

# Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for  
Mathematics

of the Ruperto-Carola University of Heidelberg  
Germany

for the degree of  
Doctor of Natural Sciences

Put forward by  
M.Sc. Songling Li  
Born in: P.R. China

Oral examination: July 4th, 2012



**From Genome-wide Probabilistic  
Functional Networks to a Model for  
Physical Chromosome Interactions**

**Referees: Prof. Dr. Deiter W. Heermann  
Prof. Dr. Michael Hausmann**



# Abstract

It's increasingly clear that the three-dimensional genome organization is dependent on the functional cell activities. In this thesis, the relationships between genome function, gene transcription activities for specific, and genome physical structure are investigated from different perspectives. From genome-wide probabilistic functional networks, we develop two different models for the three-dimensional genome architecture in *Escherichia coli* as well as in *Saccharomyces cerevisiae*.

To begin with, we explore and confirm the correlation between gene transcription and genome organization by investigating the chimeric transcripts induced by the 'transcription-induced chimerism' (TIC). Transcription-induced chimeras involve heterogeneous genes localized on different chromosomes, or on the same chromosome with a large genomic distance. When these genes are concurrently transcribed with a spatial proximity, their transcripts are more likely to be ligated as fusion products. By using bioinformatic approaches, we glean and validate chimeric transcripts from the Expressed Sequence Tag (EST) databases of human and mouse, and use them as the probe to identify physical contacts within the same chromosome or between different chromosomes. The chromosomal contact pattern extracted from the identified fusion transcripts is in agreement with the results from other independent experiments. The utilization of the chimeric transcripts in identifying chromosomal physical interactions shed light on the association between genome structure and genome function.

In a subsequent step, we postulate and test one prospective mechanism for the formation of chromosomal domains in the prokaryotic organism, *E. coli*. A genome folding model, which accounts for the role of gene transcriptional regulatory network (TRN) along with the nucleus confinement, is developed. Considering the stochastic nature of TF-promoter binding, we assume that transcription factors (TFs) and corresponding target genes (TGs) could stay in physical proximity for rapid targeting and more efficient regulation. We validate this model via numerical simulations and re-construct the ordering and the precise subnuclear distribution of the genetic loci that are experimentally screened. With this model, we contribute to a deeper understanding of the spatial chromosome organization in *E. coli*.

Last but not least, inspired by the findings from *E. coli*, we hold that a compatible interacting way between gene transcription and chromosome organization might exist in eukaryotes as well. Different from prokaryotes, eukaryotic organisms undertake their gene transcription and mRNA translation activities within different cell compartments. For this reason, we postulate a function-dependent genome structure model for budding yeast, in which we assume that genes with highly similar transcriptional control profiles might be recruited to the same subnuclear compartment enriched with specific transcription factors for their expression control. We test this idea with a simple eukaryotic organism, *S. cerevisiae*. The chromosomal interaction patterns and the folding behavior generated by this model are consistent with the experimental observations. We show that the transcriptional regulatory network has a close linkage with the genome organization in budding yeast, which is fundamental and instrumental to later studies on other more complex eukaryotes.

# Zusammenfassung

Es wird immer deutlicher, dass die dreidimensionale Struktur des Genoms von ihrer Funktion abhängt. In dieser Arbeit wird der Zusammenhang zwischen der Funktion des Genoms, Transkription von Genen für spezifische und der physikalischen Struktur des Genoms aus verschiedenen Blickwinkeln untersucht. Aus genomweiten, probabilistisch-funktionalen Netzwerken entwickeln wir verschiedene Modelle für die dreidimensionale Genomarchitektur in *Escherichia coli* und *Saccharomyces cerevisiae*.

Als Erstes erkunden und bestätigen wir die Korrelation zwischen Gentranskription und Genomorganisation durch die Untersuchung von chimären Transkripten welche durch ‘transcription-induced chromerism (TIC)’ hervorgerufen wird. TIC induzierte chimäre Transkripte beinhalten heterogene Gene welche sich auf verschiedenen Chromosomen befinden oder auf demselben Chromosom aber mit einem großen Abstand zueinander. Wenn diese Gene gleichzeitig in räumlicher Nähe zueinander transkribiert werden, so sind ihre Transkripte mit einer höheren Wahrscheinlichkeit zu einem Fusionsprodukt verbunden. Durch einen bioinformatischen Ansatz sammeln und bestätigen wir chimäre Transkripte aus Expression Sequence Tag (EST) Datenbanken für Menschen und Mäuse und benutzen sie als Testproben um räumliche Kontakte innerhalb des selben Chromosoms, oder zwischen verschiedenen Chromosomen zu identifizieren. Die chromosomalen Kontaktmuster, welche aus den fusion transcripts entnommen wurden, waren in übereinstimmung mit Resultaten aus verschiedenen Experimenten. Die Verwendung von chimären Transkripten zur Identifizierung von chromosomalen physikalischen Wechselwirkungen beleuchten die Verbindung zwischen Genomstruktur und genomische Funktion.

Danach postulieren und testen wir einen Mechanismus für die Bildung von chromosomalen Domänen im prokaryotischen Organismus *E. Coli*. Ein Genom-Faltungsmode, welches die Rolle von Gentranskriptionsnetzwerken ebenso wie die Beschränkung durch den Nukleus einbezieht, wird entwickelt. Die stochastische Natur der TF-Promotor-Bindung beachtend nehmen wir an, dass Transkriptionsfaktoren (TFs) und die entsprechenden Target Genes (TGs) für rasches Targeting und effizientere Regulierung in räumlicher Nähe zueinander bleiben könnten. Wir überprüfen das Model mittels numerischen Simulationen und re-

produzieren die Ordnung und präzise subnukleare Verteilung von genetischen Loci, welche experimentell gescreent wurden. Mit diesem Modell tragen wir zu einem tieferen Verständnis der räumlichen Organisation in *E. Coli* bei.

Schließlich vermuten wir, inspiriert durch die Ergebnisse für *E. Coli*, dass eine vergleichbare Art der Wechselwirkung zwischen Gentranskription und Genomorganisation auch in Eukaryoten existieren könnte. Anders als Prokaryoten führen eukaryotische Organismen ihre Gentranskription und mRNA Translation innerhalb von verschiedenen Zellkammern durch. Aus diesem Grund postulieren wir ein funktionsabhängiges Modell für die Genomstruktur in Backhefe, in welchem wir annehmen, dass in eukaryotischen Zellkernen, Gene mit sich stark ähnelnden Transkriptionskontrollprofilen zu den gleichen Subkammern rekrutiert werden, welche mit spezifischen Transkriptionsfaktoren für ihre Expressionskontrolle angereichert sind. Wir testen diese Idee an einem einfachen eukaryotischen Organismus, *S. cerevisiae*. Chromosomale Wechselwirkungsmuster und Faltungsverhalten welche durch dieses Modell erzeugt werden gleichen den aus Experimenten gewonnenen. Wir zeigen, dass das Transkriptionsregulationsnetzwerk eine starke Verbindung zur Genomorganisation in Backhefe besitzt, was fundamental und hilfreich für spätere Untersuchungen von anderen, komplexeren Eukaryoten ist.

# Publications Related to this Thesis

One part of this thesis has been published. The paper on chimeric ESTs has been submitted to BMC Genomics for peer review. The third paper on the list is in preparation and will be submitted soon. Information as of March 1st, 2012.

1. **S. Li**, and D. W. Heermann, Using Chimeric ESTs as the Reference to Identify Inter- and Intra-chromosomal Gene Loci Interactions. **(2012)**, *BMC Genomics*, under peer-review.

