from contact frequency to contact

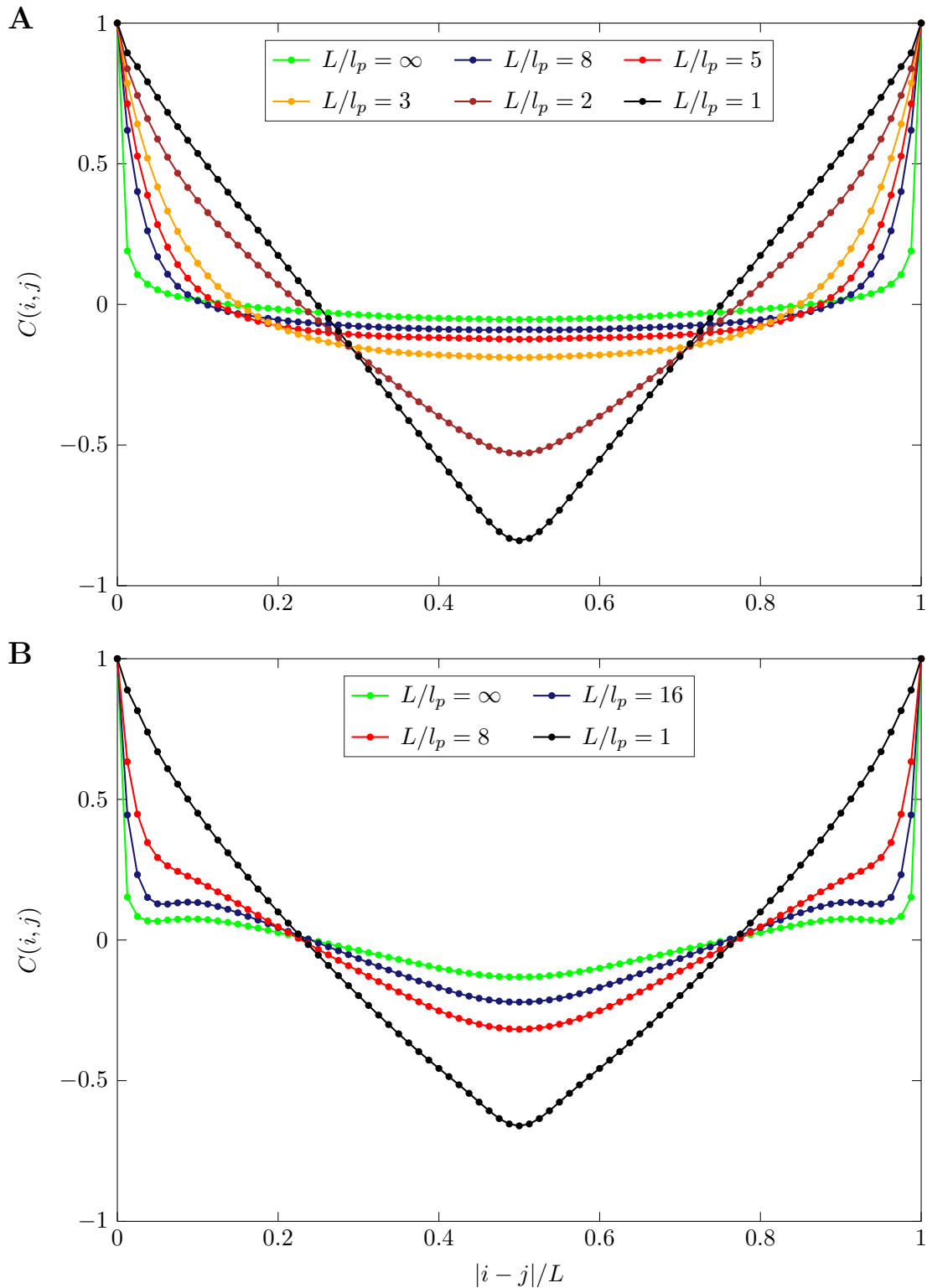


Figure 3.5: Mean bond angle correlation function $C(i, j) = \langle \mathbf{u}(i) \cdot \mathbf{u}(j) \rangle$ along the polymer backbone $|i - j| \in [0, L]$ for both **A.** an unconfined semiflexible ring polymer and **B.** a semiflexible polymer ring confined in rectangular geometry for various persistence lengths L/l_p . Both simulated polymers are composed of $N = 80$ monomers.

probability we have to average over the whole ensemble of conformations and normalize it in a subsequent step.

Contact Probability Profile

One important measure is the decay of the contact probability as a function of genomic distance which we also refer to as contact probability profile. Treating the polymer chain as a random walk, the probability of two beads n_1 and n_2 contacting each other decreases as a function of their genomic separation $|n_2 - n_1|$. More specifically, a power-law behavior is observed

$$p_c(|n_2 - n_1|) \propto |n_2 - n_1|^{-3/2}. \quad (3.2)$$

To draw a comparison between simulation and experimental data from 3C-based experiments, we consider the contact probability $p_c(|n_1 - n_2|)$ of two genomic loci n_1 and n_2 . According to equation 3.2 the contact probability of a random-walk polymer obeys a power-law behavior as a function of the length $l = |n_1 - n_2|$ and is given by $p_c(l) \propto l^{-\beta}$, $\beta = 1.5$.

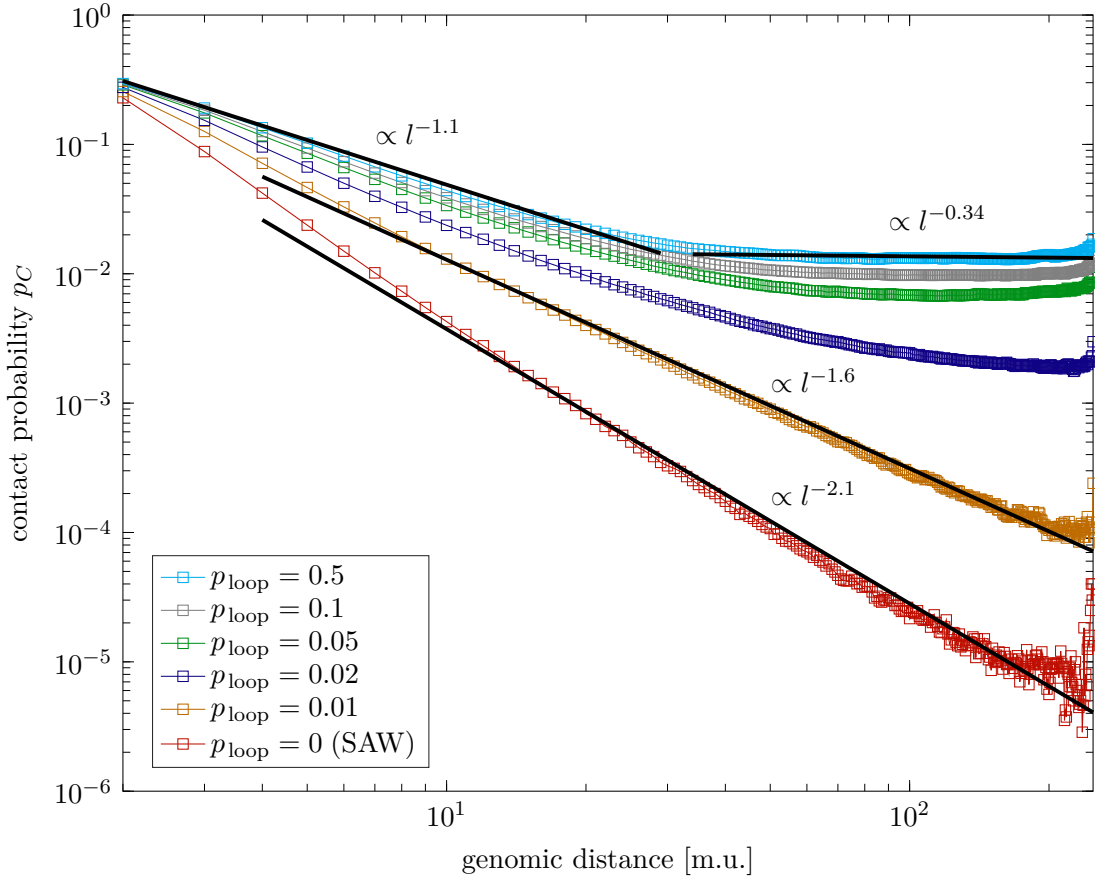


Figure 3.6: The contact probability p_c for two specific sites as a function of the genomic separation between them. Shown are the results for equilibrated polymers composed of $N = 250$ monomers and various looping probabilities including the case of the self-avoiding walk ($p = 0$) and a simple random walk. The contact probability decreases as a power-law $l^{-\beta}$ with genomic separation for separations $n \gtrsim 10$. The exponent is thereby strictly dependent on looping probability. Compared to the self-avoiding walk the co-localization probability is strongly increasing due to dynamic looping.

Self-avoiding walk polymers also show power law behavior and the exponent is given by $\beta \approx 2.1$ (see Fig. 3.6). The contact probability $p_c(l)$ of the dynamic loop model has a power-law behavior as well and, as expected, it is observed that the higher the looping probability, the smaller the exponent β (see Fig. 3.6). For higher looping probabilities p_{loop} , one has distinguish two regimes. In comparison to self-avoiding walk polymers the contact probability of distant genomic loci is increased by several orders of magnitude. Interestingly, the computed value of the exponent $\beta_1 \approx 1.1$ for a looping probability of $p_{\text{loop}} = 0.5$ and for small genomic separations (2 - 25 m.u.) is in line with the experimental observation of an exponent of $\beta_1 \approx 1.08$ in the range between 500 kbp and 7 Mbp [3]. In accordance with the dynamic loop model, the experimentally observed exponent for all genomic separations is even smaller on the genome size scale.

Contact Map

In order to visualize the contact probability of all the monomers with each other, a two-dimensional heat map, the so-called contact matrix or contact map, is appropriate. An exemplary contact map is depicted in Fig. 3.7. Contact maps encode contact probability by colors allowing for fast recognition of the prominent features. Our simulated contact maps are unbiased and do not need to be normalized since their building blocks are actually the monomers whose interactions are known.

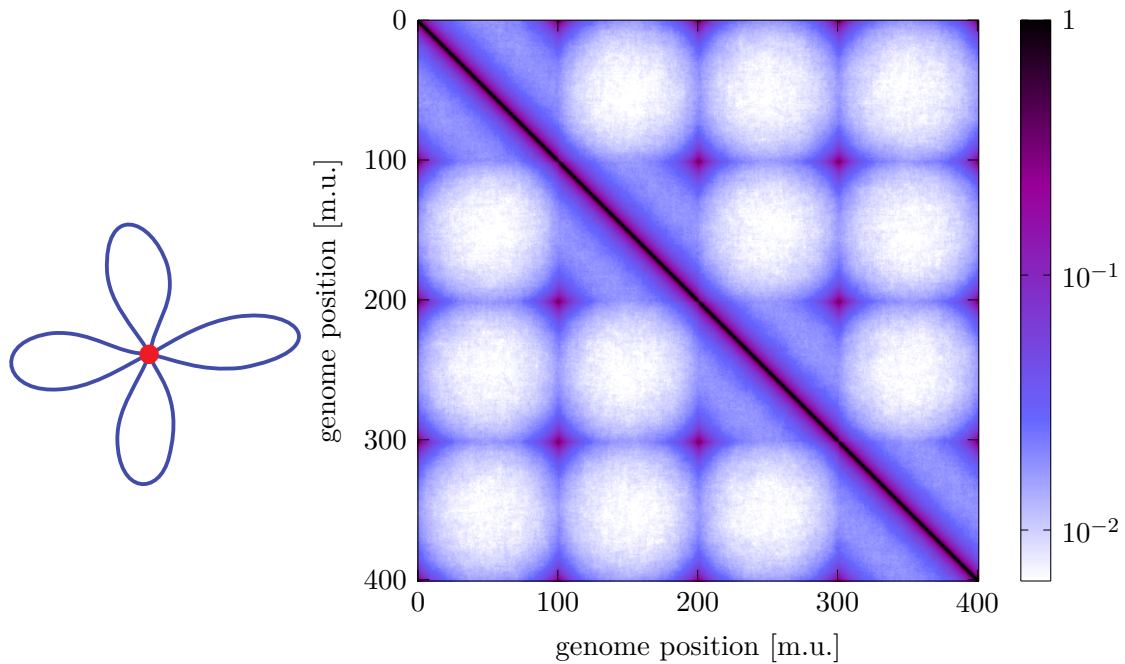


Figure 3.7: Exemplary contact map of a rosette ($N = 401$ monomers) composed of four arms (or loops), each consisting of $N = 100$ monomers.

3.3 Computational Topology

3.3.1 Simplicial Complexes and Barcodes

The linchpin of the following considerations is persistent homology, that studies the evolution – birth, life and death – of global topological features like connected components, holes

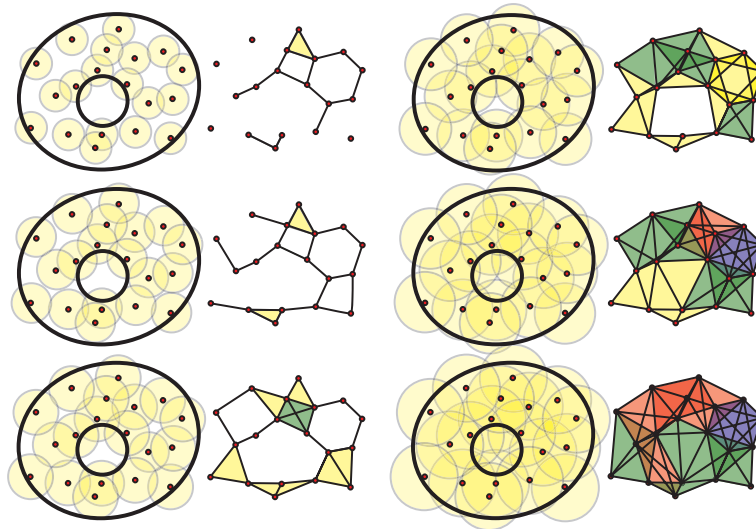


Figure 3.8: A sequence of simplicial complexes for a point cloud data set [62]. Left: Spheres around the points with radius ϵ . Right: The corresponding simplicial complex \mathcal{R}_ϵ . Different colors mark simplices of different dimensions. Upon increasing ϵ , holes appear and disappear.

and cavities. In order to determine the persistent homology of a space, the space needs to be represented as a simplicial complex at first. The simplest simplex of dimension 0 is a point. From this one can construct higher-dimensional simplices. In order to get a simplex of dimension n , add to the simplex of dimension $(n - 1)$ a new point in n -dimensional space and connect all vertices of the old simplex with the new point. Therefore a simplex of dimension 1 is a line, a triangle a 2D one, a tetrahedron a 3D one and so on, where each n -dimensional simplex has exactly $(n + 1)$ vertices. Thus, a simplicial complex is a set consisting of points, line segments, triangles, and their n -dimensional analogues (see Fig. 3.8).

The formation of simplicial complexes and their persistence can be graphically illustrated: A complex is formed by placing spheres with radius ϵ around the points and by connecting the centers of two intersecting spheres to a simplex of dimension 1.

The formation of a simplicial complex results from starting at a radius $\epsilon = \epsilon_0$ and continuously increasing it. Based on this, the barcodes can now be determined. Their length corresponds to the interval $[\epsilon_{\text{start}}, \epsilon_{\text{end}}]$, in which “inclusions” in the structure of the simplicial complex persist, where ϵ_{start} and ϵ_{end} correspond to the radii where “inclusions” occur and disappear. The barcodes H_0 of dimension 0 represent the intervals in which the points are not connected. In one dimension, periods (with respect to ϵ) are encoded in which closed polygons with more than 3 vertices exist, and barcodes of dimension 2 (H_2) indicate three-dimensional volume inclusions.

3.3.2 Hausdorff Distance

The calculation of barcodes alone is not enough, rather the challenge is to compare barcodes in order to determine their similarity. Distance measures can be used to quantify similarity or dissimilarity. Such a distance measure $d(x, y)$ of two points x and y must have certain properties [70]:

- $d(x, y) \geq 0, \quad d(x, x) = d(y, y) = 0 \iff$ The distance between two different points $x \neq y$ is *positive* and zero in the case of $x = y$.
- $d(x, y) = d(y, x) \iff$ Distance is *symmetric*.

If it also satisfies the *triangle inequality*

- $d(x, z) \leq d(x, y) + d(y, z), \iff$ The distance between two points x' and z is the shortest distance along any path.

then the distance function is also referred to as a metric.

The most well-known distance between two non-empty sets is defined as the infimum of the distances between any two of their respective points, i.e.

$$d(A, B) = \inf_{a \in A} (\inf_{b \in B} (d(a, b))), \quad (3.3)$$

where the function $d(a, b)$ is the Euclidean norm of two points $a, b \in \mathbb{R}^2$ in our case. Here, one is interested in finding the shortest of all possible distances between the elements of A and B .

With regard to the barcodes, however, we now use the Hausdorff metric as the distance measure. Let A and B be two non-empty compact subsets of a metric space M , then the Hausdorff distance is defined as follows [71]:

$$d_H(A, B) = \max \{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \}, \quad (3.4)$$

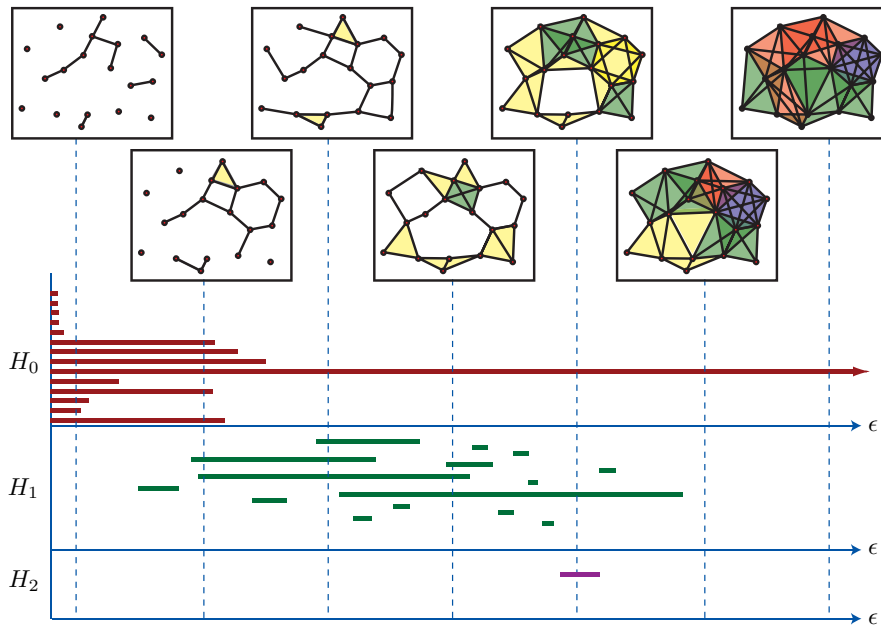


Figure 3.9: An example for barcodes [62].

where $d(a, b)$ again denotes the Euclidean norm. Hence, the Hausdorff distance is the maximum distance from one set to its nearest point in the other set (see Fig. 3.10).

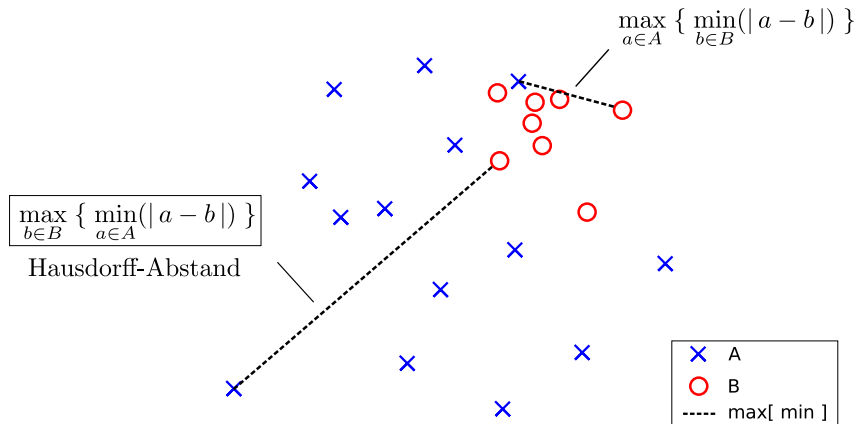


Figure 3.10: Components of the calculation of the Hausdorff distance between two 2D point clouds.

The individual “bars” of the barcodes can be viewed as points in \mathbb{R}^2 , where the x -coordinate is ϵ_{start} and the y -coordinate ϵ_{end} of the corresponding bar. The values that x and y can take are restricted to the interval $[0, R]$, where R denotes the radius of the perimeter of the structure in question. If we look at all barcodes of one dimension as a set of points, the result is a *persistence diagram*. The Hausdorff distance can now be applied to two of these *persistence diagrams*. The maximum distance that can be reached is $\sqrt{2} R$.

It should be noted that for a given set of points, as stated above, there are barcodes of various dimensions, therefore also more than one *persistence diagram*. To compare two point sets topologically with the Hausdorff metric, it only makes sense to compare the Hausdorff distance of *persistence diagrams* that represent the same dimension. In this work, the barcodes of dimension 0 and 1 were calculated. As a result, there are two different Hausdorff distances for two point sets under consideration. To obtain a single value that captures the similarity of two sets of points as a whole, and to give equal weight to each of the two dimensions, these two values are summed up. If these values are calculated for the same structures, which differ only in the order of magnitude in which they are realized, different values are obtained in spite of the same topology. Therefore it makes sense to project the calculated values onto a common scale. This can be achieved by dividing the summed values resulting from the individual *persistence diagrams* by the radius of the perimeter R of the structure under investigation. The resulting normalized value is referred to as Hausdorff distance. It should be noted that the values that this Hausdorff distance can take are generally not restricted to the interval $[0, 1]$, but to the interval $[0, 3\sqrt{2}]$. In addition, for any distance measure, as opposed to a similarity measure, a value of 0 corresponds to maximum similarity.

Chapter 4

Processing and Analysis of Hi-C Data on Bacteria

References

The results presented in this chapter are published as and adapted from

- A. Hofmann and D.W. Heermann (2018), *Processing and Analysis of Hi-C Data on Bacteria*. Bacterial Chromatin, Methods in Molecular Biology, 1837:19-31. doi: 10.1007/978-1-4939-8675-0_2.

We thank Remus T. Dame, Frédéric Crémazy and Fatema Zahra Rashid for the stimulating and fruitful discussions.

Chapter Summary

The study of the three-dimensional genome organization has recently gained much focus in the context of novel techniques for detecting genome-wide contacts using next-generation sequencing. These genome-wide chromosome conformation capture-based methods, such as Hi-C, give a deep topological insight into the architecture of the genome inside the genome in vivo. This chapter reviews the steps to process next-generation Hi-C sequencing data to generate a final contact probability map. We describe these steps using publicly available Hi-C datasets of different bacteria. We also present strategies to assess the quality of Hi-C datasets.

4.1 Introduction

In humans, nearly two meters of DNA must be folded to fit inside the micrometer-sized cell nucleus. How is DNA compacted to this level and how can it remain accessible for gene transcription, replication and repair at the same time? Novel technologies, such as “chromosome conformation capture” (3C) – based methods that map genome-wide spatial interactions along the genome have, during the last 15 years, allowed to shed light on this question. Massive improvements in the throughput of such methods produce ever-increasing amounts of data. Most of the raw data are deposited in repositories publicly available.

Mammalian interphase chromosomes are hierarchically organized [25,72]. Fluorescence microscopy and genome-wide 3C studies, such as Hi-C, have revealed inter-chromosomal compartmentalization in the form of distinct chromosome territories [3,24]. Individual chromosomes also exhibit compartmentalization to form domains [7,8,33,36]. These 3C studies indicate that eukaryotic genomes are partitioned into discrete structural units with highly increased frequency of internal contacts.

Besides studies of eukaryotic chromosomes of humans, mice and *Drosophila melanogaster*, recently, the circular chromosomes of model bacterial species such as *Caulobacter crescentus*, *Bacillus subtilis* and *Mycoplasma pneumoniae* have been shown to analogously be composed of domains with the help of Hi-C analyses [48,73–75]. Taken together, these results suggest that intra-chromosomal compartmentalization is a fundamental building block of chromosome structure of organisms.

3C was invented in 2002 by Dekker et al. [21] and allows for focused quantification of contact frequencies at selected regions. All 3C-based techniques aim to generate a two-dimensional library of three-dimensional chromosome contacts. The first step in the procedure is cross-linking of chromatin by addition of formaldehyde. This causes interacting chromatin segments to be covalently linked together. The fixed chromatin is then digested with a restriction enzyme. The ends of the fragmented cross-linked DNA are thereafter re-ligated under diluted conditions to favor intramolecular ligation of the cross-linked fragments. The ligated DNA molecules thereby form a hybrid of two DNA fragments from the two segments that were cross-linked. Next-generation sequencing is used to quantify the number of such hybrid DNA-molecules.

Hi-C is a genome-wide 3C based technology introduced by Lieberman-Aiden et al. in 2009 [3]. Realizing a genome-wide quantification of interactions, it constitutes a major breakthrough in the study of chromatin architecture. The Hi-C protocol differs from the standard 3C protocol in that there is an extra step needed before ligation. It consists of filling in the restriction digest of the chromosome with biotin-labeled nucleotides. After purification and shearing/ fragmentation of the Hi-C library, the biotin labeled material is pulled down to ensure that only ligation junctions are selected for further analysis.

“Chromosome conformation capture carbon copy” (5C) captures interactions between all restriction fragments within a selected region [55]. For example, it was used to study the spatial organization of the bacterial *Caulobacter crescentus* genome [76], the regulatory landscape of mouse X inactivation [8] and the long-range interaction landscape of gene promoters in the human genome [40]. A technique similar to Hi-C, called genome conformation capture (GCC), has been applied for mapping yeast chromosome interactions [57] as well as for studying the spatial organization of the *Escherichia coli* nucleoid [58].

All 3C-based methods, contrary to microscopy-based techniques, allow for both a more systematic and quantitative characterization of genome topology and a higher resolution. The essential drawback, however, is that the conventionally ensemble 3C-based methods are mostly performed on large populations of cells, leading to loss of information at the single-cell level.

In this chapter, we give an overview of the data analysis involved in the framework of a Hi-C experiment. Moreover, we present and discuss publicly available Hi-C datasets of bacterial genomes and present possibilities to assess and compare them in terms of data quality.

4.2 Hi-C Data Processing

This section covers the main steps involved in the data processing of a genome-wide 3C-based study. Since the focus of this chapter is mainly on genome-wide methods, such as Hi-C or GCC, methods relevant for these technologies are discussed in this section. Hi-C data processing can be subdivided into the following four main steps: (1) Mapping to the reference genome, (2) Quality control, (3) Binning and contact matrix generation, (4) Balancing (Fig. 4.1). Each of these steps is discussed in the following section.

4.2.1 Mapping to the Reference Genome

The first step of genome-wide 3C data analysis consists of mapping reads back to the reference genome. The Hi-C method quantifies an interaction by a ligation product formed between two restriction fragments. By using paired-end sequencing and mapping both ends of each paired sequence to the reference genome, the two restriction fragments in the ligation product can be determined. However, if the read length is bigger than the length of one of the restriction fragments, the mapping will not work. To solve this problem, the mapping procedure can be refined by means of an iterative mapping scheme that involves truncating reads to a smaller length prior to mapping [77]. Reads that are not aligned uniquely at both ends are then re-aligned by iteratively increasing their portions. This process is repeated until either all reads uniquely map or until the read is extended to its entirety. Only paired-end reads with both sides being uniquely mapped to the reference genome contribute to the set of Hi-C interactions. All other paired-end reads are discarded.

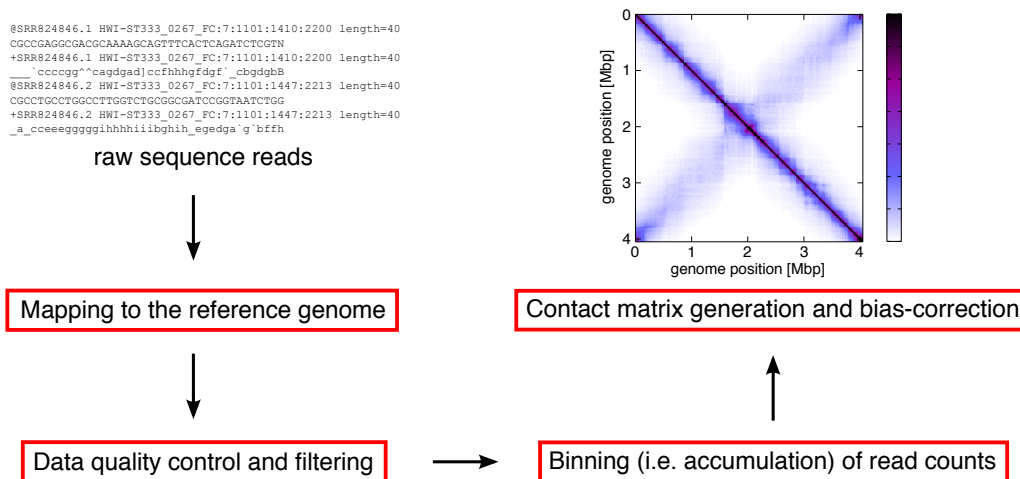


Figure 4.1: Schematic overview of the general workflow for analyzing genome-wide 3C data such as Hi-C. The input (at the top left) is raw sequencing data and as a typical output we illustrated a bias-corrected contact matrix (at the top right). The dataset used here for illustration purposes is that of the study of Le et al. [48] (SRR824846).

4.2.2 Quality Control

The next step is quality control to ensure that the aligned sequence reads are likely to be the result of proximity-based ligation of digested fragments, and that they are reflecting long-range chromatin interaction rather than just random collision. Self-circularized or un-ligated (dangling-end) products will result in reads that map with both ends on the same restriction fragment. These reads should be removed. Also reads from neighboring fragments that map to the same strand should be removed since they are likely the result of incomplete digestion. Furthermore, reads that map multiple times at the exact same location on the reference genome are often the result of biased PCR-amplification and should also be removed. Hi-C sequencing reads can be compared to randomly generated control sequencing reads, whereby the Hi-C reads should be significantly closer to the chosen restriction sites than random reads [78]. The Hi-C reads should also be in the correct orientation with respect to the restriction site.

4.2.3 Binning and Contact Matrix Generation

After the alignment of the sequence reads and quality control the next step is the construction of contact matrices of the interaction data. To produce a contact matrix, the genome is divided into equally sized loci, so called bins. The result of this aggregation of read-counts across bins is a symmetric matrix composed of interaction frequencies between bins covering the entire genome. The size of these bins used to represent the meaningful contacts between pairs of genomic loci can be referred to as the resolution of a Hi-C experiment (see Fig. 4.2 for a comparison of a contact matrix depicted at different bin sizes).

A linear increase of resolution requires a quadratic increase in total sequencing depth; the size of the bins and effectively the resolution is limited by sequencing depth. Moreover, the resolution also depends on the restriction enzyme used for the Hi-C experiment (see

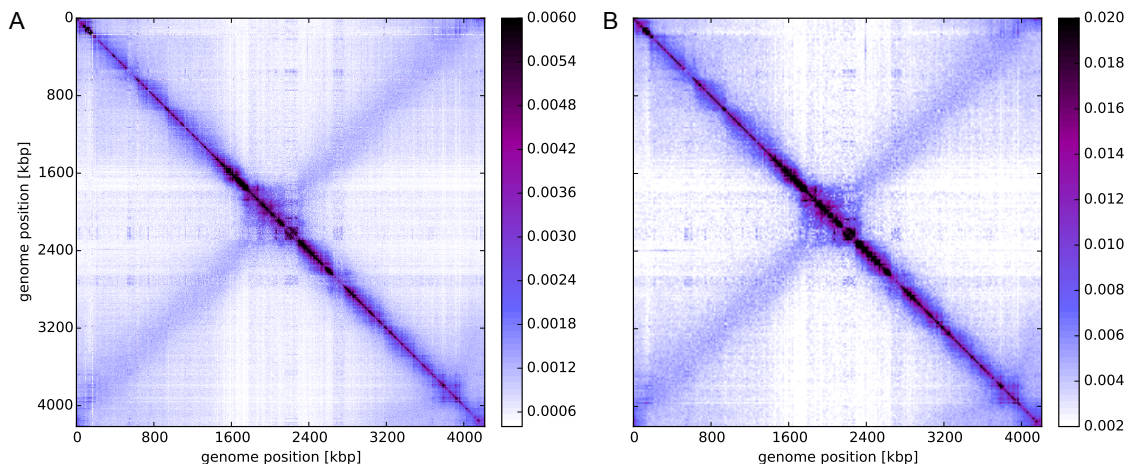


Figure 4.2: Hi-C contact map depicted at two different resolutions (**A.** 4kbp and **B.** 16kbp). The contact map of higher resolution reveals much more detailed structures, especially on the interactions within the domains along the diagonal. Data on *B. subtilis* from Marbouty [73] (SRR2772323).

section 4.3.4). To illustrate this relationship between resolution and the total number of collected read pairs, we want to highlight that for a resolution of 100 kbp in the first Hi-C study of the human genome 8.4 million reads were collected [3]. Increasing the resolution to 1 kbp required 4.9 billion reads [36]. Thus, increasing the resolution by two orders of magnitude required increasing the total number of reads by three orders of magnitudes.

4.2.4 Balancing

Besides the bias introduced by individual reads or restriction fragments, binning generates biases, as well. Yaffe and Tanay identified the origin of some of these biases, such as the non-uniform distribution of the length of restriction fragments with respect to ligation efficiency, the nucleotide composition of the genome under investigation and issues with uniquely mapping the interactions back onto the reference genome. They proposed an integrated probabilistic model [78] for eliminating these known systematic biases from the “raw” contact maps. This procedure is referred to as normalization.

Several other models for normalizing Hi-C contact maps have been proposed [77, 79, 80]. Though, these approaches do not explicitly incorporate the aforementioned biases on the grounds that it is not possible to know each and every bias. Since most of these approaches are based on the Sinkhorn-Knopp (SK) balancing algorithm [81], they can be more precisely referred to as balancing instead of normalization. Explicit bias correction and balancing yield comparable results [36].

In this section, we focus on the SK balancing algorithm [81] that transforms a symmetric non-negative matrix $\mathbf{A} = (a_{ij})$, $\mathbf{A} \in \mathbb{R}_+^{n \times n}$, into a doubly stochastic matrix $\mathbf{S} = (s_{ij})$, that is, a matrix whose rows and columns sum up to 1, i.e.,

$$\sum_i s_{ij} = \sum_j s_{ij} = 1.$$

The SK algorithm is an iterative process that consists in solving

$$\mathbf{S} = \mathbf{D}_1 \mathbf{A} \mathbf{D}_2, \quad (4.1)$$

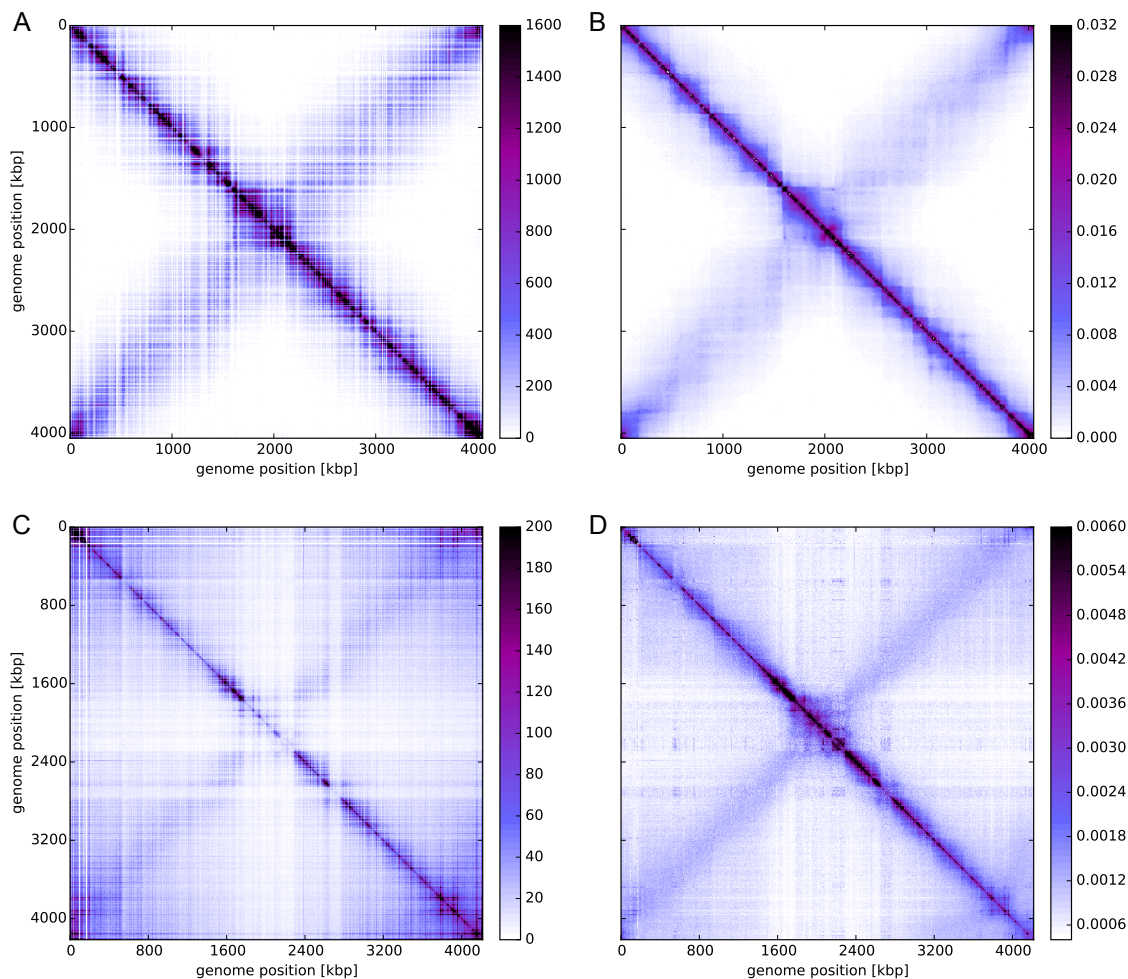


Figure 4.3: Comparison of “raw” Hi-C contact frequency matrices on the left and the respective balanced equivalents, i.e. contact probability matrices, on the right. **A, B.** Data on *C. crescentus* of the study of Le et al. [48] (SRR824846). **C, D.** Data on *B. subtilis* of the study of Marbouty et al. [73] (SRR2772323). We balanced the matrices by applying the Sinkhorn-Knopp (SK) algorithm. These two examples illustrate that the balancing procedure leads to very different results dependent on the input “raw” Hi-C contact frequency matrix. In the case of the upper matrix on *C. crescentus*, the coverage of reads is homogenous along the genome and the contact frequency matrix (**A**) already indicates a domain structure along the diagonal. In contrast, the coverage of reads is considerably more heterogeneous in the case of the lower matrix on *B. subtilis* (**C**). The region from 1.7 up to 2.3 Mbp lacks reads. Therefore the structure of the balanced contact map (**D**) in this region is rather an artifact of the balancing procedure than of biological relevance.

where \mathbf{D}_1 and \mathbf{D}_2 are unique up to a scalar factor diagonal matrices with positive main diagonal. The matrices \mathbf{D}_1 and \mathbf{D}_2 are obtained by alternatingly normalizing columns and rows of \mathbf{A} .

Applied to our situation, \mathbf{A} constitutes the raw matrix of contact frequencies, and the diagonal matrices \mathbf{D}_1 and \mathbf{D}_2 contain the biases for the bins involved in the contacts between bins i and j . \mathbf{S} is then the matrix of unbiased relative contact frequencies, which is defined such that each row and column of the upper triangular matrix sums to 1. The biases, and therefore \mathbf{S} , can be found by using the SK algorithm that converges to the solution of equation 4.1.

Regardless of the method used, it is important to compare the raw and the bias-corrected contact map to check the effect of the procedure. Fig. 4.3 depicts two such comparisons for two different Hi-C datasets. In both cases, balancing has the effect of smoothing the contact map such that no obvious high or low rows and columns remain. In extreme cases (compare the two lower maps in Fig. 4.3), large regions with a strikingly low number of captured reads are completely reorganized by balancing. The resulting contact map should be regarded with caution because it is likely to contain artifacts caused by balancing, especially in and near the aforementioned regions.

4.2.5 Concluding Remarks

It is important to be aware of the limitations of the experimental elements that impact the computational data processing. In fact, 3C experiments are never truly genome-wide since the usage of a restriction enzyme is biased to where its sites are located in the genome. Furthermore the restriction enzyme chosen has to balance, amongst others, the frequency of cutting, fragment size and size uniformity across the studied genome.

Because of the necessity of balancing contact maps constructed from 3C data, it is crucial to differentiate between raw contact maps representing all the captured contacts and their raw frequencies and the balanced contact maps representing contact probabilities. Regardless, balanced contact maps should be compared with their raw equivalent in order to check the effects of the balancing procedure.

4.3 Hi-C Data Assessment

In this section, we focus on Hi-C and other genome-wide 3C experiments performed in bacteria. We also present a procedure to assess data quality for comparison of data sets. Finally, we highlight the importance of the choice of restriction enzymes for Hi-C experiments in terms of data analysis.

4.3.1 Hi-C Data Availability

The first Hi-C study was carried out by Lieberman-Aiden et al. [3] and addressed the folding of the human genome. In 2011 Umbarger et al. [76] described the first 3C-based study, here 5C, of a bacterial genome. To date there have been several studies of bacterial genomes using genome-wide 3C-based experiments. We have summarized them in table 4.1 in chronological order. There are several datasets for each study that correspond to Hi-C libraries that differ, for example, in bacterial growth conditions, the restriction enzyme used or the sequencing method. We listed studies with publicly available Hi-C data sets.

4.3.2 How Many Reads?

Sequencing depth is decisive for the resolution of a Hi-C experiment. More precisely, it is the number of valid reads, i.e. those that remain after the filtering step discussed in section 4.3.4. Therefore, given a Hi-C dataset the relevant question to ask is: How many “valid” reads remain after the filtering process? Although in the end the absolute number of valid reads is decisive from the data processing point of view, the relative fraction of the number of valid and total reads is also very important from an experimental point of view since it determines something like the efficiency of a Hi-C experiment. There are large variations in the number of valid reads in the datasets under discussion. While it exceeds

Bacterial Species	xC	Res _{max}	RE	Datasets	Reference
<i>Caulobacter crescentus</i>	5C	13 kbp	BglIII	3	Umbarger et al. [76]
<i>Escherichia coli</i>	GCC	20 kbp	HhaI	8	Cagliero et al. [58]
<i>Caulobacter crescentus</i>	Hi-C	10 kbp	BglIII, NcoI	23	Le et al. [48]
<i>Bacillus subtilis</i>	Hi-C	4 kbp	HindIII, HpaII	15	Marbouty et al. [73]
<i>Bacillus subtilis</i>	Hi-C	10 kbp	HindIII	41	Wang et al. [74]
<i>Mycoplasma pneumoniae</i>	Hi-C	3 kbp	HindIII, HpaII	8	Trussart et al. [75]

Table 4.1: Overview of genome-wide 3C-based (xC) datasets of bacterial species for different conditions. We list the restriction enzymes (RE) that have been used in the datasets as well as the highest achieved resolution (Res_{max}) and the number of datasets that have been published.

50% for the Hi-C data on *C. crescentus* (Le et al. [48]), it amounts to less than 5% for that on *M. pneumoniae* (Trussart et al. [75]). The latter means that, in the extreme case, albeit 187 million total reads have been sequenced, only 3 million valid reads contribute to the contact map (ERR1413594). In most cases, the key limiting factor is ligation efficiency, i.e. unligated fragments (dangling-ends), self-ligated fragments (self-circles) and poor biotin label removal.

4.3.3 What Kind of Reads?

The result of a Hi-C experiment is a contact probability matrix of the binned genomic interactions. It is visualized as a heat map of relative interaction frequencies (a “contact map”) encoding the interaction data using a color map. This graphical representation

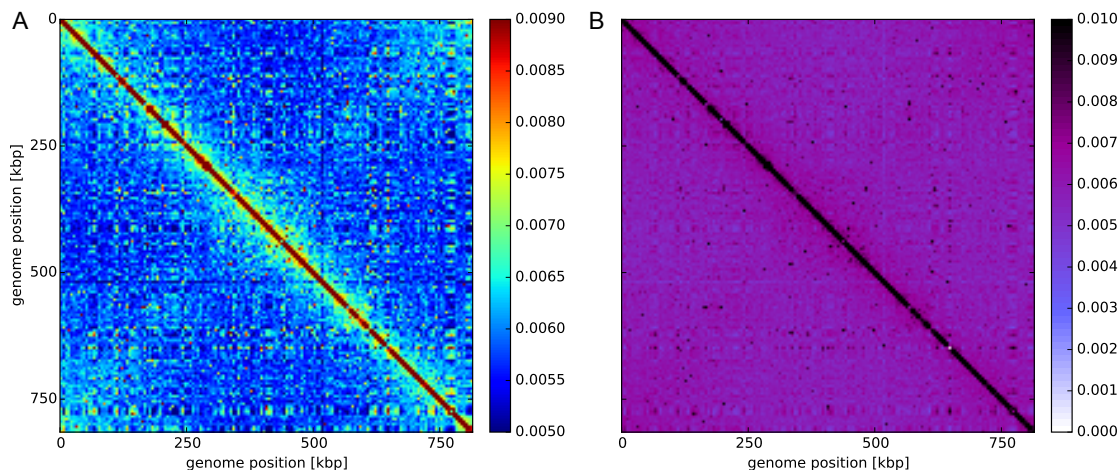


Figure 4.4: Hi-C contact map depicted for two different color maps and different lower limits of the color bar. Since the Hi-C data contains a high level of noise a lower limit on the color bar needs to be set. **A.** A rainbow color map highlights the hardly varying contact probabilities much better than **B.** a sequential colormap. Data on *M. pneumoniae* from Trussart [75] (ERR1413595).

gives the observer an immediate impression of the Hi-C data since it highlights prominent features, such as domains along the diagonal or the presence of a secondary diagonal. The graphical representation is highly dependent on the choice of the color map and on its scaling that can either be linear or logarithmic. An example of how different the same contact probability matrix can be visualized, is illustrated in Fig. 4.4. The heat map with the rainbow color map and the appropriate lower limit on the color bar on the right highlights the hardly varying contact probabilities much better than the sequential color map on the right.

For this reason, it is important to evaluate a Hi-C dataset objectively in addition to heat map visualization. One possibility is to regard the number of the captured reads as a function of the genomic distance between the interacting loci (see Fig. 4.5). In all Hi-C experiments a lot of reads are detected between neighboring genomic sites. This is, of

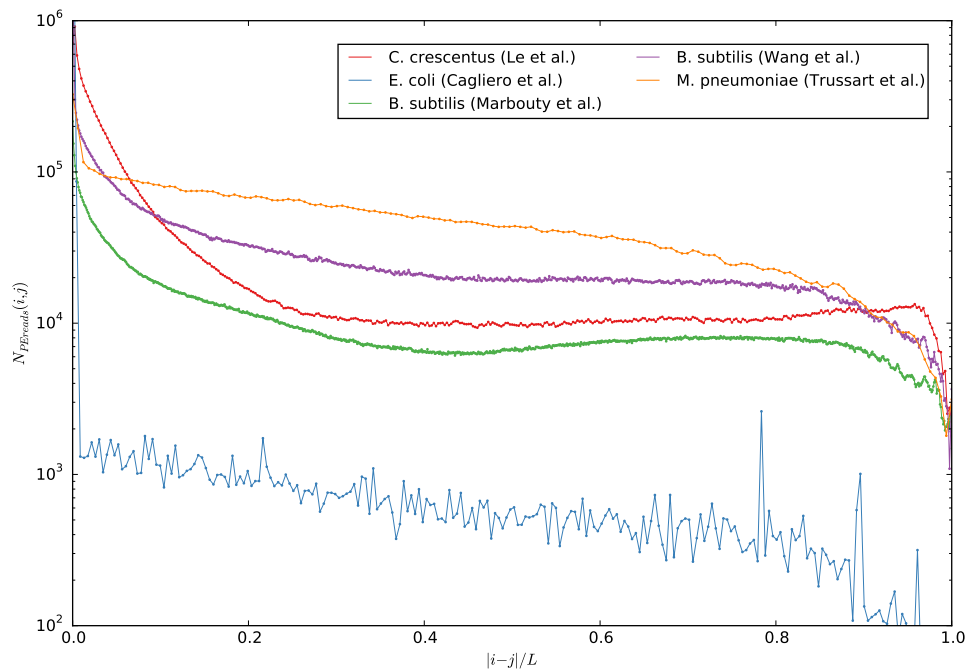


Figure 4.5: Number of valid counts between different genomic loci i and j as a function of their genomic distance (which has been normalized by the length L of the respective genome), depicted for different bacteria. Most of the counts represent interactions between neighboring or near genomic loci and with increasing genomic length between the interacting loci the counts decrease exponentially. This exponential decay is different for the various datasets. This is a consequence of the distinct three-dimensional shape of the different genomes that is reflected in distinct interaction patterns. Moreover, also experimental factors have an impact on the interaction profile as can be seen from the interaction profiles of the two *B. subtilis* Hi-C datasets. The two curves have approximately the same slope but one is shifted upwards (Wang et al.) compared to the other one (Marbouty et al.) because of the larger number of valid counts stemming from deeper sequencing of the corresponding Hi-C library. Furthermore, the read distribution also reveals what kind of reads have been captured in the Hi-C experiment, such as in the case of the GCC dataset of the study by Cagliero et al. where a sharp decline of contacts indicates that it is highly biased towards contacts of neighboring restriction fragments (99.5% of the valid read counts occur between neighboring bins).

course, due to the fact that these sites are intrinsically linked to each other. Thus, these reads do not contain relevant information. The read distribution shows how large this fraction of trivial reads is compared to more interesting short-, intermediate- and long-range interactions related to the distinct three-dimensional shape of the regarded genomes. It can therefore be used as an appropriate indicator for the quality of the reads of a given Hi-C dataset.

The read distribution also shows that the captured interactions decrease with increasing genomic distance, such that short-range interactions will typically have higher coverage and thus higher effective resolution. In the contact map this is reflected by a gradual decrease of the average interaction probability the further one moves away from the diagonal. This finding follows the intuition that topologically close loci will interact frequently assuming random motion in 3D space. Also, polymer models predict a power-law decrease of the contact probability as a function of the genomic distance [82].

4.3.4 Resolution of Hi-C Data and the Selection of Restriction Enzymes

Restriction enzymes are proteins that cut the DNA at specific sites. There is a wide range of restriction enzymes and in addition to biochemical criteria the choice of a certain restriction enzyme is also relevant for data analysis since it inevitably determines the resolution of a Hi-C map. This is due to the fact that in a Hi-C experiment interactions between genomic loci are measured in terms of restriction fragments. Thus, choosing an appropriate bin size for a given Hi-C dataset very much depends on the distribution of the lengths of the restriction fragments.

In Fig. 4.6 we have contrasted the distribution of the length of a high-frequency restriction enzyme (frequent cutter) with a low-frequency cutter (rare cutter) for the *E. coli* genome. Clearly, we could choose the frequent cutter *HhaI* in order to obtain an *E. coli* contact map of 1 kbp or even 500 bp resolution, whereas the rare cutter *BglII* would limit the maximum possible resolution to around 10 kbp. It is important to remark that the choice of a frequent cutter predetermines the need of a high sequencing depth. A summary of various restriction enzymes and the statistics of the length of the respective restriction fragments for the *E. coli* genome can be found in table 4.2.

4.4 Summary

Genome-wide chromosome conformation capture-based methods are now widely used. The resulting data, most of which is deposited in repositories publicly available, is the starting point for the computational modeling of the three-dimensional architecture of a multiplicity of genomes. Besides the results of the modeling approaches, it is also interesting to know how reliable the underlying Hi-C data is. It is not only the number of captured reads, but also the distribution of these reads which is decisive for this question. Moreover, contact maps before and after balancing can differ significantly. It is therefore necessary to compare them and understand possible deviations.

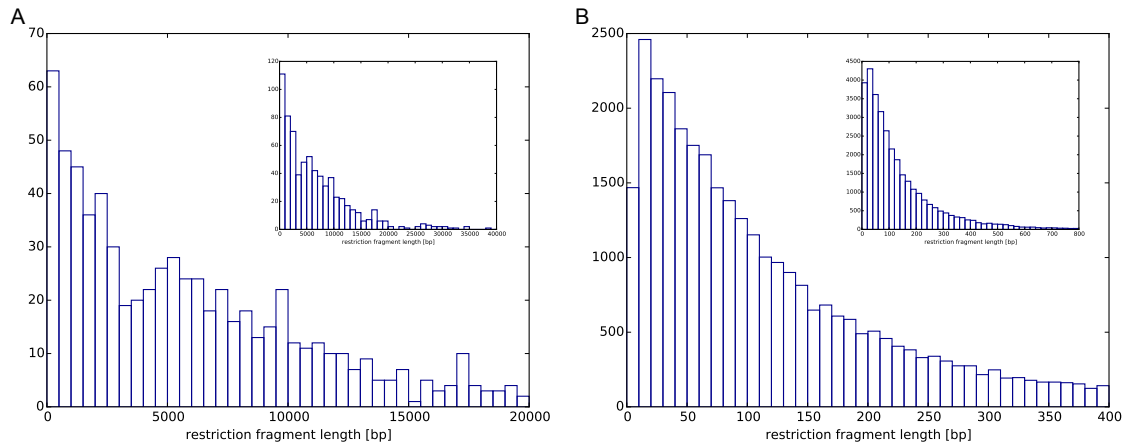


Figure 4.6: The distribution of the length of the restriction fragments for two different restriction enzymes (**A.** BglII; **B.** HhaI) using the example of the *E. coli* genome. The main graphs show the interesting part of the distribution and the inset graphs show the overall distribution. The two distributions show the huge difference between the restriction fragments generated by the frequent cutter HhaI and the rare cutter BglII as regards the total number of generated fragments as well as their typical lengths.

Restriction enzyme	Recognition site	No. of RFs	$\langle L \rangle$	\tilde{L}	L_{\min}	L_{\max}
<i>HindIII</i>	A [^] AGCTT	557	8333	4963	9	62667
<i>NcoI</i>	C [^] CATGG	614	7560	5163	21	47514
<i>BglII</i>	A [^] GATCT	702	6612	5048	16	42359
<i>MluI</i>	A [^] CGCGT	1329	3493	2444	6	33370
<i>BstYI</i>	R [^] GATCY	3191	1455	958	6	18605
<i>HpaII</i>	C [^] CGG	24312	191	122	4	3376
<i>HhaI</i>	GCG [^] C	32795	142	90	4	4100

Table 4.2: Overview of the statistics of the length of the restriction fragments for different restriction enzymes using the example of the *E. coli* genome. Listed are all restriction enzymes that have been used in the discussed Hi-C experiments. Besides its recognition site we listed the number of restriction fragments (RFs) generated by the respective restriction enzyme for the *E. coli* genome (MG1655, NC_000913.3) as reference. We also included the mean, median, smallest and biggest restriction fragment length (L).

Chapter 5

Deciphering 3D Organization of Chromosomes using Hi-C Data

References

The results presented in this chapter are published as and adapted from

- A. Hofmann and D.W. Heermann (2018), *Deciphering 3D organization of chromosomes using Hi-C data*. *Bacterial Chromatin, Methods in Molecular Biology*, 1837: 389-401. doi: 10.1007/978-1-4939-8675-0_19.

We thank Remus T. Dame, Frédéric Crémazy and Fatema Zahra Rashid for the stimulating and fruitful discussions.

Chapter Summary

In order to interpret data from Hi-C studies genome-wide contact probability maps need to be translated into models of functional 3D genome organization. Here, we first present an overview of computational methods to analyze contact probability maps in terms of features such as the level and shape of compartmentalization. Next, we describe approaches to modeling 3D genome organization based on Hi-C data.

5.1 Introduction

Following the generation of contact probability maps based on Hi-C sequencing data (see chapter 4), analysis of these maps is required to extract models of 3D organization and to establish structure-function relationships. It is important to keep in mind that Hi-C contact probability maps are matrices containing the contact probabilities of all pairs of loci at a given resolution and are conventionally visualized as heat maps. We are abbreviating contact probability maps as contact maps throughout the review.

There are two main challenges associated with the interpretation of Hi-C contact maps. First, the information contained in contact maps reflects an ensemble average of genome conformation of a large heterogeneous population of cells. Second, contact maps contain contact probabilities reflecting the probability that any pair of genomic loci co-localizes and hence characterize the neighborhood for each genomic locus. This is a major difference to imaging methods that can measure the spatial position of genomic loci in the context of the folded genome inside the cell.

There are two distinct approaches in developing spatial models of genome organization based on Hi-C contact maps. The first approach consists of hypothesis-driven modeling, which includes genome folding principles, such as DNA looping, as physical principles in the framework of polymer simulations. The properties of the ensemble of simulated polymer conformations are next compared to Hi-C data, often by comparison of experimental and simulated contact maps and defined structural features (such as domains) therein. The second approach uses the contact map as input to establish 3D structural models best fitting the experimental data.

Here, we first discuss how Hi-C contact maps can be analyzed in terms of prominent structural features and can be correlated with other types of genome-wide data. Next, we review approaches to the 3D modeling of chromosomes (see Fig. 5.1 for an overview of approaches to analyzing and modeling Hi-C contact maps).

5.2 Analyzing Hi-C Contact Maps

This section describes 1) the issue of comparing Hi-C contact maps, 2) methods for detecting certain features abundantly emerging in contact maps across different species, such as compartments, domains, loops and more complex structures and 3) the analysis of other genomic data alongside with Hi-C data.

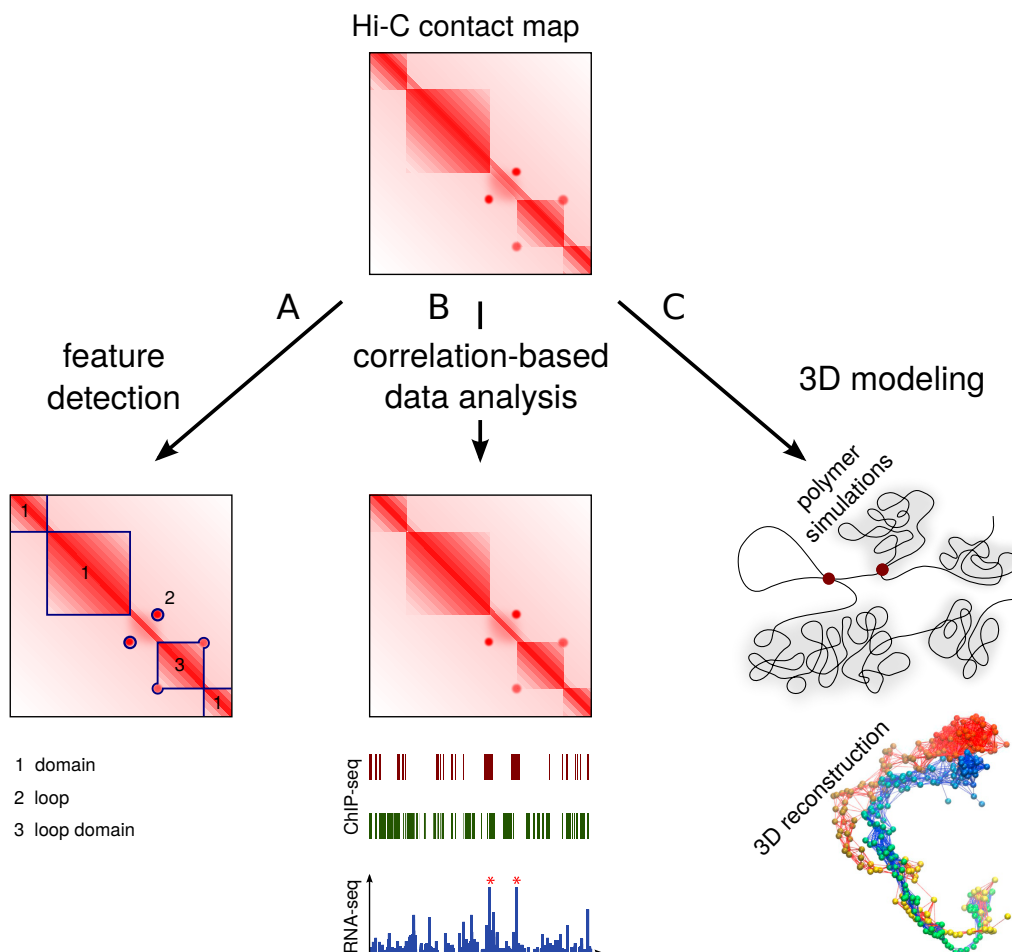


Figure 5.1: Schematic overview of approaches to analyzing and modeling Hi-C contact maps. **A.** Feature detection methods analyze the contact map in terms of prominent features, such as domains, loops and loop domains. **B.** Contact maps can be analyzed alongside with other genome-wide data, such as the expression level of genes (RNA-seq) or data on proteins being attached to the genome (ChIP-seq). **C.** Polymer simulations and 3D reconstruction methods (exemplary 3D representation adapted from [73]) aim for developing 3D models of the spatial organization of chromosomes based on Hi-C data.

5.2.1 Comparison of Contact Maps

How can the Hi-C contact maps of the same genome, but acquired under different conditions, be compared? Although qualitative differences between two contact maps can be readily detected by visual inspection, it is not straight-forward to quantify differences. The underlying mathematical challenge consists in quantifying the similarity, or rather the dissimilarity, of two matrices. However, well-known similarity measures, such as the cosine similarity S_C or the Jaccard index, take vectors as input and give as output a number between 0 and 1, where 0 indicates absolute dissimilarity and 1 absolute similarity. The same is true for statistical measures of correlation, such as the Pearson correlation coefficient ρ , that is commonly used for comparing Hi-C contact matrices and yield values between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation. The computation of these measures requires decomposition of the two-dimensional contact matrices into one-dimensional vectors row-

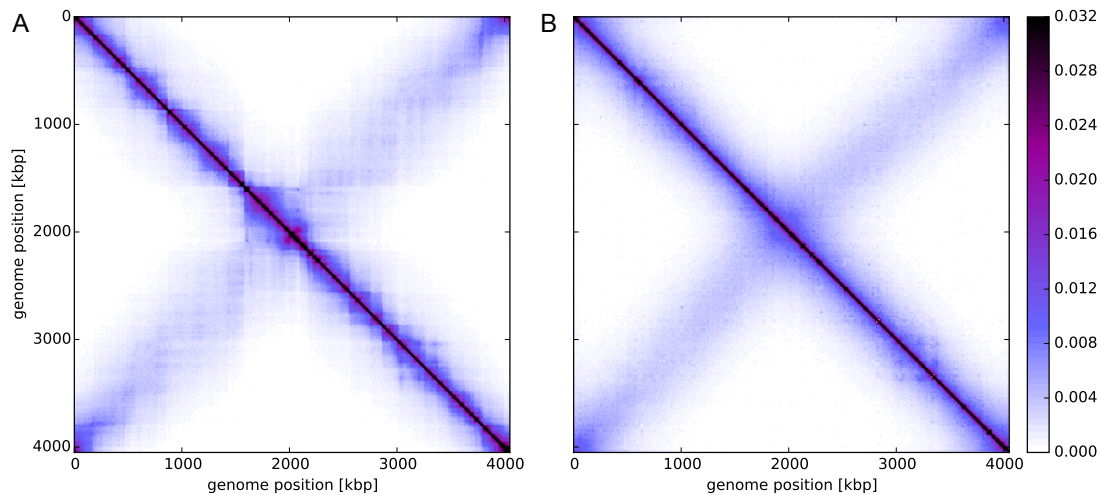


Figure 5.2: Hi-C contact probability maps of the *Caulobacter crescentus* genome (**A.** wild-type; **B.** Rifampicin treated). The color code is as follows: the darker the color, the higher the contact probability. Although there is a clear qualitative difference between the two maps, quantification of this difference is not straight-forward.

by-row.

In order to illustrate this problem, two contact maps of the *C. crescentus* chromosome [48] are shown in Fig. 5.2. In this example, the qualitative difference is the missing compartmentalization in the vicinity of the main diagonal of the contact map of Rifampicin-treated *C. crescentus*. But how to quantify this difference between the two maps? Both the Pearson correlation coefficient ($\rho = 0.97$) and the cosine similarity ($S_C = 0.98$) indicate high similarity between the two maps. This is due to the inherent emphasis of these measures on comparing global features in the matrices, such as the diagonal dominance and the presence of a secondary diagonal. By element-wise subtraction of one matrix from the other and depicting the resulting matrix as a heat map (see Fig. 5.3), it is possible to quantitatively display the differences between two contact matrices. The missing compartmentalization along the main diagonal in the Rifampicin-treated *C. crescentus* contact matrix can be recognized by the blue domains along the diagonal in this heat map.

5.2.2 Feature Detection

Eukaryotic and bacterial chromosomes have been found to be organized in compartments. In mammalian genomes domains comprise multiple length scales such as “A and B compartments” on the scale of Mbp [3] and “topologically associating domains” (TADs) on length scales ranging from 10 kbp up to 1 Mbp [7, 8]. “Chromosomal interaction domains” (CIDs) in bacteria occur on the same length scale as TADs and are considered equivalent [48, 73–75]. More complex structures of eukaryotic genomes that have been found using Hi-C experiments include loops and loop domains [36]. The observation of domains at different length scales highlights the need to be able to quantitatively characterize domains. On the basis of the detected domain structure, it is possible to compare different contact maps.

There are various different methodological approaches identifying the domain structure in Hi-C contact maps. They can be divided in two different classes depending on whether domains or domain boundaries are detected algorithmically. A first attempt at iden-

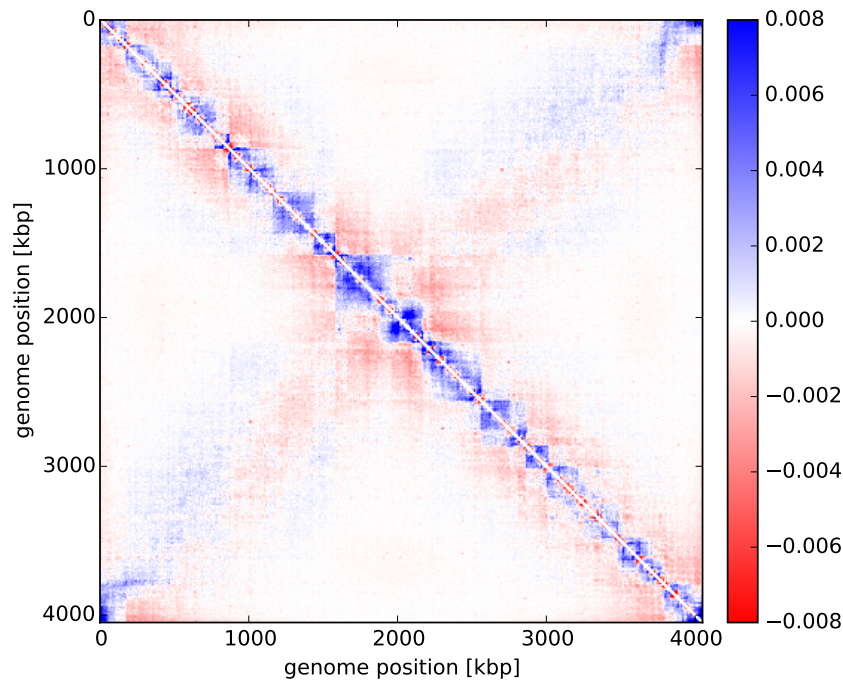


Figure 5.3: Subtracting the two Hi-C contact maps of the *Caulobacter crescentus* genome. The Rifampicin treated *C. crescentus* contact matrix has been subtracted from the wild-type one and the resulting matrix is illustrated using a heat map. A comparable contact probability in both matrices is depicted in white, enhanced contact probability in the wild-type matrix in comparison to the Rifampicin treated one in blue and the contrary case in red.

tification of TAD boundaries was presented by Dixon et al. [7] and is motivated by the observation that TADs are demarcated by regions that are biased in their interaction probability; the upstream domain boundary is preferentially interacting downstream whilst the downstream boundary is preferentially interacting upstream. This method is based on a two-step strategy. First, the 2D contact information is translated into the directionality index encoding the ratio of downstream and upstream interactions. Next, Dixon et al. argued that the directionality index can be considered as an observation of a hidden directionality bias that can be determined using a hidden Markov model and which allows the segmentation into domains. As alternative, downstream interactions can also be directly compared to upstream interactions in order to derive whether the strength of interactions are significantly stronger in one direction compared to the other. Domain boundaries correspond to positions where this preferred direction of interactions abruptly changes [48]. Lévy-Leduc et al. [83] developed a 2D model that fits a block diagonal matrix to observed contacts using maximum likelihood. In this model blocks correspond to domains. Chen et al. [84] presented a method for identifying TADs based on the interpretation of the Hi-C matrix as a weighted graph whose vertices are genomic loci and whose edge weights are contact probabilities of pairs of loci. As TADs are regions within the Hi-C matrix characterized by high internal contact probability, their identification can be translated to the problem of segmenting the graph into components with strong intra-connections and weak inter-connections. This graph partitioning is realized using spectral decomposition. Further methods use combinatorial optimization to find an optimal TAD hierarchy [85,86]. All these methods for domain detection assume that the domains are distinct, contiguous blocks of increased contact probability. In order to detect both domains and more complex

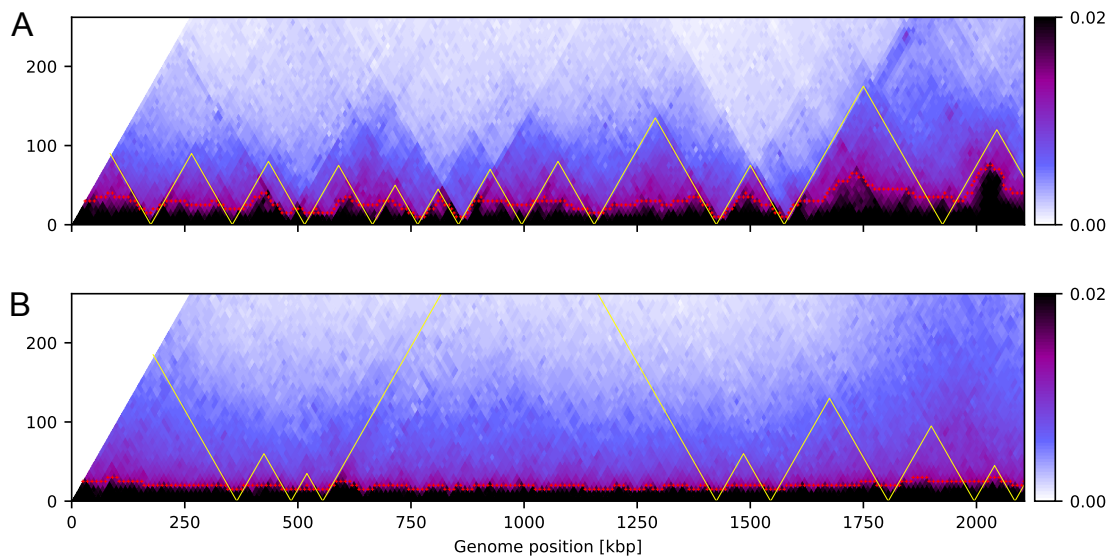


Figure 5.4: Comparison of the results of the approach using the directionality index and our probabilistic graphical model. To this end, excerpts of the 45° anti-clockwise rotated Hi-C contact maps of the *C. crescentus* chromosome shown in Fig. 5.2 are used. Contrary to the directionality index approach our method does not yield domain boundaries or rather a domain structure (yellow), but a contour (red) separating contact probabilities of a certain strength from the background. The results of our method clearly show the missing compartmentalization along the diagonal in the contact map of the Rifampicin-treated *C. crescentus* chromosome (**B**).

structures including loops or loop domains [36] without bias, we developed a probabilistic graphical model that makes no a priori assumptions on the domain structure [87]. Within this approach, the Hi-C contact matrix is analyzed using an Ising like probabilistic graphical model whose spin coupling constant is proportional to each lattice point (entry in the contact matrix). This approach is also relying on the graph theoretic interpretation of Hi-C matrices and does not yield domain boundaries, but a contour separating contact probabilities of an adjustable strength from the background. This iso-strength contour allows identification and characterization of compartments irrespective of whether or not there are contiguous domains in the form of squares.

In order to illustrate and visually compare the results of our approach and that based on the directionality index, we used the Hi-C contact maps of *C. crescentus* shown in Fig. 5.2. The contiguous domain structure computed on the basis of the directionality index as well as the iso-strength contour yielded by our method permit characterization of the compartmentalization of the wild-type *C. crescentus* chromosome (see Fig. 5.4). The iso-strength contour is nearly flat for the contact map of the Rifampicin-treated *C. crescentus* chromosome using the same model parameters. This indicates that following rifampicin perturbation this chromosome is not compartmentalized; instead the captured interactions between genomic loci decrease uniformly as their genomic distance increases. The absence of a contiguous domain structure cannot be captured by methods like those based on the directionality index.

5.2.3 Correlation-based Data Analysis

Hi-C contact information can be analyzed alongside with other genomic data such as the expression level of genes or data on proteins being attached to the genome. This

has the advantage of being able to gain insights into the correlation between the two types of information instead of only analyzing each separately. An example of such a combined analysis found that the mouse genome is organized into domains of coordinately regulated enhancers and promoters that coincide with TADs [37]. First, ChIP-seq data of RNA polymerase II indicative of active promoters and H3K4me1 as a mark for enhancers were compared across different tissues and cell types. As these signals were concordantly enriched within clusters in the genome, much like TADs in contact maps of mammalian genomes, the authors also compared both types of domains and determined that they indeed overlap. Another example of such correlation-based data analysis of ChIP-seq binding profiles and Hi-C data showed that the proteins CTCF and cohesin associate with loops that have been detected within contact maps [36]. The previously discussed domain boundaries in the contact map of the wild-type *C. crescentus* chromosome have been found to correlate with the position of highly expressed genes identified through DNA microarray analysis experiments.

We mentioned only a few exemplary studies that found interesting features of Hi-C contact maps to be correlated with other genomic data. These correlations, though not implying causality, are interesting for further specialized studies and hypothesis-driven modeling approaches since they hint at possible mechanisms of the 3D genome organization.

5.3 3D Modeling

Hi-C experiments yield information that can be interpreted using computational models of chromosome organization. There are two key strategies for building such models. The first data-driven strategy, referred to as 3D reconstruction, uses the contact probabilities as summarized in the contact map to determine an optimal structural model of the data. The second strategy aims at establishing general principles of folding for organization of chromosomes using physical principles in the framework of polymer simulations. Contrary to the first strategy, Hi-C data is not used as an input for these polymer models, but rather for validation. Here, we review several methods employing either of the two strategies. For a more complete overview of 3D reconstruction methods we refer to the review of Serra et al. [88].

5.3.1 3D Reconstruction

The goal of 3D reconstruction algorithms is to use the contact map as input to recapitulate the underlying 3D structure of a genome. In this approach 3C-based data is used to obtain spatial restraints for modeling the genome; 3D reconstruction is also known as restraint-based structure modeling. The basic concept for the reconstruction is simple: the closer two genomic loci are in 3D space, the higher the probability is that they interact. In technical terms, the assumption is that the Euclidean distance between two loci is inversely proportional to their contact probability. Following this basic notion, there are two strategies for translating the contact probabilities within the contact map into a set of 3D coordinates of loci representing the genome. In the first, optimization-based, approach the total difference between pairwise distances in the hypothesized set of 3D coordinates is minimized and the corresponding distances are inferred from the observed contact probabilities. In the second, model-based, strategy, the observed contact probabilities are assumed to follow a probability distribution from which 3D structures can be inferred.

Irrespective of the underlying strategy, these methods output either a single consensus 3D

structure or an ensemble of 3D structures. Both consensus and ensemble methods have advantages and disadvantages. Ensemble methods are biologically more plausible, because they reflect the fact that Hi-C data is obtained from an ensemble of conformations. However, the analysis of an inferred ensemble of 3D structures is not straight-forward: one option is the characterization of the ensemble average [89]; another one is to select a few structures that are representative of the diversity of the ensemble [90]. Consensus methods, in contrast, generate a single structure, which can be thought of as an visualization of the contact map and is easy to analyze. Computationally, ensemble methods are more demanding than consensus methods, because they need to sample from a very large dimensional space of candidate 3D structures.

Optimization-based Methods

“ShRec3D” [91] is a method that seeks for analytically reconstructing a consensus 3D structure. It builds upon the fact that the contact matrix can be interpreted as the adjacency matrix of a weighted graph whereby the problem is reformulated in terms of embedding a graph into Euclidean space. This problem, in turn, is well-known in the literature and can be solved using classical Multidimensional Scaling (MDS) [92]. Given a set of distances between the vertexes of a graph, this method returns an Euclidean set of coordinates. Therefore, the definition of distances between the vertexes of the graph representing the contact matrix is crucial within this framework. The authors chose the shortest path distance for this purpose, but did not show how other distance definitions, such as the resistance distance or connectivity-based distances, perform compared to that choice. “ChromSDE” [93] is a numerical method that jointly optimizes the 3D structure and a parameter that maps contact frequencies to spatial distances. The main difference to “ShRec3D” is the translation of contact frequencies to spatial distances by numerical optimization. Both methods reconstruct a consensus 3D structure. In contrast, Kalhor et al. have proposed an optimization framework that generates an ensemble of structures [89]. The idea behind this approach is to convert contact probabilities into a set of contact restraints for the 3D structures in the ensemble. However, any given contact is enforced with its contact probability, hence only in a fraction of the inferred structures in the ensemble.

Probabilistic Modeling Methods

Different from the optimization-based approaches, probabilistic modeling methods assign an uncertainty to the spatial distances between genomic loci. The observed contact frequency of two loci is typically assumed to follow a Poisson distribution [94, 95]. This accounts for the fact that 3C-based experiments detect contact frequencies among restriction fragments and, hence, count data. This approach is valid for non-genome wide input data. However, these methods are not valid for Hi-C input data as these consist of contact probabilities rather than contact frequencies among genomic loci. The Markov chain Monte Carlo (MCMC) - based method “MCMC5C” [90] is an exception in this respect since it assumes a Gaussian distribution for the input contact data; therefore it can model both Hi-C contact probabilities and other 3C-based contact frequencies. In this approach DNA is modeled as a chain of beads representing the 3D structure, which is iteratively changed using random moves that can be either accepted or rejected depending on whether the new 3D structure is more probable given the data. After a sufficient number of iterations, this MCMC scheme samples 3D structures that fit the experimental contact data. By

running many of those simulations in parallel, a large ensemble of structures is generated. Hu et al. proposed a probabilistic method called “BACH” [94] that models the contact data using a Poisson distribution. Contrary to MCMC5C, Monte Carlo methods are used in order to gradually refine an initial structure conformation and generate a consensus 3D structure. “PASTIS” [95] also models the contact data using a Poisson distribution. It uses maximum likelihood estimation of the model parameters for reconstructing the 3D structure with the highest likelihood given the observed contact data.

5.3.2 Polymer Simulations

Polymer models incorporating known or hypothesized physical or biological principles can also be used to model chromosomes. Contrary to 3D reconstruction methods, such models do not infer conformations using Hi-C data, but rather use such data to test whether generated ensembles of 3D structures agree with contact maps or key features thereof.

A polymer can be described using various properties like its average end-to-end distance or its radius of gyration. Here, we focus on two biologically relevant quantities that can be compared to experiments, such as Hi-C or microscopy imaging: 1) the contact probability between two loci as a function of their genomic distance which can be deduced from the Hi-C data; 2) the mean squared distance (MSD) of two loci as a function of their genomic distance, a quantity that can be measured, for example, by fluorescence in situ hybridization (FiSH) experiments. Both quantities are averaged over the conformational ensemble in any polymer model and over a population of cells in a Hi-C experiment. Simple polymer models include the random coil and the self-avoiding chain. The random coil is the simplest model; it is characterized by non-interacting monomers. Self-avoiding chains exhibit excluded volume interactions leading to an increased effective volume compared to the random coil. The worm-like chain or Kratky-Porod model introduces an intrinsic stiffness by associating a bending of the chain with an energy cost. Hence, it can describe semi-flexible polymers, such as double-stranded DNA. The fractal globule model [96] describes a compact polymer state that emerges during polymer condensation as a result of topological constraints. It has been reported to agree with the initial Hi-C data of the human genome [3] since it shows the same scaling behavior of the contact probability as a function of the genomic distance at a scale of $\sim 1 - 10$ Mbp [97]. However, it does not explain findings from FiSH experiments that display a leveling-off in the MSD for genomic separations above 10 Mbp [98]. In the dynamic loop (DL) model [82] the chromosomal fiber is represented as a self-avoiding chain allowed to form probabilistic intra-polymer crosslinks between non-adjacent monomers. As a consequence, loops of different size are formed. The main model parameter is the looping probability, a measure for the probability that a loop is formed between two non-adjacent monomers. The DL model explains both the scaling behavior of the contact probability and the leveling-off of the MSD on the basis of the dynamic formation and dissolution of loops. The “strings and binders switch” (SBS) model [99], which assumes diffusible factors (binders) being responsible for loop formation by linking two monomers of the polymer, is a special case of the DL model that implicitly incorporates the properties of such binders in the looping probability parameter. The polymer fiber is also modeled as a self-avoiding chain and the binding molecules are represented by Brownian particles with a certain concentration. The loop extrusion model [100] proposes that loop-extruding factors form increasingly larger loops, which are stalled by boundary elements, such as bound proteins at domain boundaries. Different from the DL and the SBS model [101, 102], it also explains the formation of domains [103, 104].

5.4 Summary

Hi-C contact maps either can be analyzed in terms of feature detection or can be used as a starting point for developing 3D models of the spatial organization of chromosomes. In the specific case of feature detection, it would be useful to have methods that allow discrimination of different possible patterns instead of detecting only a specific one. Existing methods, such as directionality index based approaches, have shown the existence of domains in eukaryotic as well as bacterial chromosomes [48, 73–75], but fail to identify loops or loop domains. There are two different ways of modeling 3D organization of chromosomes: 3D reconstruction and polymer modeling. While the first generates the most likely 3D structure given the contact data amenable to visual inspection, polymer modeling supports clarifying hypotheses of chromosomal organization deduced from contact maps. The Hi-C studies of the two bacterial chromosomes *C. crescentus* [48] and *B. subtilis* [73] exemplify the difference between hypothesis-driven modeling and 3D reconstruction. Using ShRec3D, a consensus 3D structure for the *B. subtilis* chromosome is reconstructed and then used to illustrate the point that the chromosome folds into a helicoidal shape and is organized into domains. In contrast, Le et al. deduced from their gathered Hi-C and microarray data the hypothesis that the *C. crescentus* chromosome consists of domains comprised of supercoiled DNA plectonemes and boundaries being transcription-induced. Based on this hypothesis, they performed simulations of a bottle-brush polymer with linear boundary elements and compared the simulated with the experimental contact maps.

Chapter 6

Interactive Visualization of Hi-C Data

References

The results presented in this chapter are published as and adapted from

- A. Hofmann, J. Müggenburg, Frédéric Crémazy and D.W. Heermann (2019), *Bekvaem: Integrative Data Explorer for Hi-C Data*. Journal of Bioinformatics and Genomics, 2: 11. doi: 10.18454/jbg.2019.2.11.1.

AH and JM developed the method. We thank Remus T. Dame, Fatema Z. Rashid and Gaurav Dugar for the stimulating and fruitful feedback on the application.

Chapter Summary

The browser-based visualization of Hi-C contact maps alongside complementary data tracks is a computationally challenging task and requires an efficient software implementation to run on small clients. Few software packages have yet been shared with the community to address this problem and modification of these is cumbersome.

We introduce Bekvaem that addresses these problems by using high-level Python interfaces. Wrapping several libraries for online visualizations at the front-end and the organization of large biological data sets at the server-side allows for setting up a high-performance user-defined browser visualization for Hi-C data with just a few changes in the code.

The **source code**, written in Python, of Bekvaem alongside its documentation and sample data is freely available on heiDATA [105].

A **demonstration server** is available at [106].

6.1 Introduction

Genome-wide 3C-based experiments, such as Hi-C, enable insights into the 3D genome organization at an unprecedented resolution. The analysis of Hi-C contact information alongside with other genomic data, such as ChIP-seq or RNA-seq, allows for probing correlations between genomic information instead of analyzing each separately. With the help of such a combined analysis it has been suggested that proteins shape the hierarchical organization of eukaryotes and transcription that of prokaryotes [107]. Inversely, topological associated domains (TADs) that act as barriers to restrain enhancer-promoter contacts involved in transcription regulation [108] have been identified using Hi-C and other chromatin marks are just correlated to TADs. These findings illustrate the need of an integrated as well as interactive visualization of Hi-C contact maps alongside other genomic data.

Several tools for the visualization of Hi-C and other genomic data were developed and released in the last years [109–112]. Focusing mainly on mammalian genomes, the tools often provide complete data processing libraries covering all the necessary steps from Hi-C sequencing data to the visualization of the processed data. A resulting disadvantage is that the import of files containing already generated Hi-C contact matrices is not supported and hence the flexible combination with other tools or rather existing data pipelines is not possible. The necessity of being able to program in low-level programming languages, such as C, as well as large and complex codebases are obstacles for customizing these tools.

Our aim was twofold. First, we wanted to keep the Hi-C visualization pipeline Bekvaem generic and as easy to modify as possible. Second, its usability and performance should be on a par with existing tools.

6.2 Methods

Bekvaem aims to unify all steps involved in the creation of browser-based visualization of Hi-C contact maps in the high-level language Python. By making use of Bokeh and HoloViews, graphs to the web interface are included using a syntax similar to that of 'off-line' plotting libraries such as Matplotlib or ggplot. The workflow of introducing new types of data to the Hi-C analysis is brought down to writing a class with read-in, plotting and processing methods.

The hierarchy of classes representing Hi-C data as well as other experimental data is designed around generic graph types, such as a heat map for the Hi-C contact map and a line graph for ChIP-seq and RNA-seq data. These classes handle I/O, pre-processing and live aggregation of the data sources, if applicable. Offering a broad range of powerful libraries and interfaces to many programming languages (R, C), Python handles most existing data sources and processing routines.

The bottleneck of displaying large data sets on a restricted screen resolution is addressed by just-in-time compiled regridding routines that allow live exploration of the data sets without fixed zoom levels. Individual plot elements are initialized in HoloViews, a down-stream library to Bokeh. By supplying data pipes and axes range streams to the plot containers, manipulation of the plot or elements of the user interface results in a call to the callback function of the container in which regridding and filtering operations can be placed. Dynamic loading of the requested data using cooler [113] for the Hi-C contact maps and pyBigWig [114] for continuous complementary tracks allows Bekvaem to visualize large genomes.

Meta information that is integral to an understanding of certain data tracks is accessed via Bokeh's HoverTool, a dynamic HTML widget that pops up on hover actions. Revealing information only in the user-requested region increases the quantity of displayable data while maintaining the lucidity of the visualization.

The synchronization of the server and client view is handled by the internal plotting server of Bokeh and no programming is required for its usage. Internally a document body of the browser visualization is built in Python with the Bokeh library and synchronized in

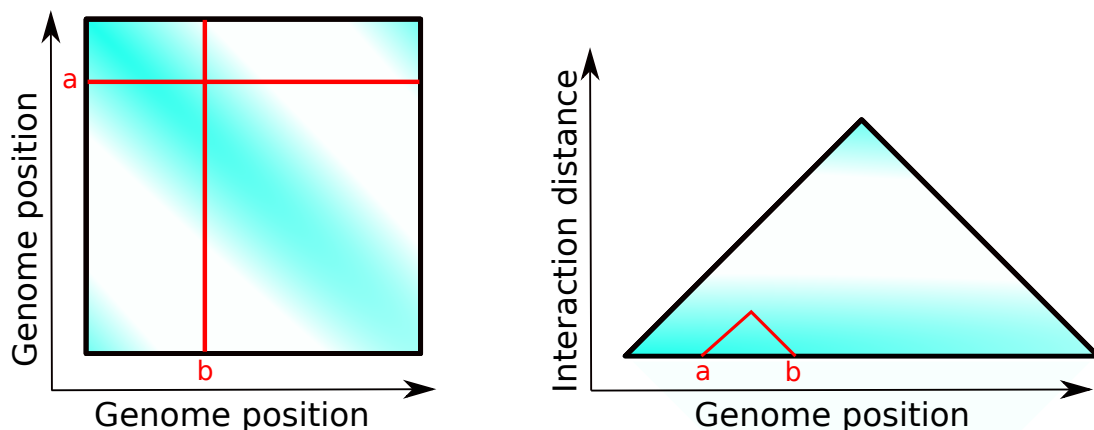


Figure 6.1: Visualization of the Hi-C matrices. The matrices can be displayed as a square heat map in which x - and y -axis correspond to positions along the genome. Due to the symmetry along the diagonal, one can also display just one of the triangular matrices of the full array. The y -axis now corresponds to the interaction distance.

a JavaScript Object Notation (JSON) format with the client side. Subsequently Bokeh's JavaScript library BokehJS renders the plots on the client side using the HTML5 Canvas element.

Due to the symmetry along the diagonal axis, all information on the interaction probabilities in a Hi-C matrix is stored in one of its triangular matrices. By applying a rotation of 45° and cropping (Fig. 6.1), the necessary space in the layout can be reduced by a factor of $\sqrt{2}$. The new y -axis is a measure of the interaction distance between two genome positions on the x -axis. Both quadratic and triangular shaped Hi-C matrices are supported by Bekvaem.

6.3 Application

Bekvaem is conceived to visualize Hi-C contact maps of any genome alongside complementary tracks. Besides Hi-C contact maps in square and triangular form, several additional data tracks of domain detection, protein binding, gene expression and gene databases were added to the pool of implemented plots.

First macros that are helpful for the data exploration were exposed via the user interface (UI). Given a data track of RNA-counts, genes can be filtered on their expression rates. Besides, the UI contains HTML widgets to interchange the data sources and adjust the color scale of the color bar of the Hi-C contact map. The user has access to markers to record positions along the genome and a pdf-export routine was written to capture the current view. Fig. 6.2 shows the output of Bekvaem using a Hi-C contact map of *M. musculus*.

The visualization can be installed and used on a local host or exposed to the general web. For the latter a reverse-proxy such as NGINX can be used to embed the visualization in a larger web application.

The Bokeh server can be embedded into the visualization in different ways. Despite being a stand-alone web server, it can be useful to build more complex architectures to enhance the configuration options or the security. In its current implementation, Bekvaem receives the request for a plot layout via a simple plain-text file. Hence the web interface can be modified in a rich fashion by this plain-text configuration file.

In the example implementation in Fig. 6.3, the configuration file for Bekvaem is generated in a slim Flask application. Flask [116] is a micro-framework to create web applications and enables the evaluation of the jQuery sortable table in the UI. The Flask servers are themselves slave application of a WSGI Gunicorn server [117] which serves as a pre-fork worker to handle multiple user requests simultaneously. Both the Bokeh Server and Gunicorn are placed behind a reverse-proxy NGINX server [118] which is exposed to the outside.

Due to the simple structure of the configuration file, Bekvaem can also be run on a local port and fed with manually generated configuration. In this constellation, the configuration can be edited via a normal text-editor.

6.3.1 Comparison with Other Hi-C Browsers

In the following the table of Yardmci and Noble [119] is modified and expanded by our software and additional tools known to us. We included only tools that are open source, have a browser interface (Juicebox via Java Web Start) and possess a local installation option.

Note that some tools offer other ways of visualizing Hi-C maps, such as local arc tracks,

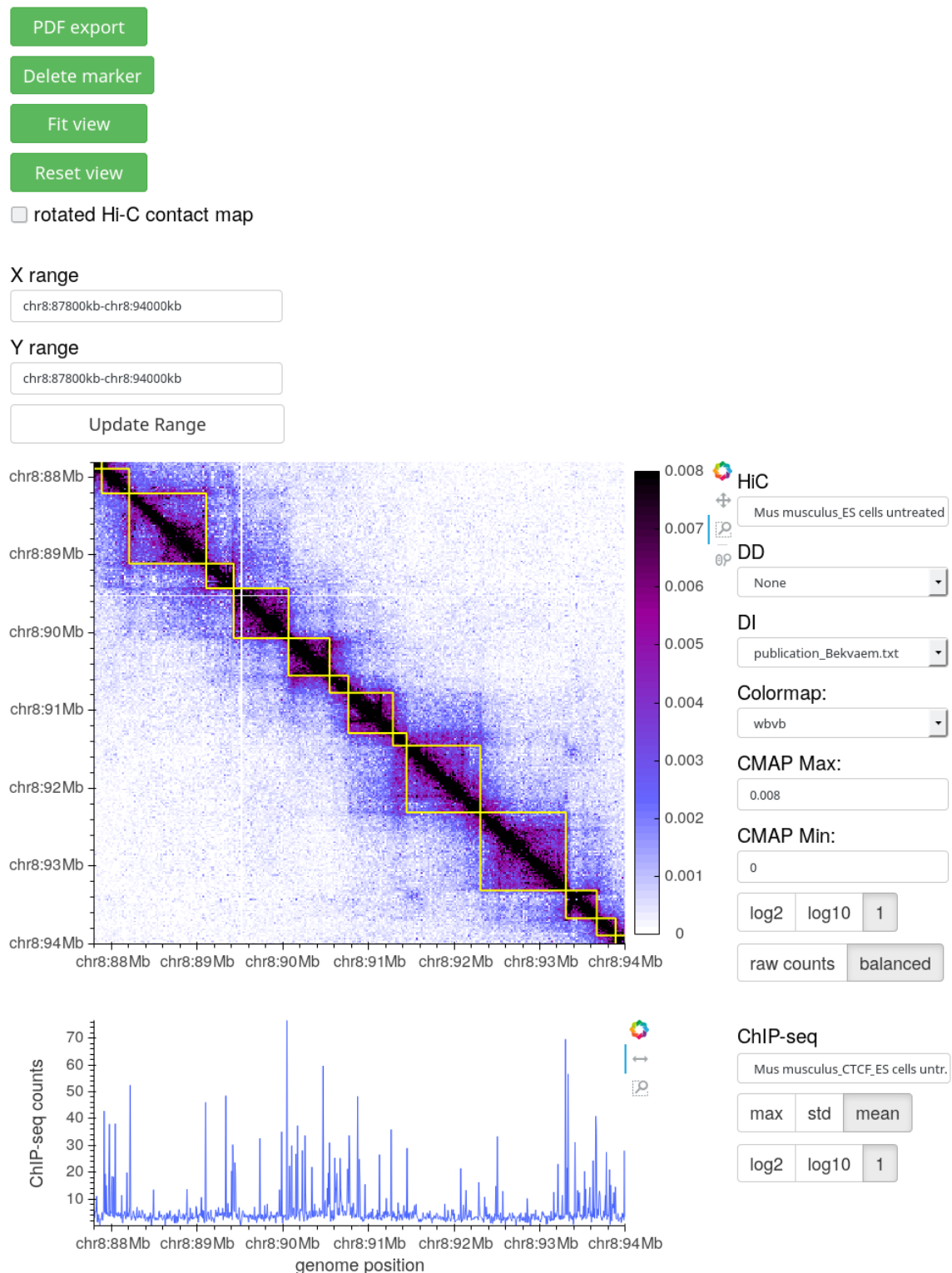


Figure 6.2: Screenshot of the web interface of Bekvaem. A 6 Mb section of a 20 kb resolution Hi-C contact map of untreated *M. musculus* embryonic stem cells overlaid with a domain detection track is shown. CTCF counts of a complementary ChIP-seq experiment are displayed alongside the Hi-C contact map. Data from Nora et al. [115].

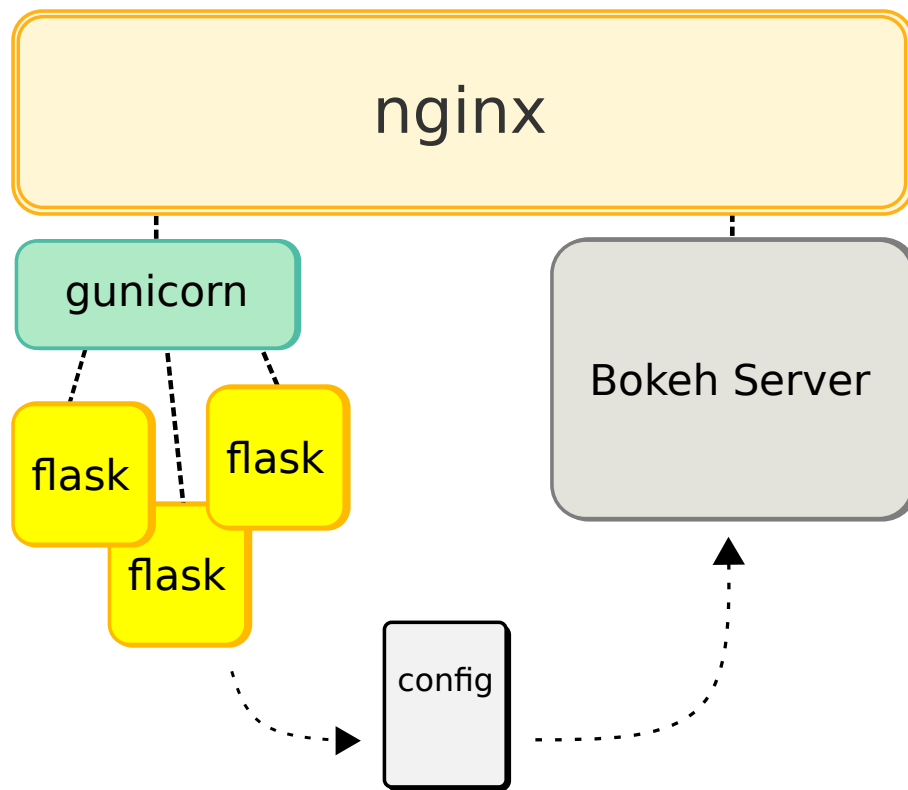


Figure 6.3: Example server architecture. Using an NGINX reverse-proxy server, users can request a specific layout of the web interface of Bekvaem by submitting a request to a flask application. A Gunicorn pre-fork server is used to handle multiple user requests simultaneously. The configuration files is read by Bekvaem running on a Bokeh server and the layout connection with the user is established through NGINX.

locus-specific circular plots and virtual 4C plots. The reader is referred to the reference or the documentations of the software for further reading. We do not deem these ways of visualization as informative as square and rotated heat map representations of Hi-C maps and withhold them in the following overview table.

Advantages of Bekvaem

To point out the easiness of adjustments that we see in Bekvaem over other existing tools, a look to holoviews' DynamicMap class is helpful.

DynamicMap objects can be understood as a container around different plot types such as line plots, heat maps or scatter plots. They can subscribe to streams and pipes with which many important user actions can be evaluated:

- 1) In Bekvaem we use RangeStreams that synchronize user-selected axes ranges of the plots with the server.
- 2) Pipes are used to handle other user input such as changes of the color scheme of the Hi-C plots. Many different objects such as strings or dictionaries can be synchronized with holoviews' pipes.

Supplemental plots like ChIP-seq tracks are synchronized through the same RangeStream that the Hi-C plot is subscribing. If a user wishes to add a new type of plot, he can follow

the procedure:

	Bekvaem	Juicebox [120]	Epigenome Browser [121]	HiGlass [122]	TADkit [123]	3D Genome Browser [124]
Intra-chromosomal square heat map	✓	✓	✗	✓	✗	✗
Inter-chromosomal square heat map	✓	✓	✗	✓	✗	✗
2D heat map features	✓	✓	✗	✓	✓	✓
Rotated local heat map	✓	✗	✓	✓	✓	✓
Multi Hi-C visualizations	✗	✓	✓	✓	✗	✓
Supplemental data visualization	✓	✓	✓	✓	✓	✓
Interactive zoom ¹	Box zoom, Scroll zoom	Scroll zoom	Box zoom, Scroll zoom	Scroll zoom	Box zoom, Scroll zoom	Scroll zoom
Browser interface	✓	✗	✓	✓	✓	✓
Java interface	✗	✓	✗	✗	✗	✗
Local installation option	✓	✓	✓	✓	✓	✓
Code lines ²	3000	41000	113000	39000	161000	37000
Main programming languages ³	Python	Java	JavaScript C	JavaScript HTML	JavaScript CSS	JavaScript CSS

¹Box zoom describes selecting a rectangular region of interest which is loaded in the full plot extent. Scroll zoom describes zooming in by a specified factor, centering e.g. the cursor position or the center of the Hi-C map.

²The lines of code were counted and mapped to the programming languages using CLOC [125]. In order to minimize influences by the documentations, test suits and CI, typical folder names [doc, docs, test, .cache] and files such as [package.json, package-lock.json] were excluded from the analysis. CLOC also discriminates comments and blank lines. Nevertheless, the line numbers are rather giving a hint of the size of the project. For all the tools except Bekavem, it cannot be ruled out that necessary source code that is stored in other locations was neglected or that small fractions of unnecessary code were included in the count. The numbers were rounded by subtracting a modulo thousand. For all programs, we used the current master branch accessed on 10/04/2019.

³Estimated on basis of the code line count. If a second language contributes a non-negligible part (i.e. at least 1000 lines) to the source code, it is noted in the above table.

- 1) Select one of holoviews plot types that is best suited to visualize the data.
- 2) Provide a callback function that takes the data of the RangeStream as input (a tuple `x_range` with start and end values of the visible axis range) and returns a holoview's plot object of the plot type that was selected in 1). Inside this routine, the user is free to call any external function to load or process the data before writing it to the plot object.
- 3) Add the DynamicMap to the list of subscribers of the RangeStream and add it to Bekvaem's layout list object.

This is the most basic scheme that allow the creation of a new type of plot. Often it is necessary that the underlying data file can be modified through the user interface. Bokeh includes different types of widgets that allow the execution of python functions on user actions. These can be added as well to Bekvaem's layout list and allow the creation of rich UI's.

Bekvaem's design is object oriented and it is advisable to design a class around one type of plot (e.g. a line plot) and inherit its methods and attributes in different child classes that represent the corresponding biological data tracks. Examplewise one can look at Bekvaem's classes `ChIP_seq` and `RNA_seq` that inherit from the class `LinePlot`. Even this single class is already versatily useable in the visualization of many different data tracks and it requires very little work to adapt the general scheme to new data tracks by e.g. overloading some of `LinePlot`'s class methods.

6.3.2 Visualization Examples

We demonstrate in Fig. 6.4 how easy the user can switch between different genomes of interest by loading the respective data files in the selection windows of Bekvaem's browser interface. In Fig. 6.5 a selected region of a Hi-C contact map of the mouse genome illustrates that Bekvaem is capable of visualizing large eukaryotic Hi-C contact matrices. There is also the possibility of exporting the current view into a PDF file (see Fig. 6.6). Bekvaem uses cooler [113] for the storage and querying of Hi-C maps and pyBigWig [114] to include BigWig tracks. While not achieving the full functionality that tools such as HiGlass have, a powerful Hi-C browser with less than 3000 lines of code is presented. For the demonstration we used publicly available data from NCBI's Gene Expression Omnibus (GEO) [4] database. The GEO data set IDs are listed alongside the information with respect to their publication in the section "Data availability" in the documentation.

6.4 Conclusion

We present a slim and interactive browser application capable of visualizing Hi-C contact maps alongside complementary data tracks. Besides Hi-C contact maps genome-wide data, such as ChIP-seq and RNA-seq, can be included in the layout. Bekvaem can be utilized for the visualization of any genomes including mammalian genomes.

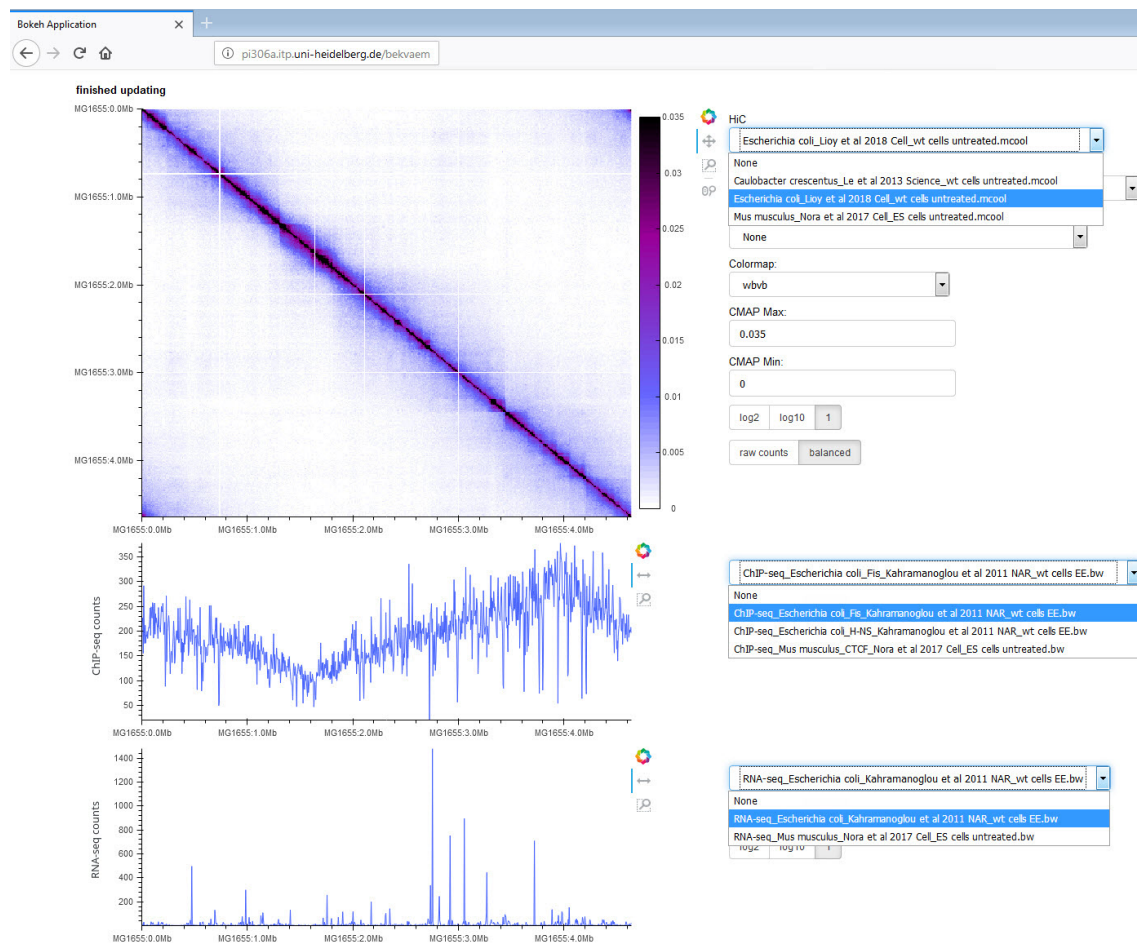


Figure 6.4: Screenshot of the web interface of Bekvaem. The user can easily switch between different genomes of interest by loading the respective data files in the selection windows of the browser interface. A Hi-C contact map of wild type *E. coli* cells is depicted. The resolution of the balanced Hi-C contact map is 10 kbp and the linear color scale ranges from a contact probability of 0 up to 0.035. The complementary data tracks show the ChIP-seq profile of Fis (top) as well as the RNA-seq profile (below) in early exponential phase. The read counts of both experiments are depicted on a linear scale. Data from Lioy et al. [47] and Kahramanoglu et al. [126].

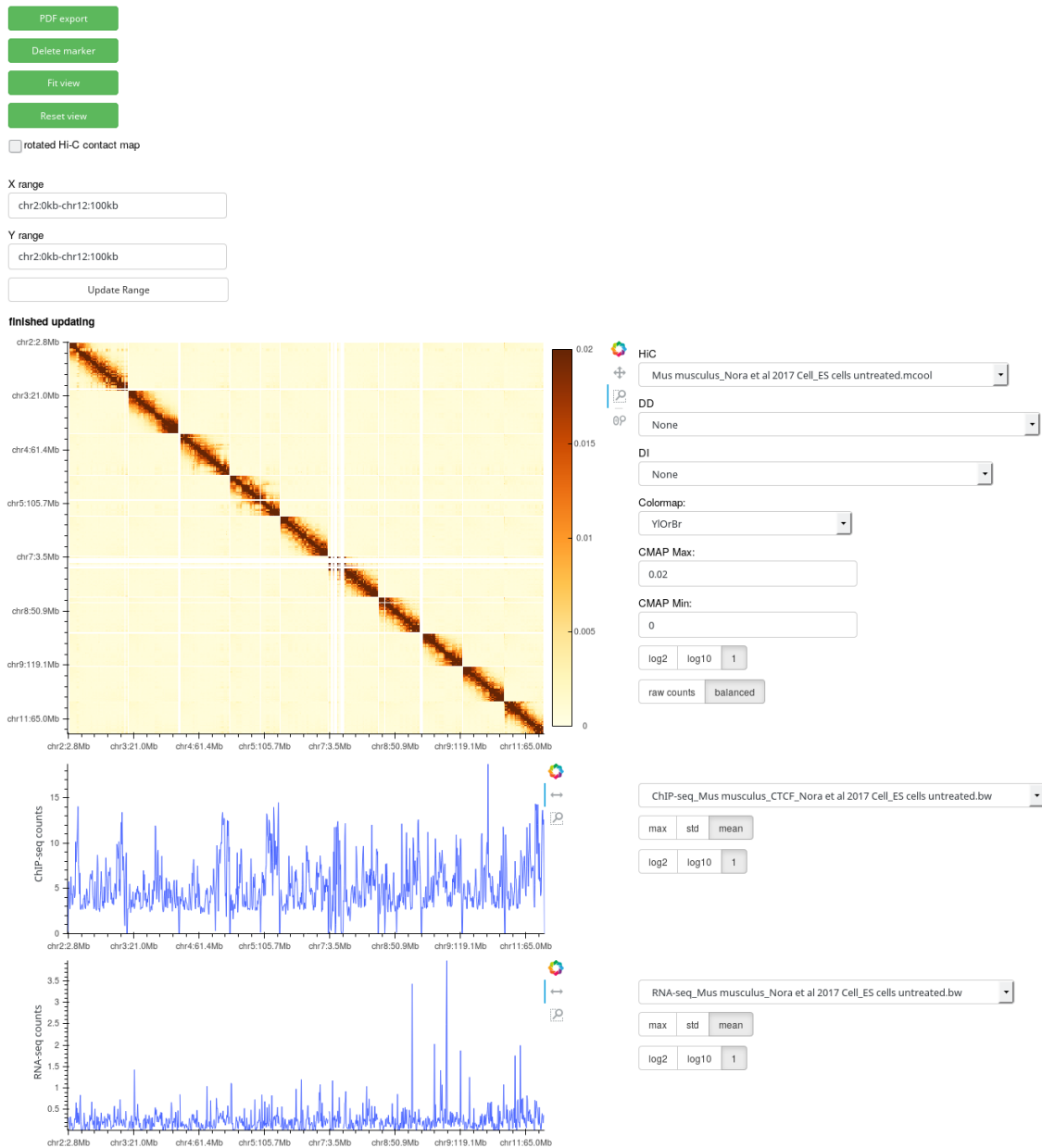


Figure 6.5: Screenshot of the web interface of Bekvaem. A heat map of the intra- and inter-chromosomal Hi-C interactions among chromosomes 2 to 11 of the mouse genome is depicted. The resolution of the balanced Hi-C contact map is 20 kbp and the linear color scale ranges from a contact probability of 0 up to 0.02. The complementary data tracks show the ChIP-seq profile of CTCF (top) as well as the RNA-seq profile (below). The read counts of both experiments were processed using a sliding window size of 1 kbp and are depicted on a linear scale. Data from Nora et al. [115].

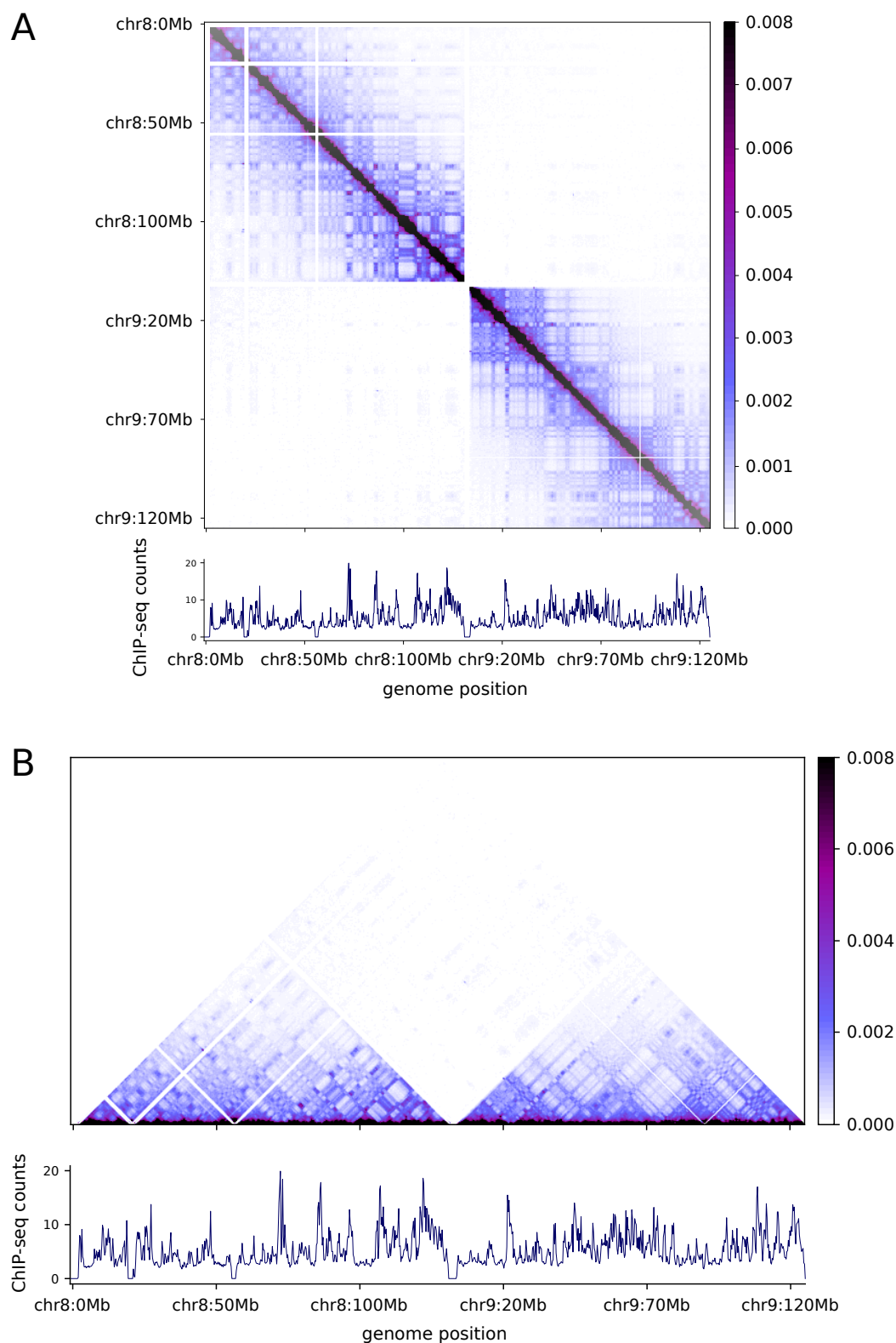


Figure 6.6: PDF export of Bekvaem. The Hi-C contact map of chromosomes 8 and 9 of the mouse genome are depicted in **A**, square and **B**, triangular shape alongside the CTCF ChIP-seq profile. The resolution of the balanced Hi-C contact map is 20 kbp and the linear color scale ranges from a contact probability of 0 up to 0.008. The ChIP-seq read counts were processed using a sliding window size of 1 kbp and are depicted on a linear scale. Data from Nora et al. [115].