2. M. Fritsche, **S. Li**, D. W. Heermann and P. A. Wiggins, A model for Escherichia coli chromosome packaging supports transcription factor-induced DNA domain formation. *Nucleic Acids Research*, **(2012)**, *40*, 972-980. **DOI:** 10.1093/nar/gkr779

3. **S. Li**, D. W. Heermann, J. M. O'Sullivan, Transcriptional Regulatory Network Shapes the Genome Structure of *Saccharomyces cerevisiae*. **(2012)**, in preparation.



# Contents

|  |           |
|--|-----------|
| <b>Acknowledgments</b>   | <b>15</b> |
| <b>1 Aim and Structure of this Thesis</b>  | <b>17</b> |
| 1.1 Intention . . . . .  | 17        |
| 1.2 Structure of this Thesis . . . . .   | 19        |
| <b>2 Genome Organization: from Structure to Function</b>                                 | <b>23</b> |
| 2.1 Genome Organization in Prokaryotic and Eukaryotic Organisms                          | 23        |
| 2.1.1 DNA: The Genetic Material . . . . .  | 23        |
| 2.1.2 Genome Organization in Bacteria . . . . .  | 25        |
| 2.1.3 Chromosome Packaging in Bacteria . . . . .   | 28        |
| 2.1.4 Genome Organization in Eukaryotes . . . . .  | 31        |
| 2.1.5 Hierarchical Genome Organization in Eukaryotes . . . . .                           | 32        |
| 2.2 Genome Function and Genome Architecture . . . . .                                    | 37        |
| 2.2.1 Gene Expression and Transcriptional Regulatory Networks                            | 37        |
| 2.2.2 Transcription Factory as a Structure Organizer . . . . .                           | 42        |
| 2.2.3 Role of Chromosome Territories . . . . .   | 45        |
| 2.2.4 Gene Expression Regulation Mediated by Specific Chromosomal Interactions . . . . . | 48        |
| 2.2.5 3D Genome Organization Revealed by 3C-based Methods                                | 50        |
| 2.2.6 Structural Alteration and Genome Malfunction . . . . .                             | 52        |
| <b>3 Introduction to Polymer Physics</b>   | <b>55</b> |
| 3.1 Describing DNA as a Polymer . . . . .  | 55        |
| 3.1.1 The Random Walk (RW) Model . . . . .   | 57        |
| 3.1.2 The Worm-like Chain Model . . . . .  | 58        |
| 3.1.3 The Self-avoiding Walk (SAW) Model . . . . .                                       | 59        |

|          |   |            |
|----------|---|------------|
| 3.1.4    | The Equilibrium Globule Model . . . . .   | 61         |
| 3.1.5    | The Fractal Globule Model . . . . .   | 62         |
| 3.1.6    | The Random Loop Model . . . . .   | 65         |
| 3.2      | Application of Polymer Models in Elucidating Genome Function  | 68         |
| 3.2.1    | Role of Transcription Factory . . . . .   | 68         |
| 3.2.2    | Non-specific Chromatin Interactions . . . . .   | 69         |
| <b>4</b> | <b>Gene Transcription-Mediated Chromosomal Interactions</b>   | <b>71</b>  |
| 4.1      | Introduction . . . . .  | 71         |
| 4.2      | Material and Methods . . . . .  | 78         |
| 4.2.1    | Data Collecting and Processing . . . . .  | 78         |
| 4.2.2    | Sequence Alignment . . . . .  | 79         |
| 4.2.3    | Chimeric Transcripts Identification . . . . .   | 79         |
| 4.2.4    | Gene Mapping . . . . .  | 80         |
| 4.2.5    | Co-expression Pattern of the Chimeric Gene Partners . .   | 81         |
| 4.2.6    | Semantic Similarities of GO Terms Associated with Chimeric<br>Gene Partners . . . . .                         | 81         |
| 4.2.7    | Comparison with Normalized Hi-C Contact Pattern . . .   | 82         |
| 4.2.8    | Inter- and Intra-chromosomal Interactions Pattern Anal-<br>ysis . . . . .                                     | 82         |
| 4.3      | Results and Discussion . . . . .  | 84         |
| 4.3.1    | Chimeric Transcripts Identification and Validation . . . .  | 84         |
| 4.3.2    | Gene Mapping Results . . . . .  | 85         |
| 4.3.3    | Co-expression Pattern of the Chimeric Gene Partners . .   | 88         |
| 4.3.4    | Semantic Similarities of GO Terms Associated with Chimeric<br>Gene Partners . . . . .                         | 88         |
| 4.3.5    | Comparison with Normalized Hi-C Contact Pattern . . .   | 90         |
| 4.3.6    | Inter-chromosomal Interaction Pattern . . . . .   | 92         |
| 4.3.7    | Intra-chromosomal Interaction Pattern . . . . .   | 94         |
| 4.4      | Conclusion . . . . .  | 96         |
| <b>5</b> | <b>Transcriptional Regulatory Network Based Chromosomal Do-<br/>main Formation in <i>Escherichia coli</i></b> | <b>101</b> |
| 5.1      | Introduction . . . . .  | 102        |
| 5.2      | Material and Methods . . . . .  | 109        |

|          |  |            |
|----------|--|------------|
| 5.2.1    | The Gene Transcriptional Regulatory Network of <i>E. coli</i><br><i>K-12</i> . . . . .                         | 109        |
| 5.2.2    | Modeling and Simulation . . . . .  | 110        |
| 5.3      | Results and Discussion . . . . .   | 115        |
| 5.3.1    | The Gene Regulatory Network as a Mechanism for Do-<br>main Formation . . . . .                                 | 115        |
| 5.3.2    | The Distribution of Chromosomal Loop Sizes . . . . .   | 122        |
| 5.4      | Conclusion . . . . .   | 124        |
| <b>6</b> | <b>Transcriptional Regulatory Network Shapes the Genome Struc-<br/>ture of <i>Saccharomyces cerevisiae</i></b> | <b>125</b> |
| 6.1      | Introduction . . . . .   | 126        |
| 6.2      | Material and Methods . . . . .   | 129        |
| 6.2.1    | Neighboring Genes Selection . . . . .  | 129        |
| 6.2.2    | Modeling and Numerical Simulation . . . . .  | 130        |
| 6.2.3    | Experimental Data . . . . .  | 135        |
| 6.2.4    | Chromosomal Interactions Analysis . . . . .  | 137        |
| 6.2.5    | Territory Analysis . . . . .   | 138        |
| 6.3      | Results and Dissuasion . . . . .   | 138        |
| 6.3.1    | Neighboring Genes Selection . . . . .  | 138        |
| 6.3.2    | Validation of the Model . . . . .  | 138        |
| 6.3.3    | Comparison of Single Chromosome Folding Pattern . . .  | 141        |
| 6.3.4    | Comparison of Inter-chromosomal Contact Pattern . . .  | 145        |
| 6.3.5    | Territory Formation . . . . .  | 153        |
| 6.4      | Conclusion . . . . .   | 155        |
| <b>7</b> | <b>Conclusions and Extensions</b>  | <b>157</b> |
| 7.1      | A Summarization of the Results . . . . .   | 157        |
| 7.2      | Extensions . . . . .   | 160        |
|          | <b>Appendix</b>  | <b>165</b> |
|          | <b>References</b>  | <b>169</b> |



# Acknowledgments

I would like to appreciate my supervisor, Prof. Dr. Dieter W. Heermann, for his support throughout my doctoral studies. His advice and instructions are invaluable to my scientific development.

I am also very grateful to my co-supervisor, Prof. Dr. Joerg Langowski, for his assistance and suggestion.

I owe many thanks to Prof. Dr. Justin M. O'Sullivan for our collaboration on yeast genome organization.

I am also very grateful to our group member, Manfred Bohn, Miriam Fritsche, Yang Zhang, Hans-Jörg Jerabek, Gabriell Mate, Benoît Knecht.

Last but not least, I gratefully acknowledges funding from the Heinz-Goetze-Foundation and the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences.



# Chapter 1

## Aim and Structure of this Thesis

### 1.1 Intention

Several decades after the identification of the DNA double helix structure by James Watson and Francis Crick [1], we have up to now decoded complete genome sequences of a plethora of organisms. The ENCODE Project has enriched our knowledge of the human genome, in particular with respect to the identification of novel functional elements [2]. With an increasing number of functional modules and novel genes identified, we become gradually aware of the fact that the linear ordering of genes along the chromosome is far from enough to get a comprehensive understanding of the genome function. Genes and regulatory modules are the essential building-blocks of the genome. With pure sequence information, however, we can hardly have a clear picture on the genome physical structure. It is evident that the three-dimensional packaging of the genome is of crucial importance to its proper function. The genetic material included in a cell nucleus is thousands time longer than the cell. After being repeatedly folded and compacted in order to fit into the limited nuclear space of the cell, the structure of the genetic material is still dynamic. The genetic information stored along the chromatin fiber can be easily accessed during the nuclear processes, such as DNA replication, DNA repair, and gene expression. What's more remarkable is that a cell nucleus can reorganize its internal structure via a large-scale chromosome movement in order to adjust gene expression in response to internal or external stimuli. The question of how could the cell nuclear functions have an impact on the three-dimensional

genome architecture rise to the surface.

Nowadays the three-dimensional genome architecture has been emphasized and investigated for a better understanding of the genome operation. While the formation of chromatin loops at several different length scales have been confirmed to be the key to the genome physical structure, it's still not clear to what degree these chromatin loops are site-specific and functionally important. To answer this question, varieties of genome structure screening methods have been devised out. For the purpose of identifying interacting genomic loci, chromosome conformation capture (3C)-based techniques [3, 4, 5] have been developed. In order to appreciate the dynamic organization of the genome, sophisticated labeling methods coupled with high-resolution microscope have been employed to scrutinize the distribution and the movement of specific genetic loci [6, 7]. At the same time, the genome architectures of different model organisms have been studied (i.e. *Escherichia coli* (*E. coli*) [7], *Saccharomyces cerevisiae* (*S. cerevisiae*) [8], and human [9]). While these methods did disclose some interesting genome architectural features, the underlying mechanism that gives rise to them and the correlation between genome architecture and genome functional aspects are still obscure. On the other hand, most recent studies have revealed the interplay between genome spatial organization and gene transcriptional activities from prokaryotic organisms [10, 11] as well as eukaryotic organisms [12]. The formation of chromatin loops mediated by the spatial co-localization of regulatory elements and regulated genes has been postulated to facilitate gene transcriptional control [11]. In addition to the role of genome architecture in facilitating gene transcription, Montero Llopis *et al.* have also revealed the use of chromatin fiber in *E. coli* as a spatial organizer to aggregate mRNAs around their encoding genetic loci [13]. Such a compartmentalization tactic could constrain the distribution of regulatory proteins within discrete sub-domains, where they can execute their regulatory functions with efficiency.

DNA in essence is a polymer. For this reason, different polymer theories in association with modeling approaches have been widely applied to the investigation of chromatin fiber packaging. And the coarse-grained approach is of particular importance in that it can describe objects with a lower-resolution by neglecting molecular details to a desired degree, and at the same time highlight

the underlying organization principles that structure the genome. In the past, more emphases have been put on the understanding of the physical rules (i.e. polymer topology, and nucleus confinement) that might give rise to the experimentally observed folding behaviors of the genome. At present, the functional aspects of the genome that might also contribute to the genome architecture call our attention, as observed from the 3C-methods based studies, in which numerous site-specific physical chromosome interactions have been detected [14]. These chromatin interactions could be cell-type specific, or development-stage specific, which reflects a function-dependent genome organization motif. An abundance of experimental evidence with regard to the genome organization published so far can not be simply explained by just considering the physical properties of the nucleic acid. Other genome architectural motifs that are in charge of the genome functional aspects need to be elucidated.

In light of the progresses and the questions above-mentioned, we aim to postulate a novel genome architecture model involving both structural and functional aspects of the genome. By means of computational modeling approach, we endeavor to explain and re-construct the experimentally disclosed genome structural features. And more importantly, we attempt to shed more light on the relationship between spatial genome organization and genome functions both in prokaryotic and in eukaryotic cells.

## 1.2 Structure of this Thesis

In this thesis, the relationships between gene transcription and genome physical structure are investigated from different perspectives and in different organisms, like the prokaryotic model organism, *Escherichia coli*, simple eukaryotic organism, *Saccharomyces cerevisiae*, and human.

In **chapter 2**, the progress of our understanding of the genome physical structure is briefly introduced. At the same time, the genome organization under several different length scales is presented. In addition, some deep insights into the association between genome organization and genome function currently achieved, especially for gene transcription activities, are presented.

In **chapter 3**, some crucial and fundamental conceptions on polymer physics are introduced. Subsequently, several different polymer models are compared. These models describe the folding behavior of a polymer from different angles. To explain the complicated DNA folding behaviors as observed in experiments, miscellaneous models have been postulated from different perspectives. Therefore the applications of various polymer models in studying the folding and packaging behavior of the DNA molecule are introduced, and some enlightening findings are underlined.

In **chapter 4**, enlightened by the role of transcription factory [15] and the postulation of ‘Transcription Induced Chimerism (TIC)’ [16, 17], we investigate the physical interactions between genes that are involved in the production of chimeric transcripts. Seeing the possibility that transcription-induced fusion transcripts consisting of fragments from heterogeneous genes might indicate the co-localization of source genes in the same on-going transcription apparatus, we genome-widely screen the Expressed Sequence Tags (ESTs) databases of human and mouse to pick out those transcription-induced chimeras. By means of bioinformatics analyses, we verify the functional relevance of these chimeric gene pairs. Moreover, by using these chimeric gene pairs as the probe to identify interactions between genes, we extract the contact patterns for inter- and intra-chromosomal interaction, which are compatible with the results from Hi-C screening experiment [18].

In **chapter 5**, we investigate the role of gene transcriptional regulatory networks (TRNs) in shaping the genome physical structure in bacteria. We suggest a prospective underlying mechanism responsible for the formation of chromosomal domains in *E. coli*. We prove that the physical constraints originated from the co-localization between transcription factor genes and controlled genes are capable of facilitating the self-organization of the *E. coli* chromosome as a series of geometrically distinguishable domains.

Eukaryotes differ from prokaryotes with regard to the way how their genetic materials are stored and expressed. In **chapter 6**, we extend our genome function-dependent model and apply it to a simple eukaryotic organism, *S.*

*cerevisiae*, from which a more complicated genome organization has been observed. We posit that genes possessing highly similar transcription control profiles could be spatially assembled for the purpose of more efficient transcriptional regulation. According to this assumption, a polymer model in association with the structural features of budding yeast nucleus is put forward. The genome organization attributes disclosed by this model are compatible with the experimental observations.

Last but not least, **chapter 7** summarizes all the results presented in this thesis, and discusses some future challenges.