Chapter 7

Domain Boundary Detection in Hi-C Maps

A Probabilistic Graphical Model Approach

References

The results presented in this chapter are adapted from

- A. Hofmann, F.Z. Rashid, F. Crémazy, R.T. Dame and D.W. Heermann (2019), *Domain Boundary Detection in Hi-C Maps: A Probabilistic Graphical Model Approach*, in preparation, to be submitted to PLoS ONE.

AH and DWH developed the method. AH performed the analysis of the Hi-C maps.

Chapter Summary

To understand the nature of a cell, one needs to understand the structure of its genome. For this purpose, experimental techniques such as Hi-C detecting chromosomal contacts are used to probe the three-dimensional genomic structure. These experiments yield topological information, consistently showing a hierarchical subdivision of the genome into self-interacting domains across many organisms. Current methods for detecting these domains using the Hi-C matrix, i.e. a doubly-stochastic matrix, are mostly based on the assumption that the domains are distinct, thus non-overlapping. For overcoming this simplification and for being able to unravel a possible nested domain structure, we developed a probabilistic graphical model that makes no a priori assumptions on the domain structure. Within this approach, the Hi-C matrix is analyzed using an Ising like probabilistic graphical model whose coupling constant is proportional to each lattice point (entry in the contact matrix). The results show clear boundaries between identified domains and the background. These domain boundaries are dependent on the coupling constant, so that one matrix yields several clusters of different sizes, which show the self-interaction of the genome on different scales.

7.1 Introduction

Early work using optical microscopy with fluorescent markers established that chromosomes are not randomly organized in the nucleus [127]. Exactly how the chromosomes are organized could not be further revealed by this method, even though multi-color experiments pushed the experimental boundary [128]. At this stage several models have been proposed how the genome is physically organized in space [82, 97, 129–133]. With the 3C technology [134] new data on 3D genome organization became available. Whereas the information coming from the microscopy experiments gives a physical relationship between points in space, i.e., Euclidean distances on single cell data, the 3C data (and later the Hi-C data [3]) yields topological information loosing the embedding into Euclidean space, i.e., only neighborhood relationships are revealed attached with a certain probability. Furthermore, the information represents an average over many cells. In a way this is very much information one would classify as of mean-field type. Thus, the challenge is to develop a model that is consistent with the mean-field result in the sense that it succeeds to re-embed the topological information into Euclidean space, i.e., geometrical information and topological information need to be reconciled.

A crucial part of this process is to identify the structures and substructures that appear in Hi-C contact maps. Most prominently are the TADs (topologically associated domains). Their defining characteristic is that the interaction frequency within domains is much higher as opposed to that across domains, i.e. the contact matrix resembles a block-diagonal matrix.

There are various different methodological approaches identifying the domain structure in Hi-C contact maps. A first attempt was presented in Dixon et al. [7] and is based on a two-step strategy. Firstly, the 2D contact information is condensed to the directionality index, a 1D measure encoding both downstream and upstream chromatin interactions. In

the second step, a hidden Markov model (HMM) is applied to this data to retrieve the segmentation into domains. Instead of a HMM, it is also possible to translate the directionality index into a test statistics in order to identify significant domain boundaries [48]. Lévy-Leduc et al. [83] developed a 2D model that fits a block diagonal matrix to observed contacts using maximum likelihood. Filippova et al. [85] use dynamic programming to find domains with maximal intra-domain contact frequency. Weinreb et al. [86] developed a method to find an optimal TAD hierarchy via dynamic programming. Chen et al. [84] present a method for detecting domains based on the spectral decomposition of the graph Laplacian of the Hi-C matrix.

Complementary to the above outlined heuristic and mostly image analysis motivated approaches, one can interpret the Hi-C data as interactions and treat them on this level. Following this idea leads naturally to probabilistic graphical models. In the following section we develop the approach.

7.2 Approach

Our main idea is to use an energy based probabilistic graphical model. In fact, we will construct a log-linear model over a Markov network. For the energy function a possible choice is to use the pair interactions (pairwise node potential) that are defined by the Hi-C contact map together with feature variables between which the interaction is defined. The energy of the pairs is symmetric. Now rather than learning parameters we sample the feature variables as a function of control parameters. If there is a strong interaction between a pair of nodes then we construct the energy function to favor the feature variables to be similar. On the other hand if there is only a weak interaction between nodes then the feature variables will be uncorrelated. Assume that the feature variables take value ± 1 . Within a domain, where the interaction is strong, the feature variables will all have nearly identical average values. Where there is a very weak or no interaction the feature variable will average to zero. Within this scheme domains can be identified by the boundary from values above a certain threshold and zero.

7.3 Methods

Let \mathbf{C} be the matrix containing the raw counts from the Hi-C experiment

$$\mathbf{C} = (c_{ij})_{i,j:1,\dots,n} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \dots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix} \quad (7.1)$$

where $c_{ij} \geq 0$ for $i, j : 1, \dots, n$. This symmetric non-negative matrix can be normal-

ized [81] such that the row/column sums in the Euclidean norm $\|\cdot\|_2$ equals one:

$$\mathbf{C} \rightarrow \eta = (\eta_{ij})_{i,j:1,\dots,n} = \begin{pmatrix} \eta_{11} & \eta_{12} & \dots & \eta_{1n} \\ \eta_{21} & \eta_{22} & \dots & \eta_{2n} \\ \vdots & \vdots & \dots & \vdots \\ \eta_{n1} & \eta_{n2} & \dots & \eta_{nn} \end{pmatrix} \quad (7.2)$$

i.e.

$$1 = \sum_{i=1}^n \eta_{ij} \text{ for all } j : 1, \dots, n. \quad (7.3)$$

For our approach, the matrix does not need to be doubly stochastic. It rather is convenient to compare later results with respect to the parameters that control coupling strengths. The model and the algorithm presented below just rests on the fact the matrix describes a network where the entries in the matrix represent values for the edges of the network.

7.3.1 The Model

We define feature variables s_i that can take on values ± 1 that are associated with the nodes of a network (of which we have $N = n^2$) which in fact has a simple square lattice structure. The network is defined by the Hi-C matrix presented above where the edges of the network are the entries of the matrix η_{ij} and the nodes carry the feature variables. For convenience we restrict ourselves here to just two features. In principle the feature set can be a set $\{0, \dots, q\}$ with $q \in \mathbf{N}$. Let $\mathbf{s} = (s_1, \dots, s_N)$ be a specific feature configuration. Based on the pair-interaction specified by the normalized Hi-C matrix and the feature configuration we specify a symmetric energy function. The idea being that if two nodes (here we restrict ourselves to nearest neighbor nodes) have a high value in the normalized Hi-C matrix then the feature variable should tend to be similar. If the next-nearest neighbor nodes have in turn similar Hi-C entries the feature would be propagated depending on a control parameter that governs the relative strength. The simplest ansatz in this direction is a log-linear model. In this scheme the probability for a specific configuration \mathbf{s} is

$$p(\mathbf{s}|\eta, \alpha, \beta) = \frac{1}{Z} e^{-\epsilon(\mathbf{s}, \eta, \alpha, \beta)} \quad (7.4)$$

with the normalization

$$Z = \sum_{\mathbf{s}} e^{-\epsilon(\mathbf{s}, \eta, \alpha, \beta)} \quad (7.5)$$

and $\epsilon(\mathbf{s}, \eta, \alpha, \beta)$ being the energy function. Assuming symmetric pairwise interaction between the nodes with the interaction given by the values of the normalized Hi-C matrix and a possible local bias we use the following form for the energy function

$$\epsilon(\mathbf{s}, \eta, \alpha, \beta) = \alpha \sum_{\langle ij \rangle} \eta_{ij} s_i s_j + \beta \sum_i \eta_{ij} s_i \quad (7.6)$$

where α and β are control parameters for the strength of the coupling between pairwise nodes (α) and β controlling the bias. Note that we restrict the pairwise interaction to

nearest-neighbor nodes, as depicted in Fig. 7.1, denoted by the symbol $\langle ij \rangle$.

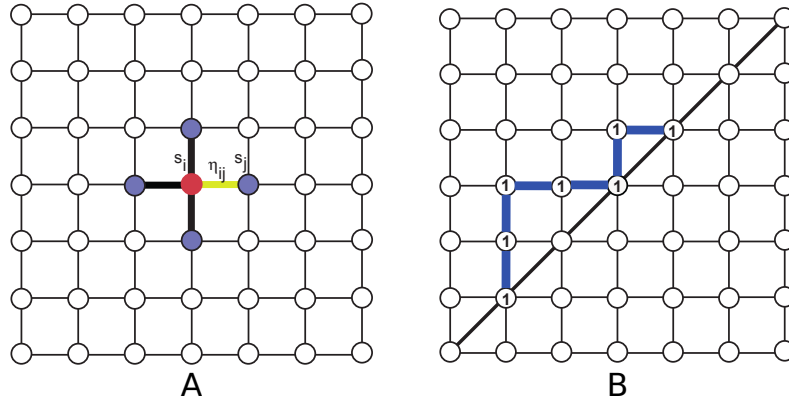


Figure 7.1: The Markov network carrying the feature variables and an interaction strength corresponding to the contact probability. **A.** The feature variables take values $s_i = \pm 1$ and the edges of the network are the entries of the contact matrix η_{ij} . **B.** Within this model, domains can be identified by the boundary from values above a certain threshold and zero.

In the above scheme we are using free boundary conditions. For nodes that have no nearest neighbors in one of the directions that interaction for this direction is taken to be zero.

7.3.2 Algorithm

Given the above model, we sample the feature variables using the Metropolis Monte Carlo method [135]. The goal is to identify the domains that have on *average* equal feature. Since strong interaction favor a like feature, i.e. the Hi-C showed a high probability for the connection between the nodes, with the parameter α we can control the relative strength of the interaction. Since this in turn influences the correlation between the nodes, large values of α will incorporate into domains of like features also nodes that have a relative lower probability of connectivity. We can thus control how much of a domain structure one wants to explore.

Algorithm 1: Hi-C Domain Structure Identification

```

initialize feature variables with feature +1
mcs ← 1
while mcs < mcsmax do
    generate a realization using Metropolis MC
    compute individual feature average
    mcs ← mcs + 1
end while
using  $\chi$  project the average feature variable to 0 or 1
delete all nodes that have at least one nearest neighbor with 0

```

At the start of the algorithm all feature variables are set to +1. Because the Hi-C interaction is non-negative, this ensures equilibrium in the sampling using MCMC [135]. The sampling is set to last up to a maximum number of iterations or terminates if the moving average of the overall feature variable has changed less than a given value.

To define the border between domains, we use a threshold c above which the average value of the feature at node i belongs to a domain, i.e. we define a characteristic function

$$\chi(i) = \begin{cases} 0, & \text{if } \langle s_i \rangle < c \\ 1, & \text{otherwise} \end{cases} \quad (7.7)$$

where $\langle s_i \rangle$ is the average value of the feature variable s . For those nodes that are not strongly connected the average in the MCMC process will tend to zero, whereas those that are strongly connected tend to +1 given the initial condition of all nodes having +1.

This yields a configuration that has only 0 or 1 for each node. To detect the boundaries we delete all nodes where at least one of the nearest neighbor nodes has feature value 0. Connecting those nodes left that have characteristic values of 1 along the diagonal define the border between the domains. With this algorithm, we are able to identify non-rectangular domains as we will shown below.

7.3.3 Validation of the Algorithm

The above outlined algorithm was tested using synthetic data. Three cases were considered. First in line is the square domain with sharp and with fuzzy boundary. The result of the domain identification is shown in the top panel of Fig. 7.2. In both cases the same control parameters $\alpha = 10000$ and $\beta = 0$ were used. In both cases the square domain is correctly identified. The dashed lines give the horizontal (vertical) identification line of the domain boundary.

The middle panel of Fig. 7.2 shows the results of the domain identification against a noisy background and have noise also inside of the domain with varying degree of intensity. In Fig. 7.2C the interaction β was slightly higher than in Fig. 7.2D.

The bottom panel shows that also non-square domains can be identified which are associated with loops in the chromosome conformations.

7.4 Results

Our method can be applied to Hi-C contact maps to detect domains, loops, loop domains and multi-scale structures.

7.4.1 Domain Detection within Hi-C Contact Maps

As discussed in the introduction, our method does not yield domain boundaries or a contiguous domain structure, but a contour that separate contact probabilities of a certain strength from the others. By identifying minima of this iso-strength contour and interpreting these as domain boundaries, domains can be described by the contour between neighboring domain boundaries as visualized in Fig. 7.3B. We used Gaussian fits in order to model domains. In order to illustrate and visually compare the results of our approach and that based on the directionality index, we used the Hi-C contact map of *C. crescentus* which is compartmentalized into CIDs [48]. The contiguous domain structure computed on the basis of the directionality index as well as the iso-strength contour yielded by our method permit characterization of the compartmentalization into CIDs of the wild-type *C. crescentus* chromosome (see Fig. 7.3A).

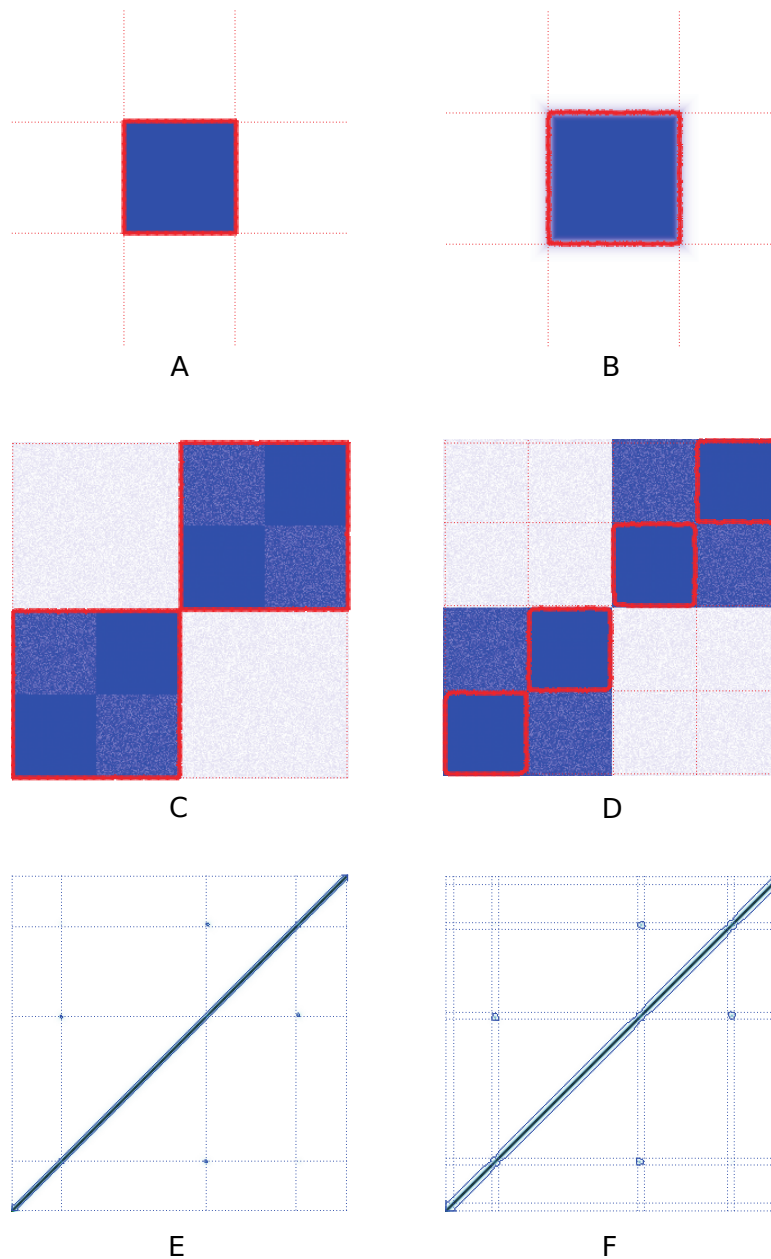


Figure 7.2: Test cases for the algorithm: **A.** simple square with no background noise and $\alpha = 10000$, **B.** simple square with diffuse boundary and no background noise $\alpha = 10000$, **C.** domains within domains with background noise $\alpha = 25000$, **D.** domains within domains with background noise, **E, F.** static loop polymer with parameters: $t = 1000$ and $t = 16000$ respectively. All results are averages over 10000 Monte Carlo Steps. Dashed lines represent the horizontal (vertical lines identified by the algorithm belonging to a domain).

7.4.2 Loop Detection within Hi-C Contact Maps

Besides simple domains like TADs or CIDs, our methodology is also able to identify loops appearing as dots of increased contact probability within Hi-C contact maps. Using Hi-C data of budding yeast cells, we show that our probabilistic graphical model correctly detects the pronounced trans interactions between the centromeres on chromosomes XIII

to XVI in the Hi-C contact map as depicted in Fig. 7.4.

7.4.3 Multi-Scale Structure Identification within Hi-C Contact Maps

The iso-strength contour yielded by our method can be adjusted to a certain scale using the strength of the coupling via the control parameters of our model. Hence, our method can be used to describe the compartmentalization of a given Hi-C contact map across multiple scales. With increasing coupling strength, the contour moves further away from the diagonal and detects larger structures. Using an exemplary excerpt of high-resolution Hi-C data of human B-lymphoblastoid cells [36], we show in Fig. 7.5 how our method identifies structures on different scales, such as loops, loop domains and complex structures.

7.5 Discussion

We have developed a probabilistic graphical model to study the domain structure visible in Hi-C contact maps. This model is based on a symmetric energy model where the interaction parameters come from the normalized entries of the contact matrix. Here the contact matrix is interpreted as a graph with $N = n^2$ nodes each node having a feature variable. Already a model where the feature variable has just two values is sufficient

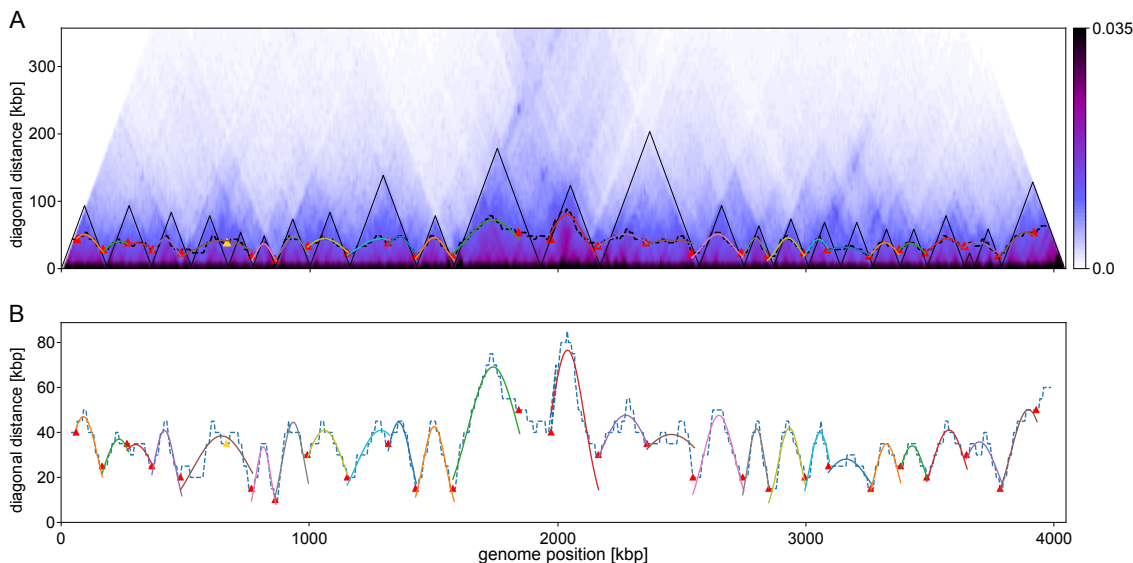


Figure 7.3: Application of our methodology to domain detection. **A.** Excerpt of the 45° anti-clockwise rotated Hi-C contact maps of the *C. crescentus* chromosome [48] and the resulting domain structure of both the approach of Le et al. using the directionality index and our probabilistic graphical model. Contrary to the directionality index approach our method does not yield domain boundaries or rather a contiguous domain structure (black), but a contour (dashed black) separating contact probabilities of a certain strength (depending on the coupling constant) from the background. This iso-strength contour allows to characterize the compartmentalization irrespective of whether or not there are contiguous domains in the form of squares. **B.** Detailed view of the iso-strength contour yielded by our method. We define domain boundaries (red triangles) as minima of the contour (dashed blue) and characterize domains by Gaussians fitted to the contour between two neighboring domain boundaries. We differ between two types of minima of the contour: Minima marked as red triangles correspond to domain boundaries and minima marked as yellow triangles indicate a boundary within a domain, i.e. a nested domain structure.

to identify synthetic domains. These domains incorporate partial noise as it would be expected in the actual contact maps. The domains themselves are set against a background of noise. The model is able to identify the noise through the average feature variable which is clearly distinct to the one in the domain. Within the domain, depending on the strength of the control parameter, the average value of the feature variable is homogeneous. This leads to the clear identification of the domain boundary as those nodes that have at least one of the nearest neighbors having a feature value different from the others. Using real Hi-C contact maps, we have showed that our method is able to identify domains like TADs or CIDs, loops and loop domains as well as multi-scale structures of complex shape.

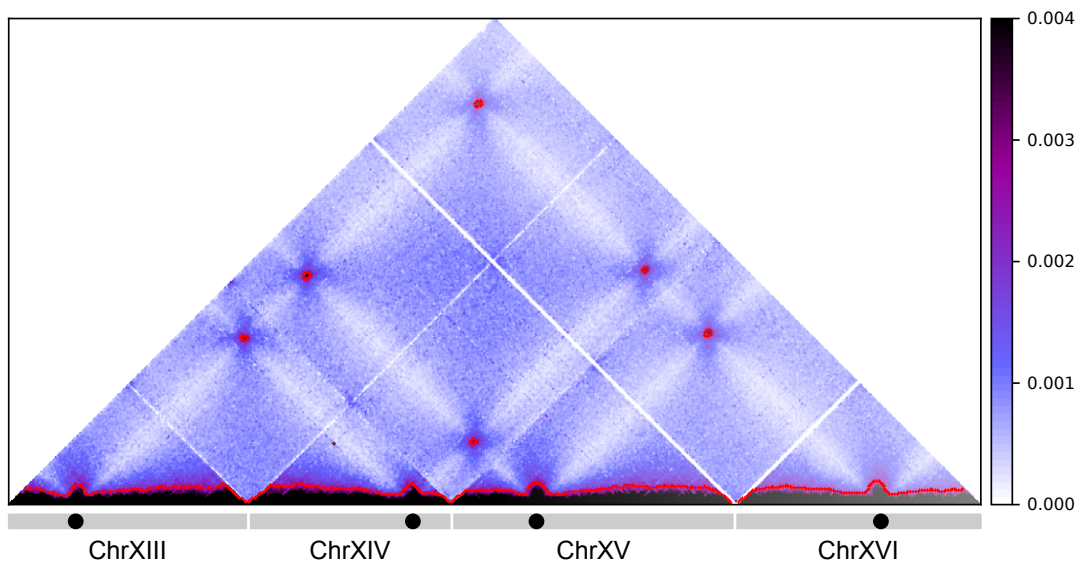


Figure 7.4: Loop detection with our approach. Using Hi-C data of budding yeast cells arrested in G1 phase [136], we show that our method correctly identifies the pronounced trans interactions between the centromeres on chromosomes XIII to XVI appearing as dots of increased contact probability in the Hi-C contact map. The positions of chromosomes XIII to XVI are illustrated as gray bars below the heat map and the locations of centromeres on each chromosome are represented by black dots. Our method yields both closed contours around the loop locations and the previously discussed iso-strength contour along the diagonal; they are illustrated by red plus signs.

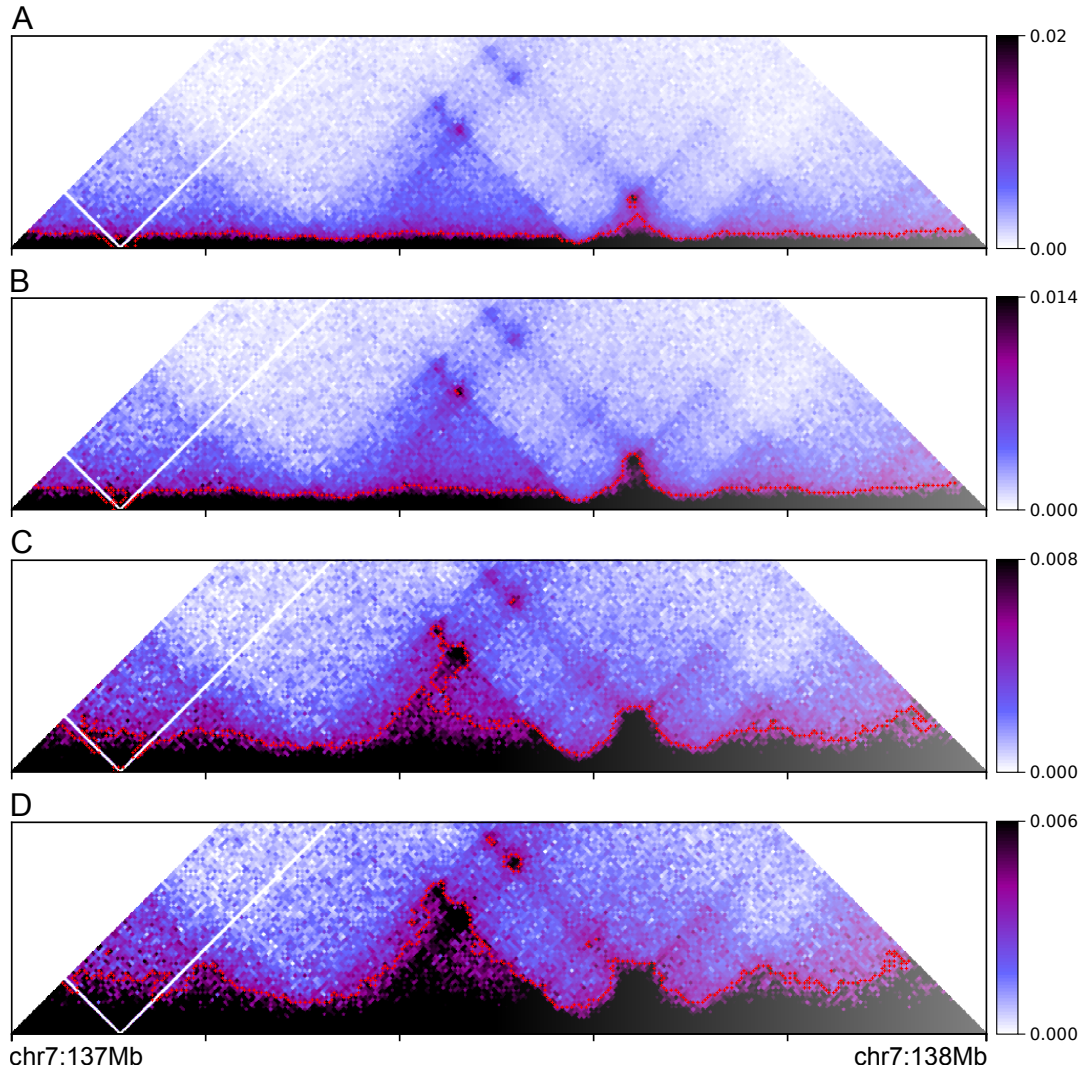


Figure 7.5: Multi-scale structure identification using our method. The contour that describes how a given Hi-C contact map is compartmentalized can be adjusted to a certain scale using the strength of the coupling via the control parameters of our model. With increasing coupling strength, the contour moves further away from the diagonal and detected structures become larger. Using an exemplary excerpt (the region between 137 and 138 Mbp of chromosome 7) of Hi-C data of human GM12878 B-lymphoblastoid cells with a resolution of 5 kbp [36], we show how our method identifies structures on different scales. **A.** On the lowest level the loop domain located at 137.65 Mbp is identified. **B.** The loop domain at 137.65 Mbp is now detected as a normal domain. Additionally, the loop at 137.45 Mbp is recognized. **C, D.** The previously detected structures become larger and further loops with lower signal intensity are identified.

Chapter 8

The Role of Loops on the Order of Eukaryotes and Prokaryotes

References

The results presented in this chapter are published as and adapted from

- A. Hofmann and D.W. Heermann (2015), *The role of loops on the order of eukaryotes and prokaryotes*. FEBS Letters, 589: 2958-2965. doi: 10.1016/j.febslet.2015.04.021.

Chapter Summary

The study of the three-dimensional organization of chromatin has recently gained much focus in the context of novel techniques for detecting genome-wide contacts using next-generation sequencing. These chromosome conformation capture-based methods give a deep topological insight into the architecture of the genome inside the nucleus. Several recent studies observe a compartmentalization of chromatin interactions into spatially confined domains. This structural feature of interphase chromosomes is not only supported by conventional studies assessing the interaction data of millions of cells, but also by analysis on the level of a single cell. We first present and examine the different models that have been proposed to elucidate these topological domains in eukaryotes. Then we show that a model which relies on the dynamic formation of loops within domains can account for the experimentally observed contact maps. Interestingly, the topological domain structure is not only found in mammalian genomes, but also in bacterial chromosomes.

8.1 Introduction

Mammalian interphase chromosomes are hierarchically organized [25, 72]. On the one hand, at the level of the nucleus, fluorescence in situ hybridization (FISH) and genome-wide chromosome conformation capture (3C) studies, such as Hi-C, have revealed an inter-chromosomal compartmentalization in the form of the formation of distinct chromosome territories [3, 24]. Individual chromosomes, on the other hand, also show a domain-like structure as observed in recent genome-wide high-resolution Hi-C and 5C studies [7, 8, 33]. These 3C-like studies indicate that eukaryotic genomes are partitioned, at the sub-megabase level, into discrete structural units with highly increased frequency of internal contacts, referred to under different terms, such as “topological domains” [7], “topologically associating domains” (TADs) [8] and “physical domains” [33]. We will stick to the term “topological domains” for these intra-chromosomal domains, within which the chromatin fiber preferentially interacts. This finding of a domain organization of individual chromosomes is not only supported by data stemming from 3C-like studies examining genomic interactions of a large population of cells, but also by an analysis of individual cells, the single-cell Hi-C methodology [59].

Besides the eukaryotic chromosomes of humans, mice and *Drosophila melanogaster*, bacterial chromosomes are also characterized by a hierarchical organization [137]. The *Escherichia coli* chromosome consists of macrodomains on the megabase scale [46, 138], which, in turn, are composed of topological domains on the smaller scale [139]. Recently, the circular chromosome of *Caulobacter crescentus*, as a further example, has been shown to be composed of topological domains with the help of an in-depth Hi-C analysis [48]. Taken together, these analogies to the organization in eukaryotes suggest that an intra-chromosomal domain structure is a fundamental building block of chromosome structure of organisms.

Although their important role in shaping the three-dimensional organization of the genome seems acknowledged, there remains the question how topological domains are

established, hence what causes the increased contact frequency within these genomic regions. One striking observation is that most identified enhancer–promoter pairs have been shown to belong to the same topological domain [37,140]. The finding that these enhancer–promoter units mostly coincide with topological domains [37], however, has to be treated with caution since the increased background of the interactions within topological domains was not taken into consideration in this analysis.

As enhancer–promoter activity is known to involve DNA loop formation [36,141,142] this hints at an important organizational role of loops [82,98,133]. In fact, there is emerging evidence that loops contribute to compartmentalization in the eukaryotic genome [36] and that a fraction of topological domains actually corresponds to loop domains that are conserved across cell types as well as species and stable against cell-to-cell variation [36]. Looped structures are thereby likely to be made up of both dynamic looping interactions [143] and a network of static loops [36]. The presence of these loops creates entropic constraints that helps maintaining chromosome structure. The role of proteins that are involved in the formation of loops, such as CTCF and cohesin, is complex, but has been established through 3C-related and chromatin immunoprecipitation studies [144,145] as well as FISH experiments [146]. However, it is controversial whether the two proteins are also involved in establishing topological domains [34,147].

In this review, we shed light on the theoretical analysis of topological domains appearing as a ubiquitous feature in contact maps based on current high-resolution Hi-C data. After the presentation of modeling approaches that appeared in the literature so far and aim to explain the appearance of topological domains, we investigate a model that is based on chromatin looping and incorporates the concept of topological domains. We conclude with a summary of the effects that loops have on the nuclear organization not only in mammalian genomes, but also in the bacterial nucleoid.

8.2 Current State of Modeling

Although topological domains have been repeatedly discovered in current high-resolution chromosome conformation capture experiments [7,8,33] as well as earlier [148] and this substructure seems to be an essential characteristics of interphase chromosomes, only little is known about their internal structure and organization. Several models have been proposed to theoretically explain the observed clusters of increased contact frequency in contact maps, none of which accounts for the essential role of loops.

The model of Benedetti et al. [149] is designed to reflect the situation where unconstrained supercoiling, referring to the over- or under-winding of the DNA double strand, acts on chromatin fibers that are sparsely attached at specific sites to nuclear granules. This model is supported by reports indicating that boundary elements of topological domains are attached to nuclear granules and, more importantly, reports indicating that chromatin fibers are supercoiled [150]. In this proposed model, individual topological domains are simulated as polymer rings. The closure is thereby essentially needed for maintaining the torsional tension introduced in order to be able to get (super-) coiled structures. Without an actual closure of the polymer chain possible torsional tension would be released through free rotation of the ends, thus a modeling of supercoiling would not be possible. However, this strategy of preventing the unangling problem comes with the price that actually one half of those supercoiled rings has to be neglected in the statistics of contacts. Additionally to the torsional potential for the purpose of introducing supercoiling into the model, it incorporates excluded volume interactions between monomeric beads

as well as a bond length and a harmonic bending potential. For mimicking the effect of high concentration of chromatin in the eukaryotic nucleus, i.e. an increased contact probability of the polymer chain, Benedetti et al. performed their simulations in cubic confinement such that the simulated chains occupied 20% of the available volume. Simulated plectonemes appear in the average contact maps of simulated chromatin fragments as compartments of increased contact frequency, thus resembling the experimental contact maps. The underlying principle of the separation of individual domains or plectonemes in this modeling approach simply follows entropic repulsion, namely, the permanently connected polymer rings repel each other, such that fixed boundaries between supercoiled regions, i.e. topological domains, arise. The supercoiling of individual rings strengthens entropic repulsion.

The idea of the “strings and binders switch” (SBS) model proposed by Barbieri et al. [99,102] is to allow for the attachment of diffusible factors (binders) to binding sites along the simulated polymer chain. The obtained polymer configurations are thus dependent on binding site distribution, binder concentration and binding affinity. The polymer fiber itself is modeled as self-avoiding polymer bead chain and the binding molecules are represented by Brownian particles with a certain concentration. A fraction of polymer sites can be bound by diffusing molecules with a certain chemical affinity. Molecules binding to more than one polymer site lead to the formation of loops. To explore the formation of chromatin globules in the SBS model, Barbieri et al. assumed a polymer containing different kinds of binding sites, i.e. specialized binding sites, each with specific affinity to one kind of binder. As a consequence, each topological domain corresponds to one specific binder. Under these conditions, the SBS model produces separate domains of increased contact frequency, though it is important to notice that the contact frequency in the appearing domains does not monotonically decrease with increasing distance from the main diagonal. This characteristic of the contact map averaging over the ensemble of simulated polymer configurations indicates that the domains are rather stiff.

In the same light of the SBS model, it was recently observed that regularly-spaced bridging in combination with a homogeneous self-adhesion interaction along a linear polymer chain can lead to a stable multi-domain configuration, hence a compartmentalization into topological domains [151].

8.3 Static Loop Domains

Inspired by the observation of thousands of loops both in a very recent high-resolution in situ Hi-C study of the human genome [36] and in earlier studies [35,40], we analyze the connection between loops and topological domains. These loops were found to link promoters and enhancers, correlate with gene activation and are conserved across cell types and species. Furthermore, it is observed that they are formed at domain boundaries and bind CTCF.

While the resolution of the Hi-C contact maps discussed in connection with the observations of topologically associating domains [7,8] is sufficient to show the existence of these distinctive clusters of high contact frequency, it does not allow for the analysis of their intrinsic structure. It was only with the high-resolution in situ Hi-C study of Rao et al. [36] that certain ends of individual topological domains were detected to be attached to each other forming simple loops or some kind of network of loops. Since loops are observed to demarcate a fraction of the boundaries of topological domains [36], we follow the terminology of Rao et al. and refer to such domains as loop domains. Being

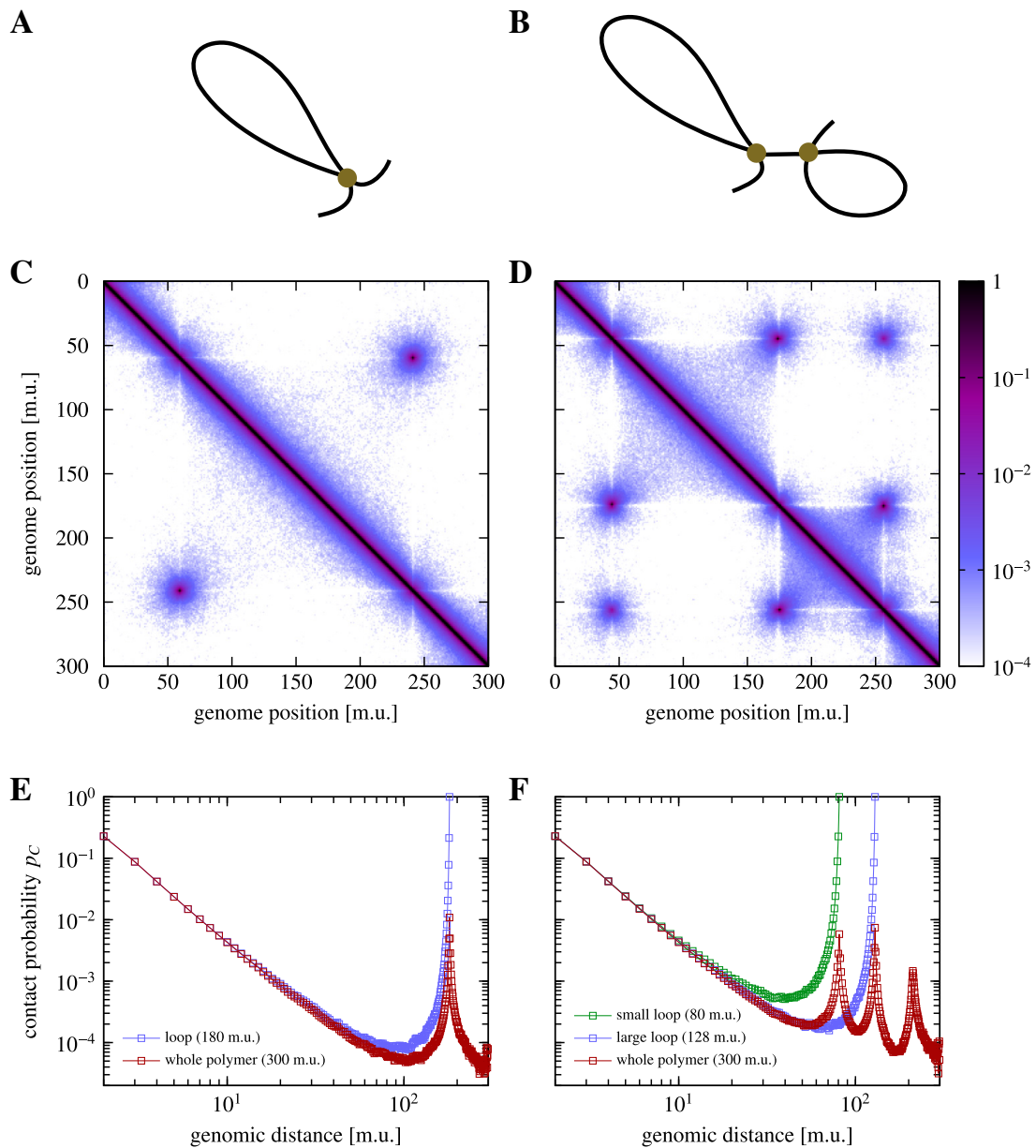


Figure 8.1: Simple loop models recapitulate the experimental observation of loop domains. **A, B.** Sketches of the loop topologies for our polymer simulations. **C.** Contact map for a simulated polymer comprised of a single static loop. The polymer ($N = 300$ monomers) is composed of a static loop modeling a topological domain with a size of $N = 180$ monomer units (m.u.). **D.** Contact map for a simulated polymer comprised of two single static loops. The polymer ($N = 300$ monomers) is composed of two static loops modeling two topological domains with sizes of $N = \{128, 80\}$ monomers, respectively. **E, F.** The contact probability profile of both polymer topologies is shown for both the individual domains (loops) and the whole conformations, respectively. The loop closures generate local maximums in the graph showing the genome-wide profile.

interested in the composition of contact maps of such loop domains, we modeled systems composed of one and two static loops (see Fig. 8.1A and B) to see whether they resemble the structures in the experimental Hi-C data. As depicted by means of the contact maps in Fig. 8.1, the presence of simple loops results in formation of sharply defined squares in the

contact map showing high intensity of contacts around their vertexes that are distal from the diagonal. These prominent peaks in the contact map reflect the fact that the border elements belonging to the same topological domains were brought together by the loop closure. The contact map of the polymer system composed of two loops of different size illustrates that neighboring loops do not interact due to entropic repulsion; a finding that has been quantified for ring polymers [67]. This feature becomes even clearer if we look at the contact probability profiles, i.e. contact probability as a function of the genomic distance (see Fig. 8.1E and F). Initially the probability of genome-wide contacts decreases with separating genomic distance. However, as this distance exceeds half of the total loop size, we observe an actual increase of the contact probability reaching a maximum at a distance that equals the loop size. The resulting “U” shape of the contact probability profiles of both individual domains and the whole polymer is due to the fact that genomic distant regions close to two border elements of the same domain are brought together by the loop closure. In fact, this prominent shape is a distinctive feature between contact probability profiles of loop domains and those of topological domains as observed in [7,8] since the latter monotonously decrease with increasing genomic distance.

8.4 Dynamic Loop Interaction within Domains

Regardless of the evidence for invariable DNA loop domains throughout the genome, the question remains on which structural principle topological domains are based. Certainly, these domains that show a strict decrease in contact probability as a function of increasing intra-domain genomic distance rather than a peaked contact probability at the corner do not correspond to invariable loops. It is, however, possible to think of these domains in terms of simple loops forming only for a certain fraction of time and stay open for the rest. Moreover, we have to bear in mind that the experimental Hi-C data are derived from a large population of cells. Hence, we deal with contact information stemming from an ensemble of cells with possible conflicting conformations on average.

Based on these considerations, it is obvious to think of loops in a dynamic fashion. Topological domains may be established through a dynamic looping mechanism as sketched in Fig. 8.2. This schematic is based on the idea that distant regulatory elements make direct contact with either the promoter or another regulatory element of the gene they control, i.e. form a loop. As indicated in the introduction, such enhancer–promoter interactions are particularly frequent within topological domains [37]. The coincidence of enhancer–promoter units with topological domains suggests that a dynamic loop domain structure underlies the topological domain structure. Analogously, loops could also dynamically form within loop domains.

A recent simulation study [152] analyzes how the looping interaction between elements in the vicinity of an enhancer–promoter pair influences their contact frequency. The simulations show that a chromatin loop, formed by elements flanking either an enhancer or a promoter, suppresses enhancer–promoter interaction, working as an insulator. In contrast, a loop formed by elements located in the region between an enhancer and a promoter, facilitates their interaction. Many enhancers, promoters, and loop-forming elements are present in a given genomic region (see Fig. 8.2), leading to a complex network of insulation and facilitation processes. Facilitation results from the effectively shortened genomic distance between enhancer and promoter due to the loop. Insulation is due to excluded volume interaction and steric exclusion by the loop. Taken altogether, loop topology influences promoter–enhancer interaction and vice versa (as depicted in Fig. 8.2).

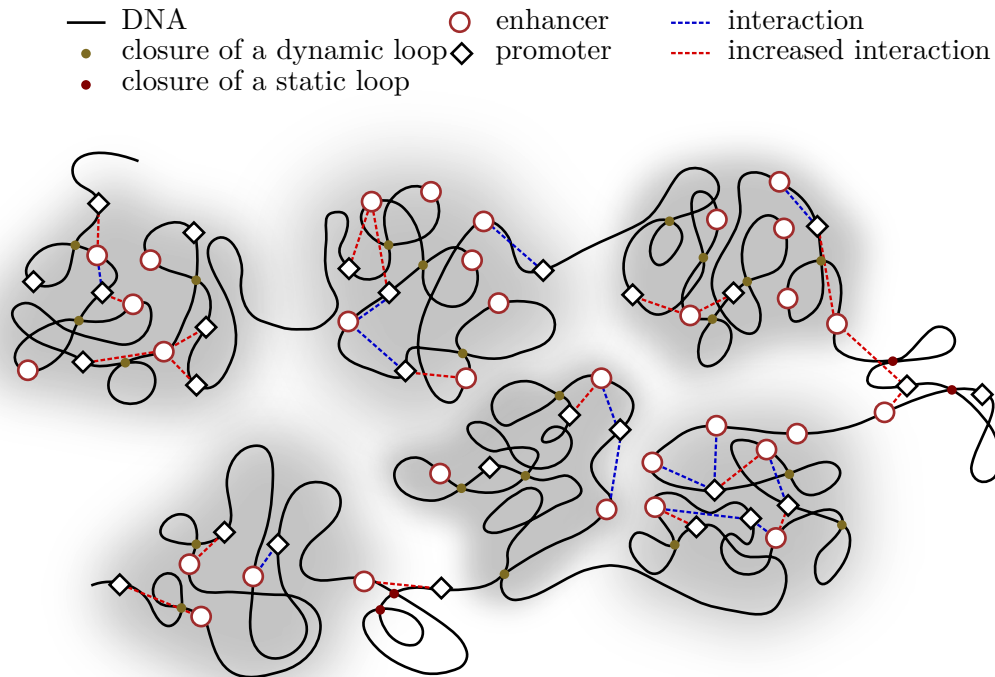


Figure 8.2: Schematic illustration of both the dynamic loop interaction within intrachromosomal domains and static loops. Promoters (black) and enhancers (red) are represented by diamonds and circles. Interactions relevant to gene expression are shown as dotted lines. Dashed red lines thereby indicate interactions enhanced by a loop as opposed to dashed blue lines that represent interactions not enhanced by a loop. Loop closures caused by certain linking proteins as well as enhancer–promoter interaction are shown as small filled circles and can be both temporary (gold-colored) and static (ruby-colored). A snapshot of the three-dimensional organization of the genome is depicted with the interactions between genomic elements. The spatial organization and the loop topology is partly subject to fluctuations that affect gene expression. However, this dynamics does not lead to a change in the organization of topological domains (shadowed areas).

We model these effects altogether by a dynamic and probabilistic loop formation within topological domains. To this end, we use a simple polymer model that has already been shown to explain the formation of distinct chromosome territories. In this dynamic loop (DL) model, the chromosomal fiber is represented as a self-avoiding (SAW) random walk polymer allowed to form probabilistic intra-polymer crosslinks between non-adjacent monomers [82]. As a consequence, loops of different size are formed. The main model parameter is the looping probability (p_{loop}), a measure for the probability that a loop is formed between two non-adjacent monomers. The dynamic formation and dissolution of loops thereby mimics the highly dynamic nature of enhancer–promoter loops as well as cell-to-cell variation that also supports variations in the loop topology. The simple example conformation is only consisting of two domains of different size as we are interested in the qualitative effects of dynamic looping on the contact probability measures rather than fitting our model to available experimental datasets. The results for the contact map and the contact probability profile are shown in Fig. 8.3 and could be fitted to those observed experimentally as it is possible to adjust the looping frequency and thus contact probability for individual domains. Moreover, by restricting the loop interaction to regularly spaced sites along the polymer chain as well as confining the interaction to certain compartments, our model adapts to the specificity binder model discussed previ-

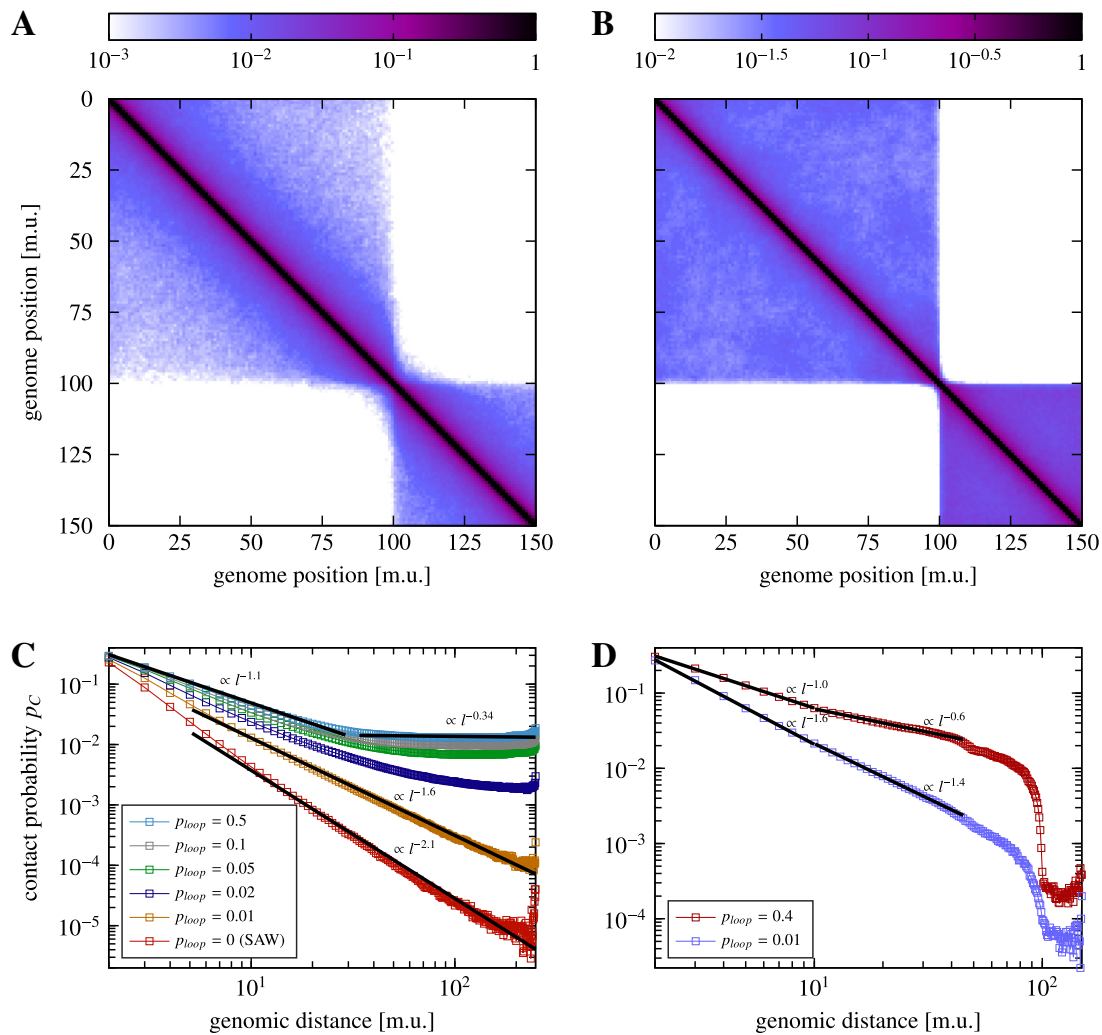


Figure 8.3: **A, B.** Contact maps of simulated polymers ($N = 150$ monomers) composed of two domains with varyingly strong dynamic looping ($p_{loop, A} = 0.01$, $p_{loop, B} = 0.4$) resemble those of experimentally observed topological domains. **C.** The contact probability p_c for two specific sites as a function of the genomic separation between them. Shown are the results for equilibrated polymers composed of $N = 250$ monomers and various looping probabilities including the case of the self-avoiding walk ($p = 0$) and a simple random walk. The contact probability decreases as a power-law $l^{-\beta}$ with genomic separation for separations $n \gtrsim 10$. As already discussed by Bohn et al. [82], the exponent is thereby strictly dependent on the looping probability. Compared to the self-avoiding walk, the co-localization probability is strongly increasing due to dynamic looping. **D.** The contact probability profiles for both polymers strictly decrease as a function of the genomic distance. The two functions can be partitioned into two regimes ($2 \leq l \leq 10$, $10 \leq l \leq 45$) where their decrease can be approximated by power laws as depicted in the graph.

ously. In fact, the SBS model, which assumes a diffusible component being responsible for loop formation by linking two monomers of the polymer, is a special case of the DL model implicitly incorporating the properties of such binders in the looping probability parameter much like the implicit water in the interaction potentials that are derived for proteins.

8.5 Effect of Loops on the Nuclear Organization

Specifically in human cells the zinc finger protein CTCF and the protein complex cohesin have been linked to the formation and maintenance of loops. CTCF has even been named the master weaver of the genome [34, 144, 147, 153, 154]. Surprisingly, however, little is known about the interaction of these proteins with DNA [155, 156].

In bacteria, nucleoid-associated proteins, such as H-NS, HU, Fis and IHF, can influence DNA structure locally by bending and wrapping DNA segments [42] as well as globally by looping [43, 157] and by providing boundaries for DNA topological domains [44]. A recent approach investigating the spatial distribution of H-NS in *E. coli* using both super-resolution microscopy and 3C provides evidence for the juxtaposition of distant DNA segments interacting with H-NS [41].

Because of their implications in the formation of loops, experiments have interfered with cohesin and CTCF as well as the zinc-finger protein family in general [34, 158, 159]. Contrary to expectation, a recent FISH study [146] shows that the chromosomes do not swell but compactify as a consequence of the depletion of these two proteins and hence a decline of loops. This observation is quite puzzling since loops are coupled with an increased level of compaction and provide a consistent framework [82, 98] for the explanation of various experiments, such as Hi-C [3, 21] as well as FISH experiments [127].

Moreover, the segregation of domains, and thus also the TADs within one chromosome can be explained within the loop framework [72, 160]. Also in *E. coli* [69, 161, 162] the segregation of chromosomes can be explained.

At least three factors influence this segregation. First, there is the repulsion between the loops [82, 163]. Indeed this is due to the entropic repulsion between the loops, i.e., based on the excluded volume of the monomers. Here entropy enters as an ordering mechanism which is a very interesting phenomenon [164–166] since with entropy one usually associates disorder. The solution to this puzzle is the change in topology in the chromosome as viewed as a polymer. Since *E. coli* is per se a circular chromosome upon replication the two chromosomes will separate [167].

Essentially due to repulsion between the loops there is a segregation between the more compact loops and those which are less compact. One can think of this as corresponding to heterochromatin and euchromatin. This segregation can be linked to the expression level of the chromosome such that those regions with high expression correspond to the not so compact loops and those with low expression to those regions with little expression [168]. Thus the chromosome is made up of domains of varying degree of loops in size and compaction.

Second, in confined space this would also be true for linear chromosomes as has been shown very convincingly by Jun and Mulder [169], at least the fact that the two linear chromosomes separate, not necessarily the internal segregation. However, what maintains the separation to a very high degree? Even though the chromosomes will separate, there is nevertheless almost always an overlap between the two chromosomes. To assist in helping and maintaining the separation the MinD proteins have been shown to play a crucial role [170].

Even on the level of the nucleus this ordering (loops that repel each other, leading to the formation of domains within the chromosome) holds true. The segregation of chromosomes in the human nucleus [24] can be explained in the framework of loops [82]. Since chromosomes in this picture are made up of loop domains within loop domains which themselves are loops clearly they repel each other. As a matter of fact the force that each chromosome exerts onto the other can be calculated [82]. Rosa and Everaers [171] have

argued on the basis of classical polymer theory that linear polymers do not mix due to the long relaxation. The point of view taken here is that the polymer is much shorter due to the loops and that the entire polymer is not linear but rather a looped ellipsoid.

Furthermore the mechanical properties [172,173] of chromosomes in metaphase depend on the loops. Specifically the local stiffness and hence the flexibility [174] is determined by the loops.

8.6 Conclusion

An important finding concerning the three-dimensional architecture of eukaryotic genomes is that individual chromosomes are compartmentalized into loops [36] and topological domains [7, 8, 33, 160], both of which depicting fundamental regulatory and structural building blocks of chromosomes that are stable between cell types. Chromatin interactions almost exclusively take place within topological domains and not across them.

Though the existence of this intra-chromosomal compartmentalization is proposed in all newly published results of 3C-like experiments, explanations from a theoretical point of view are scarce. In this review, we focused on the modeling of the experimental findings of both loop domains and topological domains, which, as opposed to the former, do not involve a closure to a loop. Loop domains can be readily simulated by statically adjusting the topology. Topological domains, on the other side, are characterized by a highly dynamic internal organization and can be modeled by assuming dynamic loop interactions accounting for this highly flexible internal structure [160]. The idea of enhancer–promoter units overlapping with these spatial domains [37] supports such an idea. Compared to the model assuming the interactions within topological domains to be due to supercoiling, our model can also explain loop domains and dynamic loop formation due to interaction between enhancers and promoters. Nevertheless, supercoiling is likely to cause further compaction of loops. The SBS model assumes that proteins bind to the chromatin fiber causing loop formation. Although quite similar to our approach, this actually needs different binders for the explanation of topological domains. Moreover, we can adjust the strength of the decrease of the contact probability as a function of the separating genomic distance.

Similarly to the findings in eukaryotic genomes, a recent study mapping the structure of the *Caulobacter crescentus* chromosome hints that bacterial genomes are also compartmentalized into topological domains of increased contact probability [48]. While it is probable that these domains are comprised of supercoiled plectonemes into a bottlebrush-like fiber for the case of the *Caulobacter* chromosome, it is possible that for other bacteria, such as *E. coli*, similar domains could be established by loop-forming proteins [41].