# Chapter 2

## Genome Organization: from Structure to Function

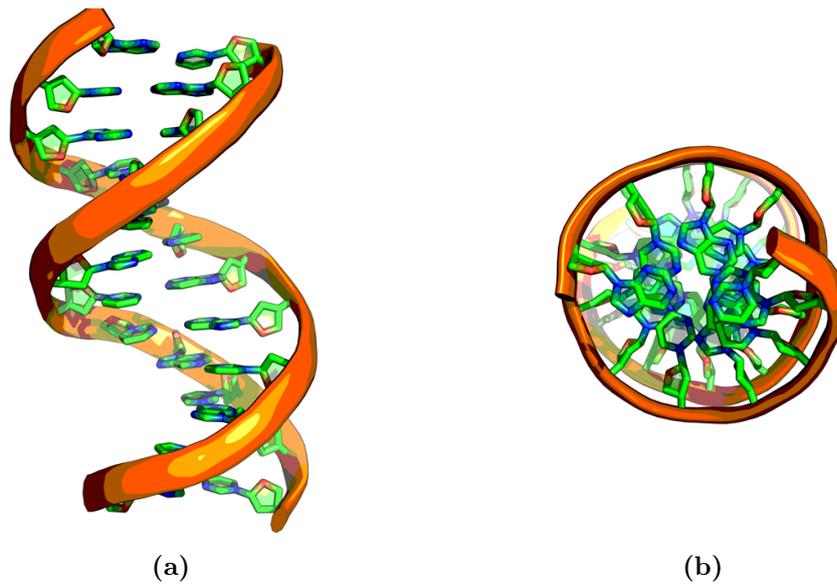
In this chapter, some biological knowledge about the genome packaging in prokaryotes and eukaryotes is introduced on the basis of our current understanding. Additionally, an evaluation of 3C-based methods to capture DNA interactions are introduced. Eventually, some evidence on the function-dependent genome architecture is discussed. For the reader, who is already familiar with these topics, can skip this chapter.

### 2.1 Genome Organization in Prokaryotic and Eukaryotic Organisms

#### 2.1.1 DNA: The Genetic Material

The central dogma of molecular biology stating that the genetic information included in the genome irreversibly flows from DNA to RNA and eventually to protein, was first postulated by Crick in 1958 [1]. No life form has been found that violates this rule, except retroviruses. As one type of macromolecule, the pivotal position of the DNA molecule as an information-carrier has been ascertained. Except for RNA viruses, deoxyribonucleic acid (DNA for abbreviation) is employed by all currently known organisms as their genetic material to reserve the information necessary for constructing structural and catalytic

apparatuses. DNA is made of four basic nucleotides, namely adenine(A), cytosine(C), guanine(G), and thymine(T) [19, 20]. It is the order of the nucleotides along the chain that encodes genetic information. A single-stranded DNA chain is composed of nucleotides aligned with the same direction and covalently connected tail by head, which give a DNA molecule a chemical polarity due to its sugar-phosphate backbone with 3' hydroxyl and 5' phosphate ends. A double-stranded DNA chain, organized in the form of nucleic acid double helix, is composed of two single-stranded chains, that are coiled round a common axis and aligned towards opposite directions with two complementary nucleotides positioned side by side and connected via non-covalent hydrogen bonds. Two nucleotides aligned on opposite complementary DNA chains is called a base pair (Figure 2.1). A naked double-stranded DNA molecule is very slender and fragile, approximately 1.0 nm in radius, and 3.4 nm for one helical turn. In living creatures, DNAs seldom exist as free polymers. In prokaryotic organisms, specific DNA-binding proteins exist, which are responsible for the condensation of the DNA molecule. In eukaryotic organisms, the resulting DNA-protein complex consisting of naked DNA and histone proteins is termed chromatin. Chromatin fibers with the help of other DNA-bound proteins can be further compacted into chromosomes. Circular chromosome is most frequently observed in prokaryotic organisms. In eukaryotic organisms, their genetic material is usually organized into several linear chromosomes. Chromosomes from various organisms or even from the same one differ tremendously in size. The longest human chromosome (Chromosome I) is around 220 Mb in length, representing about 8% of the total DNA in human cells and encodes about 4,220 genes. If the genetic material included in the human nucleus is fully relaxed, it will be in meters long. A human nucleus, by contrast, is just 1.7  $\mu\text{m}$  in diameter. As we can see, chromosomes must be well structured and extremely compacted in order to fit into the limited volume of a micrometer sized nucleus. DNA packaging is even more remarkable in the cell undergoing mitosis or meiosis, where individual chromosomes can be well visualized under a light microscope.

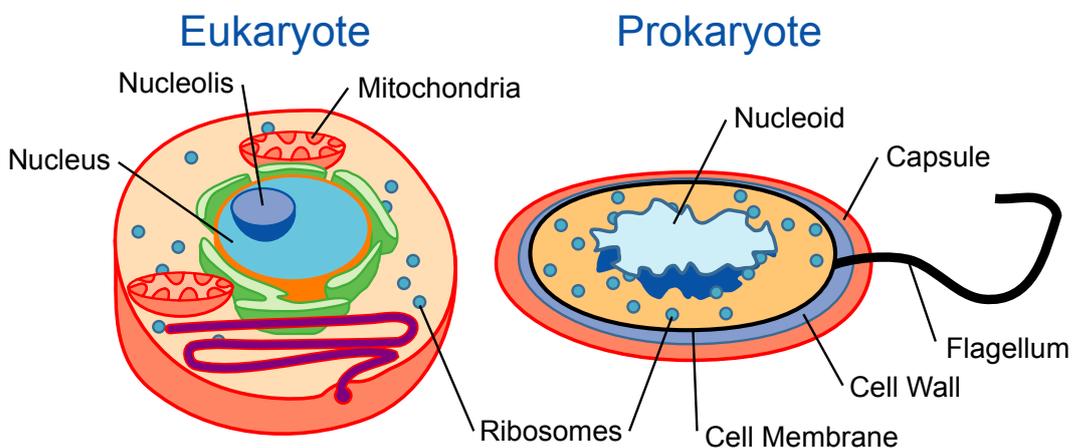


**Figure 2.1: Illustration of the double helical structure of the DNA molecule.** A fragment of ‘B-DNA’ is shown from different angles. (a) for horizontal view, and (b) for vertical view. DNA of this type is observed most frequently in living organisms. The backbone of DNA, shown as two helical ribbons in orange, are made of alternating sugar and phosphate groups. Nucleobases are connected to the backbone and extend towards interior. Other possible DNA structures include ‘A-DNA’ and ‘Z-DNA’.

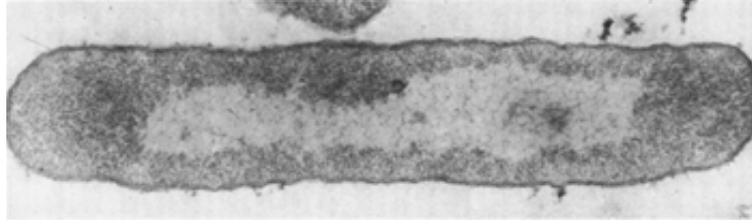
### 2.1.2 Genome Organization in Bacteria

In prokaryotic cells, genetic material and organizer proteins responsible for the DNA condensation are enclosed inside the nucleoid, a morphologically distinct structure floating inside the cell [21], as shown in Figure 2.2. Genetical material is visible as single pretty compact mass or several clumps, that occupy around one third of the cell volume (see Figure 2.3). Inside the nucleoplasm, numerous fine filaments are aggregated. Though they look randomly tangled, some dedicated structures do exist. If a bacterium cell is gently lysed, and further treated with reagents acting on RNA or proteins, the genetic material enclosed will be spread out in the form of loops of a fiber. As we can see from Figure 2.4, the *M. lysodeikticus* chromosome is organized into plenty of smaller loops and domains, which are discrete structures and independent of other domains. If a condensed chromosome is fully dissolved, an unfolded, cir-