8.7 Methods

In this study, we performed Monte-Carlo simulations using the Dynamic Loop (DL) polymer model [82] to generate chromosomal conformations. The DL model incorporates chromatin loops by using a dynamic looping mechanism of the model fiber. When two monomers come into physical proximity to each other by diffusional motion, a cross-link can be created between them with a certain probability p_{loop} , which we refer to as looping probability. In case the cross-link is formed, a lifetime drawn from a Poisson distribution with mean value τ is assigned to it. The cross-link dissolves again after this lifetime, and thus, the loop vanishes. By this dynamic mechanism, there is a constant association and

dissociation of non-adjacent monomers, resulting in loop creation and dissolution. We confined this dynamic loop formation to certain regions along the polymer chain in order to model topological domains.

In contrast to this, the topology of the backbone of the polymer is fixed during the simulation. For the polymer chains we used the well-established bond fluctuation method [65,66]. In the simulations a monomer of the polymer chain is randomly selected and, if possible, randomly moved to one of its nearest neighbors on the lattice. Excluded volume interactions are taken into account by preventing a lattice site to be occupied by more than one monomer. When simulating N monomers we define one Monte-Carlo step (MCS) to correspond to N moves, i.e. on average each monomer is translated once during a MCS.

3C-based technologies, such as Hi-C, are experimental methods that can quantify the contact frequency between different sites of the DNA molecule. Fortunately, in our simulations the contact frequency can be measured comparatively simple since we know the exact configuration of our polymer, i.e. the position of each single monomer in the three-dimensional space at each point in time. We only have to quantify the contact frequency of all pairs of monomers. By averaging over the whole ensemble of conformations and subsequent normalization we can make the step from contact frequency to contact probability p_c .

In order to generate thermodynamically equilibrated polymer conformations we used the Metropolis Monte Carlo method. Since subsequently created conformations are highly correlated, we determine, for each set of parameters, the autocorrelation function of the squared radius of gyration. Then, the integrated autocorrelation time τ_{int} is computed by applying the windowing procedure introduced by Sokal [64]. We consider two subsequent conformations as uncorrelated after $5\tau_{int}$ MCS therewith creating 10000 – 100000 independent configurations.

Further details on the simulations can be found in previous works [82,163].

Chapter 9

Insulator-like Domains Drive Genome Organization in Bacteria

References

The results presented in this chapter are adapted from

- G. Dugar, A. Hofmann, D.W. Heermann and L.W. Hamoen (2019), *Robust interaction between insulator-like domains drives genome organization in bacteria*, in preparation, to be submitted to Nature.

Chapter Summary

Chromatin insulators are DNA-protein complexes known to orchestrate genome organization in eukaryotes. Insulator elements typically comprise of clustered binding sites for a specific DNA binding protein which mediates long-range interaction with other such insulator elements. Insulator proteins like CCCTC-binding factor (CTCF) along with the Cohesin complex are the main factors involved in genome organization in vertebrates [175,176]. Insulator protein homologs are believed to be absent in bacteria, however homologs of Cohesin are present and do participate in DNA compaction [48,177,178]. Here, we developed an unbiased genome-wide approach based on differential sedimentation to identify large protein-DNA complexes, which may participate in bacterial genome organization. We show that a transcription factor, Rok [179], binds to certain specific regions of the *Bacillus subtilis* genome to form eight large DNA-protein complexes. Using chromosome conformation capture (Hi-C) and live imaging of DNA loci, we show that these insulator-like complexes robustly interact with each other over large distance, with some interactions spanning the opposite ends of the *B. subtilis* genome. These long-range interactions lead to global restructuring of the chromosome by formation of new topological domain boundaries and altering the interaction frequency within the existing domains. Upon Rok deletion, these insulator-like elements are free to engage in short-range chromosomal interactions, resulting in local restructuring. Overall, we show how a prokaryotic protein can act as a functional analog of insulator proteins previously only identified in eukaryotes.

9.1 Introduction

Chromosome capture techniques like Hi-C have revealed that all living organisms, including bacteria, form structurally organized genomes which aids in DNA compaction, replication, gene regulation and DNA segregation [2, 47, 48, 180, 181]. Chromatin insulator elements are functionally conserved from *Drosophila* to mammals. *Drosophila* harbors multiple insulator proteins, but CTCF is the only insulator protein identified in mammals. CTCF was initially identified as a transcriptional repressor, and over the last decades its role in regulation of mammalian chromatin architecture has been studied in great detail [175, 176, 182–184]. CTCF interacts with specific regions of the genome and dictates long range chromosomal interaction resulting in loop formation [7, 175]. The interaction of CTCF bound regions together with loop extrusion by the Cohesin complex leads to formation of topologically associated domains (TAD) boundaries [176, 182].

Hi-C has also revealed the presence of TAD-like structures in several bacterial species, which are called chromosomal interaction domains (CID) [47, 48, 73, 75, 181]. Although, cohesin mediated loop extrusion is involved in tethering the chromosome arms in *B. subtilis* and *C. crescentus*, this mechanism does not seem to be involved in CID formation [48, 177] and to the best of our knowledge, an insulator-like genome structuring factor has never been identified in any prokaryote.

We reasoned that unbiased identification of large protein DNA complexes can help gain insight into higher order genome organization and topological domain formation in

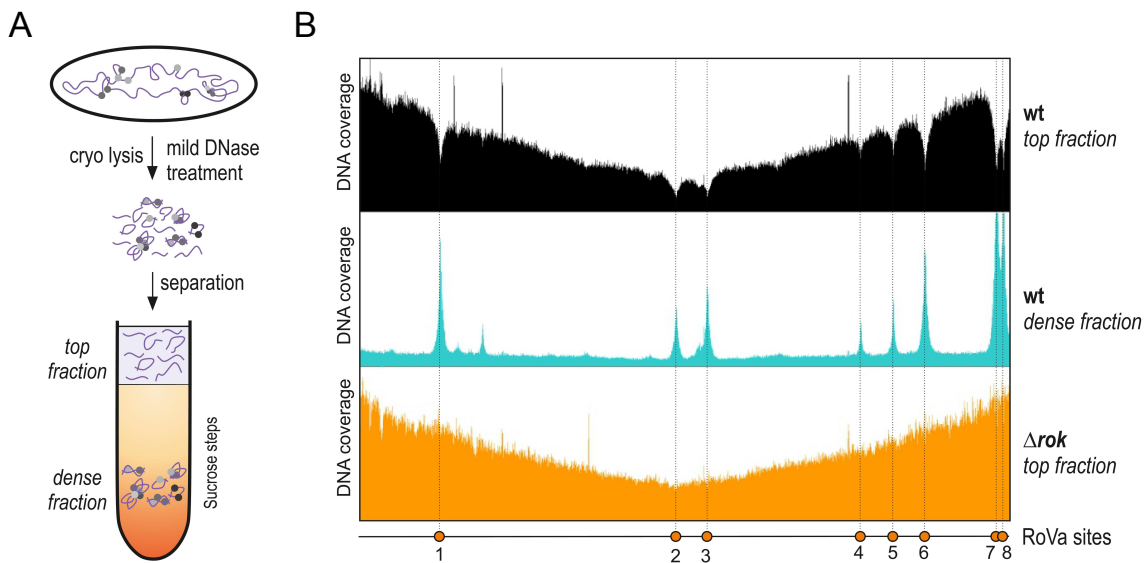


Figure 9.1: Identification of DNA associated with large complexes using SICO-seq. **A.** Large DNA-protein complexes are removed from cell lysate using ultracentrifugation over sucrose steps. The DNA retained in the cleared top fraction is sequenced. **B.** DNA coverage maps of the top fraction (black) along with the dense fraction (blue) obtained from the WT strain. DNA coverage map of the top fraction of the Δrok strain (orange).

bacteria. To test this, we developed a differential sedimentation based deep-sequencing approach to specifically identify DNA associated with large complexes. Briefly, large complexes were excluded from partially digested cell lysate of *B. subtilis* (at exponential phase) by ultracentrifugation over dense sucrose cushions. The top cleared fraction was then examined for the relative loss of DNA over the whole genome (Fig. 9.1a). Regions with relatively lower DNA coverage in the top fraction are likely to be engaged in complex formation with proteins and hence could migrate to dense sucrose fractions. Using this approach, we found 8 major valleys as local minima in DNA coverage over the whole genome of *B. subtilis* (Fig. 9.1b). These chromosomal regions could also be isolated from the sucrose-dense fraction after stabilization of cells or complexes by mild fixation (Fig. 9.1c, 9.2c). We named this sedimentation based approach to identify chromosomal complexes as SICO-seq. To find potential DNA binding proteins involved in this DNA-complex formation, we analyzed the existing chromatin immunoprecipitation (ChIP) datasets for *B. subtilis* proteins. Interestingly, we found that all 8 sites (valleys) overlapped with the binding sites for the transcription factor Rok [185]. However, Rok is known to bind more than 200 sites on the *B. subtilis* genome [185]. To examine whether Rok is indeed responsible for formation of these complexes, we performed SICO-seq with a *rok* deletion strain. As shown in Fig. 9.1b, this resulted in disappearance of all 8 valleys, and hence we named these sites - RoVa (Rok dependent Valleys).

To uncover why only a few of the Rok binding sites lead to the formation of RoVa complexes, we quantified the abundance of Rok binding motifs over the genome. Rok primarily binds A/T rich regions like other bacterial xenogeneic silencers, but using protein binding microarrays it was determined that Rok has even higher affinity for certain A/T rich motifs containing G/C residues [186]. These high affinity Rok binding motifs were found to be significantly enriched at RoVa sites when compared to other highly enriched non-RoVa Rok binding sites (Fig. 9.2a, b). As expected, high affinity Rok binding motifs

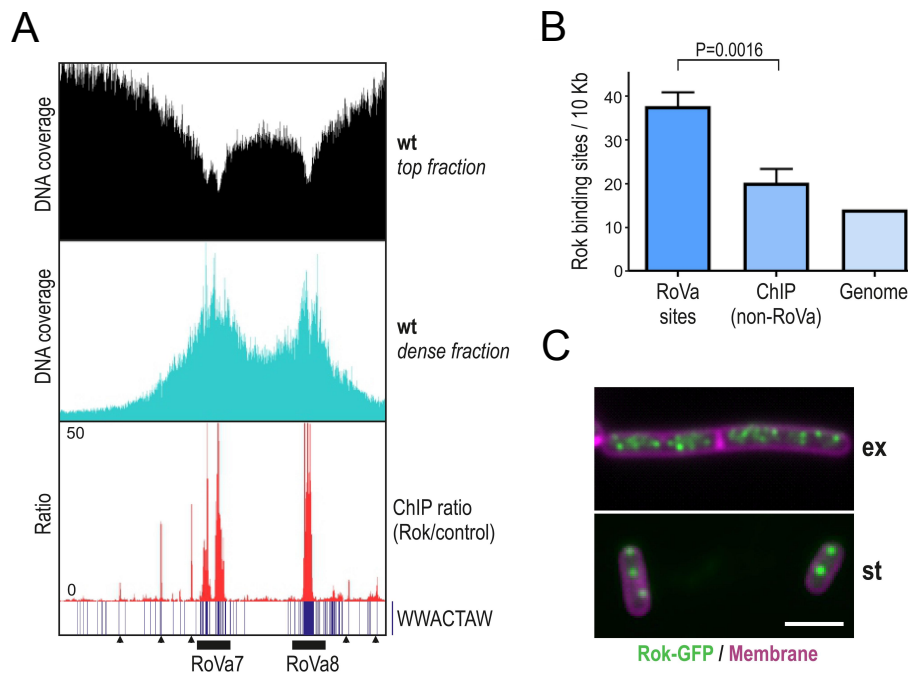


Figure 9.2: RoVa sites have higher density of Rok binding motifs and can be visualized as discrete loci. **A.** DNA coverage maps near RoVa sites 7 and 8. Individual high affinity Rok binding sites are marked using vertical lines. Other Rok enriched ChIP sites near the RoVa sites are marked with the arrow head. **B.** Quantification of high affinity Rok binding sites (WWACTAW) at the RoVa sites and non-RoVa Rok binding sites along with its average genome-wide distribution. **C.** Visualization of Rok-Gfp clusters during exponential (ex) and stationary (st) growth phases using fluorescence microscopy. Membrane was stained using Nile-red.

were also overrepresented in both RoVa and ChIP sites compared to its average abundance over the genome (Fig. 9.2b). Multiple Rok proteins can therefore associate at each RoVa site via its C-terminus DNA binding domain [187]. The N-terminus domain of Rok is also known to multimerize into a higher order oligomer [186], which may enable direct observation of RoVa complexes using fluorescent Rok-fusions. Strikingly, we observed several clusters of Rok-GFP signal spread over the bacterial cell at exponential growth phase, further supporting binding and clustering of multiple Rok proteins at each RoVa site (Fig. 9.2b). However, when cells at stationary growth phase were observed, the GFP signal from only a couple of clusters could be observed in each cell. This suggested either dissociation of Rok from some of the RoVa sites or a possibility of long-range association of distant RoVa sites (Fig. 9.2b).

To gain insight into the possible interaction between the 8 RoVa sites, we performed Hi-C [188] on WT and Δ rok strain at stationary growth phase (Fig. 9.3a). The Hi-C datasets revealed a clear secondary diagonal, representing the known juxtaposition of the two chromosome arms by SMC complexes [73,177]. Strikingly, we did observe very specific interaction between several RoVa sites (Fig. 9.3a). These specific interactions are seen as peaks of interaction in the contact matrix, and signify anchor sites for chromosome loops as routinely observed in mammalian Hi-C maps [181]. This interaction between the RoVa sites was completely absent in the Δ rok strain (Fig. 9.3a).

We also performed Hi-C at exponential and late-exponential growth phases. The contact frequency between the RoVa sites increased gradually from exponential to stationary

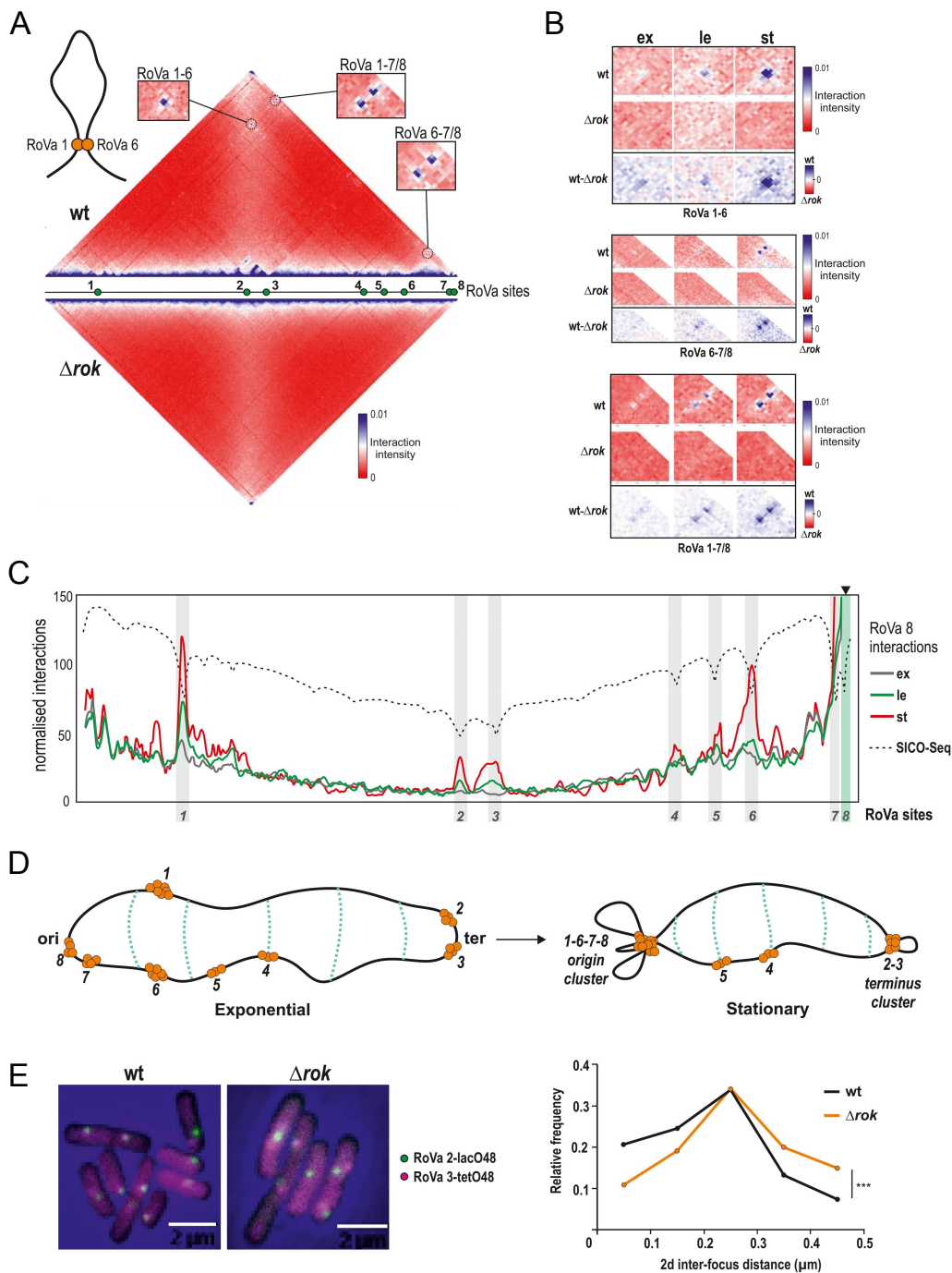


Figure 9.3: RoVa sites specifically and dynamically interact with each other. **A.** Normalized Hi-C contact maps of WT (top) and Δrok strains at stationary phase. Some of the corner peaks formed as the result of RoVa sites interaction are shown in the inset. **B.** Normalized Hi-C contact maps at RoVa interaction sites in WT and Δrok strains along with the difference plot at indicated growth phases (ex-exponential, le-late exponential, st-stationary). **C.** Virtual 4C analysis to study interactions of RoVa site 8 with the whole genome during stationary phase of wt cells. Maps are compared at different growth phases at the highlighted RoVa sites obtained using the SICO-seq coverage data. **D.** Schematic illustration of Rok dependent association of RoVa sites over growth. Green dots represent individual Rok proteins multimerized at the 8 RoVa site along the genome. Dotted light blue lines represent the SMC-mediated juxtaposition of the *B. subtilis* chromosome arms. **E.** Fluorescence microscopy based analysis of distance between RoVa sites 2 and 3 using FROS. At least 750 pairs of RoVa sites were used for the binned frequency analysis.

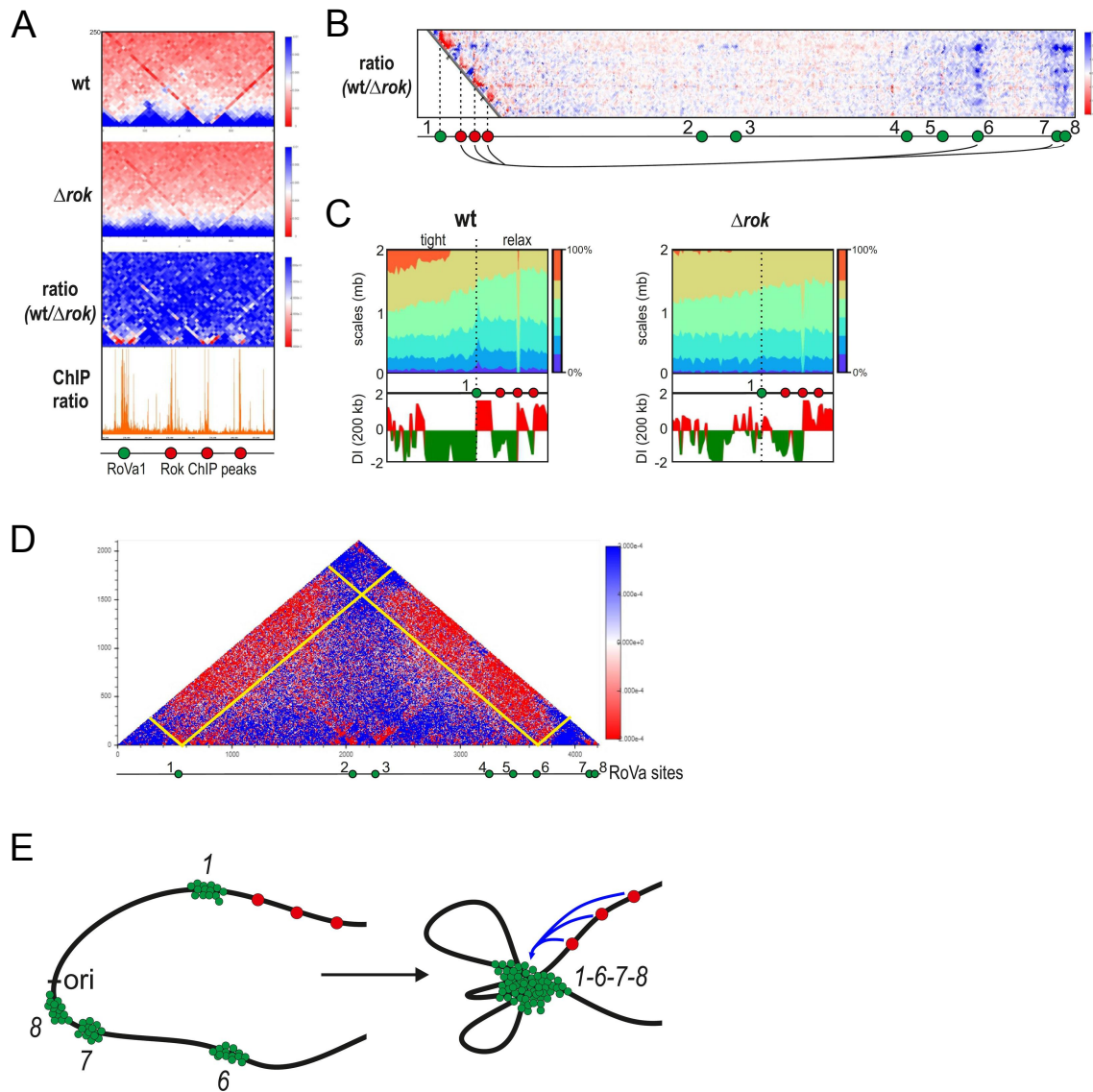


Figure 9.4: Recruitment of non-RoVa Rok ChIP sites to the origin Rok cluster. **A.** Normalized Hi-C contact maps of WT and Δrok strains near RoVa 1 along with the difference plot at stationary phase. The color scheme is changed in the difference plot to highlight bins of increased interactions (red) upon *rok* deletion. Rok ChIP data [185] (orange) is also shown along the genome below highlighting the non-RoVa Rok ChIP sites (red dots) below. **B.** Difference plot shows Rok dependent interaction of RoVa 1 and nearby non-RoVa Rok ChIP sites (red dots) with other RoVa sites. **C.** Scalogram and DI (200 kbp) analysis near the RoVA 1 shows relaxation of region between origin and RoVa 1 and changes in strength of domain boundaries upon *rok* deletion. **D.** Whole genome difference plot shows the isolation of origin Rok cluster with the rest of the genome. **E.** Illustration shows growth dependent association of RoVa sites 1,6,7 and 8 at the RoVa origin cluster and their interaction with the nearby non-RoVa Rok binding sites.

growth phases, confirming the clustering of RoVa sites over growth previously observed using fluorescence microscopy (Fig. 9.2c, 9.3b). Virtual 4C analysis with RoVa 8 site as the bait, also clearly showed a specific increase in interaction frequency with the other RoVa sites from exponential to stationary phase. Again, no changes in interaction frequency were observed in the Δ rok strain. Virtual 4C with other RoVa sites as bait also revealed that certain sets of RoVa sites showed higher interaction frequency with each other. RoVa 1-6-7-8 are relatively near the origin and interact primarily with each other to form the RoVa origin cluster, whereas RoVa 2-3 are close to the terminus and extensively interact with each other to form the RoVa terminus cluster (Fig. 9.3a, 9.3d). RoVa 4-5 on the other hand lie between the origin and terminus and shows no or relatively low interaction with other RoVa sites, respectively. RoVa 1 and 6 are more than 1 Mbp apart, but can still robustly interact with each other as part of the origin cluster (Fig. 9.3b). SMC dependent juxtaposition of the two arms in *B. subtilis* [177] may further facilitate strong interaction between such mirrored RoVa sites relative to the origin (Fig. 9.3e). RoVa sites at the origin and terminus clusters can also specifically interact with each other, albeit with much lower frequency (Fig. 9.3d). Interaction between the RoVa sites therefore depends on their relative proximity in space and the growth phase (Fig. 9.3e). Finally, we noticed that a few non-RoVa Rok-binding sites were also recruited to their nearby RoVa cluster (origin or terminus) at stationary growth phase (Fig. 9.4).

Hi-C was performed on a population of cells, and as such helps define the average chromosome interaction. To confirm that the RoVa sites indeed interact in single cells, we marked the DNA near RoVa sites 2 and 3 using two fluorescent repressor operator systems (FROS) arrays, and visualized them individually using LacI and TetR fused to YFP and CFP, respectively. The distance between RoVa 2 and 3 was found to be significantly larger in Δ rok cells compared to the WT in stationary phase, further validating the results obtained from Hi-C (Fig. 9.3b). Of note, the variation in inter-RoVa distance among cells also indicated that RoVa interactions are heterogeneous and possibly dynamic.

Next, we wanted to examine if long-range interaction between RoVa sites can also affect topological domain formation. A directionality index (DI) analysis along the *B. subtilis* genome revealed similar CID boundaries in exponential phase of WT and Δ rok strains. However, domain boundaries were rather different between the two strains at the stationary phase. Interestingly, a domain boundary was observed exactly at RoVa site 3 in stationary phase, which was found to be absent in the Δ rok strain (Fig. 9.5a). In WT strain, RoVa 3 primarily interacted with the upstream RoVa 2 site as part of the terminus cluster. However, upon Rok deletion RoVa 3 was free to interact with downstream chromosomal region leading to changes in interaction directionality and thereby disrupting the domain boundary at RoVa 3. A similar disruption in domain boundary was also observed at RoVa 1 from the origin cluster (RoVa 1-6-7-8) (Fig. 9.4c).

Hi-C also revealed that RoVa interaction within or near the existing CIDs can also lead to increase in intra-CID interaction frequency. For example, interaction and loop formation by interaction of RoVa 6 with RoVa 7/8 can lead to further compaction of the CIDs between them (Fig. 9.5a). Scalogram analysis [47] also revealed the Rok induced compaction in this region (Fig. 9.4, 9.5a). This resembles the changes in intra-TAD interaction frequency previously observed for the mammalian insulator protein CTCF [115].

Next we wondered whether direct long-range interaction between the RoVa sites changes the local short range contacts, which are observed as the primary (horizontal) diagonal in the Hi-C maps. Analysis of the changes in short-range (10-50 kbp) contact frequencies over the genome clearly revealed an increase in short-range interactions at the RoVa sites

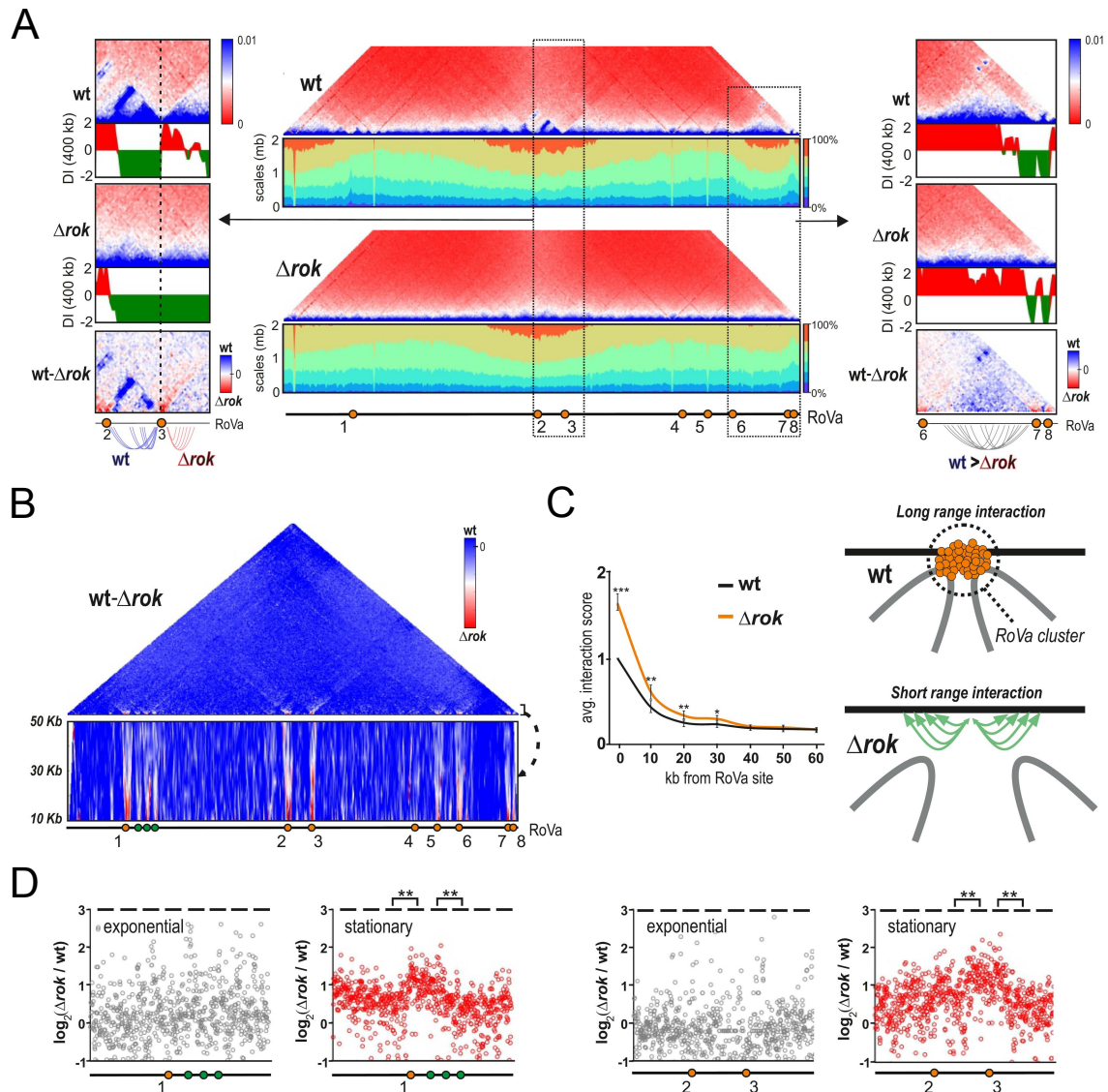


Figure 9.5: RoVa interactions impacts both global and local chromosomal architecture. **A.** Normalized Hi-C contact maps of WT and Δrok strains at stationary phase along with the scalogram representation. Scalogram displays the relative tightness along genome by plotting the percentage of the total contacts made by each bin with the neighbouring regions of increasing size. The cumulative Hi-C signal increases from dark blue (0 – 15%) to red (75% and above) and hence display tighter regions with smaller blue and larger red bars and vice-versa (center). Normalized Hi-C contact maps along with DI analysis (400 kbp) shows higher-order domain boundary formation at RoVa site 3 (left) and increased intra-domain contact frequency between RoVa sites 6 and 7/8 (right). **B.** Ratio plot of WT and Δrok strains at stationary phase, the color scheme is changed to precisely highlight bins of increased interactions (red) upon rok deletion (left). The magnified view of short-range contacts between 10 kbp and 50 kbp along with the RoVa sites (green dots) and other interacting Rok ChIP sites (red dots) is shown below. Illustration shows how elimination of long-range interactions between RoVa sites upon rok deletion increases the local short range interaction at these sites. **C.** Changes in gene expression near RoVa sites 1 and 2/3 upon rok deletion is visualized by plotting the changes in RNA abundance of 800 neighbouring genes during ex (gray) and st (red) phase. Statistical analysis on changes in transcription were determined by comparing a block of 100 genes to the adjacent blocks ($*p < 0.001$, $**p < 0.0001$).

upon Rok deletion (Fig. 9.5b). Interestingly, no changes were observed for short-range interactions at RoVa site 4, which does not participate in inter-RoVa interactions. Finally, we also found increased short range interactions at 3 non-RoVa Rok binding sites near RoVa1, all of which interact with the origin Rok cluster (RoVa1-6-7-8) at stationary growth phase. This shows that long-range interaction between RoVa sites restricts their interaction with neighbouring regions (Fig. 9.5c).

It is known that transcription levels correlates strongly with short-range interaction frequency in various bacterial species including *B. subtilis* [47]. To test whether changes in short-range contact frequency at the RoVa sites can also affect transcription, we performed RNA-seq analysis at exponential and stationary growth phases of WT and Δ rok strains. Expression of a few genes under direct repression by Rok, were upregulated upon Rok deletion in exponential phase. However, in stationary phase, hundreds of genes (within multiple operons) in and around RoVa 1 and 3 were upregulated upon Rok deletion (Fig. 9.5c). Genomic rearrangement caused by disruption in CID boundaries combined with increase in local short-range interaction at RoVa 1 and 3 upon Rok deletion are the likely factors driving changes in transcription.

9.2 Discussion

Like CTCF, Rok was also first identified as a transcription repressor of competence in *B. subtilis* [189]. Rok was later found to also regulate cell surface genes, biofilm and mobile genetic elements [179, 187, 190]. Rok is also known to activate transcription of certain genes [190], however the mechanism remains unknown. The binding specificity of Rok differentiates it from other bacterial xenogeneic silencers like H-NS and Lsr2 [186, 187]. In this study, we draw significant parallels between the known functions of eukaryotic insulator proteins and Rok. Multiple Rok protein can associate with specific regions in the genome and facilitate their interaction over large distance. Such long-range interaction can lead to formation of new domain boundaries and alter intra-domain [47, 73]. Finally, we also provide evidence of how genome rearrangement by Rok can alter the local chromosomal interactions at the RoVa sites and potentially influence gene expression, independent from its role as a transcription repressor. It is possible that other accessory proteins are involved in RoVa complex formation and their interaction.

Currently, the role of specific long-range interactions on bacterial physiology is unclear. The changing landscape of inter-RoVa interaction and loop formation over growth may link genome organization to DNA replication and segregation. Interestingly, Rok is known to associate with DnaA, initiator of chromosomal replication, when bound to DNA [185]. This interaction may link RoVa sites interactions to active replication in *B. subtilis*. Other bacterial transcription factors which can oligomerize, like GalR in *Escherichia coli*, could also participate in insulator domain formation [191]. It remains to be seen if insulator dependent long-range promoter-enhancer interaction exist in bacteria and if this can also regulate gene expression as observed in eukaryotes [184]. Differential RoVa interaction between individual cells may also create CID heterogeneity leading to altered cell fate [59, 184]. Future developments to study single cell chromosome structure [59, 183] along with the transcriptome [192] in bacteria will help provide specific answers to these questions.

9.3 Methods

9.3.1 Cell Growth

B. subtilis strains were grown LB agar plates supplemented with appropriate antibiotics: spectinomycin (150 $\mu\text{g}/\text{ml}$), erythromycin (2 $\mu\text{g}/\text{ml}$). For liquid culture, *B. subtilis* was inoculated at OD_{600} 0.005 from an overnight culture (in LB) and grown at 37°C in LB or minimal media.

9.3.2 Strain Construction

All strains were constructed using transformation and homologous recombination of an overlap PCR product. The overlap PCR product contained the antibiotic resistance gene and the given insertion or mutation between ~ 1000 bp of homologous region on either side.

9.3.3 SICO-seq

Top Fraction

B. subtilis was grown to exponential phase (OD_{600} of 0.5 – 0.8) in 200 ml of LB media. All the cells were harvested by centrifugation at $10000\times g$ for 5 minutes and the pellet was immediately frozen in liquid nitrogen. The pellets were stored at -80°C until use. The pellet was transferred to 20 or 50 ml stainless steel canister (Retsch) pre-cooled in liquid nitrogen and containing 1 ml 1 \times PBS (with protease inhibitor). The pellet was cryogenically broken using 5 rounds of disruption in TissueLyser II (20 Hz for 2 min. each round). The canister was cooled in liquid nitrogen after each round. The pulverized sample was retained from the canister and 1ml ice cold 1 \times PBS (with protease inhibitor) was added to the sample. 5 μl of 10X Fragmentase Reaction Buffer v2 and dsDNA fragmentase (NEB #M0348S) each was added to the samples and mixed by short vortexing. The samples were incubated at 30°C for 15 min to partially fragment the DNA. 50 μl of the digested lysate was collected directly for DNA extraction. The rest of the sample was then added to the top of a two-step sucrose density layers (20% and 60%) and ultra-centrifuged for 2 hrs at 30000 rpm (SW41 rotor, Beckman) and 4°C. The top fraction (100 μl) was collected for DNA extraction.

Dense Fraction

B. subtilis was grown to exponential phase (OD_{600} of 0.5 – 0.8) in 200 ml LB media and treated with rifampicin (100 μM final) for 10 minutes while shaking to abort transcription. Treatment with Rifampicin removes transcription dependent protein-DNA complexes. Cells were then fixed using 0.1% formaldehyde (final) for 10 min at room temperature and subsequently quenched with glycine. Cell pellets were stored and processed as before but this time the fraction (750 μl) at the interphase of 20% and 60% sucrose density layers was collected using a syringe introduced by puncturing the side of the tube. This fraction was diluted to 10% sucrose density using ice cold 1 \times PBS (with protease inhibitor) and again loaded on top of a two-step sucrose density layers (20% and 60%) and ultra-centrifuged for 2 hrs at 30000 rpm (SW41 rotor, Beckman) and 4°C. The dense fraction (500 μl) at the interphase of 20% and 6% sucrose density layers was collected using a syringe.

Nucleic acids were extracted from the fractions collected above (top and dense) using phenol/chloroform/isoamyl alcohol (PCI-Carl Roth #A156.3) and the nucleic acid was

concentrated using ethanol precipitation. Nucleic acid was resuspended in water and the RNA was removed using RNaseA. The DNA was again extracted and concentrated using PCI (Carl Roth #A156.3) and ethanol, respectively. The partially fragmented and purified DNA was directly used for library preparation using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB #E7645S) as per manufacturer’s instruction and subsequently subjected to sequencing using Illumina NextSeq 550.

9.3.4 RNA Isolation and RNA-seq

Total RNA was extracted from *B. subtilis* cells using hot-phenol method as previously described [193], except 5 mg/ml lysozyme was used to disrupt the *B. subtilis* cell wall. Residual DNA was removed from the total RNA using DNaseI (NEB #M0303S) digestion as per manufacturer’s instruction. Total RNA was checked on agarose gel to verify the integrity of the RNA samples by visualizing the rRNA bands. rRNA was then removed from 10 μ g of total RNA using MICROBExpress™ Bacterial mRNA Enrichment Kit (ThermoFisher #AM1905) as per manufacturer’s instruction. The RNA was extracted using PCI for RNA (Carl Roth #X985.3) and precipitated using ethanol. 100 ng of the mRNA enriched RNA was used for library preparation using NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB #E7760S) and sequenced using Illumina NextSeq 550.

9.3.5 Mapping and Visualization of SICO-seq, RNA-seq and ChIP Data

SICO-seq and RNA-seq data used in this study were generated using the experiments described above. Raw Rok ChIP data generated by Seid et. al. [185] was obtained from NCBI. All the sequencing data was processed using the open source web-based platform – Galaxy (usegalaxy.org). The quality of each dataset (FASTQ files) was firstly assessed using fastQC. All files were then trimmed using Trimmomatic (Galaxy Version 0.36.5) before mapping them to the *B. subtilis* subsp. *subtilis* str. 168 reference genome (NC_000964.3) using Bowtie2 (Galaxy version 2.3.4.2). The BAM files were converted to bigwig using the tool bamCoverage (Galaxy version 3.0.2.0) and the coverage maps were visualized using Integrated Genome Browser (IGB version 9.0.2). The RNA-seq BAM files were also used as the input for featureCounts (Galaxy Version 1.6.3) along with the respective gff3 file to quantify and compare gene expression. Normalized coverage files were used to generate SICO-seq and ChIP ratio plots directly using IGB.

9.3.6 Motif Density Analysis

The 10 kbp sequence around each RoVa site (5 kbp upstream and downstream of RoVa minima) and 8 highly enriched non-RoVa Rok ChIP sites (5 kbp from each side of ChIP peak) were isolated and quantified for presence of high affinity Rok binding motifs – “WWACTAW” identified previously [186]. The distribution of the motif along the genome was directly visualized in IGB. *B. subtilis* subsp. *subtilis* str. 168 reference genome (NC_000964.3) was used to quantify the average abundance of the motif.

9.3.7 Chromosome Capture by Hi-C

Hi-C was carried out essentially as previously described [188] with minor modifications. 2 – 5 ml culture was sequentially fixed using 80% methanol and 3% formaldehyde. Cells

were washed with ice-cold $1\times$ PBS after each step. The cells were then harvested by centrifugation and the pellets were flash frozen in liquid nitrogen. Cells were lysed using Ready-Lyse Lysozyme (Epicentre #R1802M) in $1\times$ TE buffer followed by 0.5% SDS treatment. The chromosomal DNA in the cell lysate was then digested using HindIII for 3 hrs at 37°C . The restriction ends were filled with Biotin-14-dATP, dGTP, dCTP, dTTP using DNA Polymerase I, Large (Klenow) fragment (NEB #M0210S). The chromatin was fractionated by centrifugation and subsequently the pellet was ligated using T4 DNA ligase (NEB #M0202M) overnight at 16°C . Samples were then treated with RNase A and de-crosslinked at 65°C for 6 hrs in the presence of proteinase K. The DNA was then extracted using PCI (Carl Roth #A156.3) and precipitated using ethanol. Biotin was removed from the non-ligated ends using T4 polymerase (M0203S) in the presence of dATP. The DNA was then again extracted and precipitated as before, and then fragmented using dsDNA fragmentase (NEB #M0348S) treatment for 15 min at 37°C . The fragmented DNA was used for library preparation using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB #E7645S) as per manufacturer’s instruction until adapter ligation and purification using AMPure XP beads (Beckman coulter #A63881). Biotinylated library fragments were extracted from the sample using $20\ \mu\text{l}$ of Dynabeads® MyOne™ Streptavidin T1 beads (ThermoFisher #65601) as per manufacturer’s instruction. The washed beads (with biotinylated DNA) were used for PCR library amplification (12 – 14 cycles) using NEBNext Ultra II Q5® Master Mix (NEB #M0544S). The amplified library was purified using AMPure XP beads followed by paired-end sequencing using Illumina NextSeq 550.

9.3.8 Hi-C Data Mapping and Contact Matrix

Hi-C matrices were constructed using the Galaxy HiCExplorer webserver (hicexplorer.usegalaxy.eu). Briefly, paired end reads were mapped separately to the *B. subtilis* genome (NCBI Reference Sequence NC_000964.3) using the “very sensitive” local setting mode in Bowtie2. The mapped files were used to build the contact matrix using the tool hicBuildMatrix using a bin size of 10 kbp, HindIII restriction site (AAGCTT) and (AGCT) as the dangling sequence. The contact matrix (.cool format) was then used for further analysis and visualization as described below.

9.3.9 Hi-C Data Visualization

Hi-C contact maps were assessed, compared and prepared for the illustration using the interactive Browser-based visualization tool “Bekvaem” [194] which source code is published and made available to the public [105].

Comparison of Contact Maps

First, Hi-C contact frequency matrices were normalized using the Sinkhorn-Knopp (SK) balancing algorithm [81]. In a subsequent step, the normalized contact probability matrices were compared via their difference, i.e. two contact probability matrices $A \equiv (a_{ij})$ and $B \equiv (b_{ij})$ were compared via their difference $D \equiv (d_{ij})$ by (element-wise) matrix subtraction $d_{ij} = a_{ij} - b_{ij}$. In order to be able to detect also only very small local differences, the (element-wise) logarithmic ratio was computed as $\tilde{R} \equiv \tilde{r}_{ij} = \log_2 |a_{ij} / b_{ij}|$.

Domain Detection

Domain boundaries within Hi-C contact maps were detected algorithmically using an in-house Python implementation based on the directionality index approach presented by Dixon et al. [7]. The method is motivated by the observation that domains are demarcated by regions that are biased in their interaction probability; the upstream domain boundary is preferentially interacting downstream whilst the downstream boundary is preferentially interacting upstream. This method is based on a two-step strategy. First, the 2D contact information is translated into the directionality index encoding the ratio of downstream and upstream interactions. Next, downstream interactions are compared to upstream interactions in order to derive whether the strength of interactions are significantly stronger in one direction compared to the other. Domain boundaries correspond to positions where this preferred direction of interactions abruptly changes [48].

Besides the directionality index, we also computed domains using TopDom [195], an optimal polygon algorithm that detects topological domains in a linear time. TopDom deterministically identifies domains, along with a set of statistical methods for evaluating their quality and depends on just one intuitive parameter, a window size. We used the provided TopDom implementation in R and visualized the results similar to TADbit [110] using in-house Python scripts. Detected domains are represented by gray-filled arcs and delimited by colored border symbols. The height of the depicted domains is proportional to the relative number of interactions within this domain given its size. The color code from blue to red, numbered 1-10, indicates the confidence with which the domain was identified. The y -axis displays the relative Hi-C interaction frequencies and the horizontal line at $y = 1$ indicates the expected frequency given the domain size. If the Hi-C relative interaction frequency inside the domain is higher than expected according to its size then the domain is colored in dark gray.

Scalograms

Scalograms visualizing the dispersion of the contact probability signal along the spatial scales were implemented in Python by following the description of Liroy et al. [47] on [GitHub](#). Scalograms reflect the constraints exerting on a genomic region, by revealing to which extent they “see” their flanking sequences. For each genomic position, the cumulated amount of contacts between its position and an increasing number of flanking bins is computed. As contact maps are normalized, the maximum cumulated contact signal is equal to one. Subsequently the resulting heat map using a contour line function is plotted. Within this representation, the resulting signal is divided into 5 areas, each representing 15% of the total contacts except the last one with 25% of contacts.

This visualization tool can be understood as an extension of the genomic distance law which calculates the average intra-chromosome contact probability $P(s)$ for pairs of loci separated by a genomic distance s . This contact probability decreases as a power law and can be linked to the polymeric nature of chromosomes. Scalograms allow to locally illustrate this behavior.

Virtual 4C analysis

The bait region (10 kbp bin) was used as a input for the tool `hicPlotDistVsCounts` in the Galaxy HiCEXplorer web server. The total contacts of the input bait region with all bins were quantified for a Hi-C matrix and normalized for comparison between different conditions (genotype and growth phase).

Chapter 10

Self-organized Segregation of Bacterial Chromosomal Origins

References

The results presented in this chapter are published as and adapted from

- A. Hofmann*, J. Mäkelä*, D.J. Sherratt, D.W. Heermann and S.M. Murray (2019), *Self-organized segregation of bacterial chromosomal origins*. eLife, 8: e46564. doi: 10.7554/eLife.46564.001.

AH performed the polymer simulations, SMM performed the stochastic simulations and JM conducted the experiments. We thank Remy Colin for discussions and Victor Sourjik for discussion, support and comments on the manuscript. We also thank Nathan Kuwada and Paul Wiggins for providing the raw data of their previous work.

*equal contribution

Chapter Summary

The chromosomal replication origin region (ori) of characterized bacteria is dynamically positioned throughout the cell cycle. In slowly growing *Escherichia coli*, ori is maintained at mid-cell from birth until its replication, after which newly replicated sister oris move to opposite quarter positions. Here, we provide an explanation for ori positioning based on the self-organization of the Structural Maintenance of Chromosomes complex, MukBEF, which forms dynamically positioned clusters on the chromosome. We propose that a non-trivial feedback between the self-organizing gradient of MukBEF complexes and the oris leads to accurate ori positioning. We find excellent agreement with quantitative experimental measurements and confirm key predictions. Specifically, we show that oris exhibit biased motion towards MukBEF clusters, rather than mid-cell. Our findings suggest that MukBEF and oris act together as a self-organizing system in chromosome organization-segregation and introduces protein self-organization as an important consideration for future studies of chromosome dynamics.

10.1 Introduction

The faithful and timely segregation of genetic material is essential for all cellular life. In eukaryotes the responsibility for chromosome segregation lies with a well-understood macromolecular machine, the mitotic spindle. In contrast, the mechanisms underlying bacterial chromosome segregation are much less understood mechanistically, but are just as critical for cellular proliferation [9]. The starting point for bidirectional chromosomal replication, the origin (ori), has a crucial role in chromosome organization and segregation. Not only is it duplicated and segregated first but its dynamic genomic position defines the position of other chromosomal regions with respect to the cell [10, 11].

In new-born *Escherichia coli* cells growing under relatively slow growing conditions in which initiation of replication and its completion occur within a single cell generation, the “home” position of the origin (henceforth and in the model, ori) is at mid-cell [196–198]. After replication, and consequent 10–15 min of “cohesion, arising at least in part from interlinking of the two daughter chromosomes (precatenation) [14, 199–203], duplicated origins migrate rapidly to opposite quarter positions, which become the new home positions for the remainder of the cell cycle [128, 204]. Other genomic loci migrate sequentially with similar dynamics [128].

The mechanisms that underlie ori positioning and direct newly replicated sisters to opposite cell halves remain unclear [9]. This is particularly the case in *E. coli* and its relatives, which do not carry ParABS systems that facilitate the segregation of low copy plasmids and some other bacterial chromosomes [12]. However, MukBEF, a functional homolog of ubiquitous Structural Maintenance of Chromosomes (SMC) complexes [205, 206], plays a role in *E. coli* chromosome organization-segregation. One of its functions is to recruit the type II topoisomerase Topo IV [13, 14], which is required for the timely removal of catenanes from newly replicated sister chromosomes [203]. Under slow growth conditions, MukBEF forms a small number of dynamic clusters (visualized as fluorescent foci) located at the middle or quarter positions [207, 208], in close association with ori [209], and the

splitting and movement of these foci occurs concurrently with the segregation of ori to the quarter positions [14,209]. Foci consist of on average 16 dimeric slowly-diffusing MukBEF complexes [210]. The colocalization with ori is not required for either MukBEF foci formation or positioning: depletion of Topo IV results in cells with multiple catenated ori forming a single focus at mid-cell but with multiple MukBEF clusters positioned throughout the nucleoid [14]. Thus MukBEF clusters are not necessarily assembled at or bound to ori, consistent with the lack of any sequence specificity [201]. Furthermore, restoration of Topo IV activity leads to the decatenated ori moving to the MukBEF clusters suggesting that MukBEF recruits or positions ori [14]. Consistent with this hypothesis, depletion of functional MukBEF results in ori mis-positioning that is subsequently restored upon repletion [211].

If MukBEF clusters position ori, what positions MukBEF clusters? Given that molecules in the clusters turnover continuously with a timescale of about one minute (Badrinarayanan et al., 2012b) and that MukBEF binds DNA non-specifically, how does it even form clusters? We have proposed that a self-positioning stochastic Turing pattern can explain the positioning of MukBEF clusters [212] (see Fig. 10.2, Fig. 10.3 and the methods section for a review). A Turing pattern is a spatial pattern in the concentration of a reactant in a reaction-diffusion system that arises spontaneously due to a diffusion-driven instability [213,214]. Put simply, diffusion, rather than having a homogenizing effect can actually, in combination with chemical reactions, create a spatially varying concentration profile. Such patterns are examples of self-organization, a more general term that describes any dissipative non-equilibrium energy-dependent order that arises as a result of collective non-linear interactions [215]. We used the Turing mechanism to explain the positioning of MukBEF foci and showed that a flux-balance mechanism and stochasticity work together to ensure that a specific Turing pattern is selected: short cells consistently have a single center-positioned peak in the MukBEF concentration, while longer cells have quarter positioned peaks.

With this model in hand, we now investigate how MukBEF clusters could position chromosomal origins. In particular, we address whether the self-organizing MukBEF gradient proposed in our model has the correct properties to act as an attracting gradient for ori. Additionally, it is critical that each newly replicated sister ori is recruited to a different MukBEF focus, a non-trivial requirement. We find that a self-organizing MukBEF gradient can indeed accurately reproduce the observed ori dynamics, apparent diffusion constant and drift rate. A proposed preferential loading of MukBEF within ori introduces a non-trivial interaction between MukBEF foci and oris that leads to accurate and stable partitioning as an emergent property of the system. Importantly, the model does not contain any actual directed force. MukBEF requires energy in the form of ATP to establish a self-organized gradient but it is not pulled to the middle or quarter positions by any active force. Similarly, the attraction of ori up the MukBEF gradient may be due to energetic considerations and the elastic nature of the chromosome (a DNA-relay) resulting on the macro scale in an effective (rectification) force and directed motion.

10.2 Results

10.2.1 ori is attracted towards MukBEF foci

As discussed above, perturbative experiments support the hypothesis that MukBEF clusters position oris in *E. coli* [14,211]. However, it is unclear if this hypothesis is supported by the observed colocalization of MukBEF clusters with oris in unperturbed cells. It is

possible that MukBEF clusters and *ori* could be positioned independently of one another as a result of the global organization of the chromosome with the result that they show colocalization but without their positions being correlated. To examine this possibility, we revisited the colocalization of MukBEF clusters and *oris*. We used only cells with a single *ori* focus, because, unlike cells with two *oris*, they can be grouped together without scaling by simply aligning them according to their mid-cell positions and are more amenable to statistical analysis. To enrich for such cells, we treated a strain carrying fluorescently

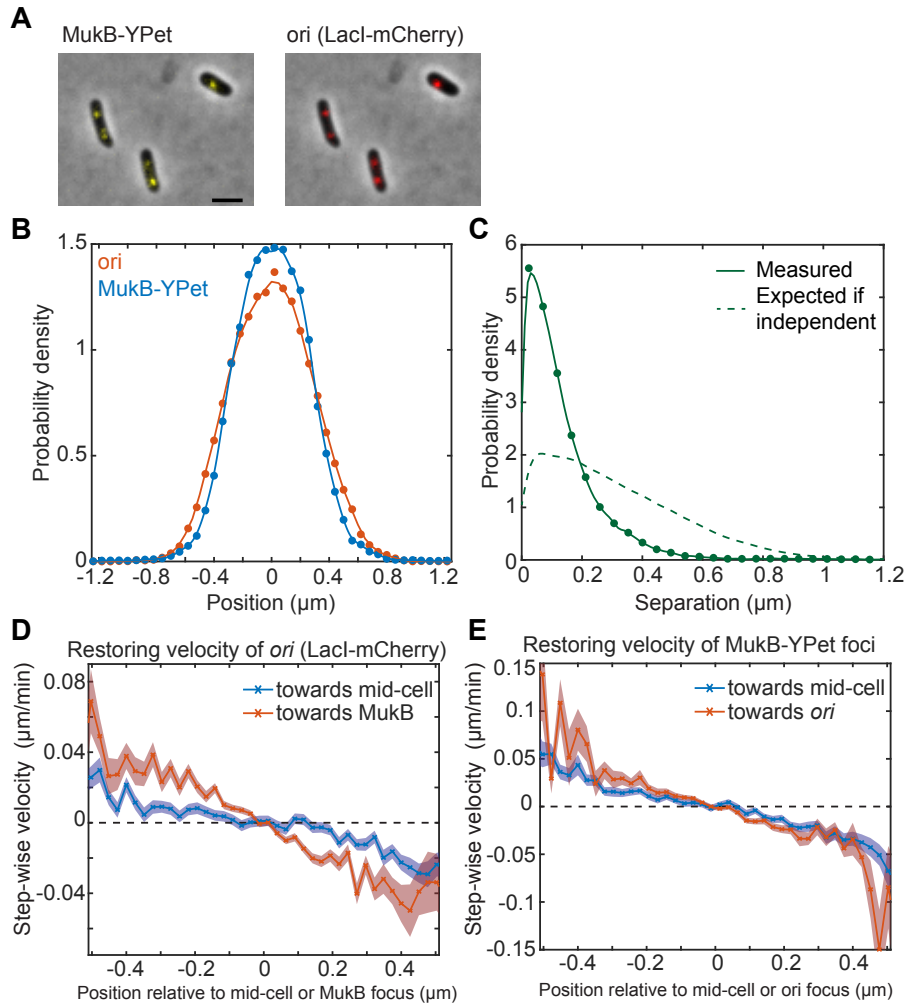


Figure 10.1: Fluorescence microscopy indicates that *ori* and MukBEF are not positioned independently of one another. A strain with FROS (LacI-mCherry) labeled *ori* and MukB-mYPet was treated with DL serine hydroxamate to obtain cells with a single non-replicating chromosome and imaged at 1 min intervals. **A.** Overlay of phase contrast and fluorescence images showing three representative cells. Bar indicates $2 \mu\text{m}$. **B.** The position distribution (along the long axis of the cell) of fluorescent foci of *ori* (red) and MukB-YPet (blue). $N = 31820$ from 952 cells tracked over up to 56 frames. Cells have a mean length of $2.2 \mu\text{m}$. **C.** The expected distribution (dashed line) of the distance between *ori* and MukB-YPet foci given that the distributions in **B** are independent. The measured distribution (circles and solid line) of separation distances from the same cells. **D.** The step-wise velocity of *ori* as a function of position relative to mid-cell (blue) and to the MukB-YPet focus (red). **E.** The step-wise velocity of MukB as a function of position relative to mid-cell (blue) and the *ori* (red). Shaded regions indicate standard error.

labeled MukB and ori (Fig. 10.1A) with DL serine hydroxamate (SHX). This structural analogue of serine triggers the stringent response thereby inhibiting DNA replication initiation [216]. We then measured the position of fluorescent foci along the long axis of the cell as has been done previously [14, 201] and found very similar distributions for ori and MukB as expected (Fig. 10.1B).