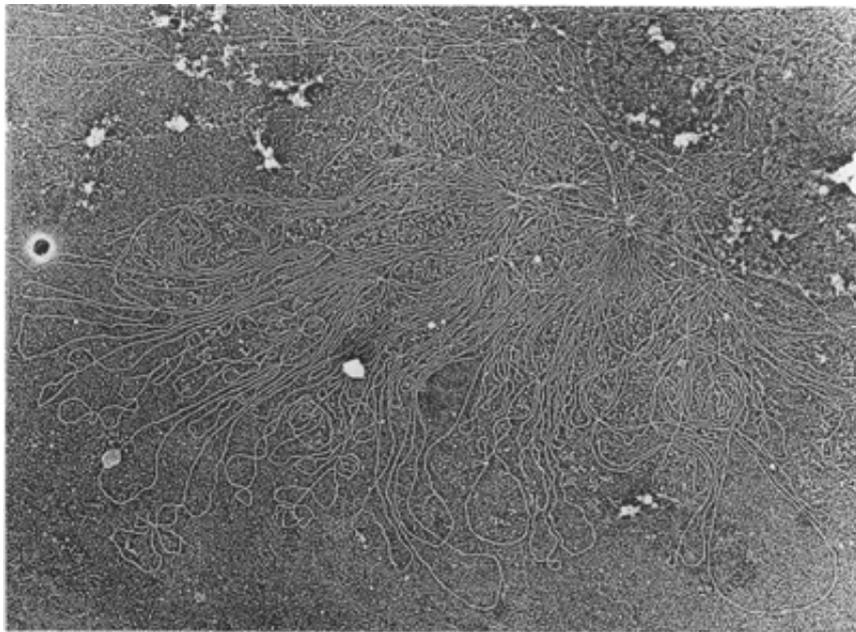
cular chromosome can be observed (see Figure 2.5). In order to achieve such a highly compact organization as illustrated in Figure 2.3, individual chromosome domains (loops) are negatively supercoiled. Some chemical reagents can destruct such an organization by intercalating between covalently bonded base pairs (i.e. ethidium bromide), or by introducing a nick in one strand of a DNA double helix (i.e. DNAase). However, tension releasing caused by the change on one domain will not affect others, since secured loop base will not permit rotation extending from one domain to others. The presence of individually independent domains in the bacterial genome allow it to differentially control the packaging status of different regions, which is crucial to the cellular activities, such as DNA duplication, DNA reparation, and gene expression.



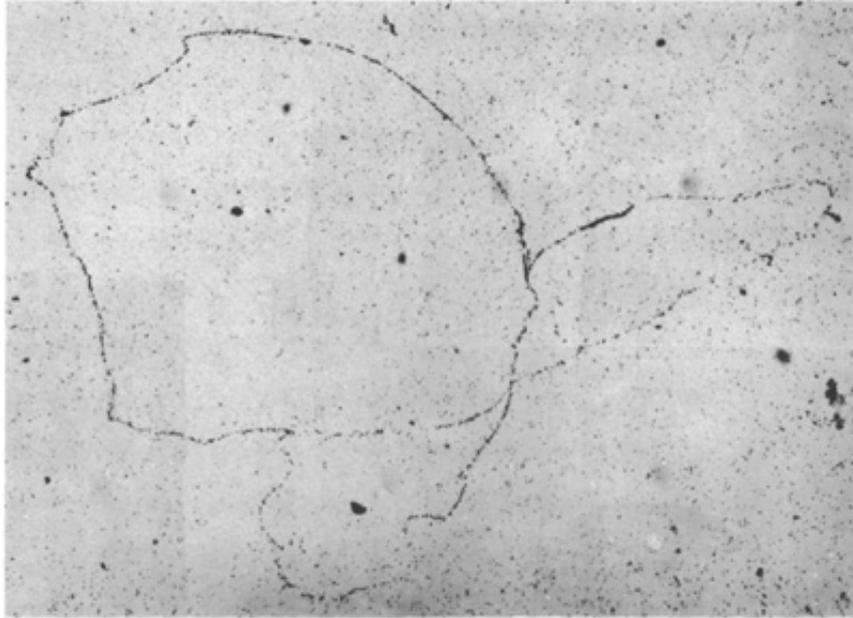
**Figure 2.2: Structural differences between eukaryotic and prokaryotic cells.** The lack of membrane-coated cell organelles is the main feature that distinguishes prokaryotic from eukaryotic cells. Genetic material is stored in the cell nucleus of eukaryotic cells. The counterpart of this structure in prokaryotic cells is named nucleoid, which is not surrounded by a nuclear membrane. Image adapted from [22].



**Figure 2.3: Bacterial nucleoid shown as a compact mass.** Nucleoplasm corresponds to the low contrast portion in this thin section of *Bacillus megaterium*. Image adapted from [21].



**Figure 2.4: Isolated protoplasm from lysed *Micrococcus lysodeikticus* with DNA filaments protruding outwards.** After being lysed, the genetic material included inside the cell is spread out in the form of loops of a fiber. Image adapted from [21].

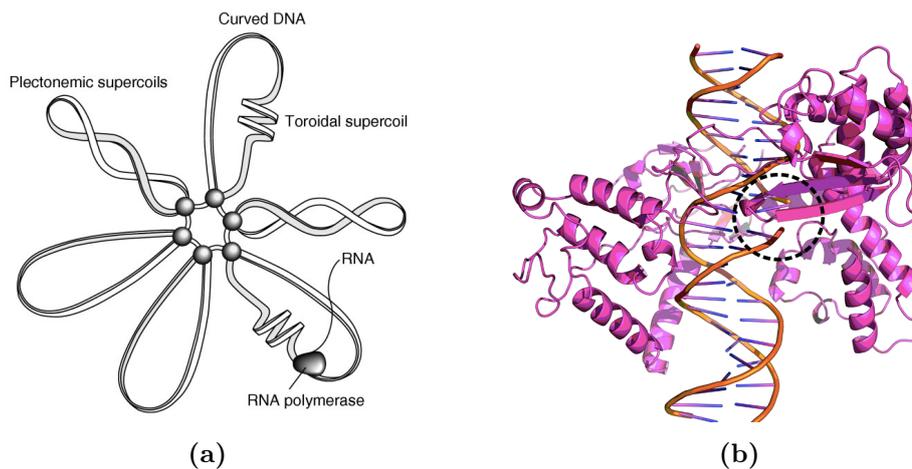


**Figure 2.5:** A fully decondensed circular genome of *E. coli*. Image adapted from [21].

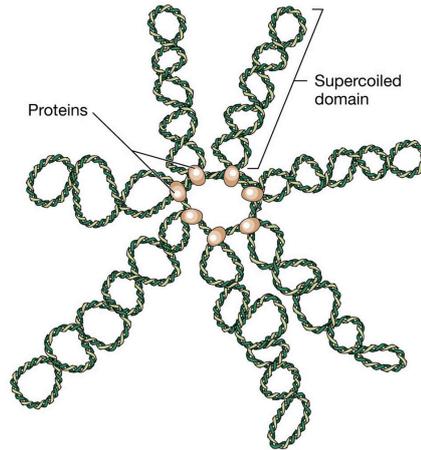
### 2.1.3 Chromosome Packaging in Bacteria

Despite the lack of histone proteins in prokaryotes, the compaction of the bacterial chromosome involves also specific DNA-binding proteins, which can introduce negative helical twists on the chromosome by twisting it in the opposite direction of the double helix. After being coiled repeatedly, the bacterial chromosome will be tremendously shrank in dimension. To gain the access of the DNA strands in some cellular processes, positive supercoils will be introduced for the ease of opening the double helix strands. Therefore helical twists on the bacterial chromosome can interchangeably be in positive- or negative-supercoiled status. The former will result in a more compacted structure, while the latter can be achieved by subtracting twists on the DNA strands. Inside the *E. coli* cells, specific enzymes named topoisomerases can act on the topology of DNA by adding or removing helical twists on the DNA strands. Two types of topoisomerases have been identified: Type I topoisomerases (as shown in Figure 2.6b), which cut only single strand of a DNA double helix, remove negative supercoils and then reanneal the double strand. The other is type II topoisomerases (i.e. DNA gyrase), which can cause a double strand break on the DNA double helix, introduce negative supercoils or remove pos-

itive supercoils. Cooperation of these topoisomerases renders a steady-state level of negative supercoiling DNA in *E. coli* [23], as shown in Figure 2.7. Another group of DNA-binding proteins, known as the nucleoid-associated proteins (NAPs), are responsible for the chromosome remodeling [24]. Some NAPs resemble the eukaryotic histone proteins in function [25]. Protein HU, a histone like protein, is able to wrap DNA into a bead-like structure [26]. Protein H-NS, a histone-like nucleoid structuring protein, prefers binding onto bent DNA sequences [27]. Some NAPs, such as FIS, H-NS and IHF, can also serve as global transcription factors [27, 28]. Cooperation of above-mentioned proteins is believed to organize the chromosome into topologically distinguishable loops (chromosomal domains) [29], as illustrated in Figure 2.7. It has been estimated that these chromatin domains in *E. coli* are sized from 10 kb [30] to 100 kb [31]. The presence of these structural proteins keeps the whole DNA molecule in a dynamic equilibrium between being twisted and unwound.



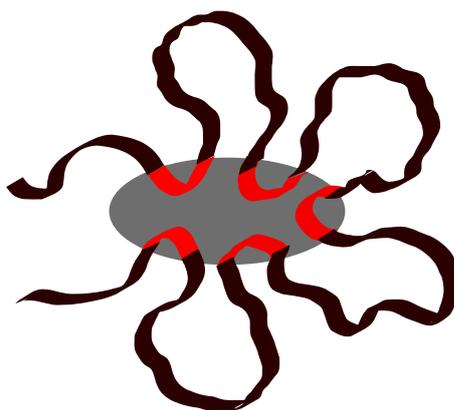
**Figure 2.6: DNA supercoiling and topoisomerase.** (a) A cartoon representation of different types of DNA supercoiling domains in the *E. coli* chromosome. Plectonemic or unrestrained supercoils are under torsional stress. Formation of toroidal or solenoidal supercoils needs the help of specific protein (like HU, the major histone-like protein in *E. coli*) by wrapping DNA around it. RNA polymerases can temporarily introduce toroidal supercoiling during transcription. Curved DNA appear more frequently at the tips of supercoils. Image (a) adapted from [32]. (b) shows a segment of DNA in combination with a type I topoisomerase molecule. The nick generated on one strand of DNA (indicated by a dashed circle) could facilitate the release of negative supercoils.



**Figure 2.7: Schematic representation of a *E. coli* chromosome with lots of supercoiled domains.** The circular *E. coli* genome is organized as multiple small domains. Image adapted from [33].

#### 2.1.4 Genome Organization in Eukaryotes

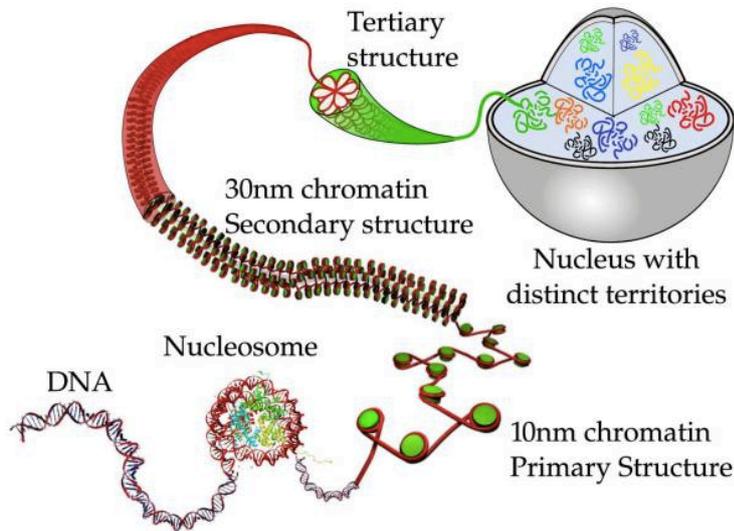
Different from prokaryotes, eukaryotic cells have a more morphologically distinguishable organization, featured by their membrane surrounded cell organelles. And a compartment named nucleus is specialized for the genetic material storage and processing, as shown in Figure 2.2. Eukaryotic nuclei are enriched with transcription apparatuses that are required for gene expression and transcripts splicing. The chromatin fiber of the cell in interphase occupies most volume of the nucleus. Chromatin fibers in the interphase cell nuclei do not simply fill up the whole volume. It has been suggested that chromatin fibers are organized with a nuclear matrix, which provides numerous docking sites for specific DNA sequences. These DNA sequences are named matrix attachment sites (MARs) or scaffold attachment sites (SARs) (Figure 2.8). MARs/SARs identified from the fragments remained on the nuclear matrix after the cleavage of nucleases indicate that they are frequently *cis* transcriptional regulators or include recognition sites of topoisomerase II, which suggests a transcription related role of the nuclear scaffold. Some MARs sequences can also behave as insulators, as reported by West *et al.* [34].



**Figure 2.8: Schematic representation of a nuclear matrix structure.** A subdomain of the nuclear matrix is represented as a gray oval. Curves in red are the sequences responsible for the attachments of the chromatin fiber to the nuclear matrix.

Chromatin in eukaryotic organisms is found in two varieties: euchromatin and heterochromatin. The former is less compact, and more flexible. Euchromatic regions include those actively expressed genes, or genes that are transcription inducible. Heterochromatic regions, by contrast, are always condensed throughout the interphase. Heterochromatin can be further split into two categories: constitutive and facultative heterochromatin. Facultative heterochromatin is an interchangeable type of chromatin, which can be transformed into euchromatin. Constitutive heterochromatin mainly composed of highly repetitive sequences is permanently condensed throughout the cell cycle. Heterochromatic regions are low in gene density and in recombination frequency. Facultative heterochromatin could function as a control mechanism by which the activities of genes included can be changed by condensing and expanding the chromatin. Eukaryotic chromosomes are usually delineated by active euchromatic and repressive heterochromatic domains. Next, we will take a close look at the molecular basis of the chromatin condensation in eukaryotic organisms.

## 2.1.5 Hierarchical Genome Organization in Eukaryotes

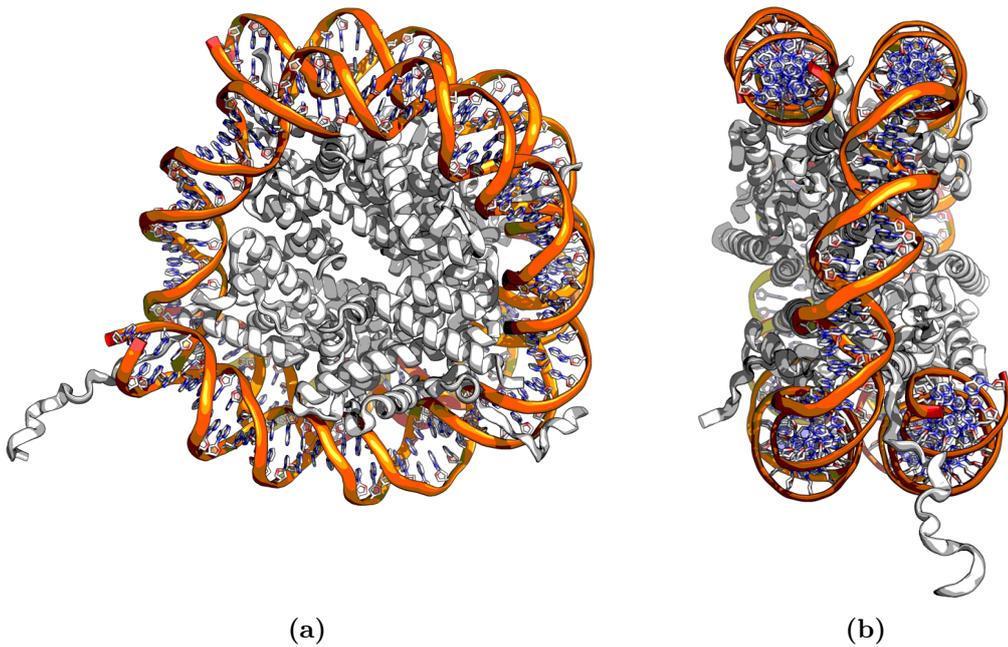


**Figure 2.9: Hierarchies of genome organization on all length scales in eukaryotes.** From naked DNA to chromosome territories. Image adapted from [35].