To investigate if MukBEF and ori are positioned independently of one another, we next compared the distribution of the measured distance between them to the distribution that would be expected if they were positioned independently (the null hypothesis). This latter distribution is obtained by randomly selecting pairs of positions from the two measured position distributions and calculating the distance between them. When we did this, we found that MukBEF foci and ori are much more colocalized than would be expected if they were positioned independently (Fig. 10.1C). This was confirmed by the relatively strong positive correlation ($r=0.8$) between MukBEF and ori positions (using a robust correlation based on the median absolute deviation [217]). Importantly, the result was not due to treatment with SHX.

If ori is indeed positioned by MukBEF, then we should be able to detect this in wild-type cells. In particular, we can measure the step-wise velocity of ori as a function of its position along the long axis of the cell. This “restoring” velocity characterizes the restoring force pulling ori back towards mid-cell. We can similarly determine the restoring velocity of ori towards MukBEF foci by measuring the step-wise velocity of ori as a function of position relative to the MukBEF focus. Comparing these two profiles, we found that ori experiences a greater restoring velocity towards the MukBEF focus than towards mid-cell (Fig. 10.1D). This indicates that ori is not attracted to mid-cell per se, rather it is more likely attracted to the MukBEF focus, which happens to be positioned at mid-cell. Hence, together with previous results, these data strongly indicate that MukBEF positions ori in *E. coli*.

We next asked whether the relationship is bi-directional i.e. is MukBEF positioning affected by ori positioning? When we examined the restoring velocity of MukBEF foci, we found that they displayed similar biases towards ori and mid-cell (Fig. 10.1E), suggesting that MukBEF foci are equally attracted to mid-cell and ori and that therefore the attraction between MukBEF and ori may be indeed be bi-directional. We will return to this result later.

The above results also confirm that ori has a special relationship with MukBEF compared to other genetic loci. This is supported by the observation that co-localization with MukBEF is strongest for ori and becomes progressively weaker for ori-distant loci [209]. What is the nature of this relationship? This has been an open question for many years, despite the application of tools such as chromatin immunoprecipitation [201] and single-molecule microscopy [210, 218]. In the following, we take an abductive reasoning approach common in theoretical physics. We make a starting assumption or *ansatz* for the nature of the MukBEF-ori relationship in order to build a computational model of the system and take confirmation of model predictions as evidence in support of this *ansatz*. In particular, we propose that MukBEF is preferentially loaded onto the chromosome within the ori region. This is the case for SMC, a distant relative of MukBEF (see [219] for a review). In *Bacillus subtilis* and other bacteria SMC is loaded onto the chromosome at *parS* sites by the protein ParB. While no analogue of ParB/*parS* has yet been discovered in *E. coli*, in the following we will focus on exploring the effect of preferential loading and examining whether it gives results that are consistent with experimental observations. We discuss other plausible scenarios in the discussion.

10.2.2 Model of ori positioning by self-organised MukBEF reproduces mid-cell positioning

As a first step in building a model of ori positioning, we incorporated the ori into our previous stochastic model of MukBEF self-organization and positioning [212], reviewed in the methods section and illustrated in Fig. 10.2A. Briefly, MukBEF exists in three states corresponding to different conformations and associations with DNA, a well-mixed cytosolic fraction and two DNA-associated states. The differing diffusion constants and nonlinear interaction between the two latter states leads to the spontaneous formation of dynamic MukBEF foci (to be understood as regions of high density) via the Turing mechanism. The positions of these foci are determined by the balancing of fluxes originating from the well-mixed cytosolic state (Fig. 10.2B). The flux of molecules (number per second) reaching the MukBEF focus is proportional to the length of the nucleoid on each side since the flux of molecules arriving from the cytosol is proportional to these lengths. Thus, if the MukBEF focus is off-center (in the case of a single focus), it experiences a differential in the incoming fluxes from either side, resulting in movement toward the equilibrium position (the center). This flux-balance mechanism was first described in the context of plasmid positioning [220–222] but is valid quite generally. Note that in this model, the chromosome is not modeled explicitly. While one could theoretically use a combined particle and polymer based approach, such simulations are not yet feasible. Rather, the action of condensins is typically implemented in polymer simulations implicitly [104, 223]. However, we are explicitly interested in the fact that MukBEF forms discrete positioned foci. We therefore take a protein-centric approach and model the chromosome implicitly but MukBEF explicitly.

We treat the ori as a diffusing particle, the movement of which is biased in the direction of increasing MukBEF concentration (Fig. 10.2C) (see methods for details). That is, the probability is higher that ori will move up the MukBEF gradient than down it. For the moment, we do not implement the effect of preferential loading at ori on MukBEF dynamics. We perform the simulations in one dimension, representing the long axis of the cell, the dimension along which positioning and segregation occur. This is justified as both ori and MukB are confined within the transverse direction to the center region of the cell: 95% of foci are within the center 40% (300 nm) of the cell width. MukBEF is also a very large molecule complex, with the arc length from its hinge domain to either of its two head domains being about 70 nm [224]. This suggests it can be treated as operating on a coarser level than individual strands of DNA. However the primary reason is due to an inherent limitation of the Reaction Diffusion Master Equation method for non-linear reactions. Every voxel (compartment) must be well-mixed and this condition is violated by non-linear reactions at small voxel volumes [225, 226]. While this can be overcome for bimolecular reactions [227–229], there is currently no such remedy for tri-molecular reactions, as present in our (and most) Turing models. As a quantification of the resulting artefacts, we measured the mean total number of species v as a function of compartment size. While the mean number is stable in one dimension, it decreases rapidly in two and three dimensions as the compartment size is decreased. One might hope for a range of compartment sizes, small enough to realize the geometry but large enough to avoid small compartment-size effects. However, no such range exists. We will therefore confine our stochastic simulations to one dimension. However, as we shall show, this limitation does not prevent us from explaining the observed experimental behavior and making falsifiable predictions.

We take initially the case of short cells ($2.5 \mu\text{m}$) with a single ori. For the moment,

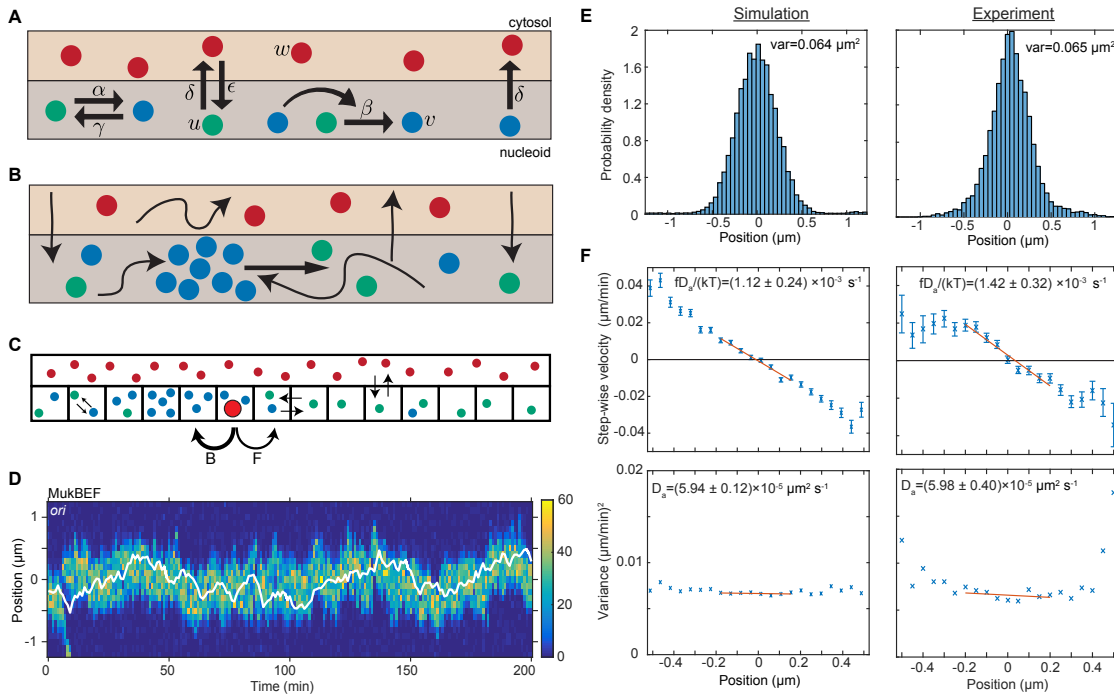


Figure 10.2: ori positioning by a self-organized protein gradient reproduces experimental results. **A.** Schematic showing the reactions of the previous MukBEF model ([212] (see methods)). Species w diffuses in the cytosol (red). Species u (green) and v (blue) diffuse on the nucleoid. Binding and species interaction are indicated by arrows. Diffusion is not shown. See the methods for a review of the model and the model parameters. **B.** Schematic showing the flux-balance mechanism. The thinner arrows represent binding/unbinding and diffusion. Species w (red) is well-mixed and therefore converts to species u (green) uniformly across the nucleoid. If a molecule of species u explores a sufficiently large region of the nucleoid before it detaches again, then the flux of u molecules reaching a high density region (focus) of species v (blue) from either side is proportional to the length of the nucleoid on either side. This difference in fluxes leads to net movement of the self-organized focus towards the position at which the fluxes balance, the mid-cell in the case of a single focus. **C.** The stochastic models is implemented using the spatial Gillespie method which discretizes the spatial dimension into compartments in which molecules react and between which molecules can diffuse. Colors label species as in **A**. The cytosolic species is taken to be well-mixed and its concentration is therefore not simulated spatially. This is the same implementation as was used previously [212]. In this work, we extend these simulations by incorporating the ori as a single diffusing particle (outlined red circle). However, unlike MukBEF its diffusion is biased, being determined by forward (F) and backward (B) jump rates that depend on the gradient of MukBEF concentration (blue circles, v) (see methods). **D.** Kymograph from a single simulation showing the number of MukBEF molecules (color scale) and the position of the ori (white line). **E F.** A comparison between the experimental data of Kuwada et al. [204] and the results of simulations in the case of a single ori and 6x preferential loading. **E.** Histograms of ori position (unscaled) along the long axis of the cell. Zero is the middle position. **F.** Mean (top) and variance (bottom) of the step-wise velocity as a function of position relative to mid-cell. Bars indicate standard error. The linear velocity profile at mid-cell is indicative of diffusion in a harmonic potential ($V(x) = \frac{1}{2}fx^2$). In such a model the variance of the step-wise velocity is independent of position. Thus we obtain the apparent diffusion constant D_a and drift rate $d_a = \frac{fD_a}{kT}$ by fitting to the central region. Bounds are 95% confidence intervals. Red lines are weighted linear fits. Simulated data are from 100 independent runs, each of 600 min duration. Experimental data are based on over 16000 data points from 377 cells. Both data sets use 1 min time-intervals. Simulations are from 2.5 μm cells, whereas experimental data is from a range of cell lengths. See methods for further details and model parameters.

we do not implement the effect of preferential loading at *ori* on MukBEF dynamics. We found, as expected, that *ori* tracks the self-organized MukBEF focus, resulting in mid-cell positioning (Fig. 10.2D and Fig. 10.3B). To more carefully examine the directed movement of *ori*, we measured the *ori* velocity as a function of position as we did previously (Fig. 10.1D). Given the self-organizing and fluctuating nature of the MukBEF gradient in our model (which is representative of the *in vivo* behavior), it was not obvious that the model would reproduce the observed relationship. However, we indeed found a similarly linear velocity profile. Furthermore, quantitative comparison of our simulations with the experimental data from Kuwada et al. [204] was carried out by fitting the data in the mid-cell region to a theoretical model of diffusion in a harmonic potential, thereby obtaining an apparent diffusion constant and drift rate. We found that by adjusting only two parameters, the *ori* diffusion constant and drift parameter, we were able to obtain good agreement with the values obtained from the experimental data. However, the resultant position distribution did exhibit somewhat fatter tails.

This fitting of the experimental data also allows us to estimate the spring-like force on the *ori*. At a distance x from mid-cell, the force is given by $F = -fx$, where f is obtained from the slope and variance of the velocity profile. At $0.2 \mu\text{m}$ from mid-cell this gives a restoring force of 0.02 pN, similar to the value measured in an *in vitro* reconstitution of a plasmid partitioning ParABS system [230]. Note that the data in Kuwada et al. are from growing cells and we use them rather than our data in Fig. 10.1 for consistency with later simulations that incorporate growth.

In the previous simulations *ori* moves up the MukBEF gradient but has no effect on the MukBEF gradient itself as we have not yet implemented that MukBEF is preferentially loaded onto the DNA within the *ori* region. In previous work, we showed that preferential loading at a fixed spatial location perturbs the positioning of the self-organized MukBEF foci due to the modified flux differential across foci [212]. In the case of a single MukBEF focus, the equilibrium position is no longer at mid-cell but somewhere between mid-cell and the location of preferential loading, depending on the strength of the loading. Thus the presence of a preferential loading site in the *ori* should lead to an effective mutual attraction between *ori* and MukBEF foci i.e. *ori* is attracted up the MukBEF gradient, while at the same time the 'home' position of the MukBEF focus is shifted towards *ori*. We expected this to increase the association between the two and reinforce mid-cell positioning.

We added preferential loading into the simulations by increasing the loading rate of MukBEF in the compartment containing the *ori* relative to the other compartments while keeping the overall loading rate unchanged. This was observed to have a suppressive effect on noise. At intermediate levels of preferential loading, the positions of both MukBEF and *ori* deviate less from mid-cell. Looking at individual simulations we could see that *ori* rarely escapes the MukBEF focus, rather the focus tracks *ori* and brings it back to the middle position. As a result *ori* only rarely undergoes diffusive excursions away from MukBEF and its home position as were observed without preferential loading. The reduction in the variance of the position distributions was reversed at higher loading ratios. We also found that preferential loading resulted in stronger colocalization of *ori* with the MukBEF focus and this led to excellent agreement with previous measurements of their separation distance. Note that this experimental data was not used to constrain the model and this agreement thus constitutes confirmation of a model prediction and support for preferential loading.

We next examined if the simulations could reproduce the observed experimental velocity profiles (Fig. 10.1D,E). We found that *ori* shows a stronger restoring velocity towards

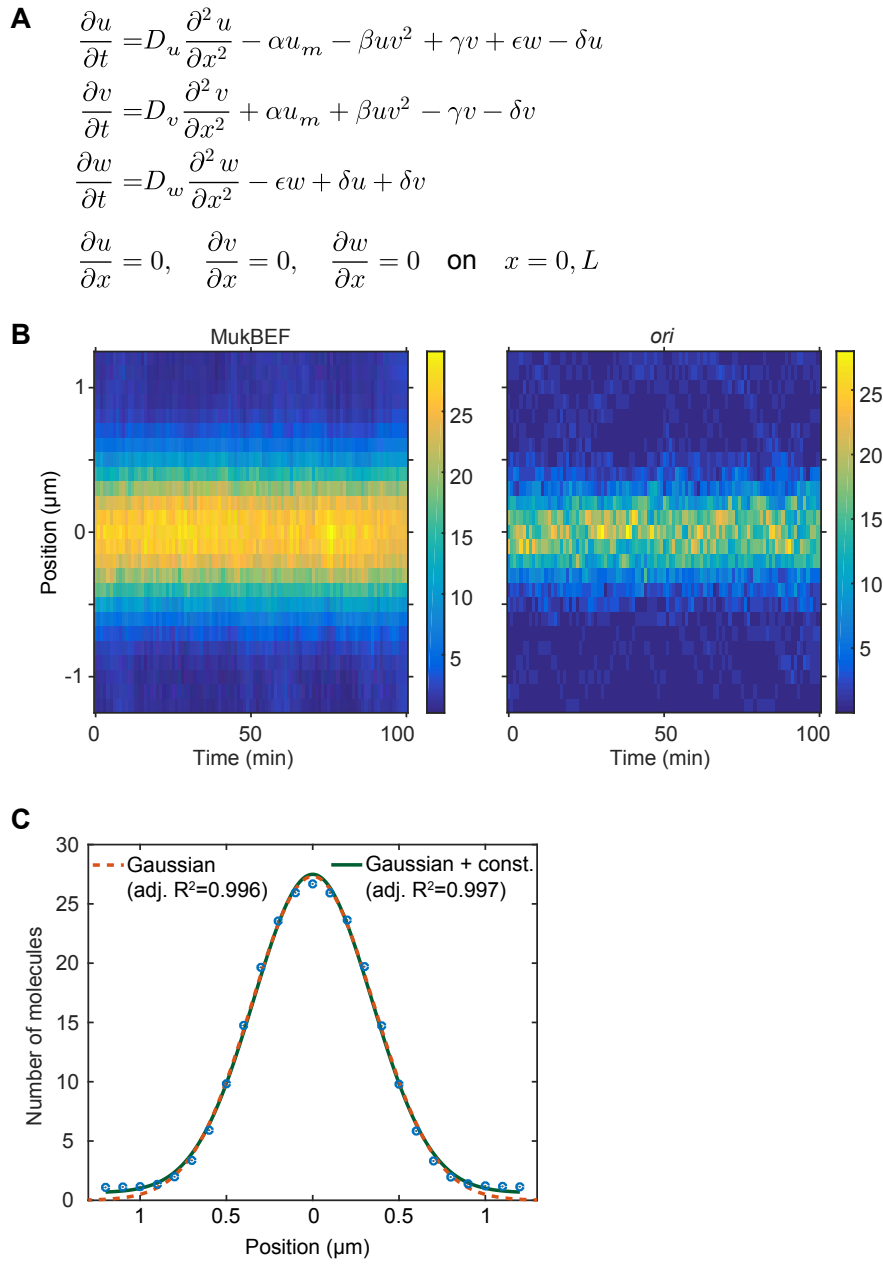


Figure 10.3: Properties of MukBEF model and *ori* positioning. **A.** The differential equations of the model in Fig. 10.2A. Given for reference and to specify the nature of the reaction terms. In this work, we exclusively use the stochastic implementation of the model (Fig. 10.2C). **B.** Average kymographs from 100 simulations showing the distributions of MukBEF and *ori* positions. **C.** The average MukBEF profile in the simulations (blue dots) is well approximated by a Gaussian.

MukBEF than mid-cell and this was independent of the level of preferential loading. The restoring velocity to mid-cell is non-zero due to the fact that MukBEF fluctuates around mid-cell. These results were to be expected as the biased motion of *ori* up the MukBEF gradient is explicitly included in the simulations. What was less clear was how the MukBEF focus would behave. We found that without preferential loading MukBEF displays a stronger restoring velocity towards mid-cell consistent with its positioning by the

flux-balance mechanism. At high loading however, it shows a stronger bias towards ori consistent with the previous observation that the attraction to ori dominates, while at intermediate levels (6x) we found a very similar restoring velocity to both ori and mid-cell as was observed in the corresponding experimental profiles (Fig. 10.1E). This is also the same level of preferential loading that leads to the most robust positioning. Thus, an intermediate level of preferential loading appears to be most consistent with experimental observations.

Given that preferential loading was found to have an effect on the apparent diffusion constant and drift rate, we needed to refit the model to the experimental ori velocity data. To do so we chose a particular value for preferential loading ratio, the one that minimized the variance, 6x (this choice will be justified in the section). We were able to find new values for the diffusion and drift parameters that lead to excellent agreement with the experimental values (Fig. 10.2F). Furthermore, the resulting distribution of the ori positions showed better agreement with the experimental distribution, with the fat tails observed in the absence of preferential loading no longer present (Fig. 10.2E).

10.2.3 Preferential loading leads to stable and accurate partitioning

While promising, the above results are not sufficient to suggest that MukBEF can explain the in vivo behavior of ori. The challenge arises after ori has been replicated. A true partitioning mechanism must ensure that each replicated ori is maintained at a different quarter position. A simple gradient based mechanism cannot, a priori, satisfy this requirement as both oris could just as easily move towards the same quarter position. Furthermore, the experimental data suggests that once oris separate they do not subsequently interchange their positions (cross paths). This ordering is essential during multi-fork replication, where the multiple oris of each segregated chromosome must be positioned to the appropriate cell half to avoid guillotining the chromosome upon cell division. To examine if the model is capable of accurate and ordered partitioning, we performed simulations with two oris in longer cells of 5 μm , in which MukBEF self-organizes into, on average, two foci, one at each quarter position. With or without preferential loading, the average profile of ori positions displayed two peaks centered on the quarter positions (Fig. 10.4D, blue line). However, we found that without preferential loading approximately half of the individual simulations have both oris near the same quarter position (Fig. 10.4A), clearly indicating that partitioning was not accurate. This was the case even though the simulations were initialized with oris at opposite quarter positions. Evidently, a model of ori simply moving up the MukBEF gradient is not sufficient to explain partitioning as the noise inherent to the system means that it can switch stochastically between partitioned and un-partitioned states.

However we found that preferential loading resulted in stable and accurate partitioning (Fig. 10.4B). A preferential loading ratio greater than six (i.e. six times more loading than elsewhere) was sufficient to ensure that one and only one ori was positioned to each quarter position and they do not interchange (Fig. 10.4C). These simulations had oris initialized at opposite quarter positions (so as to investigate the intrinsic stability of that configuration). However, the bias of the system towards the desirable quarter-positioned configuration was present when both oris were initialized at mid-cell or at random positions. While configurations with both oris associated to the same MukBEF peak occurred more frequently under these conditions, in the presence of sufficient preferential loading the system eventually and irreversibly transitions to the quarter positioned configuration.

As previously observed in simulations of short cells, we found that preferential loading

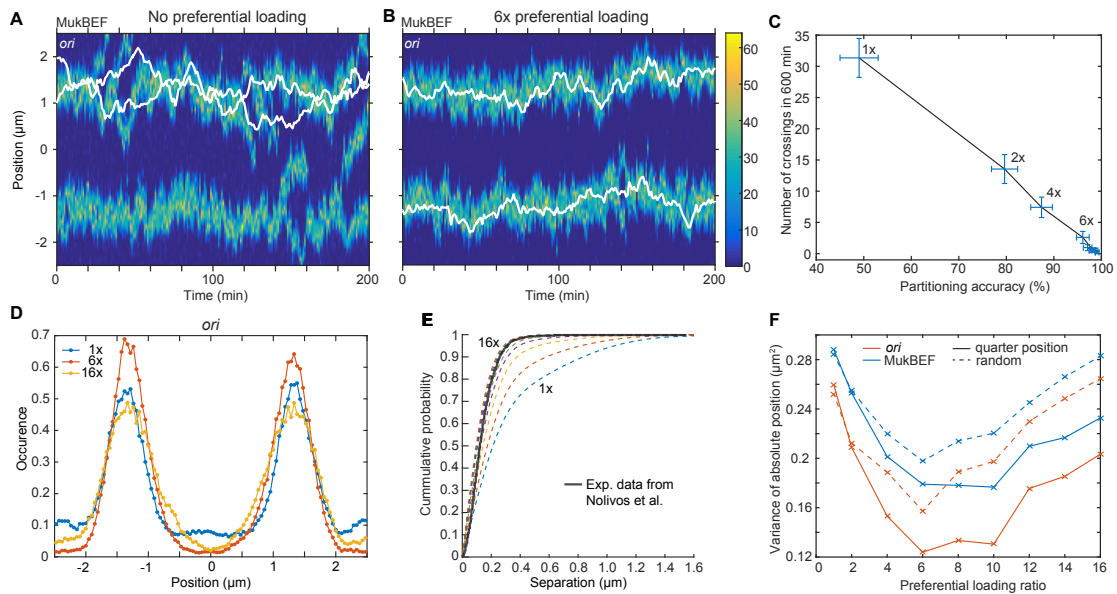


Figure 10.4: Preferential loading of MukBEF at ori leads to correct and stable partitioning. **A.** Example simulated kymograph showing two ori (white lines) diffusing around the same MukBEF peak (color scale). This occurs approximately 50% of the time. **B.** The addition of 6x preferential loading of MukBEF at ori positions results in correct partitioning of ori. The loading rate in each of the spatial compartments containing ori is six times that of the other 48 compartments. The total loading rate is unchanged. **C.** Partitioning accuracy is measured by the fraction of simulations with ori in different cell halves. Stability is measured by the number of times ori cross paths. Both partitioning accuracy and stability increase with preferential loading up to approximately 6x. Preferential loading ratios are as in **F**. Points and bars indicate mean and standard error over independent simulations. **D.** Histograms of ori positions for 1x, 6x and 16x preferential loading. Positioning is more precise at 6x than with no or 16x preferential loading. **E.** The cumulative probability distribution for the separation distance between ori and MukBEF peaks. Experimental data (black line) is from Nolivos et al. [201]. The addition of preferential loading leads to substantially better agreement. Preferential loading ratios are as in **F**. **F.** The variances of individual peaks (obtained by reflecting the data around the mid-position) have a minimum at approximately 6x preferential loading. Solid lines are from simulations with ori initially at the quarter positions, as for **A - E**. Dashed lines are from simulations with random initial ori positions. Simulations were performed for a 5 μm domain and two ori.

results in stronger colocalization of oris with MukBEF foci (Fig. 10.4E) and has a suppressive effect on noise at intermediate ratios with MukBEF foci and ori deviating less from the quarter positions (Fig. 10.4F). However, partitioning accuracy remained robust even at high preferential loading ratios (Fig. 10.4C). Looking at individual simulations, we observed that the nature of the variance was different. While the number of foci is maintained accurately at two and the foci are tightly associated to each ori (Fig. 10.4E) they are together more mobile than at intermediate ratios. Effectively, the MukBEF clusters begin to follow ori, rather than the other way around.

These results demonstrate that preferential loading of MukBEF changes the stability of the different steady states of the system. In its absence, the desirable (ori associated to opposite quarter-positioned MukBEF peaks) and undesirable (both oris associated to the same MukBEF peak) configurations have equal likelihood, as measured by partitioning accuracy (Fig. 10.4C), and the system can stochastically jump from one state to the other. As preferential loading is increased, the desirable configuration becomes more stable until

the system is found almost exclusively in that state.

10.2.4 Accurate partitioning during growth

The previous simulations were of non-growing cells and of long duration. While, they were useful to examine the intrinsic stability of the different states in order to understand why oris remain partitioned, they do not demonstrate that our model can explain how oris become partitioned within the timescale and setting of a growing cell. We therefore incorporated exponential growth and ori replication into our simulations. The former was implemented by randomly adding a new spatial compartment after every time interval corresponding to growth by one compartment length ($0.1 \mu\text{m}$). The ori was duplicated at a randomly chosen time-point obtained from an experimentally derived distribution (the mean time of duplication was 40 min into the cell cycle). After duplication, the compartment that previously contained a single ori, then contains two oris, which are free to move independently of each other (but dependent on the local MukBEF concentration).

We first examined growth in the absence of preferential loading. Similarly to what we observed previously in the simulations of non-growing cells (Fig. 10.4A), we found that duplicated oris often remained associated to the same MukBEF focus, resulting in a partitioning accuracy (defined as before as the fraction of simulations with oris in opposite cell halves) of only 25% by the end of the cell cycle (120 min). When we introduced preferential loading at ori, we found firstly that it delayed the splitting of MukBEF foci, similar to a spatially fixed loading site [212]. The feedback from ori to MukBEF, nonetheless resulted in somewhat improved (39%) partitioning. The effect was similar to what we observed in the simulations of non-growing cells. Preferential loading promotes partitioning (and colocalization) but sufficient time is required for the system to stochastically jump out of the undesirable configuration. But once it does the quarter positioned configuration is stable and does not revert back. The long runtime of the previous simulations meant the system had sufficient time to transition but this is not the case here.

Clearly, this level of partitioning accuracy is not representative of the biological situation, where 95% of cells have partitioned ori already 20 min after initial separation, with oris being separated at that point by an average distance of 33% of the cell length. We wondered whether the poor partitioning observed in the simulations could be overcome by the introduction of polymeric effects. Entropic repulsion is believed to play a role in chromosome segregation and organization [166, 169, 231–233]. In particular, newly replicated oris would experience the entropic repulsion of two closed loops. This is consistent with experimental observations, in which duplicated oris (and indeed all loci) experience an initially large segregation velocity [128, 234]. Numerical studies have demonstrated that the effective potential associated with such a repulsion has the approximate form of a (half) Gaussian in the center-of-mass separation (Bohn and Heermann, 2011, 2010a). We therefore incorporated entropic repulsion into the simulations by adding a repulsive force between oris specified by such a potential. Important to note is that this force is short-range and therefore, with a small enough value for the range, it does not affect the desirable quarter-positioned configuration but rather acts to destabilize the undesirable configuration having both oris associated to the same MukBEF focus.

This introduced two unknown parameters, the depth of the potential and its range. We performed a sweep over these parameters and measured the partitioning accuracy 20 min after ori duplication. We found that entropic repulsion on its own (i.e. no preferential loading) was not able to reproduce the observed behavior. Increasing the range of the repulsion to 400 nm, which is likely unphysical, did lead to accurate partitioning but at

the cost of ori that were too far separated, especially immediately after duplication. Furthermore, as we observed previously, without preferential loading the ori can escape from the MukBEF peaks resulting in weaker colocalization, as well as there being significant noise in the both the number and position of MukBEF peaks. Repulsion does not change these effects. As already noted, preferential loading on its own was also insufficient for accurate partitioning.

On the other hand, combining preferential loading with short-range entropic repulsion of ori, gave the properties of both and allowed the model output to move much closer to the measured partitioning accuracy and relative separation and resulted in kymographs with the experimentally observed behaviors (Fig. 10.5A-D). The short-range entropic repulsion destabilizes the undesirable configuration so that system switches, in a timely manner, to the quarter positioned configuration. As we saw for non-growing cells, preferential loading stabilizes this configuration, keeping ori in close association with their corresponding quarter-positioned MukBEF peak, thereby preventing both diffusive excursions and any attempts to return to the undesirable configuration.

While the output looks qualitatively promising, we wanted to make a quantitative comparison of ori dynamics. Therefore, we compared the simulated time-courses (using the preferential loading and the range and strength of ori repulsion suggested from the parameter sweep) with previous experimental results. To reduce the dimension of the data and accommodate the variation in cell length and cell cycle duration, we examined the dynamics of ori from the time of duplication (simulated data) or the time of initial

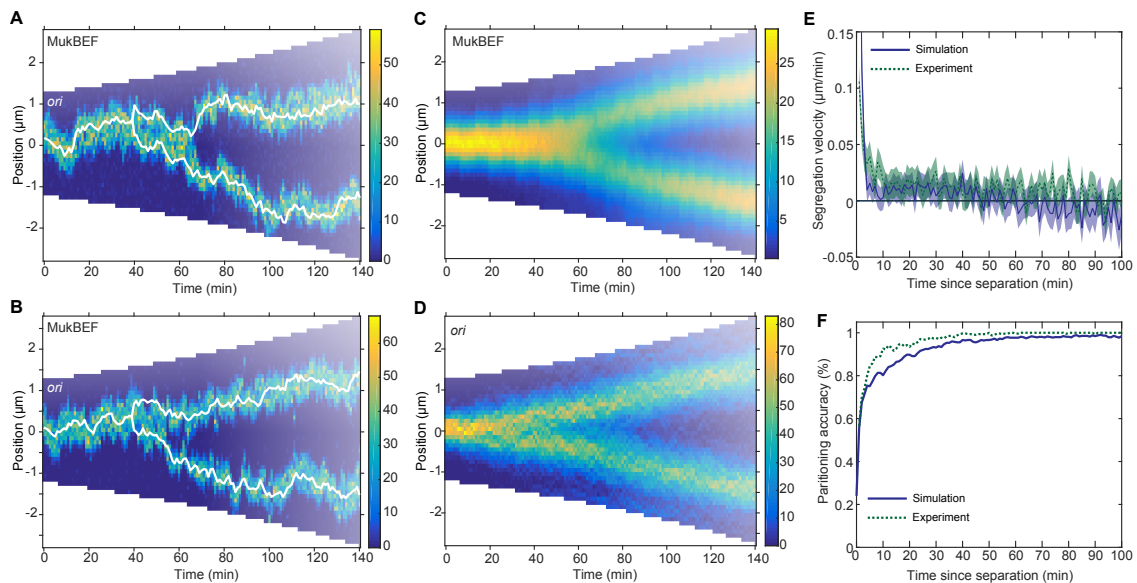


Figure 10.5: Repulsion between newly replicated ori results in realistic simulations of growing cells. **A, B.** Two example kymographs from individual simulations during exponential growth (doubling time of 120 min). in the presence of a repulsive force between ori. Shown is the number of MukBEF molecules (color scale) overlaid with the ori position (while lines). **C, D.** Average kymographs of MukBEF **C** and ori position **D**. **E, F.** Segregation velocity (the step-wise rate of change of the absolute distance between ori) **E** and partitioning accuracy **F** plotted as function of the time since ori duplication (simulations, blue) or separation (experiment, green). Experimental data is from Kuwada et al. [204]. Shading indicates 95% confidence intervals. The segregation velocity has been corrected for growth. Simulation results in (c-f) are from 450 independent simulations and use 10x preferential loading ratio and a repulsion range of 200 nm.

ori foci separation (experimental data [204]). We found good agreement between two cell length independent measures: the segregation velocity (the change of the absolute distance between oris between time points) and the partitioning accuracy (Fig. 10.5E,F). Thus, with the addition of entropic repulsion, the model is capable of reproducing the observed ori dynamics in growing cells.

10.2.5 Directed movement of ori can arise from spatially-dependent looping interactions

Our experimental data indicates, and our model assumes, that the ori experiences directed movement up the gradient constituting a MukBEF focus. How could such an attraction arise? It has previously been argued in the DNA relay and Brownian ratchet models of partition complex positioning by the ParABS system [235–238] that the elastic nature of the chromosome itself [239] can be harnessed to power directed motion of partitioning complexes. The elastic fluctuations of the chromosome allow partitioning complexes to detect local differences in ParA-ATP, a protein that tethers them non-specifically to the nucleoid. The result is that complexes move in the direction of greatest ParA-ATP concentration. However, this idea has never been tested polymerically. This is critical for migrating ori, since, unlike plasmids, the ori would experience an entropic counter force due to the polymeric nature of the chromosome. Nevertheless, we wondered whether a similar mechanism might underlie the biased movement of ori towards MukBEF foci.

In particular, we wondered whether directed movement of ori can arise due to the DNA bridging activity of MukBEF [240]. It has recently been demonstrated *in vivo* that MukBEF promotes long-range DNA interactions [47]. Given the association between MukBEF and ori, it is plausible that MukBEF preferentially forms DNA contacts involving the ori region. As such contacts would reduce the mobility of the DNA polymer, we would expect that ori would colocalize with MukBEF foci. To study this possibility, we turned to polymer simulations. We modeled the chromosome as a self-avoiding ring polymer confined in a rectangular cuboid and used the dynamic loop model (Bohn and Heermann, 2010b) to mimic the formation of DNA loops (bridges) between ori (a specific monomer of polymer chain) and distant DNA sites (other monomers) (see methods for details). As it is not computationally feasible to explicitly include the reaction-diffusion dynamics of MukBEF into the polymer simulations, we instead incorporated MukBEF implicitly via a spatially dependent looping probability along the long axis of the cuboid (nucleoid) representing the MukBEF concentration profile (Fig. 10.6A). We found that this resulted in the ori being positioned to the middle of the nucleoid, where the looping probability was greatest (Fig. 10.6B, blue line). This was in contrast to the uniform position distribution observed when a uniform looping probability was applied (red line). We also found that the positioning of ori affected the organization of the entire polymer, which took with up a left-ori/ter-right configuration (Fig. 10.7), consistent with previous results on the effect of a forced localization of ori [11].

We next asked whether the distribution of the ori arises as a time-average or whether the movement of ori is directed. When we examined the velocity of the ori as a function of the long-axis position, we found that the ori indeed experiences a restoring velocity towards mid-cell i.e. directed movement (Fig. 10.6C). We envisage this working as follows. On a short timescale, the ori fluctuates about its current “home” position. This allows it to locally sample the spatially-varying looping probability. It is then most likely to form a loop with another monomer in the direction in which the looping probability is greatest i.e. the direction of greatest MukBEF. The polymer subsequently relaxes, the ori is released

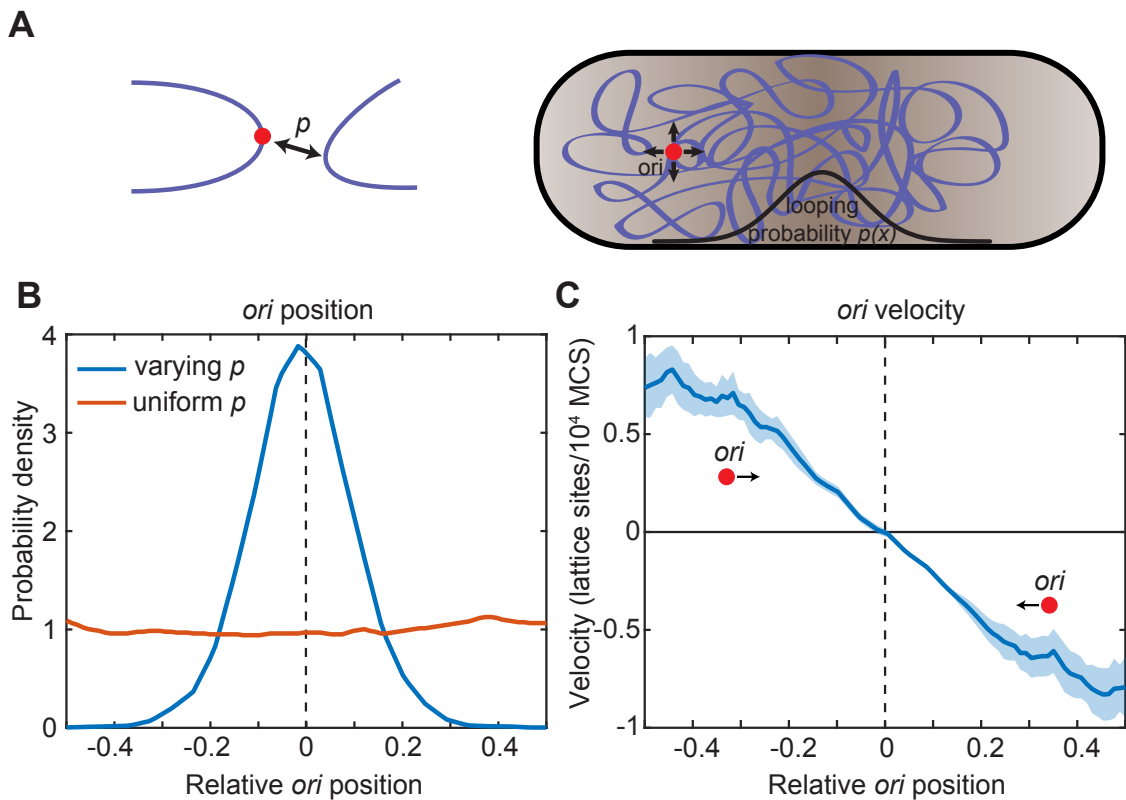


Figure 10.6: Directed movement of ori can arise from spatially-dependent looping interactions. **A.** A diagram illustrating how the elastic fluctuations of ori allow it sample the spatial looping probability distribution. It is therefore more likely to form a loop with a locus that is closer to mid-cell, where the probability of looping, p , is higher. **B.** Probability density of relative ori position along the long axis of the cuboid (aspect ratio 4:1) with (blue) and without (red) a spatially-varying looping probability (a Gaussian centered at 0 with standard deviation 0.1 in units of long-axis length; the looping probability at 0, p_{\max} , is 0.02). **C.** The mean step-wise ori velocity along the long axis as a function of relative position. Error bars indicate standard error. In **B** and **C**, the ori position was read out every 50000 Monte Carlo time-steps (MCS) and data is from 50 independent simulations with approximately 10000 data points from each.

to a new “home” position and the cycle repeats. In this way, elastic fluctuations of the polymer power the movement of ori up the gradient in the looping probability. Thus, directed movement of ori up the MukBEF gradient can plausibly arise due to a MukBEF-mediated, spatially-varying looping probability.

Finally, we examined how spatially-dependent looping affects chromosome segregation and the quarter positioning of duplicated oris. We simulated two chromosomes, initially overlapping with their oris at mid-cell, in the presence of a bi-modal looping probability distribution (with a peak at each quarter position). In the absence of looping, entropic repulsion ensures that the oris, along with the chromosomes themselves, are segregated (through not positioned [166]). However, we expected that looping of oris at the quarter positions would accelerate this separation. Indeed, we found this to be the case. Looping had a positive effect on ori segregation (Fig. 10.8D): the greater the looping, the faster ori were segregated. Furthermore, the ori are not just segregated but are positioned by the spatially varying looping probability to opposite quarter positions along the long axis of the cell in the same way as for the single chromosome case (Fig. 10.8E). Similarly, we

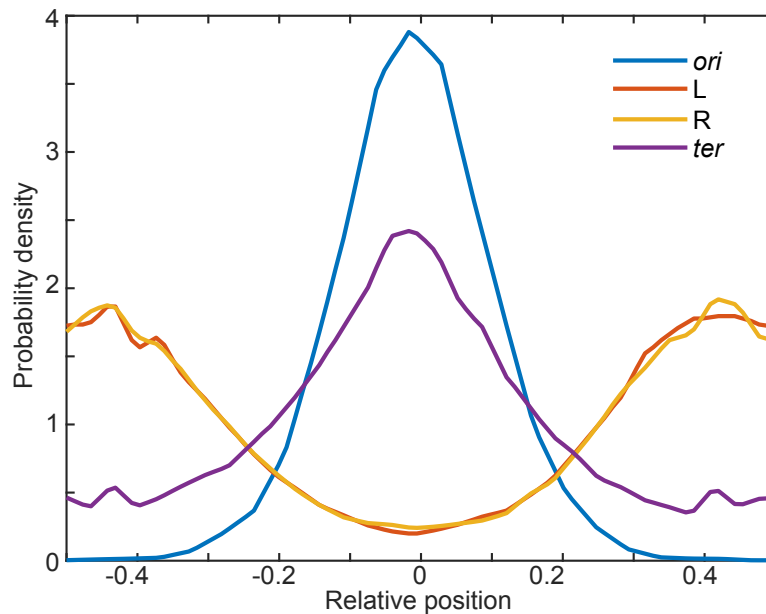


Figure 10.7: Positions of other loci. Probability density of relative *ori* (0°), *L* (-90°), *R* ($+90^\circ$), and *ter* (180°) position along the long axis of the cuboid in the presence of a spatially-varying probability for *ori* to forms loops as in Fig. 10.6B. *L* and *R* are generally positioned at the ends of the cuboid, while the *ter* is positioned roughly at mid-cell, though with a substantially broader distribution than *ori*. Data is from 50 independent simulations. Results are consistent with the histograms (from a single simulation) of Junier et al. [11], in which the *ori* is localised to mid-cell by an imposed strong harmonic potential.

found that the *oris* experience an effective restoring force around their respective quarter positions (Fig. 10.8F). The strength of this attraction (the slope of curve) increased with the frequency of looping. Note that unlike our stochastic simulations, we do not need to add repulsion between duplicated *oris* – entropic repulsion is a natural consequence of the polymer dynamics. Overall, these results indicate that a spatially-varying probability for *ori* to forms loops with other DNA (nominally due to the localized action of MukBEF) leads, in the manner of a DNA relay, to directed movement of *oris* to the locations where the probability is greatest, i.e. to the locations of MukBEF foci and, furthermore, that this can accelerate *ori* segregation.

10.3 Discussion

In this work, we have presented a quantitative explanation for positioning of the chromosomal origin of replication in *E. coli*. By analyzing the positioning and dynamics of *ori* and MukBEF foci in wild-type cells (Fig. 10.1), we first showed that *ori* are attracted towards MukBEF foci, as has been previously suggested [14,211]. We have recently argued that the positioning of MukBEF foci can be explained by a stochastic Turing and flux-balance mechanism [212]. Here, we incorporated *ori* and its interaction with MukBEF into this model and showed how self-organized MukBEF can position origins to their observed mid-cell and quarter-cell positions.

To formulate the model, we needed to specify a particular ansatz for the nature of MukBEF-*ori* relationship. Motivated by SMC in other bacteria [219] and our previous

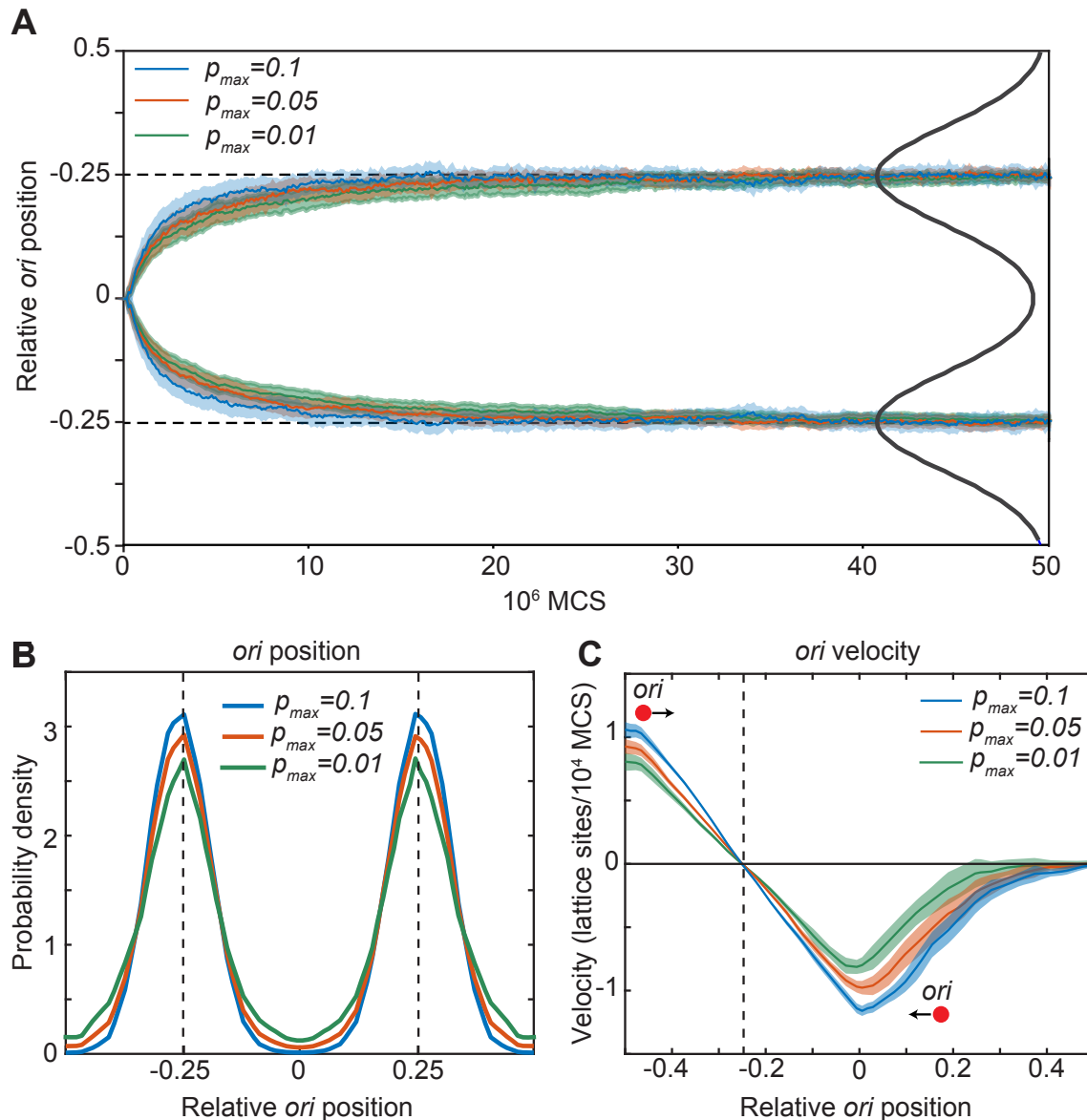


Figure 10.8: Directed movement of *ori* can arise from spatially-dependent looping interactions. **A.** The mean relative position of *ori*s along the long axis of the cuboid during segregation. Simulations were initialized with two overlapping polymers with the *ori* monomers at the middle position. We used a looping probability distribution (black line) with the shape of the sum of two Gaussians centered at the quarter positions with standard deviation 0.1 in units of long-axis length. Results for different values of the looping probability at the quarter positions, p_{max} , are shown. Data is from 500 independent simulations read out as in **C**. Shading indicates the standard error. **B.** Probability density of relative *ori* positions in simulations of two polymers described in **D** after equilibration i.e. the polymers have segregated to opposite ends of the cuboid. **C.** The mean step-wise *ori* velocity for one of the two segregated polymers. This polymer is confined to the left side of the cuboid. The *ori* experiences a restoring velocity to the approximate $-1/4$ position. The right half of the curve is due to infrequent excursions of the *ori* into the other half of the cuboid. The shaded region indicates standard error.

computational results [212], we assumed that MukBEF is preferentially loaded onto the DNA at sites within the ori. We found that the resultant feedback from ori to MukBEF led to robust ori partitioning. Preferential loading stabilizes the desirable quarter-positioned configuration, preventing stochastic switching to the undesirable configuration having both oris associated to the same MukBEF focus (Fig. 10.9A). In essence, preferential loading leads to a non-trivial mutual attraction between MukBEF and ori that results in robust association, positioning and partitioning of oris as an emergent property. We determined the ori drift and diffusion rates by fitting to the experimental ori velocity profiles (Fig. 10.2F). This also lead to excellent agreement with other experimental measurements that were not used in the fitting, namely the distributions of ori positions (Fig. 10.2E) and the MukBEF-ori separation distance (Fig. 10.4E), thereby providing further quantitative support for the model. Additionally, we found evidence of the mutual attraction between MukBEF foci

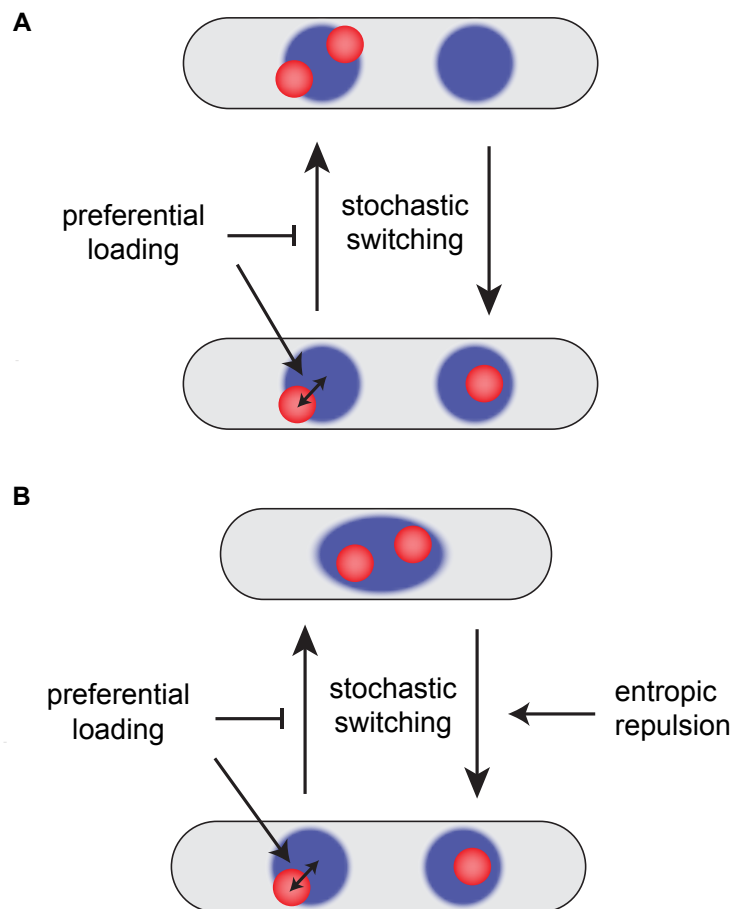


Figure 10.9: Preferential loading and entropic repulsion together lead to the observed ori dynamics. **A.** Schematic illustrating the effect of preferential loading of MukBEF at ori in the simulations of non-growing cells. In the presence of preferential loading ori (red) and MukBEF foci (blue) are strongly associated with both each other and the quarter positions. This acts to stabilize the desirable correctly partitioned configuration (bottom) over the un-partitioned one (top). In the absence of preferential loading, both configurations have equal stability (are equally likely). **B.** Short-range entropic repulsion promotes the timely separation of newly duplicated ori. Their separation promotes splitting of MukBEF foci. MukBEF-ori then move together to opposite quarter positions with preferential loading promoting their association and the stability of the quarter-positioned configuration.

and ori. MukBEF foci were found to be attracted to ori to a similar degree as their attraction to mid-cell (Fig. 10.1E), consistent with what we observed in our simulations.

The model could also reproduce the correct ori dynamics of a growing cell. This required adding entropic effects [166] to the model. Preferential loading and short-range entropic repulsion are individually not sufficient for both timely and accurate segregation and positioning. However, combined, they give very good agreement with the observed dynamics (Fig. 10.5). The repulsion between newly duplicated oris is needed to push the system out of the undesirable configuration immediately after ori duplication but is not required for the existence, stability or high colocalization of the desirable quarter-positioned state (as was seen in simulations of long cells without repulsion) (Fig. 10.9B). These properties are the result of preferential loading and the mutual interaction between oris and self-organizing MukBEF foci.

Supported by our experimental results, the model assumes that ori moves up the MukBEF gradient. What is the physics underlying this biased movement? Since MukBEF can bridge distant regions of the chromosome, it is conceivable that the MukBEF-ori relationship, however it is mediated, leads to a higher probability for MukBEF to form bridges between ori and other regions of DNA than for other genetic loci. Using polymer simulations, we showed that, combined with the elastic fluctuations of DNA, this can result in directed movement of ori up the self-organized MukBEF gradient (Fig. 10.6), similar to the DNA relay model [237, 238] proposed for ParABS-based positioning. However, the situation here is different in that the protein gradient is not generated entirely by the ori itself (partition complex in the case of ParABS). In this sense, it is similar to the proposed bulk segregation of chromosomes by membrane-based protein gradients [170]. The proposed mechanism leads to directed movement of oris to the positions of greatest looping (bridging) probability - the middle or quarter-cell positions according to the distribution of MukBEF, as well as accelerated entropic segregation of duplicated ori.

It is interesting to compare our results to a previous study of how macro-domain formation and positioning affect chromosome organization [11]. It was found that a macrodomain formed by spatially independent condensation of the ori region led to it being pushed to the poles of the cell. The authors therefore needed to additionally impose the mid-cell localization of the ori macrodomain. In our case, the mid-cell location is marked by MukBEF and the increased looping that it induces suffices to keep the ori region at that location. Since MukBEF foci are self-positioned (as explained by our stochastic model), no external determinants of location are imposed.

10.4 Predictions

Our model assumes that MukBEF is loaded onto the chromosome at positions within the ori region. However, there are other plausible hypotheses for the MukBEF-ori relationship. In general, we expect that any relationship that induces a mutual attraction between MukBEF foci and ori would result in similar dynamics. Indeed, initial simulations have indicated one possibility is that the ori region acts as a “stop” site for translocating MukBEF complexes. Therefore, the fundamental prediction of our model is not necessarily that MukBEF is loaded at sites within the ori region, as for SMC, but rather that whatever the nature of the MukBEF-ori relationship, it is such that it leads to an effective mutual attraction between ori and MukBEF foci.

In any case, the specificity of ori must be specified, directly or indirectly, by some sequence (or sequences) within the ori region. We know that the actual site of replication

initiation, *oriC*, is not responsible because moving it to another location on the chromosome does not affect *ori* positioning [41]. If this unknown “centromeric” sequence were inserted into a plasmid lacking a partitioning system, then we would predict that the resulting plasmids would colocalize with MukBEF foci just like the *ori*, and thereby be maintained in the absence of their own partitioning system. However, the challenge lies in identifying the centromeric site as it may be some distance (tens of kb) from *oriC* as is the case for the *parS* sites in *B. subtilis*.

The focus of this work was on *ori* positioning in slow-growing cells. However, an important question and one about which has received comparably little attention, is how *ori* are positioned during faster growth in which cells have multiple replication forks. Youngren et al. have examined the positioning of several genetic loci for the case of four replication forks, i.e. two to four *ori* [233]. They found that cells are born with two quarter positioned *ori*, that after replication, move to the quarter positions of each cell half. While polymeric effects and bulk chromosome segregation likely play an important role in this behavior, we nonetheless wondered whether our simplified model could recapitulate these results. Taking into account the higher copy number of MukBEF in faster growing cells [241], we found very consistent *ori* dynamics (Fig. 10.10). At birth, the two *ori* are quarter positioned, while after replication they migrate to the quarter positions of each cell half. Similar to the slow growth case, we observed that *ori* and MukBEF remain in tight association and MukBEF splitting is coincident with *ori* separation, so that there are approximately as many MukBEF foci as *ori*. This behavior requires the aforementioned higher copy number. Without it, the number of MukBEF is not substantially different from the slow-growing case and hence correct positioning was not observed. These predictions could be tested in the future by examining MukBEF in these cells as well as the effect of modulating MukBEF expression on the number of foci and on *ori* positioning.

10.5 Outlook

Overall the agreement with the experimental data is very promising given the simplistic nature of the model and that we did not perform a systemic fitting of the parameters to the experimental data (we fit only the *ori*-related parameters – see methods). Nevertheless the depth of the comparison is beyond what has been achieved previously for origin positioning in other bacterial systems. Hence, we suggest that this approach warrants further consideration and that protein self-organization may have an unappreciated role in chromosome organization.

More generally, the idea of dynamically controlling the positioning and splitting of a Turing pattern is interesting from a mathematical point of view and may be applicable to other unrelated systems. Indeed, a major aspect of the “robustness problem” of Turing patterning is the sensitivity of splitting to model parameters, domain size and stochastic effects [242]. The non-trivial coupling to *ori* in our model, as well as the self-positioning nature of the pattern [212], goes some way towards mitigating this sensitivity.