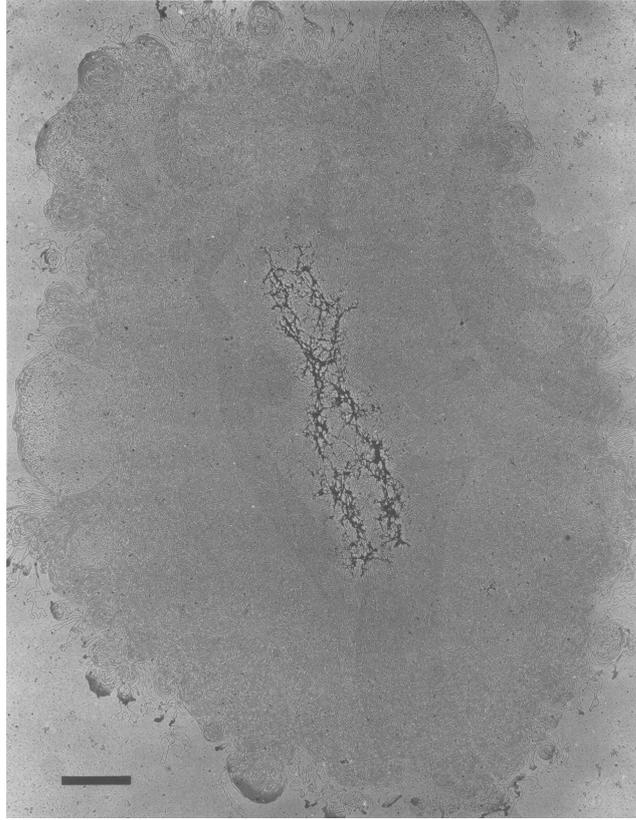
### i) 11-nm chromatin fiber

The basic structural unit of the eukaryotic chromatin is the nucleosome, which consists of a segment of 146 bp long core DNA wound two turns around the outside surface of an octamer of histone proteins (2 copies of each H2A, H2B, H3, and H4), as illustrated in Figure 2.10. Nucleosome monomers are connected via a linker DNA into a bead-like structure [36, 37, 38, 39]. Histone H1 is localized at the place where DNA enters and leaves the nucleosome, and controls the angle. Histone proteins are highly conserved among eukaryotes, which indicates an important role in genome organization. A naked DNA is repeatedly twisted around numerous histone protein complexes to form a chromatin fiber. This process provides the very first compaction of the DNA molecule (a packing ratio of approximately 6) [19, 20]. Figure 2.11 shows the conformation of a histone-depleted chromosome from HeLa cells, which has an internal protein scaffold surrounded by a halo of DNA. DNA organization at this level is much better characterized. After this process, a naked DNA is compacted into a 11-nm chromatin fiber composed of a string of nucleosomes connected by linker DNAs. The structure of the 11-nm chromatin fiber is still dynamic, and subject to changes by the chromatin-remodelling complexes. Moreover,

histone proteins do not merely wind the chromatin fiber, it can also affect the chromatin structure via recruiting variant copies of subunits or through the modifications on the amino residues of themselves. Each histone core protein has an unstructured N-terminal tail, which can be simultaneously modified by acetylation, methylation, or phosphorylation at several critical residues (the so-called histone modification codes) [40]. Histone modification is epigenetic in that it can influence the cellular phenotype without altering the genetic code. In this case, it is the structure of the chromatin fiber rather than the genetic code that determines the expression of genes, and subsequently the phenotype. Therefore both the underlying DNA sequence and the structure of the chromatin fiber can cause heritable changes in gene expression or cellular phenotype.



**Figure 2.10: A 3D illustration of the basic structure of a nucleosome.** In eukaryotes, the first level of DNA compaction is completed by winding a naked DNA around a series of histone protein cores. A histone core complex is composed of two copies each of four different core histones, shown as a gray clump in the inner part of the nucleosome.



**Figure 2.11: Metaphase chromosome from HeLa cell with histone depleted.** The central protein scaffold is surrounded by a halo of DNA loops. Image adapted from [41].

### **ii) 30-nm chromatin fiber**

In the next round of compaction, a series of nucleosomes in combination with other much less well-characterized structural proteins are coiled into a helical stack as a fiber with a diameter of around 30 nm. The chromosomal compaction under this length scale is more controversial. Some researchers are suspicious of its existence [42, 43]. 30-nm fibers are suggested to be the basic visible structures of chromatin from very gently lysed nuclei visualized by an electron microscope. However, the question of how are the nucleosomes arranged on top of each other to form the 30-nm fiber is still not well answered. Evidence shows that the core histone tail domains, two H4 tails in particular, play a critical role in aggregating the nucleosomes into the 30-nm chromatin fiber [44]. Histone H1, also known as the linker histone, has also been elucidated to be able to change the path of the DNA as it exists from the nucleosome, and to compact the nucleosomal DNA by interlocking neigh-

boring nucleosomes [45]. Different structural models, such as zigzag ribbon models [46, 47, 48, 49, 50] and helical solenoid models [51, 52, 53], have been postulated to explain the chromosomal compaction at this length scale. So far we still can not reach an agreement on the chromatin organization under this length scale.

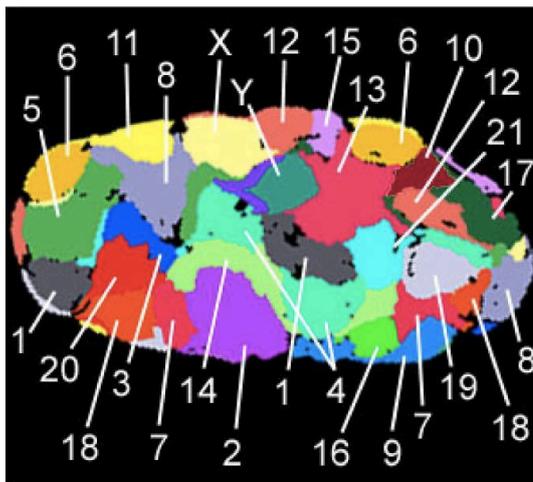
### **iii) Scales above 30-nm**

From a mathematical calculation, we can clearly see that if a 30-nm chromatin fiber is further assembled into a higher-order chromatin structure, it will give an overall packing ratio of around 1000 for euchromatin, and a ratio of around 10,000 for heterochromatin. In spite of the fact that the chromatin folding above the length scale of 30-nm is indispensable to the compaction of the 30-nm chromatin fiber into the confined space of a cell [35, 45], our understanding of the higher-order chromatin structure is still very limited at the moment. To address this question, the organization of the interphase chromatin fiber into loops has been postulated by Bohn and Heermann [54, 55, 56, 57]. According to this model, a chromosome is viewed as under on all scales looped polymer. This model is able to explain FISH experiments [54] and the experimental results of Lieberman [18].