As noted earlier, the cubic reaction present in our model leads to a compartment-size effect in dimensions higher than one. One way to overcome this limitation would be to use particle-based simulation methods, which have recently been extended to higher order reactions [243]. However, this would likely involve a substantial increase in computation time, which would make it more challenging to perform multiple runs and parameter sweeps like we have done here. It may rather be possible to reformulate the model in terms of only bimolecular reactions. This would not only be more chemically realistic only

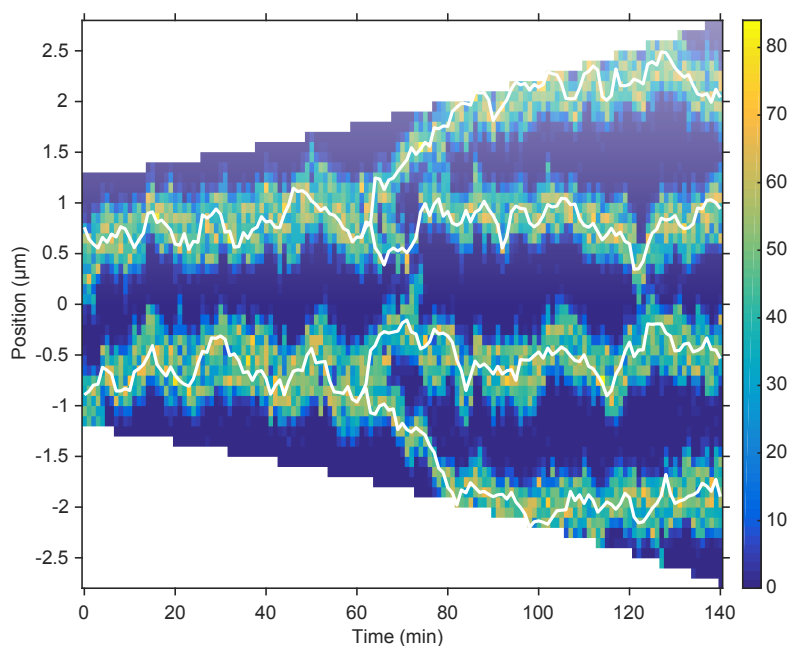


Figure 10.10: The model qualitatively reproduces the observed ori dynamics during multi-fork replication. An example kymograph showing multi-fork replication. Results qualitatively agree with Youngren et al. [233]. All parameters are as in Fig. 10.5, except for the total number of MukBEF molecules, which was increased from 520 nM to 1000 nM, broadly in line with previous measurements [241].

but would also allow for higher-dimension lattice based simulations.

For future work, combining particle and polymer simulations, at least to whatever extent is feasible, may provide a deeper understanding of the system. In particular, MukBEF has a major role in chromosome organization and facilitates long-range DNA interactions [47]. Like other SMC complexes, MukBEF may act by extruding loops of DNA [47], and/or be involved in stabilizing them [244]. Furthermore, MatP, which binds to matS sites in the replication terminus region, interacts with MukBEF, displacing it from that region of the chromosome [201] and thereby restricting long-range DNA interactions between the terminus region and other regions of the chromosome [47]. Both effects may help position this region at mid-cell, while simultaneously encouraging the co-localization of ori with MukBEF [201]. Thus, the role of MatP may need to be incorporated into future models. Lastly, with sufficient computing power, a fitting over more model parameters should be possible and may allow the quantitative evaluation of different hypotheses for the MukBEF-ori interaction. Overall, we envisage that the study of protein self-organization in the context of chromosome dynamics has a promising future.

10.6 Materials and Methods

10.6.1 Review of the Model

We briefly summarize the underlying model for MukBEF self-organization [212]. The general model scheme consists of three “species” (Fig. 10.2A). Species u and v exist on the surface (the nucleoid), while species w exists in the bulk (the cytosol). Species v diffuses slower than u . In the context of MukBEF, v is the basic functional complex, the ATP-

bound dimer of dimer, in a state in which it has entrapped multiple strands of DNA and is therefore relatively immobile. Species u represents the dimer of dimers in state in which it is non-specifically associated to DNA but without being immobilized e.g. it has not entrapped any DNA strands. These two species interconvert with the latter becoming the former at a basal rate α and cooperatively at a rate β , while the former becomes the latter at a rate γ . The cytosolic state w is the ATP-unbound dimer state which converts to and from the DNA associated states with linear rates ϵ and δ respectively. The differential equations describing the model are given in Fig. 10.3A. Parameter values are specified below.

This system exhibits Turing pattern formation i.e. the differing diffusion rates and the reactions of the system are such that a diffusion-driven instability can occur that leads to the spontaneous formation of spatially varying concentration profiles. The two states u and v generate the Turing pattern while the cytosolic state w is responsible for positioning the pattern. The latter is required to be well-mixed and it is so because it interconverts with the Turing species u and v on a sufficiently slow timescale. Note that more generally this state is not required to be cytosolic, only well-mixed. The kymographs and distributions shown in this work are of v (which we simply refer to as MukBEF). See [212] for further details and a detailed justification of the model.

10.6.2 Stochastic Simulations

Stochastic simulations were performed in C++. We used the Gillespie method (also called the Stochastic Simulation Algorithm) [245,246] to obtain exact realizations of the Reaction Diffusion Master Equation (RDME) as described previously [212] but with some changes for efficiency and the addition of simulated ori. We replaced the binary tree search of the enhanced direct method [247] with a 2D search as proposed by Mauch and Stalzer [248] and switched from 32-bit uniform random numbers (using the Ziggurat method) to 64-bit numbers (using `std::mt19937_64`) to ensure enough significant digits to accurately sample reactions occurring with very low relative rates (namely, ori diffusion). As before, the spatial domain (the long axis of the cell) is divided into discrete compartments, each having a width of $h = 0.1$ μm and between which the species can diffuse (Fig. 10.2C). The cytosolic state is well-mixed and is therefore treated implicitly. The system state was read out every 60s for consistency with the experimental procedure. For simulations with growth, the simulation was paused after every time-duration that corresponded to growth by one compartment. An additional (empty) compartment was then inserted at a random position and the volume and total number of molecules (via the cytosolic fraction) were increased, maintaining the same overall concentration.

The simulations were extended from those of the previous work by the addition of ori. We treated the ori as an additional diffusing species (with only one to four copies as appropriate) and implemented its biased diffusion up the MukBEF gradient, its duplication and its repulsion from other ori as follows.

Diffusion of ori Subject to the MukBEF Gradient

We assume that MukBEF is linearly related to the potential surface experienced by ori. This is the simplest choice and is supported by the agreement of the resultant linear velocity profile with the experimental one. Furthermore, since any symmetric potential is approximately quadratic around its minimum (up to third order due to symmetry), we would in any case expect a linear velocity profile and since we have the best statistics

near the minimum (the MukBEF peak), we would likely not be able distinguish other relationships between the MukBEF concentration and the potential it generates.

Given a linear relationship, the drift ori experience is based on the derivative of the local MukBEF concentration. We use jump rates (the rate at which ori jump between neighboring compartments, illustrated in Fig. 10.2C) derived by Wang, Peskin and Elston [249]. The forward and backward jump rates from compartment i to $i + 1$ and $i - 1$ respectively are

$$F_{i,i+1} = \frac{D_{\text{ori}}}{h^2} \frac{\alpha_{i,i+1}}{1 - e^{-\alpha_{i,i+1}}} \quad (10.1)$$

$$B_{i,i-1} = \frac{D_{\text{ori}}}{h^2} \frac{\alpha_{i,i-1}}{e^{\alpha_{i,i-1}} - 1} \quad (10.2)$$

where the dimensionless quantity $\alpha_{i,j} = \frac{\mu(\nu_j - \nu_i)}{h}$ is, up to a factor, the difference in the MukBEF concentration between the compartments (i is the number of molecules of slowly diffusing MukBEF species in compartment i), D_{ori} is the diffusion constant and μ is the drift parameter determining the strength of the attraction up the MukBEF gradient. We use the difference in the slowly-diffusing species only as MukBEF clusters have been shown in vivo to consist only of this state (Badrinarayanan et al., 2012b). This form for the jump rates respects detailed balance since the exchange between two neighboring compartments balances i.e. $F_{i,i+1} = B_{i+1,i}$. The derivation of these rates relies on the assumptions that, within individual compartments, the probability density for ori is at steady state and that the MukBEF gradient is approximately linear. Both of these requirements can be satisfied for sufficiently small compartment widths. However, it is not feasible to decrease the compartment width much below $0.1 \mu\text{m}$ due to the increased computational cost. Yet, the often sharp MukBEF profile (at a fixed moment in time) suggested that shorter compartment widths might be required. We therefore introduced sub-compartments within every compartment but only for ori positions. This approach has previously been applied to stochastic Turing patterns [250] but here we apply it to a “non-Turing” species (ori). Each compartment was divided into an odd number of sub-compartments and the MukBEF concentration was linearly interpolated across sub-compartments. The jump rates between sub-compartments were then defined as above. Performing simulations for different numbers of sub-compartments, we found that the apparent diffusion constant and drift rate (see below and Fig. 10.2) stabilized with greater than approximately 5 sub-compartments. The apparent diffusion was approximately 40% higher without sub-compartments. Since higher numbers of sub-compartments carried very little computational cost, we chose an arbitrary but relatively high value of 21 sub-compartments for the simulations presented in this work in order to be confident that there are no sub-compartment-size dependent effects.

Duplication of ori

The timing of ori replication was chosen randomly in each simulation by picking a duplication length from a normal distribution with mean 3 m and coefficient of variation 0.16 (based on the distribution of ori-foci splitting length of the data in Kuwada et al. [204]). As we use a fixed growth range ($2.5\text{-}5 \mu\text{m}$) and doubling time (120 min), we truncate the distribution to this range. This duplication length is then converted to a duplication time via the exponential relationship between cell length and time. The simulation is paused when

it reaches this time, the ori is duplicated (remaining within the same sub-compartment) and the simulation continued. Note that we do not mimic cohesion of newly replicated stands so that what we refer to in the simulation as ori duplication actually more closely corresponds to initial ori separation in vivo, which occurs 10-15 minutes after replication initiation.

ori repulsion

As discussed in the text, newly duplicated ori are likely to experience, for entropic reasons, a repulsive force between them [169]. Numerical studies have shown that the corresponding potential has the qualitative form of a Gaussian in the center-of-mass separation (Bohn and Heermann, 2011, 2010a). We assume that we are in the overdamped regime such that the separation velocity ν_s due to this force is proportional to the force. We therefore have the form $\nu_s = k d e^{-\frac{1}{2}(\frac{d}{\sigma})^2}$, where d is the separation between ori, σ is the range and k is the strength.

Parameters

All parameters of the core MukBEF model are given in table 10.1 and are as previously described and justified [212], except for the total species concentration C , which is increased by 30% to 520 nM but which is still within the experimentally justified range [210, 241]. This was done for compatibility of the MukBEF splitting time with the lower range of cell lengths used in this work (2.5 μm - 5 μm), which were chosen to more closely match the range of the experimental data in Kuwada et al. [204]. The cell volume ($V = 1.25 \times 10^{-15}$ L at birth (2.5 μm)) was taken to scale linearly with length and is required to convert the parameter β to the appropriate dimensions for use in the stochastic simulations. With the above total concentration and cell volume, there are 391 simulated molecules at birth. The remaining (ori-related) parameters were chosen by comparison with experimental data as described below and are given in table 10.1A,B,C.

Initialization of Simulations

Initial concentrations were set to the integer homogeneous configuration closest to the deterministic homogeneous state. Unless stated otherwise, single ori were initially placed at mid-cell, while in simulations starting with two ori, they were placed at the quarter positions. Simulations were first run for 30 min to equilibrate and then read out every 1 min (chosen to match the experimental data).

10.6.3 Apparent ori Diffusion Constant and Drift Rate

To be able to quantitatively compare the experimental and simulated data, we needed quantitative descriptors of the ori dynamics. We compared both data sets to a theoretical model of particle diffusion in a harmonic potential $U = \frac{1}{2}fx^2$ over an infinite 1D domain. Given a particle at position x_0 , the probability density that it is at position x at time δt later is [251]

$$p(x, \delta t | x_0) = \sqrt{\frac{f/kT}{2\pi S}} \exp \left[-\frac{f/kT}{2S} (x - x_0 e^{-\delta t/\tau})^2 \right], \quad (10.3)$$

Parameter	Value	
Common parameters		
α	0.5 s^{-1}	
β	$1.5 \times 10^{-4} \text{ nM}^{-2} \text{ s}^{-1}$	
γ	3.6 s^{-1}	
δ	$\log(2) / 50 \text{ s}^{-1}$	
ϵ	3δ	
D_u	$0.3 \mu\text{m}^2 \text{ s}^{-1}$	
D_v	$0.012 \mu\text{m}^2 \text{ s}^{-1}$	
V (volume at length $2.5 \mu\text{m}$)	$1.25 \times 10^{-15} \text{ L}$	
C	520 nM	
Additional parameters		
A.	D_{ori}	$5.4 \times 10^{-5} \mu\text{m}^2 \text{ s}^{-1}$
	μ	$0.026 \mu\text{m}$
B.	D_{ori}	$5.1 \times 10^{-5} \mu\text{m}^2 \text{ s}^{-1}$
	μ	$0.052 \mu\text{m}$
C.	k	1 s^{-1}
	σ	200 nm

Table 10.1: Additional parameters: **A.** obtained by fitting the model without preferential loading to the data of Kuwada et al. and used in Fig. 10.2D, Fig. 10.4, **B.** obtained by fitting the model with 6x preferential loading to the data of Kuwada et al. and used in Fig. 10.2E,F, Fig. 10.5, **C.** (together with those in (B)) used in Fig. 10.5 for ori repulsion with 10x preferential loading.

where $S = 1 - e^{-2\delta t/\tau}$ and $\tau = \frac{kT}{fD}$. From this, it is straightforward to calculate the expected value and variance of the step-wise velocity $\nu := \frac{x-x_0}{\delta t}$:

$$\mathbb{E}[\nu] = \frac{e^{-\delta t/\tau} - 1}{\delta t} x_0 \approx -\frac{x_0}{\tau} \quad (10.4)$$

$$\text{Var}[\nu] = \frac{D\tau}{\delta t^2} (1 - e^{-2\delta t/\tau}) \approx \frac{2D}{\delta t}, \quad (10.5)$$

where the second equality holds for $2\delta t/\tau \ll 1$ (the full expression is used when fitting). Note the expected value of the step-wise velocity depends linearly on position, while the variance is independent of position. This is observed in both experiments and simulations close within the neighborhood of the ori “home” position. We therefore use the measured slope of the velocity relationship and its variance to determine an apparent diffusion constant D_a and drift rate $d_a = \frac{1}{\tau} = \frac{fD_a}{kT}$. Linear fitting (Fig. 10.2) was performed using the fit function in Matlab with the inverse square of the standard errors as weights.

10.6.4 ori Drift and Diffusion Parameters

To search for parameter values for the ori diffusion constant (D_{ori}) and the strength of attraction towards MukBEF (k) that gave agreement between the measured apparent diffusion constants (D_a) and drift rates (d_a), we performed simple parameters sweeps. For the initial fitting, we chose D_{ori} to be a percentage of the experimentally measured diffusion constant D_a , ranging from 70% to 110% in 5% intervals, while the drift parameter μ was ranged from 0.5 to 2.5 times a nominal value of $0.026 \mu\text{m}$ (in steps of 0.5). The combination giving the best agreement was $D_{\text{ori}} = 0.9 D_a = 5.4 \times 10^{-5} \mu\text{m}^2\text{s}^{-1}$ and $\mu = 0.026 \mu\text{m}$. These values were used for the simulations shown in Fig. 10.2D, Fig. 10.3 and Fig. 10.4. To produce Fig. 10.2E,F, we performed the same parameter sweep in the presence of 6x preferential loading and found that the best agreement was obtained with $D_{\text{ori}} = 0.85 D_a = 5.1 \times 10^{-5} \mu\text{m}^2\text{s}^{-1}$ and $\mu = 2 \times 0.026 \mu\text{m}$. It should be noted that given the stochastic nature of the simulations, even with 100 runs of 600 min each, there was quite some variability in the data. These parameters were used for subsequent simulations with growth.

10.6.5 Entropic Repulsion of ori

We performed a parameter sweep of the strength and range of the ori repulsion and the preferential loading ratio. The range σ was varied over the values 50, 100, 200, 300, 400 nm, while the strength k was varied over 0.2, 1, 5, 25, 125 s^{-1} . We performed 450 independent simulations of a growing cell as described and measured the partitioning accuracy and relative separation of ori 20 min after ori duplication. We repeated this for different preferential loading ratios.

10.6.6 Polymer Simulations

In order to investigate the interplay between MukBEF and ori positioning within the nucleoid, we used a coarse-grained lattice polymer (Bohn and Heermann, 2010b). Within this framework, the DNA is described as a self-avoiding ring polymer that is confined in an elongated cuboid with an aspect ratio of 4:1 comparable to that of the *E. coli* nucleoid.

Using a ring polymer composed of 464 monomers, we chose a lattice of size $22 \times 22 \times 88$ that leads to a system density (monomer-to-volume ratio) of around 10%. Dynamic looping interactions were enabled between one specific monomer (ori) and distant monomers. For simplicity, we did not include loop formation between any two arbitrary sites but we do not expect this to change the nature of our results other than giving a homogeneous background of looping events. The looping probability is set to be dependent on the spatial position of ori along the long axis of the nucleoid and is drawn from a Gaussian distribution centered around mid-cell with a standard deviation of 8.8 lattice units. A lifetime of 10000 Monte-Carlo steps (MCS) was assigned to each loop. 50 independent Monte-Carlo trajectories were used to sample the dynamics of the system. In each simulation, 10000 polymer conformations were recorded, one every 50000 MCS. The initial position of ori was varied in each simulation in order to uniformly cover the long axis of the cuboid. In the simulations of the two chromosomes, each chromosome was modeled as a ring polymer composed of 232 monomers in a cuboid of the same size. Hence, the system density of 10% stayed the same compared to the single-chromosome simulations. The spatially-varying probability for looping between both either of the two oris and a distant monomers of any of the two polymers were drawn from the superposition of two Gaussian distributions centered around the two quarter positions of the cuboid with a standard deviation of 8.8 lattice units. 500 independent Monte-Carlo trajectories were used to sample the dynamics of the two polymer system. The simulations were initialized with two overlapping polymers with the two oris at the center of the cuboid.

10.6.7 Experiments

Strain SN192 (AB1157 lacO240-hyg at ori1, tetO240-gen at ter3, Plac-lacI-mCherry-frt at leuB, Plac-tetR-mCerulean-frt at galK, mukB-mYPet-frt) [201] was grown in M9 minimal medium supplemented with 0.2% glycerol and required amino acids (threonine, leucine, proline, histidine and arginine—0.1 mg ml⁻¹) at 30 °C. Cells were grown O/N, diluted 1000-fold and grown to an A₆₀₀ of 0.05-0.2. Unlike longer cells with two (quarter positioned) ori foci, cells with a single (mid-cell localized) ori focus, can be analyzed together in absolute, rather than scaled, coordinates by simply measuring foci positions relative to mid-cell (as we did in Fig. 10.2 for the dataset of Kuwada et al. [204]). We therefore, unless otherwise indicated, treated cells with DL serine hydroxamate (SHX) (Sigma-Aldrich, S4503) to a final concentration of 1 mg/ml. During the treatment, cells do not initiate a new round of replication, but complete any ongoing rounds [216]. To allow sufficient time for replication to complete to termination, cultures were grown for 3 h in the presence of SHX (generation time ~170 min). Finally, cells were spotted onto an M9-glycerol 1% agarose pad with the growth medium on a slide for imaging.

Time-lapse movies were acquired on a Nikon Ti-E inverted fluorescence microscope equipped with a perfect focus system, a 100 × NA 1.4 oil immersion objective, an sCMOS camera (Hamamatsu Flash 4), a motorized stage, and a 30 °C temperature chamber (Okolabs). Fluorescence images were automatically collected at 1-min intervals for 56 minutes using NIS-Elements software (Nikon) and an LED excitation source (Lumencor SpectraX). Exposure times were 150 ms for mCherry, and 100 ms for mYPet using 50% LED intensity. Phase contrast images were also collected at 1-min intervals for cell segmentation.

Alignment of frames, cell segmentation and linking of cells in consecutive frames were performed using SuperSegger [252]. To ensure that we considered only cells with a single ori, we subsequently filtered the dataset as follows. For any cells that had two ori foci on

any frame, we kept only the frames before two foci were first detected. This reduced the dataset of SHX treated from 1431 to 952 cells. We then used only frames with exactly one ori focus and one MukB focus. This resulted in 31820 data points (cell-frame combinations). Similar results were obtained if we further restricted the data set to cells of similar length. The same data was used to generate the step-wise velocity profiles (Fig. 10.1D,E) but as two consecutive frames are required this reduced the data set to 26226 data points. Linear fitting (Fig. 10.2) was performed using the fit function with the inverse square of the standard errors as weights.

Analysis, fitting and plotting were performed in MATLAB (Mathworks Inc.)

Chapter 11

Topological Data Analysis for Super-Resolution Localization Microscopy

A New Approach

References

The results presented in this chapter are published as and adapted from

- A. Hofmann*, M. Krufczik*, D.W. Heermann and M. Hausmann (2018), *Using Persistent Homology as a New Approach for Super-Resolution Localization Microscopy Data Analysis and Classification of γ H2AX Foci/Clusters*. International Journal of Molecular Sciences, 19: 2263. doi: 10.3390/ijms19082263.

AH analyzed the data and implemented the new methodology. MK performed the cell experiments and helped in the development of the new approach.

*equal contribution

Chapter Summary

DNA double strand breaks (DSB) are the most severe damages in chromatin induced by ionizing radiation. In response to such environmentally determined stress situations, cells have developed repair mechanisms. Although many investigations have contributed to a detailed understanding of repair processes, such as homologous recombination repair or non-homologous end-joining, the question is not sufficiently answered, how a cell decides to apply a certain repair process at a certain damage site, since all different repair pathways could simultaneously occur in the same cell nucleus. One of the first processes after DSB induction is phosphorylation of the histone variant H2AX to γ H2AX in a given surroundings of the damaged locus. Since the spatial organization of chromatin is not random, it may be conclusive that the spatial organization of γ H2AX foci is also not random and contributes to accessibility of special repair proteins to the damaged site and thus to the following repair pathway at this given site. The aim of this article is to demonstrate a new approach to analyze repair foci by their topology in order to obtain a cell independent method of categorization. During the last decade, novel super-resolution fluorescence light microscopic techniques have enabled new insights into genome structure and spatial organization on the nano-scale in the order of 10 nm. One of these techniques is single molecule localization microscopy (SMLM) with which the spatial coordinates of single fluorescence molecules can precisely be determined and density and distance distributions can be calculated. This method is an appropriate tool to quantify complex changes of chromatin and to describe repair foci on the single molecule level. Based on the pointillist information obtained by SMLM from specifically labeled heterochromatin and γ H2AX foci reflecting the chromatin morphology and repair foci topology, we have developed a new analytical methodology of foci or foci cluster characterization, respectively, by means of persistence homology. This method allows for the first time a cell independent comparison of two point distributions (here the point distributions of two γ H2AX clusters) with each other of a selected ensemble and to give a mathematical measure of their similarity. In order to demonstrate the feasibility of this approach, cells were irradiated by low LET radiation with different doses and the heterochromatin and γ H2AX foci were fluorescently labeled by antibodies for SMLM. By means of our new analysis method we were able to show that the topology of clusters of γ H2AX foci can be categorized depending on the distance to heterochromatin. This method opens up new possibilities to categorize spatial organization of point patterns by parameterization of topological similarity.

11.1 Introduction

DNA double-strand breaks (DSBs) can be induced by ionizing radiation and are known to be the most severe damages in the genome of a cell nucleus. The amount of DSBs simultaneously occurring is dependent on the radiation dose, the LET of radiation, the cell type, the radio-sensitivity of the cell, etc. Recent investigations have shown that DNA damaging is accompanied by an instant spatial reorganization of chromatin at and around the damaged site [253–255] and an activation of the repair machinery. One of the first steps of chromatin modification after DSB induction is phosphorylation of the histone variant

H2AX [256] to γ H2AX within a given neighborhood of the damaged site [257–259]. Such foci seem to “tag” the locations of damaged DNA for the recruitment of proteins that are starting and processing the follow-up repair [260, 261]. At that point a decision about the next procedure has to be made by the cell [262]. Several factors like cell cycle state, functional activity of genes, break position along the DNA sequence, temporal state of DNA compaction, number of simultaneously occurring DSBs, etc., are known to influence this decision and the consequences for a cell nucleus and the genome [262–264].

At a very first glimpse, the cell has to decide between fast or error-free repair for each DSB within the first minutes after damaging by irradiation. One choice may be homologous recombination repair (HRR) [263], which is a rather slow but error-free repair process. HRR needs an intact DNA sequence template of the homologous chromosome, along which a complementary strand can be reconstructed. In contrast to HRR, non-homologous end joining (NHEJ), a very frequently used repair process, may cause errors in the DNA base sequence but works much faster than HRR. Several specific proteins process the broken DNA ends by strand resection and re-connection of the broken ends at appropriately complimentary bases. HRR and NHEJ are the most often chosen pathways (for review see [264, 265]). HRR may be sometimes suppressed within repetitive DNA units if the damaged DNA side is not relocated to the heterochromatin periphery. In these cases single-strand annealing (SSA) takes place instead [266]. It has been also shown that especially in cases of irradiation at higher doses (> 2 Gy) and consequently more DSBs, the conventional NHEJ (c-NHEJ) may fail at some breakage sites and an alternative NHEJ process (a-NHEJ) is applied, which is a slow and error-prone repair process [265, 267].

On the one hand, HRR may be the first choice and preferentially used to keep the genome as much preserved as possible. Only if HRR is insufficient (e.g. due to too many DSBs at higher doses) then NHEJ saves the situation since it is much faster. Recently, it has been shown that in G1 resection dependent NHEJ is possible which seems to be different from resection processes in HRR [268]. On the other hand in G2, most DSBs are repaired by NHEJ. So some people believe that NHEJ is always the first choice and only in those cases where NHEJ fails HRR saves the repair [269].

Each of these different repair processes requires a different cascade of proteins that are time dependent recruited and de-recruited during the repair process and are responsible for DNA strand end clipping and processing, end-to-end fixation, or correct sequence re-association [270, 271]. Although many steps of DNA strand processing and its relevant proteins are known and the interaction of proteins during the different pathways are often well understood, the question as to what makes up the cell’s decision for a certain pathway at a certain damage site remains insufficiently answered. Considering all the major factors that influence the repair pathway choice and the quickness of the cell response may suggest a still unknown or not sufficiently understood central mechanism behind the pathway choice. This mechanism should work at each damaged side individually. This means that physical as well as topological parameters of the DNA strand break environment may determine the repair pathway choice together with epigenetic conditions [260, 261]. This assumption has been recently supported by investigations showing that radio-sensitivity can be modulated by chromatin remodeling in daughter cells of irradiated samples [272]. Assuming that the genome architecture and the architecture of repair complexes on the micro- and especially on the nano-scale become important for a repair focus region, not only novel techniques for a detailed analysis of spatial foci organization are required but also methods to categorize foci or sub-foci (clusters) and to compare each focus/cluster with each other independently of the cell or cell nucleus. Nano-scaled analysis has rea-

soned several transmission electron and super-resolution light microscopic studies in order to elucidate the spatio-temporal internal organization of repair foci and their chromatin surroundings with molecular resolution [255, 273–281]. Recently, it has been shown by super-resolution light microscopy that γ H2AX foci are built up by clusters that form nano-foci with different repair activities [275, 280] and that inside these nano-foci repair proteins are well organized [275, 276, 281] whereas the chromatin environment is interacting in a characteristic arrangement [280, 282, 283]. In addition, it has been shown that after radiation exposure and DNA damaging, Alu heteroduplexes may undergo Alu/Alu recombination into a single chimeric Alu element by NHEJ [284]. This may reason a dose dependent accessibility of ALU-sequence specific oligonucleotides (17mer uniquely binding to the ALU consensus sequence) as detected by SMLM [285, 286].

During the last years, it has been demonstrated that single molecule localization microscopy (SMLM) [287] is an appropriate technique to elucidate conformations of molecular arrangements and their functional relevance in cell nuclei, cytosol, and on cell membranes [253, 280–283, 285, 286, 288–290]. An embodiment of SMLM [291] as being used in this article, applies standard fluorescent dyes for specific labeling that can be switched between spectral “on” and “off” states [292, 293] to spatial separation of molecules (“reversible photo-bleaching”). From a reversible dark state, the fluorescent molecules randomly return to the emission state and cause blinking events that can be separated from a continuously fluorescent background. Each position of an emitting fluorophore is represented by an Airy disc and can precisely be located as the center-of-mass (barycentre) of such a disc. This also allows the precise calculation of spatial distances between fluorescent molecules in the 10 nm regime [286, 294, 295]. Using the matrix of the coordinates of fluorescent tags, all acquired positions can be visualized by an artificial “pointillist”, super-resolution image. In the images representing the point distribution, the effective resolution is only depending on the localization precision [295]. Moreover, the images can also encode results of distance analysis evaluations or density measurements.

However, localization data sets (e.g. labeling molecules of γ H2AX or methylation sides of heterochromatin like H3K9me3) consist of tens or even some ten thousands of individual point coordinates and their visualization and analysis is a separate challenge, since a point pattern does not automatically reveal a characteristic conformation or shape. In that way SMLM data fundamentally differ from conventional microscope images. While a conventional microscope provides an image with contours resolved with a scale of the order of 100 nm, SMLM is only producing a coordinate matrix with the positions of the fluorophores in the nanometer range. Such a pointillist representation requires a new approach to extract the relevant conformational information in such a way that the point distribution is unequivocally transferred into a certain shape or better topology that may be also maintained under different perspectives and different deformations. This requires quantitative analysis using mathematical concepts.

Approaches for a quantitative point density, distance, or cluster analysis exist for SMLM [275, 280, 280, 281, 285]. The analysis is restricted in scale to a certain order of magnitude and does not consider shape deformation. Quantifications on several orders of magnitude (for example in the range of a few nanometers up to several hundred nanometers) are hardly possible and cannot be easily compared according to typical characteristics. In order to overcome these restrictions, a novel mathematical approach is presented here, which analyzes SMLM data with methods of persistent homology [62]. This has the advantage that both, the geometric and the topological properties of given point distributions are considered [296] and a parameter-free quantification of the structural arrangement

of a point distribution over several orders of magnitude is possible. Thus, the accuracy achieved by state-of-the-art SMLM can be used not only for a point pattern analysis but also for a structural analysis of molecular arrangements. The point distributions and thus the underlying structures (e.g. heterochromatin distributions or γ H2AX foci/clusters) can now be directly compared independently of a cell nucleus whereby both nano-scale and micro-scale level differences are considered. Mors theory and set theory allow a quantitative comparison of two point distributions and a categorization according to a similarity measure. This higher degree of abstraction compared to image visualization achieves a higher degree of information and functionally relevant insights.

In order to demonstrate the power of this new mathematical approach for SMLM data, a proof of principle has been applied to analyze and categorize clusters of γ H2AX repair foci according to their structure and chromatin vicinity. The packaging degree of the DNA has consequences for the repair process. This is especially true for the densely packed heterochromatin because the damaged DNA has to be histone free for the repair and must also be accessible for the repair protein complexes [297–299]. It has been shown that DSBs in the heterochromatin region are usually repaired at the border of heterochromatic chromatin regions [253, 254, 266, 300] whereby the methylation degree typical for heterochromatin remains unchanged. Re-organization within heterochromatic regions is necessary to make the damage accessible for repair proteins. Therefore the proximity to heterochromatin was the parameter that was correlated to the internal topology by means of the topological data analysis (TDA). The topological representation of each focus was compared to each other and the degree of similarity was determined.

11.2 Results

In the following the mathematical approach is described with the aim to show the requirements for the γ H2AX foci/cluster analysis.

11.2.1 SMLM Data Processing

The raw data obtained by SMLM consist of a stack of about 2000 image frames acquired in a continuous time series. The blinking events registered above the fluorescent background are used to compute the exact positions of the individual fluorophores. Tests during former experiments [280, 281, 286] have revealed that the intensity of the maximum of a point signal must be at least four times higher than the background intensity to get registered as an event. The double of the average background intensity is subtracted from each pixel. The intensity barycenter μ and the associated standard deviation σ of each signal are determined using a two-dimensional Gaussian function f . Finally, the localization precision $\Delta\mu$ can be calculated that depends, among others, on the specificity and accuracy of labeling, on the number of detected photons q_i and the background intensity N_B . A more detailed description can be found in [285, 286, 290]. The results of the data acquisition are matrices containing the coordinates for each measured fluorescent point and the localization precision of each point. From such a matrix, data can be evaluated and structures can be interpreted as well as a pointillist image can be produced. As being an artificial image, results of data processing can also be used to code the image points.

In Fig. 11.1, representative images are shown comparing localization microscopy results with microscopic visualization. Here, an example of an irradiated cell is shown. We also had a look at untreated cells where only background or very few γ H2AX clusters of the same size as in irradiated samples are visible. It should be mentioned that for normal

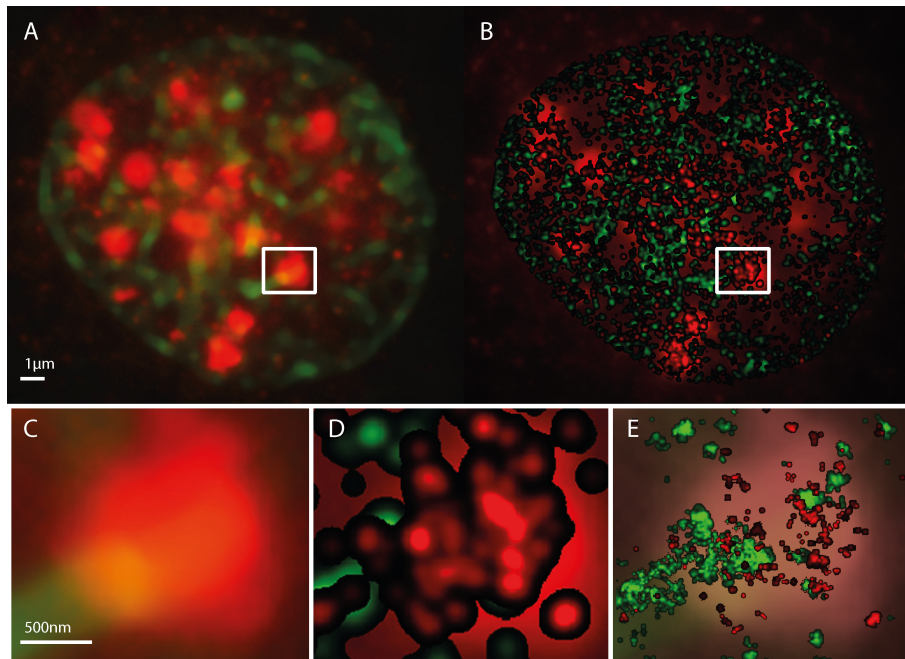


Figure 11.1: Microscopy images of H3K9me3 (green) and γ H2AX (red) immunostaining in a SkBr3 cell nucleus 30 min after irradiation with 1 Gy, 6 MeV photons. **A.** conventional widefield microscopy image. **B.** SMLM image represented as a density image where the brightness of a point refers to the number of next neighbor points. In the background the conventional widefield microscopy image of γ H2AX fluorescence is shown. The SMLM image is created from the coordinate matrix where the pixel geometry and intensity is stored in pixel values. After determination of the pixel size in nm, the coordinates and distances can be evaluated. In a fixed radius R around each coordinate, it is determined how many further coordinates are within this radius. This value is coded in the intensity of the coordinate point. In order to emphasize contiguous structures, each coordinate with an assigned value greater than zero is the starting point of a Gaussian distribution with a given sigma. The sum of all Gaussian distributions then represents the intensity distribution of the Gaussian-filtered density image (example: pixel size = 10 nm/pixel, Radius $R = 1000$ nm, Gaussian filter $\sigma = 50$ nm). **C.** Magnified insert from figure **A**. **D.** Magnified insert of figure **B**. **E.** Figure **C** superposed by the standard localization image. This supposing image is directly created from the localization data. In-homogeneities and sub-structures within the γ H2AX cluster are visible. Again a pixel size has to be determined. Here, every coordinate is starting point of a Gaussian distribution with the localization precision as sigma (pixel size = 10nm/pixel).

light microscopy, the image is reflecting the visual impression whereby for SMLM, the image is the result of data processing. This means that in SMLM data can be analyzed without an image and thus data analysis is independent of any image and from procedures of computer image analysis.

11.2.2 γ H2AX Cluster Recognition and Cluster Classification

SKBr3 is known being a cell line of increased radio-resistance with well separated γ H2AX foci also after radiation treatment with doses of several Gy. It has been shown that such foci can be separated into functionally relevant sub-foci or clusters of about the same size independent from the dose applied [275, 280, 281]. This has also been found for the cell nuclei analyzed here (see Fig. 11.2). Depending on the dose applied, the clusters are relaxing during the repair time (for details see [280]). Therefore cluster formation in

γ H2AX foci was determined at 30 min post irradiation, i.e., at an early repair time, but for different doses (see Materials and Methods).

A cluster analysis software was applied to the localization data of γ H2AX foci. The algorithm identifies points referring to a cluster within all the localization data according to user-defined parameters that were iteratively determined. These parameters are the minimum number of neighboring fluorescence signals within a defined radius around each labeling point and this given radius. A labeling point is identified as a member of a cluster if at least a minimum number of points are located within the predefined radius. The remaining points are identified as outside cluster points (“noise-points”). If two cluster points have a smaller distance than the given radius, they belong to the same cluster. All points whose distance to a cluster point is smaller than the radius also belong to the cluster. This cluster search algorithm is called “Density-Based Spatial Clustering of Applications with Noise” (DBSCAN) [301]. In this case, this allows the identification of γ H2AX-dense regions. Here, a radius of 200 nm and a minimum point number of 50 was used to identify “ γ H2AX clusters”. These clusters are not identical with γ H2AX foci obtained by diffraction limited wide-field imaging. It has been shown that the foci are subdivided into several sub-units which are compatible to the clusters described here [275,280]. Furthermore, the two parameters for the cluster recognition have been varied in order to ensure that the results of the subsequent analysis discussed in the following are not crucially dependent on this choice of parameters. Fig. 11.2 depicts boxplots of the γ H2AX cluster sizes of the respective irradiation doses.

After cluster recognition, the centers of the γ H2AX clusters were computed with the Surveyors Area Formula [302]. These centers were used as the center points of increasing circular shells. The heterochromatin density in these shells was computed (Fig. 11.3)

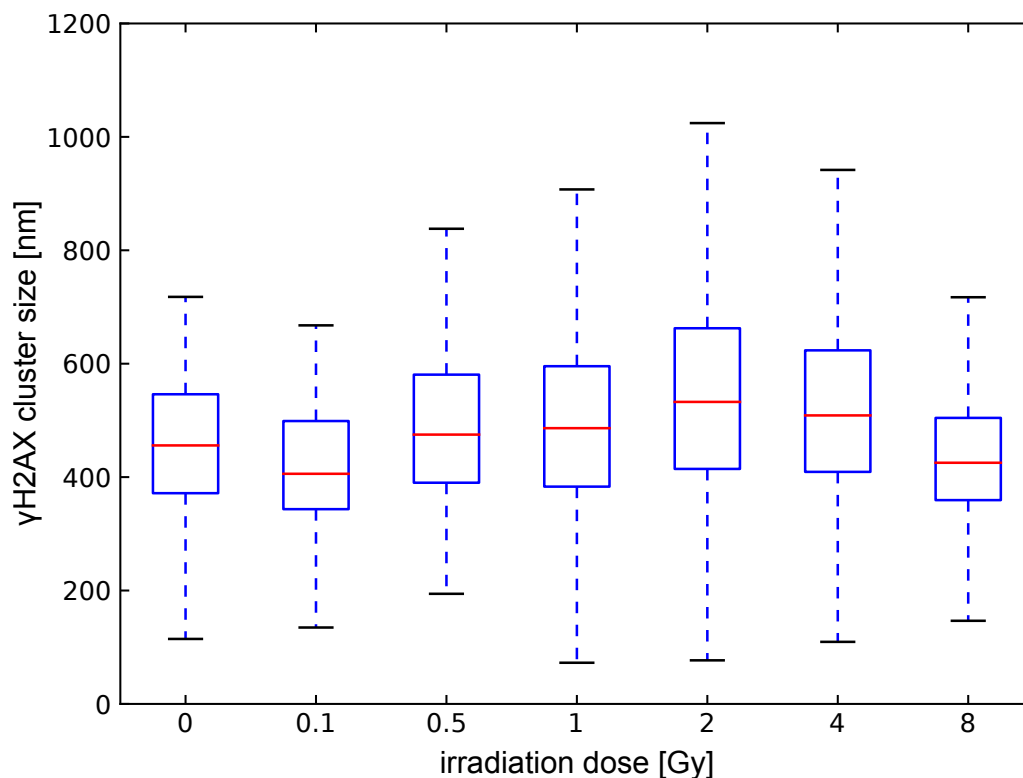


Figure 11.2: Boxplots of the γ H2AX cluster sizes of the respective irradiation doses.

(heterochromatin density equals the number of heterochromatin points per area of the corresponding shell).

Based on the heterochromatin density distribution tagged by antibodies against the histone modification H3K9me3, the γ H2AX clusters obtained from several cell nuclei exposed to different radiation doses were divided into heterochromatin-associated ones (HC) and non-heterochromatin-associated ones (nHC) clusters. For this purpose, the local maxima of the density distribution were determined (see also Fig. 11.3B); thereby the expansion of the circular shells necessary for determination of the density distribution affects this classification. A step size being too small (e.g. 1 nm) leads to a distribution with very many local maxima and minima since the density can change significantly in such small circular shells. Due to the small extent, it is possible that there is no point in a circular shell and a certain number of points in the nearest one. Due to the small surface area of the circular shells, considerable jumps are produced, which are methodically caused and only represent the real density distribution poorly. However, a step size being too large results in less separable characteristic peaks of the mapped distribution. After some iteration, a step size of 25 nm was chosen for density analysis. This step size ensures that the major local maxima are recognizable within a density distribution of heterochromatin points without losing the characteristics of the distribution.

For the applications shown here on SkBr3 cells, a minimum for the amplitude of a local maximum of at least $A_{min} = 2.5 \times 10^4$ points/nm² was determined in pre-experiments of automatic cluster search. The local maxima should have been spaced by at least 150 nm from each other in order to avoid overlapping maxima. Clusters with a local heterochromatin density equal or higher than A_{min} within a radius $R = 250$ nm were assigned as HC clusters. If the distance of the density maximum to the center of the cluster was larger

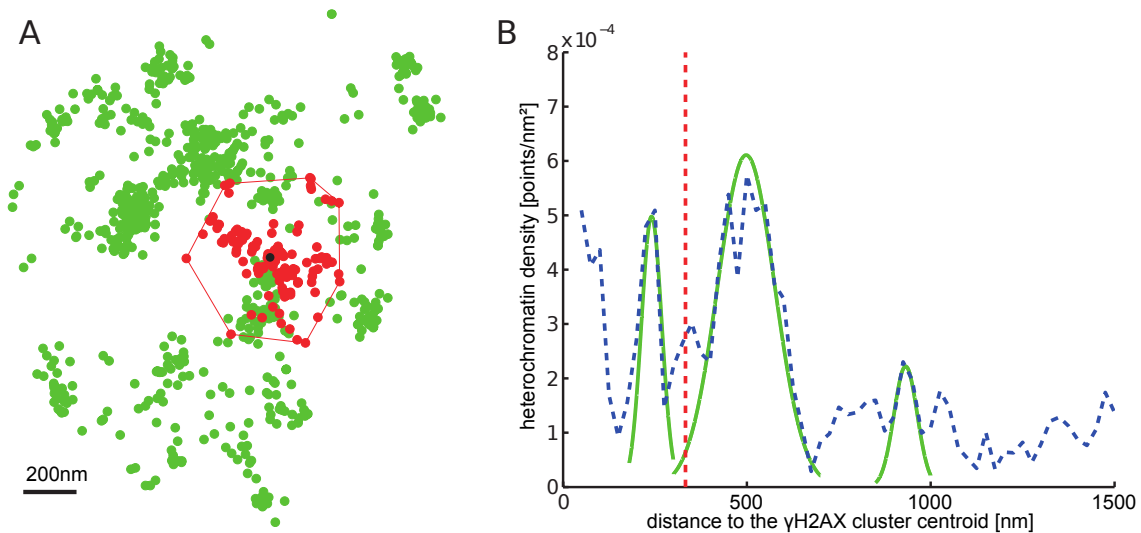


Figure 11.3: Density distribution around γ H2AX cluster. **A.** Schematic representation of a γ H2AX cluster (red) recognized by DBSCAN, the center of gravity (black) and the heterochromatin (green) around it. **B.** Density distribution of heterochromatin around the γ H2AX cluster center of gravity (blue). This density distribution is used to classify whether a cluster is heterochromatin-associated. For this purpose, the local maxima are approximated by Gaussian functions (green). The amplitude of these maxima is then compared with the predetermined threshold value. The red line represents half the mean square distance of the convex hull of the cluster (i.e. a kind of “radius” of the cluster).

than 250 nm, the clusters were assigned as nHC clusters. $R = 250$ nm is a suitable discrimination threshold when analyzing heterochromatin densities at the border of γ H2AX clusters in HeLa cells (for details see [280]). It must be noted that the central range of a γ H2AX cluster, i.e. a radius from 0 to 50 nm is excluded from the determination of the heterochromatin density distribution in order not to over-estimate no or a few labeling points on a very small central area. Such densities can only lead to extreme values, i.e. either zero (if a cluster does not contain any heterochromatin labeling point) or a large number (if some points are on a very small area in the cluster). Furthermore, it is assumed that in the center of a cluster the major repair activity takes place so that these labeling points may be due to relaxed and highly compacted heterochromatin.

11.2.3 Topological Analysis of the Clusters

SMLM data of H2AX phosphorylation were analyzed by means of persistent homology, a method for computing topological features at different spatial resolutions. In particular, the structure of each γ H2AX cluster was characterized by its so called α -shape. In the following, we introduce the computational strategy.

Barcodes as a Representation for a Pointillist Structure

A major principle to characterize the meaning of “topology” or “topological analysis” is to record properties of structures (depicted in a pointillist manner) which are invariant under certain deformations of the object. Mathematically these deformations correspond to continuous transformations of the topological space defined by the structures. Deformations which might fragment the structures are excluded. In the following, the attention will be focused on two quantifiable properties: a) the number of components which are independent from each other in such sense that connections between points only exist within the respective components; b) the number of holes of the structures inside the components. In algebraic topology, these properties are called the Betti numbers for zero dimensional and one dimensional simplicial complexes, respectively. They turn out to be very important topological invariants which help to distinguish between different topological spaces.

By comparing these quantities to two objects, it can be decided whether they have the same topology or not. Localization microscopy images are actually point-sets defined by the location of the fluorophores. Thus an appropriate method is required by which components and holes can be defined. In order to accomplish this, the point-set is converted into an object as described by the following procedure: First, the point-set is defined by the coordinates of blinking labeling points. In the next step, a geometric relationship among the points is defined by growing spheres of radius α around each of them. Whenever two spheres mutually embed each-other’s center, these centers of the growing spheres are connected by an edge. Points connected in that way are considered to belong to the same component. Any two points which are connected by a path through the existing edges are in the same component. Increasing the radii of the spheres, further points are reached connecting two previously disjoint components. Thus one can follow how the number of components is changing as a function of the increasing radius α . This means that each point is a separate component at the beginning, whereas for an increasing radius being large enough, each point is connected with each other. At the end of the procedure a single component is remaining.

The definition of holes also stems from this process. In order to build a solid, beside points and lines, face building blocks are required. For this, the simplest polygon, the

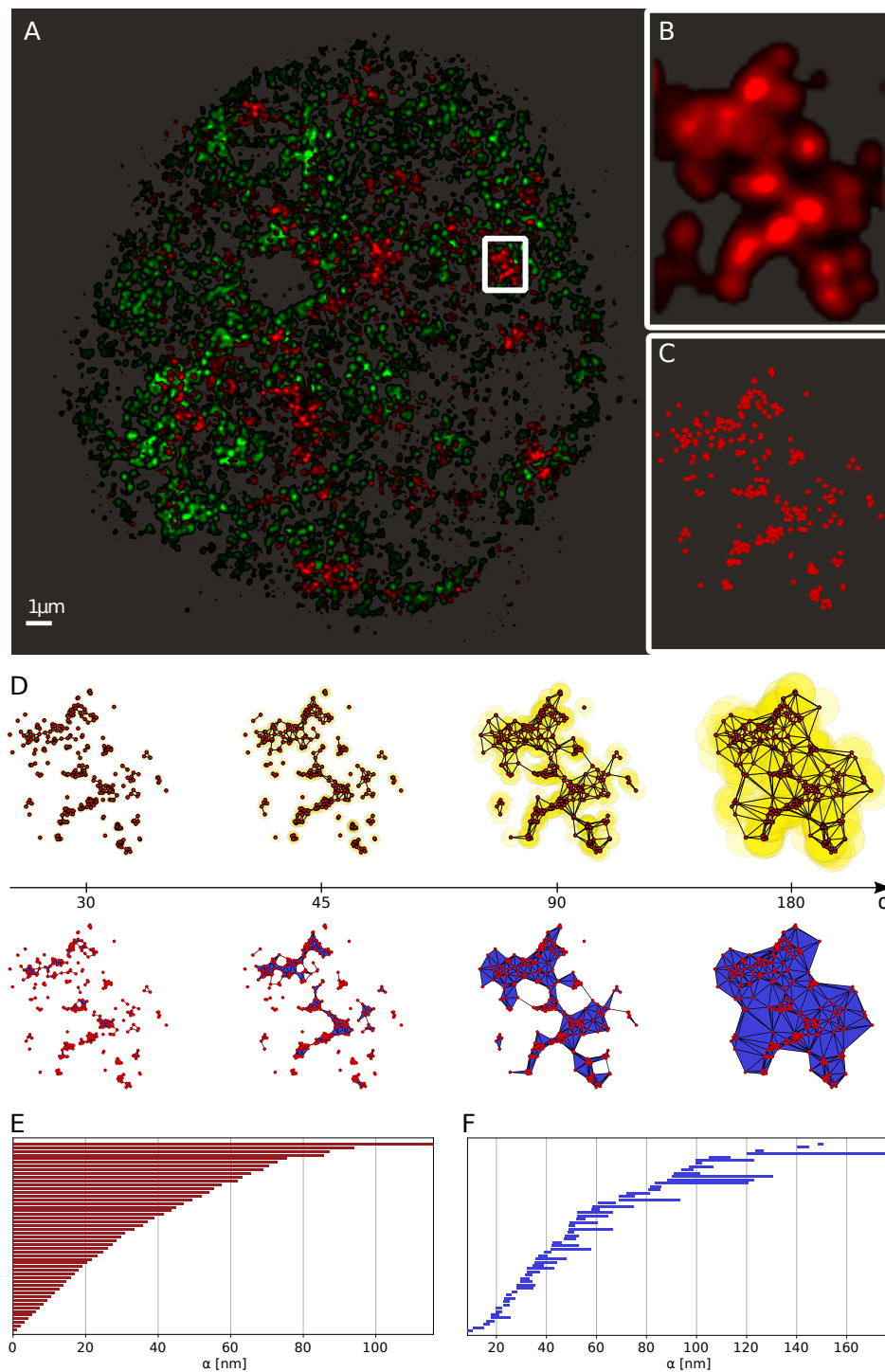


Figure 11.4: From the SMLM data to the barcode representation. **A.** Full view of heterochromatin (green; labeled by antibodies against H3K9me3) and γ H2AX labeling (red) of a SkBr3 cell nucleus 30 min after irradiation with 2 Gy depicted as a density image. **B.** Zoom-in into the marked γ H2AX cluster. **C.** Scatter plot of the marked γ H2AX cluster (every point represents a detected fluorophore). **D.** Components of the α -shape filtration of the marked γ H2AX cluster exemplarily depicted at $\alpha = 30, 45, 90, 180$ nm (left to right). As the growing spheres mutually embed the center of each-other the corresponding centers are connected by an edge (as shown in the upper row). Whenever a triangle is formed, it is included in the solid as a face element (illustrated in the lower row). **E.** Barcodes of dimension 0 (Betti number) corresponding to connected components. **F.** Barcodes of dimension 1 (Betti number) corresponding to holes.

triangle is appropriate. Whenever three edges form a triangle, not only the edges but the face of the triangle is considered. The described procedure is presented in Fig. 11.4D for a particular γ H2AX cluster. Once the surfaces are defined, the holes are counted. In fact, it is possible to register their number and the number of components for every separate value of the radius α .

In [62] an approach is presented, how all components and holes can be summarized in a compact way. The presented approach allows the representation as “barcodes” to track the formation and disappearance of components and holes as the value of α increases and thus independently of a fixed value of α . An example is shown in Fig. 11.4E and 11.4F. The beginning of a line in the barcode representation shows at which value of the radius α , the component or hole has arisen, and the end of the line for which value of α it has disappeared as a result of association to a larger component. All red bars start at zero, because at a value of $\alpha = 0$, all points are unconnected and for that reason each represents its own component. Whenever two points are joint to a line (or three points to a triangle), the two (or three) points are combined in the newly created component “line” (or “triangle”), and hence, the associated red bar ends. By further joining points with increasing value of α , the lines, triangles and holes shown in Fig. 11.4D are created. The lifetime of a hole is represented in Fig. 11.4F by blue bars. A bar starts when the hole is created and ends when it is completely filled.

The barcode thus represents an image of the examined structure on all scales. The creation and dissolution of complexes on a small scale is recorded alongside the lifetime of complexes on larger scales. As a consequence, in the case of chromatin structure or γ H2AX cluster structures, barcodes contain information about components and holes in both the nanometer and micrometer scale ranges. This compact and illustrative representation also allows selection of specific substructures, such as all components and holes that exist in the range of α_1 to α_2 .

As illustrated, the characterization of point structures by barcodes opens up new possibilities to analyze and categorize clustered structures in cell nuclei. Although, the barcode representation of a point-set might appear at least at first glance as confusing as the point-set itself, the possibility to compare different sets of barcodes and to define parameters describing their similarity is a significant advance in the analysis of point structures.

Similarity of Barcodes

In [296] an approach is presented that can be used as a measure of the similarity S of barcodes. The similarity of two barcodes A and B of given dimension, which are comprised of bars a and b , respectively, is then represented by:

$$S(A, B) = \frac{1}{|A| + |B|} \left[\sum_{a \in A} \underbrace{\sup_{b \in B} \frac{|a \cap b|}{|a \cup b|}}_C + \sum_{b \in B} \underbrace{\sup_{a \in A} \frac{|a \cap b|}{|a \cup b|}}_D \right] \quad (11.1)$$

$J(a, b) = |a \cap b| / |a \cup b|$ represents the Jaccard index [303], which is a measure of the similarity of two bars. The result is a value between 0 and 1, where a value of 0 means no overlap of the two bars and two identical bars have a value of 1. The similarity measure for barcodes is described by the formula $S(A, B)$. The part marked C in equation $S(AB)$ states that for every bar a the bar b is searched, for which the Jaccard index $J(a, b)$ is maximized. This is repeated for each a and summed up. Analogously, for each bar b the bar a is searched, for which the Jaccard index $J(a, b)$ is maximized. Again, the results for

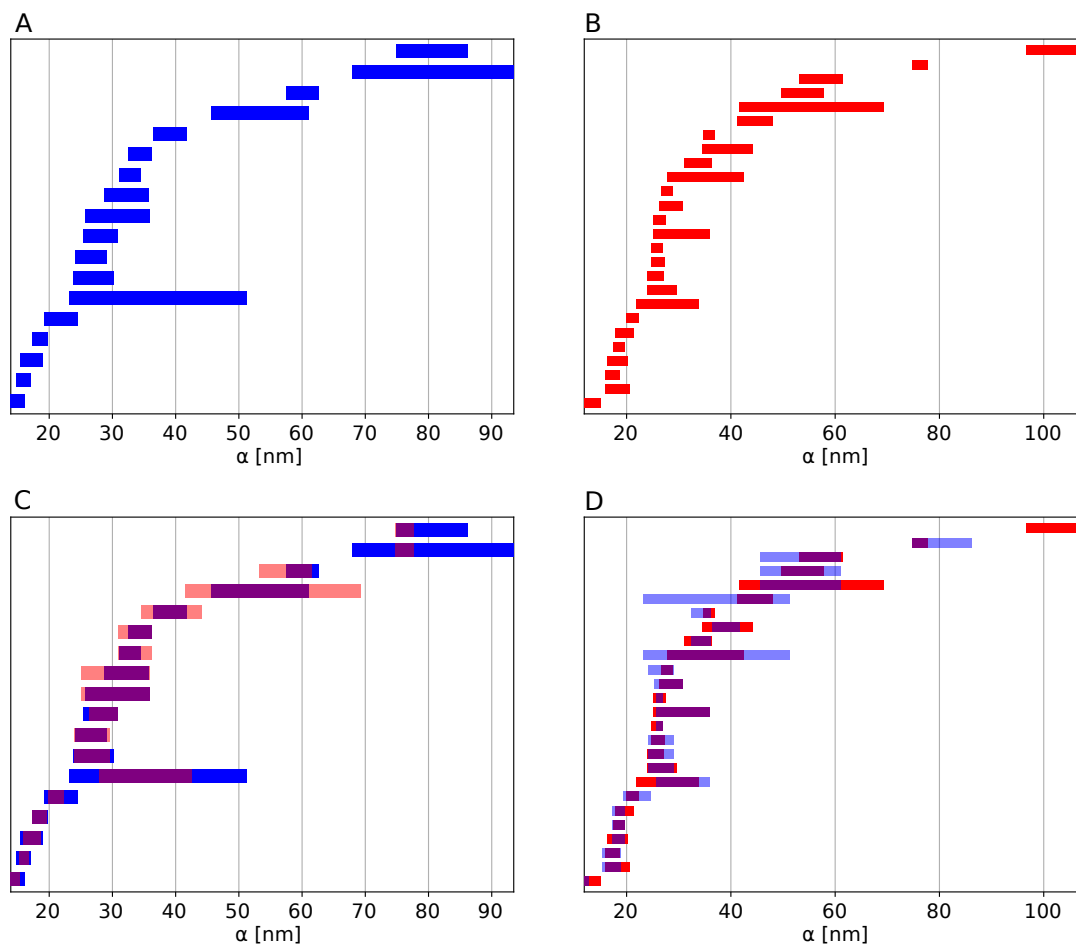


Figure 11.5: Example of the result of the barcode similarity measure. **A, B.** For each bar of barcode A (blue), the bar from B (red) is sought for which the Jaccard index is maximal. **C.** The Jaccard index represents the extent to which two bars overlap (pink). This corresponds to C in equation $S(A, B)$. These values are summed for each bar of A . **D.** Analogously, for each bar from B the bar from A is sought, for which the Jaccard index becomes maximal. This corresponds to D in equation $S(B, A)$. These values are in turn summed for the bar of B . The sum of these two subtotals is divided by the number of bars in both barcodes. The result $S(A, B)$ or $S(B, A)$ quantifies the similarity of the barcodes A and B in terms their overlap. Here, two barcodes of dimension 1 (holes) with a high similarity are depicted, i.e. the overlap of the two sets of barcodes is high. For comparison, we illustrate two dissimilar barcodes of dimension 1 in Fig. 11.6.

the individual bars b are summed up.

The resulting sums are then added up and divided by the total number of bars of the two barcodes. As a result of the division, the similarity measure $S(A, B)$ can vary between 0 and 1. An illustrative description based on two example barcodes with a high similarity is shown in Fig. 11.5. For comparison, we show two example barcodes with a low similarity in Fig. 11.6.

The similarity of barcodes of different dimensions can be defined as the average of the individual similarity values.

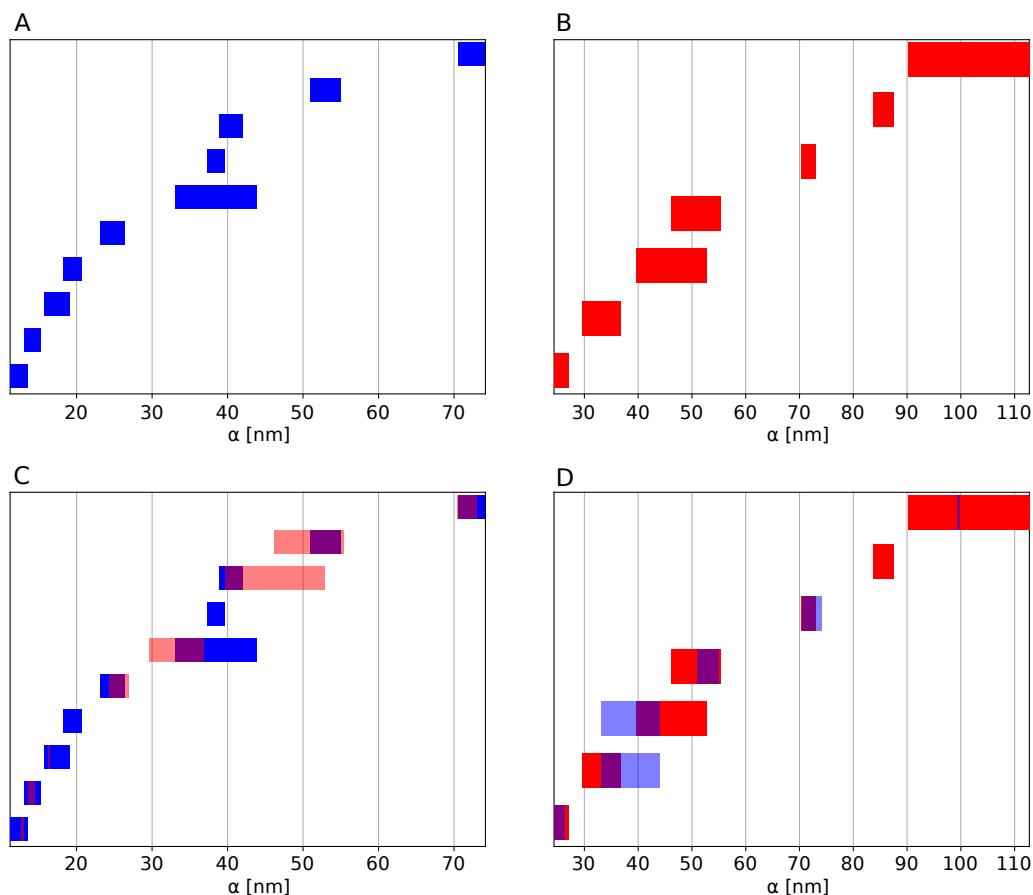


Figure 11.6: Example of the result of the barcode similarity measure (analogous to Fig. 11.5). **A, B.** The two barcodes *A* (blue) and *B* (red) of dimension 1 (holes) have a low similarity. **C, D.** Therefore, the overlaps of the two sets of barcodes is low.

Proof-of-Concept Experiments

From the γ H2AX clusters classified, 200 HC clusters and 200 nHC clusters were selected by determining those with the highest and lowest heterochromatic densities, respectively. This number of clusters was chosen because this group is large enough to avoid statistical outliers, but small enough to visually check microscopy images as obtained for each cluster from the point matrices and, if necessary, the corresponding density distribution.

The HC- and nHC-associated γ H2AX clusters were examined according to their topological similarity as defined by the overlap measure described above. For comparison, the similarity between the two groups of γ H2AX clusters was also determined in terms of density and size. Since the clusters are polygonal, the root mean square of the distances of the points of the convex hull has been defined as a measure of cluster size.

To enable a comparison of all γ H2AX clusters, the values of the topological similarity measure are depicted in a heat map (Fig. 11.7). The arrangement of the heat maps is as follows: The upper left quarter compares the HC clusters with each other, the lower right the nHC clusters. The upper right and lower left quarters each compare nHC with HC clusters. The arrangement of the individual HC and nHC clusters is random. The spectrum of the color bar of all the heat maps ranges from red representing dissimilarity between the analyzed clusters to blue depicting similarity. In this representation, it is then easy to detect certain patterns, such as similarity of all the foci clusters, outliers or

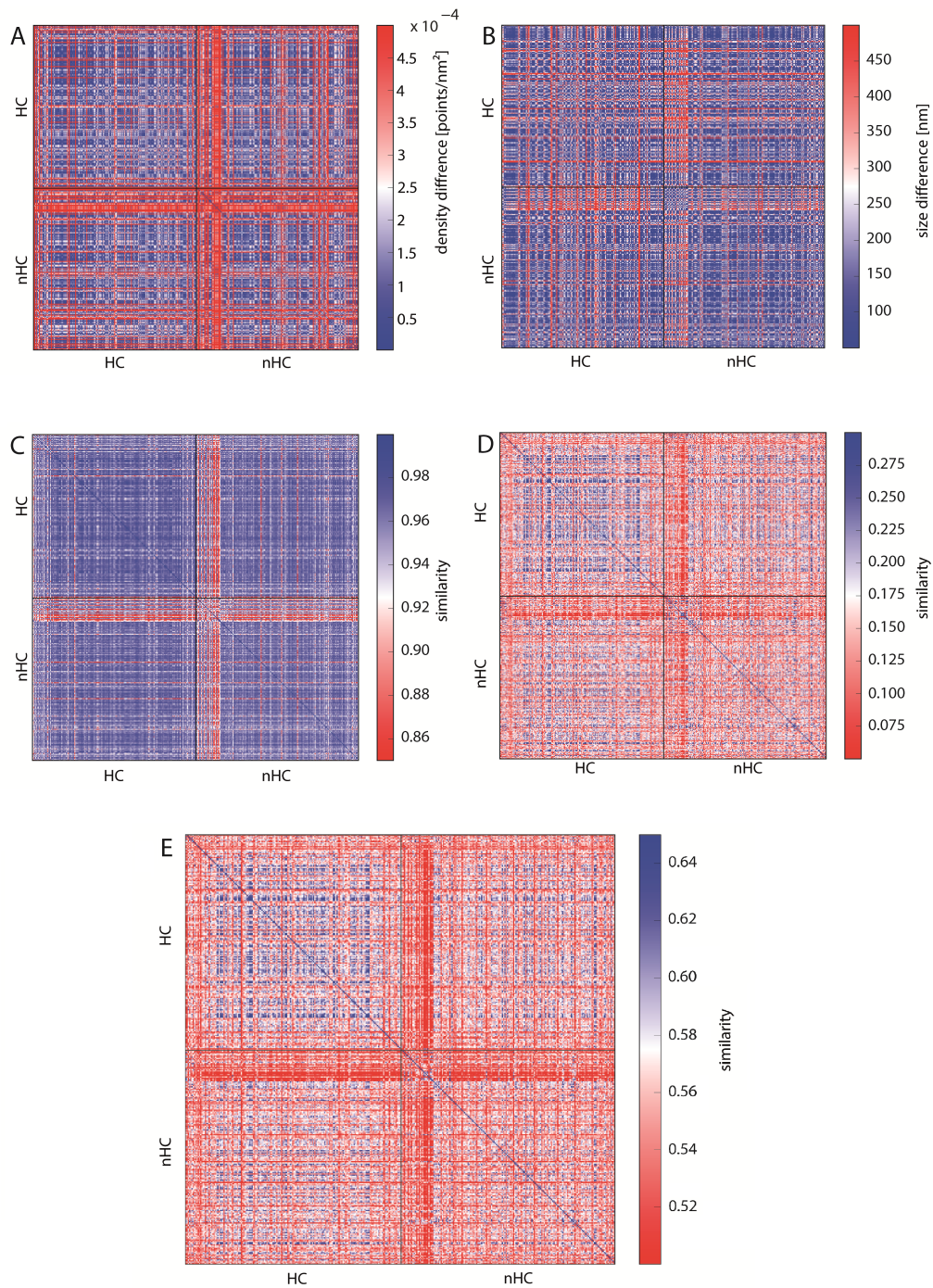


Figure 11.7: Heat maps depicting the similarity of HC- and nHC-associated γ H2AX clusters according to **A.** cluster density, **B.** cluster size indicating a high similarity (see also Fig. 11.2), **C.** topological similarity of the connected components (“lines”), **D.** topological similarity of the “holes”, **E.** average topological similarity. The spectrum of the color bar of all the heat maps ranges from red representing dissimilarity between the analyzed clusters to blue depicting similarity.

distinct areas of increased similarity or rather dissimilarity. As a measure of the similarity or rather dissimilarity of density (or size), the difference in density (or size) of two clusters was used. A small difference here means a great similarity.

In Fig. 11.7A the difference in point densities between the analyzed clusters is presented. There is no clear difference between the clusters assigned to HC or nHC with most of the clusters showing similar density. However, there is a set of nHC clusters which are highly dissimilar to all the other clusters. This is shown by the red cross in the middle of the heat map and can be explained by the fact that these nHC clusters are significantly smaller than the median cluster size (see Fig. 11.7B and Fig. 11.2). Small clusters have a higher density than the other cluster. Analogously, Fig. 11.7B does not reveal a difference between HC and nHC cluster. Hence, in terms of density and size of the foci, the proximity to heterochromatin does not appear to affect differences for the clusters.

Fig. 11.7C-E show the similarity of the clusters using the topological similarity measure presented above. Here, cluster structures are characterized by barcodes for connected components and holes (Fig. 11.4 shows the actual barcodes for one γ H2AX-cluster). Surprisingly, the clusters show nearly identical results for the components of interconnected points (Fig. 11.7C). Once again, the crosses are visible, which refer to extremely small clusters as it has already been explained above in the heat maps depicting the density. These aberrant clusters may be sorted out and further analyzed by repair protein staining. Maybe these clusters refer to such components that do not contain any repair activity, as it has recently been shown by Natale et al. [275]. But this should be tasks for further investigations.

Comparing the barcodes of dimension 1 representing holes (Fig. 11.7D), the HC assigned clusters show a higher similarity in contrast to the nHC clusters, although the topological similarity values are very low. This, however, can be adjusted by using the average similarity for components and holes (Fig. 11.7E). After this averaging, HC assigned clusters show a clear similarity whereas the nHC assigned clusters do not. This means that by topological analysis the HC clusters may be discriminated as those clusters of high topological similarity. The proximity of γ H2AX clusters to heterochromatin seems to have a significant measurable impact on its structure. Interestingly, the nHC and HC clusters are more similar than the nHC foci themselves. It can be clearly seen that, on the one hand, the proximity to heterochromatin influences the structure of the foci, but on the other hand that there are other γ H2AX cluster influencing factors, otherwise the similarity between nHC and HC foci would be unexplainable.

11.3 Discussion

DNA double strand repair uses fascinating mechanisms that have been developed during evolution towards two different directions fast and error tolerable or slow and exact [264–267]. After having induced a DSB by ionizing radiation, chromatin re-arranges and H2AX phosphorylation occurs in the damage environment [253, 254, 280] within a few minutes accompanied by the recruitment of proteins specific for a certain repair mechanism. During the last decades modern techniques applied in radiation biology and radiation biophysics, have offered detailed insights into the protein interactions and cascades along the different repair pathways [260, 261]. These investigations have completed our understanding about repair processes and boundary conditions that favor repair towards either end-joining processes like NHEJ or recombination processes like HRR. The better our understanding has become the more the question becomes urgent how a cell can decide

which repair pathway should be the appropriate one at a certain damage site. Cells can simultaneously use all repair pathways in a cell nucleus at different damaged sites.

The repair pathway choice could be random for instance. This, however, is not convincing since it has been shown that whenever it is functionally relevant for cell survival a fast repair process is addressed.

Assuming a non-random pathway choice at a given damaged site raises the question for a fast, easy and therefore always functionally available, and everywhere implemented mechanism for the cell's decision. Beyond several epigenetic approaches, people have started to discuss whether such a mechanism may be encoded in the architecture of chromatin around the damaged site [(key note) lectures and discussions at the joint ERRS and GBS conference 2017 in Essen, Germany]. This would, however, require deeper insights into the internal structural organization of a repair focus of a typical order of size of about the resolution limit of a light microscope (about 200 nm).

Recent applications of electron-microscopy [277,278] and super-resolution light microscopy like SMLM, STED or GSDIM [253,255,275,276,280–283,285,286] have demonstrated that it is feasible to study single molecular arrangements within a repair focus. With improving resolution of microscopy and data evaluation of structures on the meso- and nano-scale, the question for best suited analysis parameters and potentially useful classification criteria of repair foci and damaged chromatin sites has become important.

Here, we have introduced a rather unconventional approach for SMLM data analysis of γ H2AX foci and their chromatin environment. This approach makes use of the advantage that SMLM data can be evaluated without image production and image processing [286]. This novel approach combines a geometrical evaluation based on Ripley's distance and cluster analysis with persistence homology for similarity classification of repair cluster loci. Although the mathematical principles behind this approach are well established, it is the first time that topology has been used as biologically relevant criteria. This may allow to circumvent locally occurring deformations in the analysis and to extract a parameter pattern that is scale independent and can categorize repair foci into structural classes. Here we have demonstrated a very first proof-of-concept experiment, in which we could show that the category of HC associated γ H2AX clusters are highly similar in terms of both topology and geometry whereas nHC associated clusters are completely dissimilar. This topological similarity was independent of the irradiation doses. However, only one early repair time was considered. In future experiments other later repair times may also be considered in order to find out whether a change in topology occurs during repair.

In addition clusters that do not fit in size could be ruled out also by the topological similarity measure. Here, however, the practicability of this method has been demonstrated; therefore, the foci selected by the presented method have not been sorted out. The number of 400 clusters used for this analysis has been large enough that outliers, such as the mentioned foci, are not significant. On the other hand 400 clusters are manageable by interactive control of the experiment.

The aim of this article was to demonstrate the methodological approach. In future experiments systematic studies for further parameters like other chromatin types (e.g. euchromatin, ALU sequence regions etc.) in the environment or assignment to the follow-up proteins in the repair pathway (e.g. MRE11, Ku70, Ku80, 53BP1, Rad51 etc.) are necessary in order to understand the correlation of γ H2AX clusters and clusters formed by further recruited proteins during repair. Furthermore the application to other cell types, different repair times and radiation types (e.g. high LET ions, α -particles, β -particles etc.) would contribute to a conclusive knowledge of pathway choice and the correlation to

structure and topology. This will be subject of next years' investigations.

11.4 Materials and Methods

11.4.1 Sample Preparation

For the experiments SKBr3 cells were used, a well established and well characterized cell line in breast cancer research [304]. It has the advantage of fast growing and usually not reaching a complete confluency (about 80% only) so that localization microscopy in the culture dishes can be performed with less background and more precision [305].

As described in details [285] SkBr3 cells were grown on coverslips until about 80% confluency. The cells were washed in 1x phosphate-buffered saline buffer (PBS) with MgCl₂ (0.901 mM)/CaCl₂ (0.493 mM) for 5 min and fixed in 4% formaldehyde (in 1x PBS + Mg/Ca, freshly prepared from paraformaldehyde) for 10 min at 37°C. After washing three times in 1 PBS + Mg/Ca for 5 min, the cells were incubated in 0.2% Triton-X in 1x PBS + Mg/Ca 3 min for permeabilisation followed by additional washing three times and blocking in 2% BSA in 1 PBS + Mg/Ca for 30 min.

For labeling of heterochromatin antibodies against H3K9me₃ were used. H3K9me₃ is traditionally associated with non-coding parts of the genome. Recent investigations [306,307] have shown that H3K9me₃ is a key player in repressing lineage-inappropriate genes and shielding them from transcription. In contrast to other heterochromatin markers constitutive heterochromatin and tissue specific inactivated sites can be highlighted.

Incubation with the primary rabbit anti-histone H3 methylation side antibody (anti histone H3 tri methyl K9 - CHIP grade; Abcam plc, Cambridge, UK; concentration: 1.4 mg/L) in a humidified chamber at 37°C for 30 min and washing three times in 1 PBS + Mg/Ca on a shaker for 5 min was then followed by incubation with the secondary goat anti-rabbit IgG H&L (Alexa Fluor® 488) (Abcam plc, Cambridge, UK; concentration 4 mg/L) () in a humidified chamber at 37°C for 30 min and washing three times in 1x PBS + Mg/Ca on a shaker for 5 min. The specimen was again fixed in 2% formaldehyde at 37°C for 10 min and washed three times in 1 PBS + Mg/Ca on a shaker for 5 min. Labeling quality was checked by experiments using the secondary antibody without specific primary antibody.

Incubation with the second primary mouse anti-phospho-histone H2A.X (Ser139) antibody, clone JBW301 (Merck Chemicals; concentration 2 mg/L) in a humidified chamber at 4°C for 12 h and washing three times in 1x PBS + Mg/Ca on a shaker for 5 min was then followed by incubation with the secondary goat anti-mouse (Alexa Fluor® 568) antibody (Thermo Fisher Scientific; concentration 4 mg/L) in a humidified chamber at 37°C for 30 min and washing three times in 1x PBS + Mg/Ca on a shaker for 5 min.

Finally the chromatin was counterstained with 4',6-DiAmidin-2-PhenylIndol (DAPI; Sigma Aldrich) for 5 min and after washing twice in 1x PBS + Mg/Ca on a shaker for 5 min embedded in 15 µL ProlongGold embedding medium (ThermoFisher Scientific, Massachusetts, USA, ProLong Gold Antifade Mountant, P36930). After sealing the specimen can be stored at 4°C.

11.4.2 Single Molecule Localization Microscopy (SMLM)

The microscope was built at the Light Microscopy Facility of the German Cancer Research Center in Heidelberg and is described in detail in several publications [278,279,282]. For the experiments described here two lasers (excitation at 491 nm and 561 nm) were used

for the excitation of fluorescently labeled antibodies (green H3K9me3 for heterochromatin; red for γ H2AX). The laser intensity was 3 KW/cm² (491 nm) and 5 KW/cm² (561 nm); homogeneous illumination is important for localization microscopy because blinking of dye molecules due to reversible photo-bleaching is dependent on the laser intensity. A 100x oil immersion lens with a numerical aperture of NA = 1.46 is used. The emission light captured by the objective lens is imaged onto an EmCCD-camera. The exposure time was 100 ms per frame. Two thousand frames were captured in each channel.

11.4.3 Sample Irradiation

After culturing SkBr9 cells were irradiated by 6 MeV photons (dose rate of 3 Gy/min) at a medical linear accelerator (ARTISTE LB35) with doses of 0.1, 0.5, 1, 2, 4 and 8 Gy. SMLM images were acquired 30 min after the irradiation. For the data analysis the γ H2AX foci were selected irrespective of the dose with the aim to show that the considered foci properties are independent of the dose.

11.5 Conclusions

SMLM opens new perspectives into chromatin architecture from the micro- to the nano-scale and detailed insights into molecular arrangements as repair foci. Recent investigations have shown the advantages of this method for radiation research and cell biophysics. So far microscopic analyses are usually based on images and are applied after image processing. SMLM is not necessarily dependent on image processing since the result of data acquisition is a coordinate matrix of points precisely localized. This principle difference to so far mostly established procedures allows the application of novel data evaluation procedures and mathematical concepts. In this article, the geometrical analysis of γ H2AX foci towards sub-clusters was combined with persistent homology in order to classify clusters according to their heterochromatin distance. Topological characteristics of γ H2AX clusters were compared independently from cell nuclei and doses applied just according to the neighborhood to heterochromatin. The detected pointillist pattern has been transferred into barcode representations of connected components and holes (Betti numbers of dimension 0 and 1) and a similarity measure has been applied leading to a similar category of clusters associated to heterochromatin and a dissimilar category of clusters not associated to heterochromatin. This proof-of-concept approach opens up new possibilities for SMLM and for a rigorous comparison of point distributions obtained for compatible objects like repair foci and a measure of their similarity.

Chapter 12

Conclusion and Outlook

12.1 Short Summary of the Results

The motivation of the work presented in this thesis is to investigate a wide range of complex biological systems varying in their dynamics, length scales, structural complexity and functional purpose. The aim of this thesis is to analyze complex and big data from state-of-the-art experiments and to develop physical models in order to expand the current knowledge on structure-function relationships in the context of eukaryotic and bacterial cells.

Hi-C Data Analysis and Visualization Genome-wide chromosome conformation capture-based methods like Hi-C are now widely used. The resulting data, most of which is deposited in publicly accessible databases, is the starting point for the computational modeling of the three-dimensional architecture of a multiplicity of genomes. We showed how essential it is to know how reliable the underlying Hi-C data is. It is not only the number of captured reads, but also the distribution of these reads which is decisive for this question. Moreover, contact maps before and after balancing can differ significantly. It is therefore necessary to compare them and understand possible deviations before interpreting the results.

Data visualization is key to the interpretation of Hi-C experiments. We therefore developed a slim and interactive browser application capable of visualizing Hi-C contact maps alongside complementary data tracks. Besides Hi-C contact maps genome-wide data, such as ChIP-seq and RNA-seq, can be included in the layout. The application can be utilized for the visualization of any genomes including mammalian genomes. We have designed the tool in such a way that it can be both installed on a local host and exposed to general web via a reverse-proxy such as NGINX which has the advantage that the visualization interface can be embedded in a larger web application.

Hi-C Data Modeling In order to interpret data from Hi-C studies genome-wide contact probability maps need to be translated into models of functional 3D genome organization. Here, we first presented an overview of computational methods to analyze contact probability maps in terms of features such as the level and shape of compartmentalization. Next, we described approaches to modeling 3D genome organization based on Hi-C data. There are two different ways of modeling 3D organization of chromosomes: 3D reconstruction and polymer modeling. While the first generates the most likely 3D structure given the contact data amenable to visual inspection, polymer modeling supports clarifying hypotheses of chromosomal organization deduced from contact maps. We exemplified the difference between both approaches by using two case studies.

Domain Boundary Detection in Hi-C maps To interpret the results of Hi-C experiments it is essential to understand the structures within contact maps. To this end, we have developed a probabilistic graphical model to study the domain structure visible in Hi-C contact maps. This model is based on a symmetric energy model where the interaction parameters come from the normalized entries of the contact matrix. Here the contact matrix is interpreted as a graph.

Domain Formation by Dynamic Looping Though the existence of this intrachromosomal compartmentalization is proposed in all newly published results of Hi-C experiments, explanations from a theoretical point of view are scarce. We focused on the modeling of the experimental findings of both loop domains and topological domains, which, as opposed to the former, do not involve a closure to a loop. Loop domains can be readily simulated by statically adjusting the topology. Topological domains, on the other side, are characterized by a highly dynamic internal organization and can be modeled by assuming dynamic looping interactions accounting for this highly flexible internal structure. Besides eukaryotic genomes, bacterial chromosomes are also found to be compartmentalized into topological domains of increased contact probability that could potentially be explained by our model as well.

3D Bacterial Chromosome Organization We employed the role of loops on the 3D organization of bacterial genomes using Hi-C and live cell imaging of DNA loci of the *B. subtilis* genome. By forming insulator-like complexes, the DNA binding protein Rok loops the *B. subtilis* genome over large distances altering the overall genome structure. This biological mechanism is similar to insulator dependent long-range promoter-enhancer interaction in eukaryotes and shows the applicability of our dynamic loop formation model for prokaryotic genomes.

Self-organized Segregation of *E. coli* Replication Origins In a subsequent step, we have presented a quantitative explanation for positioning of the chromosomal origin of replication in *E. coli*. By analyzing the positioning and dynamics of ori and MukBEF foci in wild-type cells, we first showed that ori are attracted towards MukBEF foci. We could show how the self-organization of MukBEF complexes can position origins to their observed mid-cell and quarter-cell positions. We found excellent agreement with quantitative experimental measurements and confirm key predictions. In particular, we showed that oris exhibit biased motion towards MukBEF clusters, rather than mid-cell. Our findings suggest that MukBEF and oris act together as a self-organizing system in chromosome organization-segregation and introduces protein self-organization as an important consideration for future studies of chromosome dynamics.

Topological Data Analysis for Localization Microscopy In the final project, we have introduced a fundamentally new approach for super-resolution localization microscopy data analysis. This approach makes use of the advantage that the microscopy data can be evaluated without image production and image processing. It combines a geometrical evaluation based on distance and cluster analysis with persistence homology for similarity classification of repair cluster loci. Although the mathematical principles behind this approach are well established, it is the first time that topology has been used as biologically relevant criteria. The method is independent on a certain set of parameters

and inherently multi-scale and can categorize repair foci into structural classes. Here we have demonstrated a very first proof-of-concept experiment, in which we could show that the category of heterochromatin associated repair foci are highly similar in terms of both topology and geometry whereas clusters that are not associated with heterochromatin are completely dissimilar. This topological similarity was independent of the irradiation doses. Our developed method opens up new possibilities to categorize spatial organization of point patterns by parameterization of topological similarity.

12.2 Outlook

Hi-C and Eukaryotic Genome Organization 3D genome organization of eukaryotic cells and functional implications thereof have emerged as subjects undergoing intense study across many disciplines. State-of-the-art methods, such as Hi-C and super-resolution microscopy, have revealed that eukaryotic genomes are hierarchically organized into large compartments on the megabase scale consisting of topologically associated domains (TADs) on the kilobase scale. In vertebrates, the transcription factor CTCF forms loops or loop domains and those establish the 3D architecture of the genome together with compartmental domains. The dynamic formation and dissolution of chromatin loops may be responsible for establishing enhancer-promoter interactions and may introduce stochasticity into the transcription process. These considerations highlight the importance of the 3D organization of the genome and suggest that it is both a determinant and a consequence of its function.

Improvements to the Hi-C technology and decreasing sequencing costs led to an increase in the amount of information on genomic interactions and an enhanced resolution. Further development of the Hi-C technology towards nucleosome resolution [308,309] has the potential to understand the structure-function relationship on the molecular level [310]. Our developed methods for the visualization and comparison of Hi-C contact maps as well as the identification of fine structures such as loops within contact maps are important building blocks for this endeavor.

As Hi-C data is gathered using populations of millions of cells, Hi-C contact maps condense the average information of captured interactions among multiple genomic loci and hence do not include any information about cell-to-cell variability or dynamics. Indeed, the results of a recent study tracking the dynamics of CTCF loops and chromosomal domains from thousands of single-cell Hi-C contact maps show a substantial variation of 3D genome organization between individual nuclei [311]. Currently, single-cell Hi-C approaches are limited by genome coverage, and thus the achievable resolution for single-cell contact maps [311–313]. At the same time, imaging methods provide the resolution for the visualization of single-cell domain structures [183]. Coupled to computational modeling imaging and genomics technologies have the potential to reveal novel insights into the spatially and dynamically organized 3D structure of the genome inside cells [314].

The loop extrusion model [103,104] is actually very similar to our dynamic loop model: complexes of the proteins CTCF and cohesin bind to the DNA and form progressively larger loops or, in other words, extrude loops. However, there is no experimental confirmation for the loop extrusion process for cohesin in mammalian cells yet and there are many open questions concerning, among others, its energy consumption, DNA translocation speed and the relationship between transcription and SMC complex movement. Although current work in vitro [315] and in bacterial systems using similar SMC complexes [316] is promising, this model will only be accepted if direct evidence for the extrusion process is

obtained in the context of CTCF loops in vivo.

Thanks to massive improvements in throughput, ever-increasing amounts of genomic data are being produced and pose a serious challenge to data processing and subsequent analysis. Besides handling large volumes of sequencing data, the combination of many different types of genomic data is also a challenging task. The integration of bioinformatics tools into highly scalable and high-performance computational platforms, such as Apache Spark [317] or Apache Hadoop [318], is a possible solution to face these challenges [319]. The interactive visualization of large Hi-C contact maps is a prime example where a distributed general-purpose cluster-computing framework like Apache Spark offers significant advantages in the real-time handling of large data volumes.

Bacterial Genome Organization and Segregation Bacterial genomes are organized by nucleoid-associated DNA-binding proteins (NAPs) [42, 45, 47], DNA supercoiling [48, 139] and transcriptional regulatory interactions [167, 239, 320, 321]. In chapters 8 and 9, we have examined the impact of looping interactions on 3D genome organization and have investigated the role of the NAP Rok in driving dynamic domain formation by long-range interactions in *B. subtilis* using a combined approach of Hi-C and super-resolution live cell imaging.

Like CTCF, Rok was first identified as a transcriptional repressor protein in *B. subtilis* [189]. Although, Rok is known to activate transcription of certain genes [190], the mechanism remains unknown and thus represents an interesting topic for future research. Since Rok can alter the local chromosomal interactions at the RoVA sites, it could potentially influence gene expression, independent from its role as a transcription repressor. Furthermore, it is possible that other accessory proteins are involved in RoVa complex formation and their interaction. Our findings raise the question whether insulator dependent long-range promoter-enhancer interaction exist in bacteria and whether this can also regulate gene expression as observed in eukaryotes [184]. Answers to these open questions, could finally elucidate the role of the transcriptional regulatory network [322, 323] on the overall folding of the *B. subtilis* genome.

Beyond that, it is interesting to investigate the influence of NAPs on bacterial chromosome segregation during cell division. Although a wide spectrum of proteins and mechanisms have been suggested to facilitate chromosome segregation, there is no consensus solution to the problem. In chapter 10, we have contributed new insights into this issue in the case of chromosome segregation in *E. coli*. We showed how self-organization of the NAP MukBEF leads to positioning of the origin of replication at mid-cell which enables the proper initiation of replication. In our polymer simulations, we did not model MukBEF molecules explicitly, but incorporated them implicitly via a spatially dependent looping probability along the long axis of the nucleoid representing the MukBEF concentration profile. For future work, combining particle and polymer simulations, at least to whatever extent is feasible, may provide a deeper understanding of the system. For example, the incorporation of the NAP MatP, which binds to *matS* sites in the replication terminus region, could yield new knowledge. Because of its interaction with MukBEF, it displaces the latter from the terminus region [201] and hence restricts long-range DNA interactions between the terminus region and other regions of the chromosome [47]. Both effects may help position this region at mid-cell, while simultaneously encouraging the co-localization of *ori* with MukBEF [201].

Topological Data Analysis As super-resolution single molecule localization microscopy (SMLM) techniques for high throughput data acquisition [324] and sample labeling [325] improve, both automated and robust analytical methods are needed. The mathematical field of topological data analysis provides a powerful framework for structural analysis of SMLM data. The purpose of persistent homology is to describe the topological structure within pointillist datasets [62] and it can thus be applied to the analysis of SMLM data. We have developed a novel method and demonstrated its applicability to analyze and categorize DNA repair foci by means of topology. Our methodology can be applied to a wide range of topologically interesting questions providing fundamentally new insights into biological nano-structure inaccessible with existing tools. A future challenge for our method is the application to 3D super-resolution localization images. In the context of 3D datasets, topological features of dimension 2 are enclosed voids (as reminder: topological features of dimension 0 correspond to the number of connected components in the complex, topological features of dimension 1 are holes or loops). Furthermore, it would also be interesting to extend our method similar to [326] in order to incorporate the possibility of persistence-based clustering [327].

Acknowledgments

First, and foremost, I would like to thank my advisor, Prof. Dieter W. Heermann for the opportunity to do this PhD, for his guidance and his constant support throughout my time at the Institute for Theoretical Physics despite his function as Vice-Rector for International Relations.

I also want to thank Prof. Heinz Horner for kindly agreeing to act as a referee for this thesis. I would also like to thank Prof. Michael Hausmann and Prof. Kurt Roth for accepting to be members of my examination committee.

My special thanks go to Dr. Remus T. Dame for his ongoing professional assistance during my PhD as my second supervisor and the great support subject to our collaboration.

Many thanks go to Dr. Frédéric Crémazy for the many insightful discussions on our Hi-C experiments as well as on all kind of different topics. I am especially grateful for his friendship.

Thanks also go to Fatema Zahra Rashid, James Hancock, David Grainger and Ramon van der Valk for our fruitful collaboration.

This work was funded by a grant from the International Human Frontier Science Program Organization (RGP0014/2014). I also appreciate support by the Institute for Theoretical Physics and travel funding by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp).

Many thanks go to Sonja Bartsch, Melanie Steiert and Cornelia Merkel for managing all bureaucratic obstacles that pop up in the German university environment.

I would like to thank Dr. Jiyong Jia, Prof. Nestor Oiwa, Jörg Eisele, Dr. Lei Liu, Dr. Gabriell Maté, Dr. Yang Zhang, Wei Xiong, Dr. Fei Xing, Kunhe Li, Jan Müggenburg, Min Chu for the excellent atmosphere in our research group and our many conversations on all kind of different topics.

I am also grateful for the support of my family.

Last but not least, I am most grateful to my wife Katerina for her patience, encouragement and proofreading.

Conference/Workshop Participation

I have participated in the following conferences and workshops:

- Joint Workshop of Nankai University and Heidelberg University “Big Data”, November 11 – 14, 2015, Nankai University, Tianjin, China [talk]
- Annual Colloquium of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, November 30 – December 1, 2015, Altleiningen [poster]
- Conference “biophychrom2016 – The Biology and Physics of Bacterial Chromosome Organization”, September 5 – 8, 2016, Collège de France, Paris, France [poster]
- Annual Colloquium of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, December 1 – 2, 2016, Speyer [poster]
- 7th Workshop on Monte Carlo Methods and Big Data, February 10 – 11, 2017, Institute for Theoretical Physics, Heidelberg [talk]
- Joint Meeting of the Biophysical Society 61st Annual Meeting, February 11 – 15, 2017, New Orleans, Louisiana [poster]
- Annual Colloquium of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, November 30 – December 1, 2017, Altleiningen [poster]
- Conference “biophychrom2018 – The Biology and Physics of Bacterial Chromosome Organization”, June 4 – 6, 2018, Golden Tulip, Leiden, The Netherlands [poster]

Bibliography

- [1] National Human Genome Research Institute (NHGRI): Human Genome Project. URL: <https://www.genome.gov/human-genome-project>.
- [2] Esteban Toro and Lucy Shapiro. Bacterial Chromosome Organization and Segregation. *Cold Spring Harb. Perspect. Biol.*, 2(2):a000349, 2010. doi:10.1101/cshperspect.a000349.
- [3] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragooczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, 2009. doi:10.1126/science.1181369.
- [4] GEO. URL: <https://www.ncbi.nlm.nih.gov/geo/>.
- [5] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, 14(7):679–685, 2017. doi:10.1038/nmeth.4325.
- [6] M. Jordan Rowley and Victor G. Corces. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, 19(12):789–800, 2018. doi:10.1038/s41576-018-0060-8.
- [7] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. doi:10.1038/nature11082.
- [8] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012. doi:10.1038/nature11049.
- [9] Anjana Badrinarayanan, Tung B.K. Le, and Michael T. Laub. Bacterial Chromosome Organization and Segregation. *Annu. Rev. Cell Dev. Biol.*, 31(1):171–199, 2015. doi:10.1146/annurev-cellbio-100814-125211.
- [10] Stéphane Duigou and Frédéric Boccard. Long range chromosome organization in *Escherichia coli*: The position of the replication origin defines the non-structured

- regions and the Right and Left macrodomains. *PLOS Genetics*, 13(5):e1006758, 2017. doi:10.1371/journal.pgen.1006758.
- [11] Ivan Junier, Frédéric Boccard, and Olivier Espéli. Polymer modeling of the E. coli genome reveals the involvement of locus positioning and macrodomain structuring for the control of chromosome conformation and segregation. *Nucl. Acids Res.*, 42(3):1461–1473, 2014. doi:10.1093/nar/gkt1005.
- [12] Jonathan Livny, Yoshiharu Yamaichi, and Matthew K. Waldor. Distribution of Centromere-Like parS Sites in Bacteria: Insights from Comparative Genomics. *J. Bacteriol.*, 189(23):8693–8703, 2007. doi:10.1128/JB.01239-07.
- [13] Yinyin Li, Nichole K. Stewart, Anthony J. Berger, Seychelle Vos, Allyn J. Schoeffler, James M. Berger, Brian T. Chait, and Martha G. Oakley. Escherichia coli condensin MukB stimulates topoisomerase IV activity by a direct physical interaction. *Proc. Natl. Acad. Sci.*, 107(44):18832–18837, 2010. doi:10.1073/pnas.1008678107.
- [14] Emilien Nicolas, Amy L. Upton, Stephan Uphoff, Olivia Henry, Anjana Badrinarayanan, and David Sherratt. The SMC Complex MukBEF Recruits Topoisomerase IV to the Origin of Replication Region in Live Escherichia coli. *mBio*, 5(1), 2014. doi:10.1128/mBio.01001-13.
- [15] Quanli Wang, Jarad Niemi, Chee-Meng Tan, Lingchong You, and Mike West. Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy. *Cytom. A*, 77A(1):101–110, 2010. doi:10.1002/cyto.a.20812.
- [16] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: An open-source platform for biological-image analysis. *Nat. Methods*, 9(7):676–682, 2012. doi:10.1038/nmeth.2019.
- [17] Loris Nanni, Alessandra Lumini, and Sheryl Brahnem. Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intell. Med.*, 49(2):117–125, 2010. doi:10.1016/j.artmed.2010.02.006.
- [18] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. doi:10.1038/nature03001.
- [19] D E Comings. The rationale for an ordered arrangement of chromatin in the interphase nucleus. *Am. J. Hum. Genet.*, 20(5):440–460, 1968.
- [20] G. A. Rappold, T. Cremer, H. D. Hager, K. E. Davies, C. R. Müller, and T. Yang. Sex chromosome positions in human interphase nuclei as studied by in situ hybridization with chromosome specific DNA probes. *Hum. Genet.*, 67(3):317–325, 1984. doi:10.1007/BF00291361.
- [21] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002. doi:10.1126/science.1067799.

- [22] J T Finch and A Klug. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci.*, 73(6):1897–1901, 1976.
- [23] Eden Fussner, Mike Strauss, Ugljesa Djuric, Ren Li, Kashif Ahmed, Michael Hart, James Ellis, and David P. Bazett-Jones. Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Rep.*, 13(11):992–996, 2012. doi:10.1038/embor.2012.139.
- [24] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, 2(4):292–301, 2001. doi:10.1038/35066075.
- [25] Johan H. Gibcus and Job Dekker. The Hierarchy of the 3D Genome. *Mol. Cell*, 49(5):773–782, 2013. doi:10.1016/j.molcel.2013.02.011.
- [26] Xiaoli Weng and Jie Xiao. Spatial organization of transcription in bacterial cells. *Trends Genet.*, 30(7):287–297, 2014. doi:10.1016/j.tig.2014.04.008.
- [27] Peter Fraser. Transcriptional control thrown for a loop. *Curr. Opin. Genet. Dev.*, 16(5):490–495, 2006. doi:10.1016/j.gde.2006.08.002.
- [28] F. J. Iborra, A. Pombo, D. A. Jackson, and P. R. Cook. Active RNA polymerases are localized within discrete transcription "factories" in human nuclei. *J. Cell Sci.*, 109(6):1427–1436, 1996.
- [29] Cameron S. Osborne, Lyubomira Chakalova, Karen E. Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A. Mitchell, Susana Lopes, Wolf Reik, and Peter Fraser. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.*, 36(10):1065–1071, 2004. doi:10.1038/ng1423.
- [30] Heidi Sutherland and Wendy A. Bickmore. Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, 10(7):457–466, 2009. doi:10.1038/nrg2592.
- [31] Sonya Martin and Ana Pombo. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res.*, 11(5):461–470, 2003. doi:10.1023/A:1024926710797.
- [32] Jennifer A. Mitchell and Peter Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev.*, 22(1):20–25, 2008. doi:10.1101/gad.454008.
- [33] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, 148(3):458–472, 2012. doi:10.1016/j.cell.2012.01.010.
- [34] Jessica Zuin, Jesse R. Dixon, Michael I. J. A. van der Reijden, Zhen Ye, Petros Kolovos, Rutger W. W. Brouwer, Mariëtte P. C. van de Corput, Harmen J. G. van de Werken, Tobias A. Knoch, Wilfred F. J. van IJcken, Frank G. Grosveld, Bing Ren, and Kerstin S. Wendt. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci.*, 111(3):996–1001, 2014. doi:10.1073/pnas.1317788111.

- [35] Fulai Jin, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013. doi:10.1038/nature12644.
- [36] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, 2014. doi:10.1016/j.cell.2014.11.021.
- [37] Yin Shen, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V. Lobanenko, and Bing Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012. doi:10.1038/nature11243.
- [38] Elizabeth M. Blackwood and James T. Kadonaga. Going the Distance: A Current View of Enhancer Action. *Science*, 281(5373):60–63, 1998. doi:10.1126/science.281.5373.60.
- [39] Chunhui Hou and Victor G. Corces. Throwing transcription for a loop: expression of the genome in the 3D nucleus. *Chromosoma*, 121(2):107–116, 2011. doi:10.1007/s00412-011-0352-7.
- [40] Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012. doi:10.1038/nature11279.
- [41] Wenqin Wang, Gene-Wei Li, Chongyi Chen, X. Sunney Xie, and Xiaowei Zhuang. Chromosome organization by a nucleoid-associated protein in live bacteria. *Science*, 333(6048):1445–1449, 2011. doi:10.1126/science.1204697.
- [42] Remus T. Dame, Mariliis Tark-Dame, and Helmut Schiessel. A physical approach to segregation and folding of the caulobacter crescentus genome. *Mol. Microbiol.*, 82(6):1311–1315, 2011. doi:10.1111/j.1365-2958.2011.07898.x.
- [43] Remus T. Dame. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol. Microbiol.*, 56(4):858–870, 2005. doi:10.1111/j.1365-2958.2005.04598.x.
- [44] Martijn S. Luijsterburg, Malcolm F. White, Roel van Driel, and Remus Th. Dame. The major architects of chromatin: Architectural proteins in bacteria, archaea and eukaryotes. *Crit. Rev. Biochem. Mol. Biol.*, 43(6):393–418, 2008. doi:10.1080/10409230802528488.
- [45] Shane C. Dillon and Charles J. Dorman. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.*, 8(3):185–195, 2010. doi:10.1038/nrmicro2261.
- [46] Michèle Valens, Stéphanie Penaud, Michèle Rossignol, François Cornet, and Frédéric Boccard. Macrodome organization of the escherichia coli chromosome. *EMBO J.*, 23(21):4330–4341, 2004. doi:10.1038/sj.emboj.7600434.

- [47] Virginia S. Lioy, Axel Cournac, Martial Marbouty, Stéphane Duigou, Julien Mozziconacci, Olivier Espéli, Frédéric Boccard, and Romain Koszul. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell*, 172(4):771–783.e18, 2018. doi:10.1016/j.cell.2017.12.027.
- [48] Tung B. K. Le, Maxim V. Imakaev, Leonid A. Mirny, and Michael T. Laub. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, 342(6159):731–734, 2013. doi:10.1126/science.1242059.
- [49] Emanuela V. Volpi and Joanna M. Bridger. FISH glossary: an overview of the fluorescence in situ hybridization technique. *BioTechniques*, 45(4):385–409, 2008. doi:10.2144/000112811.
- [50] Elzo de Wit and Wouter de Laat. A decade of 3C technologies: Insights into nuclear organization. *Genes Dev.*, 26(1):11–24, 2012. doi:10.1101/gad.179804.111.
- [51] Job Dekker. The three 'C' s of chromosome conformation capture: Controls, controls, controls. *Nat. Methods*, 3(1):17–21, 2006. doi:10.1038/nmeth823.
- [52] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, 14(6):390–403, 2013. doi:10.1038/nrg3454.
- [53] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat. Genet.*, 38(11):1348–1354, 2006. doi:10.1038/ng1896.
- [54] Harmen J. G. van de Werken, Gilad Landan, Sjoerd J. B. Holwerda, Michael Hoichman, Petra Klous, Ran Chachik, Erik Splinter, Christian Valdes-Quezada, Yuva Öz, Britta A. M. Bouwman, Marjon J. A. M. Verstegen, Elzo de Wit, Amos Tanay, and Wouter de Laat. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods*, 9(10):969–972, 2012. doi:10.1038/nmeth.2173.
- [55] Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D. Green, and Job Dekker. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10):1299–1309, 2006. doi:10.1101/gr.5571506.
- [56] Mark A. Umbarger. Chromosome conformation capture assays in bacteria. *Methods*, 58(3):212–220, 2012. doi:10.1016/j.ymeth.2012.06.017.
- [57] C. D. M. Rodley, F. Bertels, B. Jones, and J. M. O’Sullivan. Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet. Biol.*, 46(11):879–886, 2009. doi:10.1016/j.fgb.2009.07.006.
- [58] Cedric Cagliero, Ralph S. Grand, M. Beatrix Jones, Ding J. Jin, and Justin M. O’Sullivan. Genome conformation capture reveals that the *Escherichia coli* chromosome is organized by replication and transcription. *Nucl. Acids Res.*, 41(12):6058–6071, 2013. doi:10.1093/nar/gkt325.

- [59] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013. doi:10.1038/nature12593.
- [60] Melissa J. Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G. Y. Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N. Ariyaratne, Vinsensius B. Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K. D. Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V. Desai, Jane S. Thomsen, Yew Kok Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G. Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T. Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009. doi:10.1038/nature08497.
- [61] Lusy Handoko, Han Xu, Guoliang Li, Chew Yee Ngan, Elaine Chew, Marie Schnapp, Charlie Wah Heng Lee, Chaopeng Ye, Joanne Lim Hui Ping, Fabianus Mulawadi, Eleanor Wong, Jianpeng Sheng, Yubo Zhang, Thompson Poh, Chee Seng Chan, Galih Kunarso, Atif Shahab, Guillaume Bourque, Valere Cacheux-Rataboul, Wing-Kin Sung, Yijun Ruan, and Chia-Lin Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638, 2011. doi:10.1038/ng.857.
- [62] Robert Ghrist. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, 45:61–75, 2008. doi:10.1090/S0273-0979-07-01191-3.
- [63] Afra J. Zomorodian. *Topology for Computing*. Cambridge University Press, Cambridge, 2005.
- [64] Alan D. Sokal. Monte carlo methods in statistical mechanics: Foundations and new algorithms. In Cecile DeWitt-Morette, Pierre Cartier, and Antoine Folacci, editors, *Functional Integration: Basics and Applications*, number 361 in NATO ASI Series, pages 131–192. Springer, 1997.
- [65] I. Carmesin and Kurt Kremer. The bond fluctuation method: a new effective algorithm for the dynamics of polymers in all spatial dimensions. *Macromolecules*, 21(9):2819–2823, 1988. doi:10.1021/ma00187a030.
- [66] H. P. Deutsch and K. Binder. Interdiffusion and self-diffusion in polymer mixtures: A monte carlo study. *J. Chem. Phys.*, 94(3):2294, 1991. doi:10.1063/1.459901.
- [67] Manfred Bohn, Dieter W. Heermann, Odilon Lourenço, and Claudette Cordeiro. On the influence of topological catenation and bonding constraints on ring polymers. *Macromolecules*, 43(5):2564–2573, 2010. doi:10.1021/ma902623u.
- [68] Hsiao-Ping Hsu, Wolfgang Paul, and Kurt Binder. Standard Definitions of Persistence Length Do Not Describe the Local “Intrinsic” Stiffness of Real Polymer Chains. *Macromolecules*, 43(6):3094–3102, 2010. doi:10.1021/ma902715e.

- [69] Miriam Fritsche, Songling Li, Dieter W. Heermann, and Paul A. Wiggins. A model for escherichia coli chromosome packaging supports transcription factor-induced DNA domain formation. *Nucl. Acids Res.*, 40(3):972–980, 2011. doi:10.1093/nar/gkr779.
- [70] P. Willett, J.M. Barnard, and G.M. Downs. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 38:983–996, 1998. doi:10.1021/ci9800211.
- [71] Jeff. Henrikson. Completeness and Total Boundedness of the Hausdorff Metric. MIT Undergraduate Journal of Mathematics, 1999.
- [72] Wendy A. Bickmore and Bas van Steensel. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell*, 152(6):1270–1284, 2013. doi:10.1016/j.cell.2013.02.001.
- [73] Martial Marbouty, Antoine Le Gall, Diego I. Cattoni, Axel Cournac, Alan Koh, Jean-Bernard Fiche, Julien Mozziconacci, Heath Murray, Romain Koszul, and Marcelo Nollmann. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cell*, 59(4):588–602, 2015. doi:10.1016/j.molcel.2015.07.020.
- [74] Xindan Wang, Tung B. K. Le, Bryan R. Lajoie, Job Dekker, Michael T. Laub, and David Z. Rudner. Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes Dev.*, 29(15):1661–1675, 2015. doi:10.1101/gad.265876.115.
- [75] Marie Trussart, Eva Yus, Sira Martinez, Davide Baù, Yuhei O. Tahara, Thomas Pengo, Michael Widjaja, Simon Kretschmer, Jim Swoger, Steven Djordjevic, Lynne Turnbull, Cynthia Whitchurch, Makoto Miyata, Marc A. Marti-Renom, Maria Lluch-Senar, and Luís Serrano. Defined chromosome structure in the genome-reduced bacterium *Mycoplasma pneumoniae*. *Nature Communications*, 8:14665, 2017. doi:10.1038/ncomms14665.
- [76] Mark A. Umbarger, Esteban Toro, Matthew A. Wright, Gregory J. Porreca, Davide Baù, Sun-Hae Hong, Michael J. Fero, Lihua J. Zhu, Marc A. Marti-Renom, Harley H. McAdams, Lucy Shapiro, Job Dekker, and George M. Church. The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. *Mol. Cell*, 44(2):252–264, 2011. doi:10.1016/j.molcel.2011.09.010.
- [77] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, 9(10):999–1003, 2012. doi:10.1038/nmeth.2148.
- [78] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, 43(11):1059–1065, 2011. doi:10.1038/ng.947.
- [79] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, 2012. doi:10.1186/1471-2164-13-436.

- [80] Wenyuan Li, Ke Gong, Qingjiao Li, Frank Alber, and Xianghong Jasmine Zhou. Hi-Corrector: A fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*, 31(6):960–962, 2015. doi:10.1093/bioinformatics/btu747.
- [81] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, 1967.
- [82] Manfred Bohn and Dieter W. Heermann. Diffusion-Driven Looping Provides a Consistent Framework for Chromatin Organization. *PLoS ONE*, 5(8):e12218, 2010. doi:10.1371/journal.pone.0012218.
- [83] Celine Lévy-Leduc, M. Delattre, T. Mary-Huard, and S. Robin. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, 30(17):i386–i392, 2014. doi:10.1093/bioinformatics/btu443.
- [84] Jie Chen, Alfred O. Hero, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, July 2016. doi:10.1093/bioinformatics/btw221.
- [85] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9:14, 2014. doi:10.1186/1748-7188-9-14.
- [86] Caleb Weinreb and Benjamin J. Raphael. Identification of hierarchical chromatin domains. *Bioinformatics*, 32(11):1601–1609, June 2016. doi:10.1093/bioinformatics/btv485.
- [87] Dieter W. Heermann, Andreas Hofmann, and Eva Weber. Domain Boundary Detection in Hi-C Maps: A Probabilistic Graphical Model Approach. *arXiv:1703.03656 [q-bio]*, 2017.
- [88] François Serra, Marco Di Stefano, Yannick G. Spill, Yasmina Cuartero, Michael Goodstadt, Davide Baù, and Marc A. Marti-Renom. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Letters*, 589(20, Part A):2987–2995, 2015. doi:10.1016/j.febslet.2015.05.012.
- [89] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, 2012. doi:10.1038/nbt.2057.
- [90] Mathieu Rousseau, James Fraser, Maria A. Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1):414, 2011. doi:10.1186/1471-2105-12-414.
- [91] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3D genome reconstruction from chromosomal contacts. *Nat. Methods*, 11(11):1141–1143, 2014. doi:10.1038/nmeth.3104.
- [92] Forrest W. Young. *Multidimensional Scaling: History, Theory, and Applications*. Psychology Press, 2013.

- [93] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data. *J. Comput. Biol.*, 20(11):831–846, 2013. doi:10.1089/cmb.2013.0076.
- [94] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S. Liu. Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS Comput. Biol.*, 9(1):e1002893, 2013. doi:10.1371/journal.pcbi.1002893.
- [95] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014. doi:10.1093/bioinformatics/btu268.
- [96] A. Yu Grosberg, S. K. Nechaev, and E. I. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. France*, 49(12):2095–2100, 1988. doi:10.1051/jphys:0198800490120209500.
- [97] Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, 19(1):37–51, 2011. doi:10.1007/s10577-010-9177-0.
- [98] Julio Mateos-Langerak, Manfred Bohn, Wim de Leeuw, Osdilly Giromus, Erik M. M. Manders, Pernette J. Verschure, Mireille H. G. Indemans, Hincó J. Gierman, Dieter W. Heermann, Roel van Driel, and Sandra Goetze. Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci.*, 106(10):3812–3817, 2009. doi:10.1073/pnas.0809501106.
- [99] Mariano Barbieri, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Josée Dostie, Ana Pombo, and Mario Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci.*, 109(40):16173–16178, 2012. doi:10.1073/pnas.1204799109.
- [100] Elnaz Alipour and John F. Marko. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucl. Acids Res.*, 40(22):11202–11212, 2012. doi:10.1093/nar/gks925.
- [101] Andreas Hofmann and Dieter W. Heermann. The role of loops on the order of eukaryotes and prokaryotes. *FEBS Letters*, 589(20, Part A):2958–2965, 2015. doi:10.1016/j.febslet.2015.04.021.
- [102] Mario Nicodemi and Ana Pombo. Models of chromosome structure. *Curr. Opin. Cell Biol.*, 28:90–95, 2014. doi:10.1016/j.ceb.2014.04.004.
- [103] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.*, 112(47):E6456–E6465, 2015. doi:10.1073/pnas.1518552112.
- [104] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.*, 15(9):2038–2049, 2016. doi:10.1016/j.celrep.2016.04.085.

- [105] Bekvaem Source Code. URL: <https://doi.org/10.11588/data/KGY0S6>.
- [106] Bekvaem Demonstration Server. URL: <http://pi306a.itp.uni-heidelberg.de/bekvaem>.
- [107] Tom Sexton and Giacomo Cavalli. The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*, 160(6):1049–1059, 2015. doi:10.1016/j.cell.2015.02.040.
- [108] Martin Franke, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, 2016. doi:10.1038/nature19800.
- [109] James T. Robinson, Douglass Turner, Neva C. Durand, Helga Thorvaldsdóttir, Jill P. Mesirov, and Erez Lieberman Aiden. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst.*, 6(2):256–258.e1, 2018. doi:10.1016/j.cels.2018.01.001.
- [110] François Serra, Davide Baù, Mike Goodstadt, David Castillo, Guillaume J. Filion, and Marc A. Marti-Renom. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Computational Biology*, 13(7):e1005665, 2017. doi:10.1371/journal.pcbi.1005665.
- [111] Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobel, Jacob M. Lubert, Scott B. Ouellette, Alaleh Azhir, Nikhil Kumar, Jeewon Hwang, Soohyun Lee, Burak H. Alver, Hanspeter Pfister, Leonid A. Mirny, Peter J. Park, and Nils Gehlenborg. Higlass: Web-based visual exploration and analysis of genome interaction maps. *bioRxiv*, 2017. doi:10.1101/121889.
- [112] Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, Mayank N. K. Choudhary, Yun Li, Ming Hu, Ross Hardison, Ting Wang, and Feng Yue. The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.*, 19(1):151, 2018. doi:10.1186/s13059-018-1519-9.
- [113] cooler. URL: <https://github.com/mirnylab/cooler>.
- [114] pyBigWig. URL: <https://github.com/deeptools/pyBigWig>.
- [115] Elphège P. Nora, Anton Goloborodko, Anne-Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. Targeted degradation of ctfc decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930 – 944.e22, 2017. doi:10.1016/j.cell.2017.05.004.
- [116] Flask. URL: <https://palletsprojects.com/p/flask/>.
- [117] Unicorn. URL: <https://unicorn.org/>.

- [118] NGINX. URL: <https://www.nginx.com/>.
- [119] Galip Gürkan Yardımcı and William Stafford Noble. Software tools for visualizing Hi-C data. *Genome Biology*, 18(1):26, 2017. doi:10.1186/s13059-017-1161-y.
- [120] Juicebox. URL: <https://www.aidenlab.org/juicebox/>.
- [121] Epigenome Browser. URL: <https://epigenomegateway.wustl.edu/>.
- [122] HiGlass. URL: <https://higlass.io/>.
- [123] TADkit. URL: <http://sgt.cnag.cat/3dg/tadkit/>.
- [124] 3D Genome Browser. URL: <http://promoter.bx.psu.edu/hi-c/>.
- [125] CLOC. URL: <https://github.com/AlDanial/cloc>.
- [126] Christina Kahramanoglou, Aswin S. N. Seshasayee, Ana I. Prieto, David Ibberson, Sabine Schmidt, Jurgen Zimmermann, Vladimir Benes, Gillian M. Fraser, and Nicholas M. Luscombe. Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucl. Acids Res.*, 39(6):2073–2091, 2011. doi:10.1093/nar/gkq934.
- [127] Mariliis Tark-Dame, Roel van Driel, and Dieter W. Heermann. Chromatin folding – from biology to polymer models and back. *J. Cell Sci.*, 124(6):839–845, 2011. doi:10.1242/jcs.077628.
- [128] Julie A. Cass, Nathan J. Kuwada, Beth Traxler, and Paul A. Wiggins. *Escherichia coli* Chromosomal Loci Segregate from Midcell with Universal Dynamics. *Biophys. J.*, 110(12):2597–2609, 2016. doi:10.1016/j.bpj.2016.04.046.
- [129] P. Hahnfeldt, J. E. Hearst, D. J. Brenner, R. K. Sachs, and L. R. Hlatky. Polymer models for interphase chromosomes. *Proc. Natl. Acad. Sci.*, 90(16):7854–7858, 1993. doi:10.1073/pnas.90.16.7854.
- [130] H. Yokota, G. van den Engh, J. E. Hearst, R. K. Sachs, and B. J. Trask. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol.*, 130(6):1239–1249, 1995. doi:10.1083/jcb.130.6.1239.
- [131] R. K. Sachs, G. van den Engh, B. Trask, H. Yokota, and J. E. Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proc. Natl. Acad. Sci.*, 92(7):2710–2714, 1995. doi:10.1073/pnas.92.7.2710.
- [132] H. Yokota, M. J. Singer, G. J. van den Engh, and B. J. Trask. Regional differences in the compaction of chromatin in human G0/G1 interphase nuclei. *Chromosome Res*, 5(3):157–166, 1997. doi:10.1023/A:1018438729203.
- [133] Dieter W. Heermann. Physical nuclear organization: loops and entropy. *Curr. Opin. Cell Biol.*, 23(3):332–337, 2011. doi:10.1016/j.ceb.2011.03.010.
- [134] Job Dekker. A closer look at long-range chromosomal interactions. *Trends Biochem. Sci.*, 28(6):277–280, 2003. doi:10.1016/S0968-0004(03)00089-6.

- [135] Kurt Binder and Dieter Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Graduate Texts in Physics. Springer-Verlag, 5 edition, 2010. doi: 10.1007/978-3-642-03163-2.
- [136] Stephanie Andrea Schalbetter, Anton Goloborodko, Geoffrey Fudenberg, Jon-Matthew Belton, Catrina Miles, Miao Yu, Job Dekker, Leonid Mirny, and Jonathan Baxter. SMC complexes differentially compact mitotic chromosomes according to genomic context. *Nat. Cell Biol.*, 19(9):1071–1080, 2017. doi:10.1038/ncb3594.
- [137] Remus T. Dame, Olga J. Kalmykova, and David C. Grainger. Chromosomal macrodomains and associated proteins: Implications for DNA organization and replication in gram negative bacteria. *PLoS Genet.*, 7(6):e1002123, 2011. doi: 10.1371/journal.pgen.1002123.
- [138] Hironori Niki, Yoshiharu Yamaichi, and Sota Hiraga. Dynamic organization of chromosomal DNA in escherichia coli. *Genes Dev.*, 14(2):212–223, 2000. doi: 10.1101/gad.14.2.212.
- [139] Lisa Postow, Christine D. Hardy, Javier Arsuaga, and Nicholas R. Cozzarelli. Topological domain structure of the escherichia coli chromosome. *Genes Dev.*, 18(14):1766–1779, 2004. doi:10.1101/gad.1207504.
- [140] Andrea Smallwood and Bing Ren. Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.*, 25(3):387–394, 2013. doi: 10.1016/j.ceb.2013.02.005.
- [141] Bas Tolhuis, Robert-Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and interaction between hypersensitive sites in the active β -globin locus. *Mol. Cell*, 10(6):1453–1465, 2002. doi:10.1016/S1097-2765(02)00781-5.
- [142] Takanori Amano, Tomoko Sagai, Hideyuki Tanabe, Yoichi Mizushima, Hiromi Nakazawa, and Toshihiko Shiroishi. Chromosomal dynamics at the shh locus: Limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, 16(1):47–57, 2009. doi:10.1016/j.devcel.2008.11.011.
- [143] Yoko Ito, Raffaella Nativio, and Adele Murrell. Induced DNA demethylation can reshape chromatin topology at the IGF2-h19 locus. *Nucl. Acids Res.*, 41(10):5290–5302, 2013. doi:10.1093/nar/gkt240.
- [144] Jennifer E. Phillips and Victor G. Corces. CTCF: Master weaver of the genome. *Cell*, 137(7):1194–1211, 2009. doi:10.1016/j.cell.2009.06.001.
- [145] Jennifer E. Phillips-Cremins, Michael E. G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S. K. Bell, Chin-Tong Ong, Tracy A. Hookway, Changying Guo, Yuhua Sun, Michael J. Bland, William Wagstaff, Stephen Dalton, Todd C. McDevitt, Ranjan Sen, Job Dekker, James Taylor, and Victor G. Corces. Architectural protein subclasses shape 3d organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, 2013. doi:10.1016/j.cell.2013.04.053.
- [146] Mariliis Tark-Dame, Hansjoerg Jerabek, Erik M. M. Manders, Dieter W. Heermann, and Roel van Driel. Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: Insight into chromatin folding by polymer modelling. *PLoS Comput. Biol.*, 10(10):e1003877, 2014. doi:10.1371/journal.pcbi.1003877.

- [147] Yuanyuan Li, Weichun Huang, Liang Niu, David M. Umbach, Shay Covo, and Leping Li. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. *BMC Genomics*, 14(1):1–12, 2013. doi:10.1186/1471-2164-14-553.
- [148] Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.*, 8(2):104–115, 2007. doi:10.1038/nrg2041.
- [149] Fabrizio Benedetti, Julien Dorier, Yannis Burnier, and Andrzej Stasiak. Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucl. Acids Res.*, 42(5):2848–2855, 2014. doi:10.1093/nar/gkt1353.
- [150] Catherine Naughton, Nicolaos Avlonitis, Samuel Corless, James G. Prendergast, Ioulia K. Mati, Paul P. Eijk, Scott L. Cockroft, Mark Bradley, Bauke Ylstra, and Nick Gilbert. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.*, 20(3):387–395, 2013. doi:10.1038/nsmb.2509.
- [151] Vittore F. Scolari and Marco Cosentino Lagomarsino. Combined collapse by bridging and self-adhesion in a prototypical polymer model inspired by the bacterial nucleoid. *Soft Matter*, 11(9):1677–1687, 2015. doi:10.1039/C4SM02434F.
- [152] Boryana Doyle, Geoffrey Fudenberg, Maxim Imakaev, and Leonid A. Mirny. Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Comput. Biol.*, 10(10):e1003867, 2014. doi:10.1371/journal.pcbi.1003867.
- [153] Laura Baranello, Fedor Kouzine, and David Levens. CTCF and cohesin cooperate to organize the 3d structure of the mammalian genome. *Proc. Natl. Acad. Sci.*, 111(3):889–890, 2014. doi:10.1073/pnas.1321957111.
- [154] Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay, and Suzana Hadjur. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.*, 10(8):1297–1309, 2015. doi:10.1016/j.celrep.2015.02.004.
- [155] Anton V. Persikov and Mona Singh. An expanded binding model for cys2his2 zinc finger protein-DNA interfaces. *Phys. Biol.*, 8(3):035010, 2011. doi:10.1088/1478-3975/8/3/035010.
- [156] Lei Liu and Dieter W. Heermann. The interaction of DNA with multi-cys2his2 zinc finger proteins. *J. Phys.: Condens. Matter*, 27(6):064107, 2015. doi:10.1088/0953-8984/27/6/064107.
- [157] Ramon A. van der Valk, Jocelyne Vreede, Frédéric Crémazy, and Remus T. Dame. Genomic looping: A key principle of chromatin organization. *J. Mol. Microbiol. Biotechnol.*, 24(5-6):344–359, 2014. doi:10.1159/000368851.
- [158] Vlad C. Seitan, Andre J. Faure, Ye Zhan, Rachel Patton McCord, Bryan R. Lajoie, Elizabeth Ing-Simmons, Boris Lenhard, Luca Giorgetti, Edith Heard, Amanda G.

- Fisher, Paul Flicek, Job Dekker, and Matthias Merckenschlager. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.*, 23(12):2066–2077, 2013. doi:10.1101/gr.161620.113.
- [159] Sevil Sofueva, Eitan Yaffe, Wen-Ching Chan, Dimitra Georgopoulou, Matteo Vietri Rudan, Hegias Mira-Bontenbal, Steven M Pollard, Gary P Schroth, Amos Tanay, and Suzana Hadjur. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.*, 32(24):3119–3129, 2013. doi:10.1038/emboj.2013.237.
- [160] Hansjoerg Jerabek and Dieter W. Heermann. Expression-dependent folding of interphase chromatin. *PLoS ONE*, 7(5):e37525, 2012. doi:10.1371/journal.pone.0037525.
- [161] Axel Cournac and Jacqueline Plumbridge. DNA looping in prokaryotes: Experimental and theoretical approaches. *J. Bacteriol.*, 195(6):1109–1119, 2013. doi:10.1128/JB.02038-12.
- [162] Pascal Reiss, Miriam Fritsche, and Dieter W. Heermann. Looped star polymers show conformational transition from spherical to flat toroidal shapes. *Phys. Rev. E*, 84(5):051910, 2011. doi:10.1103/PhysRevE.84.051910.
- [163] Manfred Bohn and Dieter W. Heermann. Repulsive Forces Between Looping Chromosomes Induce Entropy-Driven Segregation. *PLoS ONE*, 6(1):e14428, 2011. doi:10.1371/journal.pone.0014428.
- [164] Davide Marenduzzo, Cristian Micheletti, and Peter R. Cook. Entropy-driven genome organization. *Biophys. J.*, 90(10):3712–3721, 2006. doi:10.1529/biophysj.105.077685.
- [165] Peter R. Cook and Davide Marenduzzo. Entropic organization of interphase chromosomes. *J. Cell Biol.*, 186(6):825–834, 2009. doi:10.1083/jcb.200903083.
- [166] Suckjoon Jun and Andrew Wright. Entropy as the driver of chromosome segregation. *Nat. Rev. Microbiol.*, 8(8):600–607, 2010. doi:10.1038/nrmicro2391.
- [167] Miriam Fritsche and Dieter W. Heermann. Confinement driven spatial organization of semiflexible ring polymers: Implications for biopolymer packaging. *Soft Matter*, 7(15):6906–6913, 2011. doi:10.1039/C1SM05445G.
- [168] Mariliis Tark-Dame, Martijn S. Luijsterburg, Dieter W. Heermann, and Roel van Driel. Understanding genome function: Quantitative modeling of chromatin folding and chromatin-associated processes. In Karsten Rippe, editor, *Genome Organization and Function in the Cell Nucleus*, pages 535–555. Wiley-VCH Verlag GmbH & Co. KGaA, 2011.
- [169] Suckjoon Jun and Bela Mulder. Entropy-driven spatial organization of highly confined polymers: Lessons for the bacterial chromosome. *Proc. Natl. Acad. Sci.*, 103(33):12388–12393, 2006. doi:10.1073/pnas.0605305103.
- [170] Barbara Di Ventura, Benoît Knecht, Helena Andreas, William J. Godinez, Miriam Fritsche, Karl Rohr, Walter Nickel, Dieter W. Heermann, and Victor Sourjik. Chromosome segregation by the Escherichia coli Min system. *Mol. Syst. Biol.*, 9(1):686, 2013. doi:10.1038/msb.2013.44.

- [171] Angelo Rosa and Ralf Everaers. Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.*, 4(8):e1000153, 2008. doi:10.1371/journal.pcbi.1000153.
- [172] John F. Marko. Supercoiled and braided DNA under tension. *Phys. Rev. E*, 55(2):1758–1772, 1997. doi:10.1103/PhysRevE.55.1758.
- [173] Yang Zhang, Sebastian Isbaner, and Dieter W. Heermann. Mechanics of sister chromatids studied with a polymer model. *Front. Phys.*, 1:16, 2013. doi:10.3389/fphy.2013.00016.
- [174] Dieter W. Heermann. Mitotic chromosome structure. *Exp. Cell Res.*, 318(12):1381–1385, 2012. doi:10.1016/j.yexcr.2012.03.027.
- [175] Chin-Tong Ong and Victor G. Corces. CTCF: An architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, 15(4):234–246, 2014. doi:10.1038/nrg3663.
- [176] Vania Parelho, Suzana Hadjur, Mikhail Spivakov, Marion Leleu, Stephan Sauer, Heather C. Gregson, Adam Jarmuz, Claudia Canzonetta, Zoe Webster, Tatyana Nesterova, Bradley S. Cobb, Kyoko Yokomori, Niall Dillon, Luis Aragon, Amanda G. Fisher, and Matthias Merkenschlager. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell*, 132(3):422–433, 2008. doi:10.1016/j.cell.2008.01.011.
- [177] Xindan Wang, Hugo B. Brandão, Tung B. K. Le, Michael T. Laub, and David Z. Rudner. Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science*, 355(6324):524–527, 2017. doi:10.1126/science.aai8982.
- [178] Tatsuya Hirano. Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell*, 164(5):847–857, 2016. doi:10.1016/j.cell.2016.01.033.
- [179] Mark Albano, Wiep Klaas Smits, Linh T. Y. Ho, Barbara Kraigher, Ines Mandic-Mulec, Oscar P. Kuipers, and David Dubnau. The Rok Protein of Bacillus subtilis Represses Genes for Cell Surface and Extracellular Functions. *J. Bacteriol.*, 187(6):2010–2019, 2005. doi:10.1128/JB.187.6.2010-2019.2005.
- [180] Xindan Wang, Paula Montero Llopis, and David Z. Rudner. Organization and segregation of bacterial chromosomes. *Nat. Rev. Genet.*, 14(3):191–203, 2013. doi:10.1038/nrg3375.
- [181] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Sci. Adv.*, 5(4):eaaw1668, 2019. doi:10.1126/sciadv.aaw1668.
- [182] Suzana Hadjur, Luke M. Williams, Natalie K. Ryan, Bradley S. Cobb, Tom Sexton, Peter Fraser, Amanda G. Fisher, and Matthias Merkenschlager. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–413, 2009. doi:10.1038/nature08079.
- [183] Bogdan Bintu, Leslie J. Mateo, Jun-Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N. Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), 2018. doi:10.1126/science.aau1783.

- [184] Gang Ren, Wenfei Jin, Kairong Cui, Joseph Rodrigez, Gangqing Hu, Zhiying Zhang, Daniel R. Larson, and Keji Zhao. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol. Cell*, 67(6):1049–1058.e6, 2017. doi:10.1016/j.molcel.2017.08.026.
- [185] Charlotte A. Seid, Janet L. Smith, and Alan D. Grossman. Genetic and biochemical interactions between the bacterial replication initiator DnaA and the nucleoid-associated protein Rok in *Bacillus subtilis*. *Mol. Microbiol.*, 103(5):798–817, 2017. doi:10.1111/mmi.13590.
- [186] Bo Duan, Pengfei Ding, Timothy R. Hughes, William Wiley Navarre, Jun Liu, and Bin Xia. How bacterial xenogeneic silencer rok distinguishes foreign from self DNA in its resident genome. *Nucl. Acids Res.*, 46(19):10514–10529, 2018. doi:10.1093/nar/gky836.
- [187] Wiep Klaas Smits and Alan D. Grossman. The Transcriptional Regulator Rok Binds A+T-Rich DNA and Is Involved in Repression of a Mobile Genetic Element in *Bacillus subtilis*. *PLoS Genet.*, 6(11):e1001207, 2010. doi:10.1371/journal.pgen.1001207.
- [188] Frédéric G. Crémazy, Fatema-Zahra M. Rashid, James R. Haycocks, Lisa E. Lamberte, David C. Grainger, and Remus T. Dame. Determination of the 3D Genome Organization of Bacteria Using Hi-C. In Remus T. Dame, editor, *Bacterial Chromatin: Methods and Protocols*, Methods in Molecular Biology, pages 3–18. Springer, 2018. doi:10.1007/978-1-4939-8675-0_1.
- [189] Tran Thu Hoa, Pablo Tortosa, Mark Albano, and David Dubnau. Rok (YkuW) regulates genetic competence in *Bacillus subtilis* by directly repressing comK. *Mol. Microbiol.*, 43(1):15–26, 2002. doi:10.1046/j.1365-2958.2002.02727.x.
- [190] Ákos T. Kovács and Oscar P. Kuipers. Rok Regulates yuaB Expression during Architecturally Complex Colony Development of *Bacillus subtilis* 168. *J. Bacteriol.*, 193(4):998–1002, 2011. doi:10.1128/JB.01170-10.
- [191] Zhong Qian, Emilios K. Dimitriadis, Rotem Edgar, Prahathees Eswaramoorthy, and Sankar Adhya. Galactose repressor mediated intersegmental chromosomal connections in *Escherichia coli*. *Proc. Natl. Acad. Sci.*, 109(28):11336–11341, 2012. doi:10.1073/pnas.1208595109.
- [192] Antoine-Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell RNA-seq: Advances and future challenges. *Nucl. Acids Res.*, 42(14):8845–8860, 2014. doi:10.1093/nar/gku555.
- [193] Gaurav Dugar, Alexander Herbig, Konrad U. Förstner, Nadja Heidrich, Richard Reinhardt, Kay Nieselt, and Cynthia M. Sharma. High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple *Campylobacter jejuni* Isolates. *PLoS Genet.*, 9(5):e1003495, 2013. doi:10.1371/journal.pgen.1003495.
- [194] Andreas Hofmann, Jan Müggenburg, Frédéric Crémazy, and Dieter W. Heermann. Bekvaem: Integrative Data Explorer for Hi-C Data. *J. Bioinform. Gen.*, 2(11), 2019. doi:10.18454/jbg.2019.2.11.1.

- [195] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xi-anhong Jasmine Zhou. TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucl. Acids Res.*, 44(7):e70–e70, 2016. doi:10.1093/nar/gkv1505.
- [196] Henrik J. Nielsen, Jesper R. Ottesen, Brenda Youngren, Stuart J. Austin, and Flemming G. Hansen. The Escherichia coli chromosome is organized with the left and right chromosome arms in separate cell halves. *Mol. Microbiol.*, 62(2):331–338, 2006. doi:10.1111/j.1365-2958.2006.05346.x.
- [197] Hironori Niki, Yoshiharu Yamaichi, and Sota Hiraga. Dynamic organization of chromosomal DNA in Escherichia coli. *Genes & Dev.*, 12:213–223, 2000. doi:10.1101/gad.14.2.212.
- [198] Xindan Wang, Xun Liu, Christophe Possoz, and David J. Sherratt. The two Escherichia coli chromosome arms locate to separate cell halves. *Genes Dev.*, 20(13):1727–1731, 2006. doi:10.1101/gad.388406.
- [199] Mohan C. Joshi, Aude Bourniquel, Jay Fisher, Brian T. Ho, David Magnan, Nancy Kleckner, and David Bates. Escherichia coli sister chromosome separation includes an abrupt global transition with concomitant release of late-splitting intersister snaps. *Proc. Natl. Acad. Sci.*, 108(7):2765–2770, 2011. doi:10.1073/pnas.1019593108.
- [200] Henrik J. Nielsen, Yongfang Li, Brenda Youngren, Flemming G. Hansen, and Stuart Austin. Progressive segregation of the Escherichia coli chromosome. *Mol. Microbiol.*, 61(2):383–393, 2006. doi:10.1111/j.1365-2958.2006.05245.x.
- [201] Sophie Nolivos, Amy L. Upton, Anjana Badrinarayanan, Julius Müller, Katarzyna Zawadzka, Jakub Wiktor, Amber Gill, Lidia Arciszewska, Emilien Nicolas, and David Sherratt. MatP regulates the coordinated action of topoisomerase IV and MukBEF in chromosome segregation. *Nat. Commun.*, 7(1):1–12, 2016. doi:10.1038/ncomms10466.
- [202] Rodrigo Reyes-Lamothe, Christophe Possoz, Olessia Danilova, and David 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.
- [203] Xindan Wang, Rodrigo Reyes-Lamothe, and David J. Sherratt. Modulation of Escherichia coli sister chromosome cohesion by topoisomerase IV. *Genes Dev.*, 22(17):2426–2433, 2008. doi:10.1101/gad.487508.
- [204] Nathan J. Kuwada, Keith C. Cheveralls, Beth Traxler, and Paul A. Wiggins. Mapping the driving forces of chromosome structure and segregation in Escherichia coli. *Nucl. Acids Res.*, 41(15):7370–7377, 2013. doi:10.1093/nar/gkt468.
- [205] Sophie Nolivos and David Sherratt. The bacterial chromosome: Architecture and action of bacterial SMC and SMC-like complexes. *FEMS Microbiol Rev*, 38(3):380–392, 2014. doi:10.1111/1574-6976.12045.
- [206] Valentin V. Rybenkov, Viridiana Herrera, Zoya M. Petrushenko, and Hang Zhao. MukBEF, a Chromosomal Organizer. *J. Mol. Microbiol. Biotechnol.*, 24(5-6):371–383, 2014. doi:10.1159/000369099.

- [207] Tanneke Den Blaauwen, Arne Lindqvist, Jan Löwe, and Nanne Nanninga. Distribution of the Escherichia coli structural maintenance of chromosomes (SMC)-like protein MukB in the cell. *Mol. Microbiol.*, 42(5):1179–1188, 2001. doi:10.1046/j.1365-2958.2001.02691.x.
- [208] Katsufumi Ohsumi, Mitsuyoshi Yamazoe, and Sota Hiraga. Different localization of SeqA-bound nascent DNA clusters and MukF–MukE–MukB complex in Escherichia coli cells. *Mol. Microbiol.*, 40(4):835–845, 2001. doi:10.1046/j.1365-2958.2001.02447.x.
- [209] Olessia Danilova, Rodrigo Reyes-Lamothe, Marina Pinskaya, David Sherratt, and Christophe Possoz. MukB colocalizes with the oriC region and is required for organization of the two Escherichia coli chromosome arms into separate cell halves. *Mol. Microbiol.*, 65(6):1485–1492, 2007. doi:10.1111/j.1365-2958.2007.05881.x.
- [210] Anjana Badrinarayanan, Rodrigo Reyes-Lamothe, Stephan Uphoff, Mark C. Leake, and David J. Sherratt. In Vivo Architecture and Action of Bacterial Structural Maintenance of Chromosome Proteins. *Science*, 338(6106):528–531, 2012. doi:10.1126/science.1227126.
- [211] Anjana Badrinarayanan, Christian Lesterlin, Rodrigo Reyes-Lamothe, and David Sherratt. The Escherichia coli SMC Complex, MukBEF, Shapes Nucleoid Organization Independently of DNA Replication. *J. Bacteriol.*, 194(17):4669–4676, 2012. doi:10.1128/JB.00957-12.
- [212] Seán M. Murray and Victor Sourjik. Self-organization and positioning of bacterial protein clusters. *Nat. Phys.*, 13(10):1006–1013, 2017. doi:10.1038/nphys4155.
- [213] A. J. Koch and H. Meinhardt. Biological pattern formation: From basic mechanisms to complex structures. *Rev. Mod. Phys.*, 66(4):1481–1507, 1994. doi:10.1103/RevModPhys.66.1481.
- [214] Alan Mathison Turing. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. B. Biol. Sci.*, 237(641):37–72, 1952-08-14. doi:10.1098/rstb.1952.0012.
- [215] Julianne D. Halley and David A. Winkler. Consistent concepts of self-organization and self-assembly. *Complexity*, 14(2):10–17, 2008. doi:10.1002/cplx.20235.
- [216] Daniel J. Ferullo, Deani L. Cooper, Hayley R. Moore, and Susan T. Lovett. Cell cycle synchronization of Escherichia coli using the stringent response, with fluorescence labeling assays for DNA content and replication. *Methods*, 48(1):8–13, 2009. doi:10.1016/j.ymeth.2009.02.010.
- [217] G. L. Shevlyakov. On Robust estimation of a correlation coefficient. *J. Math. Sci.*, 83(3):434–438, 1997. doi:10.1007/BF02400929.
- [218] Pawel Zawadzki, Mathew Stracy, Katarzyna Ginda, Katarzyna Zawadzka, Christian Lesterlin, Achillefs N. Kapanidis, and David J. Sherratt. The Localization and Action of Topoisomerase IV in Escherichia coli Chromosome Segregation Is Coordinated by the SMC Complex, MukBEF. *Cell Rep.*, 13(11):2587–2596, 2015. doi:10.1016/j.celrep.2015.11.034.

- [219] Stephan Gruber. SMC complexes sweeping through the chromosome: Going with the flow and against the tide. *Curr. Opin. Microbiol.*, 42:96–103, 2018. doi:10.1016/j.mib.2017.10.004.
- [220] Robert Ietswaart, Florian Szardenings, Kenn Gerdes, and Martin Howard. Competing ParA Structures Space Bacterial Plasmids Equally over the Nucleoid. *PLoS Comput. Biol.*, 10(12):e1004009, 2014. doi:10.1371/journal.pcbi.1004009.
- [221] Seán M. Murray and Martin Howard. Center Finding in *E. coli* and the Role of Mathematical Modeling: Past, Present and Future. *J. Mol. Biol.*, 431(5):928–938, 2019. doi:10.1016/j.jmb.2019.01.017.
- [222] Takeshi Sugawara and Kunihiro Kaneko. Chemophoresis as a driving force for intracellular organization: Theory and application to plasmid partitioning. *Biophysics*, 7:77–88, 2011. doi:10.2142/biophysics.7.77.
- [223] Christiaan A. Miermans and Chase P. Broedersz. Bacterial chromosome organization by collective dynamics of SMC condensins. *J. Royal Soc. Interface*, 15(147):20180495, 2018. doi:10.1098/rsif.2018.0495.
- [224] Kyoko Matoba, Mitsuyoshi Yamazoe, Kouta Mayanagi, Kosuke Morikawa, and Sota Hiraga. Comparison of MukB homodimer versus MukBEF complex molecular architectures by electron microscopy reveals a higher-order multimerization. *Biochem. Biophys. Res. Commun.*, 333(3):694–702, 2005. doi:10.1016/j.bbrc.2005.05.163.
- [225] Samuel A. Isaacson. The Reaction-Diffusion Master Equation as an Asymptotic Approximation of Diffusion to a Small Target. *SIAM J. Appl. Math.*, 70(1):77–111, 2009. doi:10.1137/070705039.
- [226] Samuel A. Isaacson and Charles S. Peskin. Incorporating Diffusion in Complex Geometries into Stochastic Chemical Kinetics Simulations. *SIAM J. Sci. Comput.*, 28(1):47–74, 2006. doi:10.1137/040605060.
- [227] Radek Erban and S. Jonathan Chapman. Stochastic modelling of reaction–diffusion processes: Algorithms for bimolecular reactions. *Phys. Biol.*, 6(4):046001, 2009. doi:10.1088/1478-3975/6/4/046001.
- [228] Stefan Hellander and Linda Petzold. Reaction rates for a generalized reaction-diffusion master equation. *Phys. Rev. E*, 93(1):013307, 2016. doi:10.1103/PhysRevE.93.013307.
- [229] Samuel A. Isaacson. A convergent reaction-diffusion master equation. *J. Chem. Phys.*, 139(5):054101, 2013. doi:10.1063/1.4816377.
- [230] Anthony G. Vecchiarelli, Keir C. Neuman, and Kiyoshi Mizuuchi. A propagating ATPase gradient drives transport of surface-confined cellular cargo. *Proc. Natl. Acad. Sci.*, 111(13):4880–4885, 2014. doi:10.1073/pnas.1401025111.
- [231] Youngkyun Jung, Juin Kim, Suckjoon Jun, and Bae-Yeun Ha. Intrachain Ordering and Segregation of Polymers under Confinement. *Macromolecules*, 45(7):3256–3262, 2012. doi:10.1021/ma2025732.

- [232] Elena Minina and Axel Arnold. Induction of entropic segregation: The first step is the hardest. *Soft Matter*, 10(31):5836–5841, 2014. doi:10.1039/C4SM00286E.
- [233] Brenda Youngren, Henrik Jörk Nielsen, Suckjoon Jun, and Stuart Austin. The multi-fork Escherichia coli chromosome is a self-duplicating and self-segregating thermodynamic ring polymer. *Genes Dev.*, 28(1):71–84, 2014. doi:10.1101/gad.231050.113.
- [234] Thomas J. Lampo, Nathan J. Kuwada, Paul A. Wiggins, and Andrew J. Spakowitz. Physical Modeling of Chromosome Segregation in Escherichia coli Reveals Impact of Force and DNA Relaxation. *Biophys. J.*, 108(1):146–153, 2015. doi:10.1016/j.bpj.2014.10.074.
- [235] Longhua Hu, Anthony G. Vecchiarelli, Kiyoshi Mizuuchi, Keir C. Neuman, and Jian Liu. Brownian Ratchet Mechanism for Faithful Segregation of Low-Copy-Number Plasmids. *Biophys. J.*, 112(7):1489–1502, 2017. doi:10.1016/j.bpj.2017.02.039.
- [236] Longhua Hu, Anthony G. Vecchiarelli, Kiyoshi Mizuuchi, Keir C. Neuman, and Jian Liu. Directed and persistent movement arises from mechanochemistry of the ParA/ParB system. *Proc. Natl. Acad. Sci.*, 112(51):E7055–E7064, 2015. doi:10.1073/pnas.1505147112.
- [237] Hoong Chuin Lim, Ivan Vladimirovich Surovtsev, Bruno Gabriel Beltran, Fang Huang, Jörg Bewersdorf, and Christine Jacobs-Wagner. Evidence for a DNA-relay mechanism in ParABS-mediated chromosome segregation. *eLife*, 3:e02758, 2014. doi:10.7554/eLife.02758.
- [238] Ivan V. Surovtsev, Manuel Campos, and Christine Jacobs-Wagner. DNA-relay mechanism is sufficient to explain ParA-dependent intracellular transport and patterning of single and multiple cargos. *Proc. Natl. Acad. Sci.*, 113(46):E7268–E7276, 2016. doi:10.1073/pnas.1616118113.
- [239] Paul A. Wiggins, Keith C. Cheveralls, Joshua S. Martin, Robert Lintner, and Jané Kondev. Strong intranucleoid interactions organize the Escherichia coli chromosome into a nucleoid filament. *Proc. Natl. Acad. Sci.*, 107(11):4991–4995, 2010. doi:10.1073/pnas.0912062107.
- [240] Zoya M Petrushenko, Yuanbo Cui, Weifeng She, and Valentin V Rybenkov. Mechanics of DNA bridging by bacterial condensin MukBEF in vitro and in singulo. *EMBO J.*, 29(6):1126–1135, 2010. doi:10.1038/emboj.2009.414.
- [241] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, 157(3):624–635, 2014. doi:10.1016/j.cell.2014.02.033.
- [242] Philip K. Maini, Thomas E. Woolley, Ruth E. Baker, Eamonn A. Gaffney, and S. Seirin Lee. Turing’s model for biological pattern formation and the robustness problem. *Interface Focus*, 2(4):487–496, 2012. doi:10.1098/rsfs.2011.0113.
- [243] Mark B. Flegg. Smoluchowski Reaction Kinetics for Reactions of Any Order. *SIAM J. Appl. Math.*, 76(4):1403–1432, 2016. doi:10.1137/15M1030509.

- [244] Rupesh Kumar, Małgorzata Grosbart, Pearl Nurse, Soon Bahng, Claire L. Wyman, and Kenneth J. Mariani. The bacterial condensin MukB compacts DNA by sequestering supercoils and stabilizing topologically isolated loops. *J. Biol. Chem.*, 292(41):16904–16920, 2017. doi:10.1074/jbc.M117.803312.
- [245] Daniel T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem.*, 58(1):35–55, 2007. doi:10.1146/annurev.physchem.58.032806.104637.
- [246] Daniel T. Gillespie, Andreas Hellander, and Linda R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *J. Chem. Phys.*, 138(17):170901, 2013. doi:10.1063/1.4801941.
- [247] Michael A. Gibson and Jehoshua Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A*, 104(9):1876–1889, 2000. doi:10.1021/jp993732q.
- [248] Sean Mauch and Mark Stalzer. Efficient Formulations for Exact Stochastic Simulation of Chemical Systems. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8(1):27–35, 2011. doi:10.1109/TCBB.2009.47.
- [249] Hongyun Wang, Charles S. Peskin, and Timothy C. Elston. A Robust Numerical Algorithm for Studying Biomolecular Transport Processes. *J. Theor. Biol.*, 221(4):491–511, 2003. doi:10.1006/jtbi.2003.3200.
- [250] Yang Cao and Radek Erban. Stochastic Turing Patterns: Analysis of Compartment-Based Approaches. *Bull. Math. Biol.*, 76(12):3051–3069, 2014. doi:10.1007/s11538-014-0044-6.
- [251] M. Doi and S. F. Edwards. *The Theory of Polymer Dynamics*. International Series of Monographs on Physics. Oxford University Press, 1988.
- [252] Stella Stylianidou, Connor Brennan, Silas B. Nissen, Nathan J. Kuwada, and Paul A. Wiggins. SuperSegger: Robust image segmentation, analysis and lineage tracking of bacterial cells. *Mol. Microbiol.*, 102(4):690–700, 2016. doi:10.1111/mmi.13486.
- [253] Yang Zhang, Gabriell Máté, Patrick Müller, Sabina Hillebrandt, Matthias Krufczik, Margund Bach, Rainer Kaufmann, Michael Hausmann, and Dieter W. Heermann. Radiation Induced Chromatin Conformation Changes Analysed by Fluorescent Localization Microscopy, Statistical Physics, and Graph Theory. *PLoS ONE*, 10(6):e0128555, 2015. doi:10.1371/journal.pone.0128555.
- [254] Burkhard Jakob, Jörn Splinter, Sandro Conrad, Kay-Obbe Voss, Daniele Zink, Marco Durante, Markus Löbrich, and Gisela Taucher-Scholz. DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin. *Nucl. Acids Res.*, 39(15):6489–6499, 2011. doi:10.1093/nar/gkr230.
- [255] Guanghai Du, Guido A. Drexler, Werner Friedland, Christoph Greubel, Volker Hable, Reiner Krücken, Alexandra Kugler, Laura Tonelli, Anna A. Friedl, and Günther Dollinger. Spatial Dynamics of DNA Damage Response Protein Foci along the Ion Trajectory of High-LET Particles. *Radiat. Res.*, 176(6):706–715, 2011. doi:10.1667/RR2592.1.

- [256] Noel F. Lowndes and Geraldine W.-L. Toh. DNA Repair: The Importance of Phosphorylating Histone H2AX. *Curr. Biol.*, 15(3):R99–R102, 2005. doi:10.1016/j.cub.2005.01.029.
- [257] Markus Löbrich, Atsushi Shibata, Andrea Beucher, Anna Fisher, Michael Ensminger, Aaron A. Goodarzi, Olivia Barton, and Penny A. Jeggo. H2AX foci analysis for monitoring DNA double-strand break repair: Strengths, limitations and optimization. *Cell Cycle*, 9(4):662–669, 2010. doi:10.4161/cc.9.4.10764.
- [258] Emmy P. Rogakou, Duane R. Pilch, Ann H. Orr, Vessela S. Ivanova, and William M. Bonner. DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. *J. Biol. Chem.*, 273(10):5858–5868, 1998. doi:10.1074/jbc.273.10.5858.
- [259] Emmy P. Rogakou, Chye Boon, Christophe Redon, and William M. Bonner. Megabase Chromatin Domains Involved in DNA Double-Strand Breaks in Vivo. *J. Cell Biol.*, 146(5):905–916, 1999. doi:10.1083/jcb.146.5.905.
- [260] Martin Falk, Michael Hausmann, Emilie Lukasova, Abin Biswas, Georg Hildenbrand, Marie Davidkova, Evgeny Krasavin, Zdenek Kleibl, Iva Falkova, Lucie Jezkova, Lenka Stefancikova, Jan Sevcik, Michal Hofer, Alena Bacikova, Pavel Matula, Alla Boreyko, Jana Vachelova, Anna Michaelidesova, and Stanislav Kozubek. Determining Omics Spatiotemporal Dimensions Using Exciting New Nanoscopy Techniques to Assess Complex Cell Responses to DNA Damage: PART A - Radiomics. *Crit. Rev. Eukaryot. Gene Express.*, 24(3), 2014. doi:10.1615/CritRevEukaryotGeneExpr.2014010313.
- [261] Martin Falk, Michael Hausmann, Emilie Lukasova, Abin Biswas, Georg Hildenbrand, Marie Davidkova, Evgeny Krasavin, Zdenek Kleibl, Iva Falkova, Lucie Jezkova, Lenka Stefancikova, Jan Sevcik, Michal Hofer, Alena Bacikova, Pavel Matula, Alla Boreyko, Jana Vachelova, Anna Michaelidesova, and Stanislav Kozubek. Determining Omics Spatiotemporal Dimensions Using Exciting New Nanoscopy Techniques to Assess Complex Cell Responses to DNA Damage: Part B - Structuromics. *Crit. Rev. Eukaryot. Gene Express.*, 24(3), 2014. doi:10.1615/CritRevEukaryotGeneExpr.v24.i3.40.
- [262] Raphael Ceccaldi, Beatrice Rondinelli, and Alan D. D’Andrea. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol.*, 26(1):52–64, 2016. doi:10.1016/j.tcb.2015.07.009.
- [263] Sonali Bhattacharjee and Saikat Nandi. Choices have consequences: The nexus between DNA repair pathways and genomic instability in cancer. *Clin. Transl. Med.*, 5(1):45, 2016. doi:10.1186/s40169-016-0128-z.
- [264] Emil Mladenov, Simon Magin, Aashish Soni, and George Iliakis. DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation. *Sem. Cancer Biol.*, 37-38:51–64, 2016. doi:10.1016/j.semcancer.2016.03.003.
- [265] Agnes Schipler and George Iliakis. DNA double-strand-break complexity levels and their possible contributions to the probability for error-prone processing and repair pathway choice. *Nucl. Acids Res.*, 41(16):7589–7605, 2013-09-01. doi:10.1093/nar/gkt556.

- [266] Katerina Tsouroula, Audrey Furst, Melanie Rogier, Vincent Heyer, Anne Maglott-Roth, Alexia Ferrand, Bernardo Reina-San-Martin, and Evi Soutoglou. Temporal and Spatial Uncoupling of DNA Double Strand Break Repair Pathways within Mammalian Heterochromatin. *Mol. Cell*, 63(2):293–305, 2016. doi:10.1016/j.molcel.2016.06.002.
- [267] Rositsa Dueva and George Iliakis. Alternative pathways of non-homologous end joining (NHEJ) in genomic instability and cancer. *Transl. Cancer Res.*, 2(3):163–177–177, 2013. doi:10.21037/1152.
- [268] Ronja Biehs, Monika Steinlage, Olivia Barton, Szilvia Juhász, Julia Künzel, Julian Spies, Atsushi Shibata, Penny A. Jeggo, and Markus Löbrich. DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. *Mol. Cell*, 65(4):671–684.e5, 2017. doi:10.1016/j.molcel.2016.12.016.
- [269] Atsushi Shibata, Sandro Conrad, Julie Birraux, Verena Geuting, Olivia Barton, Amani Ismail, Andreas Kakarougkas, Katheryn Meek, Gisela Taucher-Scholz, Markus Löbrich, and Penny A Jeggo. Factors determining DNA double-strand break repair pathway choice in G2 phase. *EMBO J.*, 30(6):1079–1092, 2011. doi:10.1038/emboj.2011.27.
- [270] Emil Mladenov, Simon Magin, Aashish Soni, and George Iliakis. DNA Double-Strand Break Repair as Determinant of Cellular Radiosensitivity to Killing and Target in Radiation Therapy. *Front. Oncol.*, 3, 2013. doi:10.3389/fonc.2013.00113.
- [271] Markus Löbrich and Penny Jeggo. A Process of Resection-Dependent Nonhomologous End Joining Involving the Goddess Artemis. *Trends Biochem. Sci.*, 42(9):690–701, 2017. doi:10.1016/j.tibs.2017.06.011.
- [272] Ping Wang, Dexiao Yuan, Fei Guo, Xiaoyan Chen, Lin Zhu, Hang Zhang, Chen Wang, and Chunlin Shao. Chromatin remodeling modulates radiosensitivity of the daughter cells derived from cell population exposed to low- and high-LET irradiation. *Oncotarget*, 8(32):52823–52836, 2017. doi:10.18632/oncotarget.17275.
- [273] Dylan A. Reid, Sarah Keegan, Alejandra Leo-Macias, Go Watanabe, Natasha T. Strande, Howard H. Chang, Betul Akgol Oksuz, David Fenyo, Michael R. Lieber, Dale A. Ramsden, and Eli Rothenberg. Organization and dynamics of the non-homologous end-joining machinery during DNA double-strand break repair. *Proc. Natl. Acad. Sci.*, 112(20):E2575–E2584, 2015. doi:10.1073/pnas.1420115112.
- [274] Dylan A. Reid, Michael P. Conlin, Yandong Yin, Howard H. Chang, Go Watanabe, Michael R. Lieber, Dale A. Ramsden, and Eli Rothenberg. Bridging of double-stranded breaks by the nonhomologous end-joining ligation complex is modulated by DNA end chemistry. *Nucl. Acids Res.*, 45(4):1872–1878, 2017. doi:10.1093/nar/gkw1221.
- [275] Francesco Natale, Alexander Rapp, Wei Yu, Andreas Maiser, Hartmann Harz, Annina Scholl, Stephan Grulich, Tobias Anton, David Hörl, Wei Chen, Marco Durante, Gisela Taucher-Scholz, Heinrich Leonhardt, and M. Cristina Cardoso. Identification of the elementary structural units of the DNA damage response. *Nat. Commun.*, 8(1):1–18, 2017. doi:10.1038/ncomms15760.

- [276] Judith Reindl, Stefanie Girst, Dietrich W. M. Walsh, Christoph Greubel, Benjamin Schwarz, Christian Siebenwirth, Guido A. Drexler, Anna A. Friedl, and Günther Dollinger. Chromatin organization revealed by nanostructure of irradiation induced H2AX, 53BP1 and Rad51 foci. *Sci. Rep.*, 7(1):1–11, 2017. doi:10.1038/srep40616.
- [277] Yvonne Lorat, Christina U. Brunner, Stefanie Schanz, Burkhard Jakob, Gisela Taucher-Scholz, and Claudia E. Rübe. Nanoscale analysis of clustered DNA damage after high-LET irradiation by quantitative electron microscopy – The heavy burden to repair. *DNA Repair*, 28:93–106, 2015. doi:10.1016/j.dnarep.2015.01.007.
- [278] Yvonne Lorat, Sara Timm, Burkhard Jakob, Gisela Taucher-Scholz, and Claudia E. Rübe. Clustered double-strand breaks in heterochromatin perturb DNA repair after high linear energy transfer irradiation. *Radiother. Oncol.*, 121(1):154–161, 2016. doi:10.1016/j.radonc.2016.08.028.
- [279] Aleksander T. Szczurek, Kirti Prakash, Hyun-Keun Lee, Dominika J. Żurek Biesiada, Gerrit Best, Martin Hagmann, Jurek W. Dobrucki, Christoph Cremer, and Udo Birk. Single molecule localization microscopy of the distribution of chromatin using Hoechst and DAPI fluorescent probes. *Nucleus*, 5(4):331–340, 2014. doi:10.4161/nuc1.29564.
- [280] Michael Hausmann, Emma Wagner, Jin-Ho Lee, Gerrit Schrock, Wladimir Schaufler, Matthias Krufczik, Franziska Papenfuß, Matthias Port, Felix Bestvater, and Harry Scherthan. Super-resolution localization microscopy of radiation-induced histone H2AX-phosphorylation in relation to H3K9-trimethylation in HeLa cells. *Nanoscale*, 10(9):4320–4331, 2018. doi:10.1039/C7NR08145F.
- [281] Marion Eryilmaz, Eberhard Schmitt, Matthias Krufczik, Franziska Theda, Jin-Ho Lee, Christoph Cremer, Felix Bestvater, Wladimir Schaufler, Michael Hausmann, and Georg Hildenbrand. Localization Microscopy Analyses of MRE11 Clusters in 3D-Conserved Cell Nuclei of Different Cell Lines. *Cancers*, 10(1):25, 2018. doi:10.3390/cancers10010025.
- [282] Margund Bach, Claudia Savini, Matthias Krufczik, Christoph Cremer, Frank Rösl, and Michael Hausmann. Super-Resolution Localization Microscopy of -H2AX and Heterochromatin after Folate Deficiency. *Int. J. Mol. Sci.*, 18(8):1726, 2017. doi:10.3390/ijms18081726.
- [283] Jan Philipp Eberle, Alexander Rapp, Matthias Krufczik, Marion Eryilmaz, Manuel Gunkel, Holger Erfle, and Michael Hausmann. Super-Resolution Microscopy Techniques and Their Potential for Applications in Radiation Biophysics. In Holger Erfle, editor, *Super-Resolution Microscopy: Methods and Protocols*, Methods Mol. Biol., pages 1–13. Springer, 2017. doi:10.1007/978-1-4939-7265-4_1.
- [284] Maria E. Morales, Travis B. White, Vincent A. Streva, Cecily B. DeFreece, Dale J. Hedges, and Prescott L. Deininger. The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. *PLoS Genet.*, 11(3):e1005016, 2015. doi:10.1371/journal.pgen.1005016.
- [285] Matthias Krufczik, Aaron Sievers, Annkathrin Hausmann, Jin-Ho Lee, Georg Hildenbrand, Wladimir Schaufler, and Michael Hausmann. Combining Low Temperature Fluorescence DNA-Hybridization, Immunostaining, and Super-Resolution

- Localization Microscopy for Nano-Structure Analysis of ALU Elements and Their Influence on Chromatin Structure. *Int. J. Mol. Sci.*, 18(5):1005, 2017. doi:10.3390/ijms18051005.
- [286] Michael Hausmann, Nataša Ilić, Götz Pilarczyk, Jin-Ho Lee, Abiramy Logeswaran, Aurora Paola Borroni, Matthias Krufczik, Franziska Theda, Nadine Waltrich, Felix Bestvater, Georg Hildenbrand, Christoph Cremer, and Michael Blank. Challenges for Super-Resolution Localization Microscopy and Biomolecular Fluorescent Nano-Probing in Cancer Research. *Int. J. Mol. Sci.*, 18(10):2066, 2017. doi:10.3390/ijms18102066.
- [287] Christoph Cremer and Barry R. Masters. Resolution enhancement techniques in microscopy. *Eur. Phys. J. H*, 38(3):281–344, 2013. doi:10.1140/epjh/e2012-20060-1.
- [288] Patrick Müller, Niels A. Lemmermann, Rainer Kaufmann, Manuel Gunkel, Daniel Paech, Georg Hildenbrand, Rafaela Holtappels, Christoph Cremer, and Michael Hausmann. Spatial distribution and structural arrangement of a murine cytomegalovirus glycoprotein detected by SPDM localization microscopy. *Histochem. Cell Biol.*, 142(1):61–67, 2014-07-01. doi:10.1007/s00418-014-1185-2.
- [289] Manfred Bohn, Philipp Diesinger, Rainer Kaufmann, Yanina Weiland, Patrick Müller, Manuel Gunkel, Alexa von Ketteler, Paul Lemmer, Michael Hausmann, Dieter W. Heermann, and Christoph Cremer. Localization Microscopy Reveals Expression-Dependent Parameters of Chromatin Nanostructure. *Biophys. J.*, 99(5):1358–1367, 2010. doi:10.1016/j.bpj.2010.05.043.
- [290] Götz Pilarczyk, Ines Nesnidal, Manuel Gunkel, Margund Bach, Felix Bestvater, and Michael Hausmann. Localisation Microscopy of Breast Epithelial ErbB-2 Receptors and Gap Junctions: Trafficking after -Irradiation, Neuregulin-1, and Trastuzumab Application. *Int. J. Mol. Sci.*, 18(2):362, 2017. doi:10.3390/ijms18020362.
- [291] Christoph Cremer, Rainer Kaufmann, Manuel Gunkel, Sebastian Pres, Yanina Weiland, Patrick Müller, Thomas Ruckelshausen, Paul Lemmer, Fania Geiger, Sven Degenhard, Christina Wege, Niels A. W. Lemmermann, Rafaela Holtappels, Hilmar Strickfaden, and Michael Hausmann. Superresolution imaging of biological nanostructures by spectral precision distance microscopy. *Biotechnol. J.*, 6(9):1037–1051, 2011. doi:10.1002/biot.201100031.
- [292] P. Lemmer, M. Gunkel, Y. Weiland, P. Müller, D. Baddeley, R. Kaufmann, A. Urich, H. Eipel, R. Amberger, M. Hausmann, and C. Cremer. Using conventional fluorescent markers for far-field fluorescence localization nanoscopy allows resolution in the 10-nm range. *J. Microsc.*, 235(2):163–171, 2009. doi:10.1111/j.1365-2818.2009.03196.x.
- [293] P. Lemmer, M. Gunkel, D. Baddeley, R. Kaufmann, A. Urich, Y. Weiland, J. Reymann, P. Müller, M. Hausmann, and C. Cremer. SPDM: Light microscopy with single-molecule resolution at the nanoscale. *Appl. Phys. B*, 93(1):1, 2008. doi:10.1007/s00340-008-3152-x.
- [294] Russell E. Thompson, Daniel R. Larson, and Watt W. Webb. Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophys. J.*, 82(5):2775–2783, 2002. doi:10.1016/S0006-3495(02)75618-X.

- [295] Hendrik Deschout, Francesca Cella Zancacchi, Michael Mlodzianoski, Alberto Diaspro, Joerg Bewersdorf, Samuel T. Hess, and Kevin Braeckmans. Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods*, 11(3):253–266, 2014. doi:10.1038/nmeth.2843.
- [296] Gabriell Máté, Andreas Hofmann, Nicolas Wenzel, and Dieter W. Heermann. A topological similarity measure for proteins. *BBA Biomembranes*, 1838(4):1180–1190, 2014. doi:10.1016/j.bbamem.2013.08.019.
- [297] Yi-Li Feng, Ji-Feng Xiang, Na Kong, Xiu-Jun Cai, and An-Yong Xie. Buried territories: Heterochromatic response to DNA double-strand breaks. *Acta Biochim. Biophys. Sin.*, 48(7):594–602, 2016. doi:10.1093/abbs/gmw033.
- [298] Michael J. Kruhlak, Arkady Celeste, Graham Dellaire, Oscar Fernandez-Capetillo, Waltraud G. Müller, James G. McNally, David P. Bazett-Jones, and André Nussenzweig. Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *J. Cell Biol.*, 172(6):823–834, 2006. doi:10.1083/jcb.200510015.
- [299] Ali Sak, Dennis Kübler, Kristina Bannik, Michael Groneberg, and Martin Stuschke. Dependence of radiation-induced H2AX phosphorylation on histone methylation: Evidence from the chromatin immunoprecipitation assay. *Int. J. Radiat. Biol.*, 91(4):346–353, 2015. doi:10.3109/09553002.2015.997895.
- [300] Taehyun Ryu, Brett Spatola, Laetitia Delabaere, Katherine Bowlin, Hannah Hopp, Ryan Kunitake, Gary H. Karpen, and Irene Chiolo. Heterochromatic breaks move to the nuclear periphery to continue recombinational repair. *Nat. Cell Biol.*, 17(11):1401–1411, 2015. doi:10.1038/ncb3258.
- [301] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Min. Knowl. Discov.*, 2(2):169–194, 1998. doi:10.1023/A:1009745219419.
- [302] Bart Braden. The Surveyor’s Area Formula. *Coll. Math. J.*, 17(4):326–337, 1986. doi:10.1080/07468342.1986.11972974.
- [303] P Jaccard. Etude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaud. Sci. Nat.*, 37:547–579, 1901.
- [304] Linda W. Engel and Nathaniel A. Young. Human Breast Carcinoma Cells in Continuous Culture: A Review. *Cancer Res.*, 38:4327–4339, 1978.
- [305] Marc Lacroix and Guy Leclercq. Relevance of Breast Cancer Cell Lines as Models for Breast Tumours: An Update. *Breast Cancer Res. Treat.*, 83(3):249–289, 2004. doi:10.1023/B:BREA.0000014042.54925.cc.
- [306] Justin S. Becker, Dario Nicetto, and Kenneth S. Zaret. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. *Trends Genet.*, 32(1):29–41, 2016. doi:10.1016/j.tig.2015.11.001.
- [307] Robin C. Allshire and Hiten D. Madhani. Ten principles of heterochromatin formation and function. *Nat. Rev. Mol. Cell Biol.*, 19(4):229–244, 2018. doi:10.1038/nrm.2017.119.

- [308] Tsung-Han S. Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and Oliver J. Rando. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, 162(1):108–119, 2015. doi:10.1016/j.cell.2015.05.048.
- [309] Tsung-Han S. Hsieh, Geoffrey Fudenberg, Anton Goloborodko, and Oliver J. Rando. Micro-C XL: Assaying chromosome conformation from the nucleosome to the entire genome. *Nat. Methods*, 13(12):1009–1011, 2016. doi:10.1038/nmeth.4025.
- [310] Masae Ohno, Tadashi Ando, David G. Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs. *Cell*, 176(3):520–534.e25, 2019. doi:10.1016/j.cell.2018.12.014.
- [311] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, 2017. doi:10.1038/nature23001.
- [312] Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O’Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluís Morey, Miriam Sansó, Matthieu G. S. Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, 2017. doi:10.1038/nature21429.
- [313] Ilya M. Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikue Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017. doi:10.1038/nature21711.
- [314] Job Dekker, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, Stavros Lomvardas, Leonid A. Mirny, Clodagh C. O’Shea, Peter J. Park, Bing Ren, Joan C. Ritland Politz, Jay Shendure, and Sheng Zhong. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017. doi:10.1038/nature23884.
- [315] Mahipal Ganji, Indra A. Shaltiel, Shveta Bisht, Eugene Kim, Ana Kalichava, Christian H. Haering, and Cees Dekker. Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384):102–105, 2018. doi:10.1126/science.aar7831.
- [316] Hugo B. Brandão, Xindan Wang, Payel Paul, Aafke A. van den Berg, David Z. Rudner, and Leonid A. Mirny. RNA polymerases as moving barriers to condensin loop extrusion. *bioRxiv*, page 604280, 2019. doi:10.1101/604280.
- [317] Apache Spark. URL: <https://spark.apache.org/>.
- [318] Apache Hadoop. URL: <https://hadoop.apache.org/>.
- [319] Runxin Guo, Yi Zhao, Quan Zou, Xiaodong Fang, and Shaoliang Peng. Bioinformatics applications on Apache Spark. *GigaScience*, 7(8), 2018. doi:10.1093/gigascience/giy098.

- [320] Sarath Chandra Janga, Heladia Salgado, and Agustino Martínez-Antonio. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucl. Acids Res.*, 37(11):3680–3688, 2009. doi:10.1093/nar/gkp231.
- [321] Ruth Hershberg, Esti Yeger-Lotem, and Hanah Margalit. Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet.*, 21(3):138–142, 2005. doi:10.1016/j.tig.2005.01.003.
- [322] José P. Faria, Ross Overbeek, Ronald C. Taylor, Neal Conrad, Veronika Vonstein, Anne Goelzer, Vincent Fromion, Miguel Rocha, Isabel Rocha, and Christopher S. Henry. Reconstruction of the Regulatory Network for *Bacillus subtilis* and Reconciliation with Gene Expression Data. *Front. Microbiol.*, 7, 2016. doi:10.3389/fmicb.2016.00275.
- [323] Semen A. Leyn, Marat D. Kazanov, Natalia V. Sernova, Ekaterina O. Ermakova, Pavel S. Novichkov, and Dmitry A. Rodionov. Genomic Reconstruction of the Transcriptional Regulatory Network in *Bacillus subtilis*. *J. Bacteriol.*, 195(11):2463–2473, 2013. doi:10.1128/JB.00140-13.
- [324] Anne Beghin, Adel Kechkar, Corey Butler, Florian Levet, Marine Cabillic, Olivier Rossier, Gregory Giannone, Rémi Galland, Daniel Choquet, and Jean-Baptiste Sibarita. Localization-based super-resolution imaging meets high-content screening. *Nat. Methods*, 14(12):1184–1190, 2017. doi:10.1038/nmeth.4486.
- [325] Abdullah O. Khan, Victoria A. Simms, Jeremy A. Pike, Steven G. Thomas, and Neil V. Morgan. CRISPR-Cas9 Mediated Labelling Allows for Single Molecule Imaging and Resolution. *Sci. Rep.*, 7(1):1–9, 2017. doi:10.1038/s41598-017-08493-x.
- [326] Jeremy A. Pike, Abdullah O. Khan, Chiara Pallini, Steven G. Thomas, Markus Mund, Jonas Ries, Natalie S. Poulter, and Iain B. Styles. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. *Bioinformatics*, pages 1–8, 2019. doi:10.1093/bioinformatics/btz788.
- [327] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-Based Clustering in Riemannian Manifolds. *J. ACM*, 60(6):41:1–41:38, 2013. doi:10.1145/2535927.