### **iv) Chromosomes organized as discrete territories**

On an even larger length scale, chromosome painting via chromosome-specific DNA-probes coupled with improved fluorescent *in-situ* hybridization (FISH) has demonstrated that individual chromosomes are organized as distinct chromosome territories [6, 58], as shown in Figure 2.12. By a much less characterized mechanism, chromatin fibers are packaged into higher-ordered chromatin architectures as discrete entities. From this study, we can clearly see that chromatin fibers do not just simply fill up the nucleus volume. They actually have a non-random interphase chromatin arrangement. For each single chromosome, it can be further split into multiple self-organized chromosome sub-territories. On this length scale, the chromatin loop model [54, 56, 57] works again in explaining the segregation of chromosomes into territories due to entropic repulsion between the loops [59] as well as the shape of the chromosome territories. Thus loops provide a unified model for the higher-ordered

organization of the nucleus.



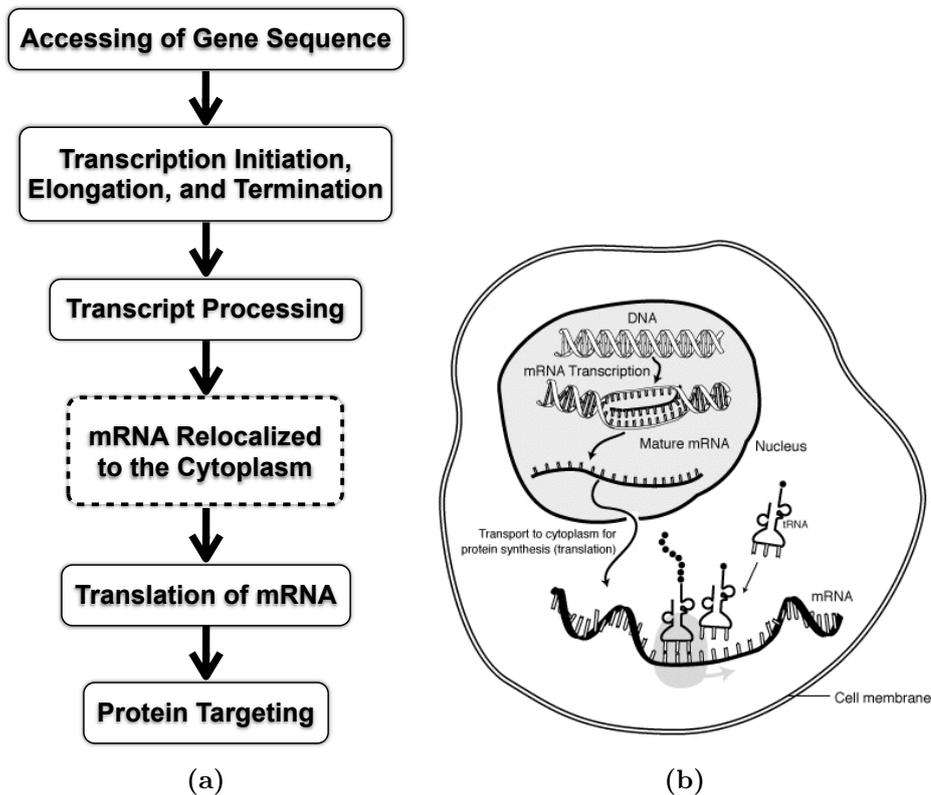
**Figure 2.12: Whole chromosome painting in a human fibroblast nucleus.** 24 human chromosomes are labeled with different fluorescent colors to show the organization of individual chromosomes as distinct territories. Image adapted from [6].

## 2.2 Genome Function and Genome Architecture

### 2.2.1 Gene Expression and Transcriptional Regulatory Networks

The gene transcription schema utilized in prokaryotic and eukaryotic organisms are illustrated in Figure 2.13. To begin with, the promoter sequence of a gene [19, 20] is unwound for the accessibility to the apparatus for gene expression, like transcription factor, RNA polymerase, and so on. Three major classes of RNA polymerases (from type I to type III) present in eukaryotic cells are responsible for the expression of genes under different categories (protein coding genes or RNA coding genes [19, 20]). RNA polymerases are necessary for synthesizing RNA chains by using DNA genes as templates. The genetic codes included in the sequence of a gene are continuously read and transcribed into a complementary premature RNA sequence. After the transcription termination, premature RNAs are further processed into mature RNAs. mRNAs of eukaryotes contain multiple exons interrupted by introns. In order to get an mRNA molecule that yields a working protein, the cell needs to trim out the introns and then stitch the exons together [60]. In eukaryotes, mature

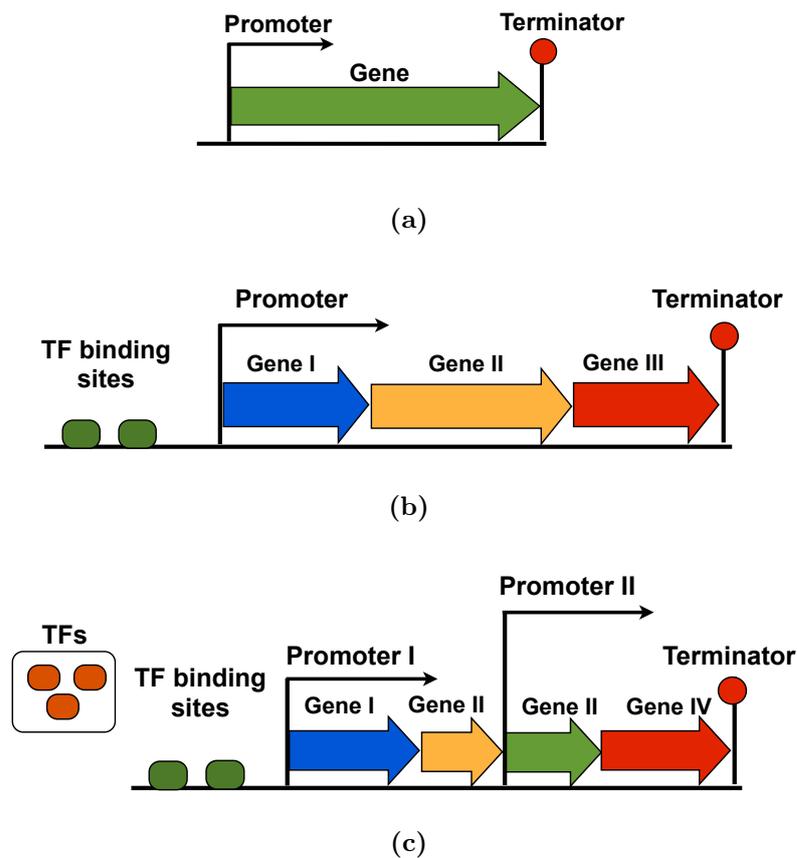
mRNAs are further exported from the cell nucleus to the cytoplasm in order to be transcribed into proteins. Ribosomes enriched in the cytoplasm are the sites where mRNAs are translated into functional proteins. Regulation of gene expression can actually occur at every step. But most frequently, the regulation is emphasized on the stage of gene expression initialization mediated by transcription factors. For this purpose, general transcription factors (or called basal transcription factors) are aggregated at the promoter region of a gene and assembled with RNA polymerases to initiate the gene transcription. Transcription factors can play a dual role: either as gene expression activators or repressors. Multiple transcription factors can work in coordination to achieve a multi-variant control. Target DNA sequences of transcription factors can be localized either on the upstream or downstream of genes. Distant regulatory elements which are several Mb far away from the target genes have also been identified [55].



**Figure 2.13: Gene transcription and translation.** (a) It is specific to eukaryotic organisms that mRNAs are transported from the cell nucleus to the cytoplasm and translated by the ribosome, a step indicated by dashed border lines. In prokaryotes, transcription and translation are more tightly linked in space. (b) The whole life cycle of an mRNA in a eukaryotic cell. Image adapted from [61].

To facilitate the regulation of gene expression, the linear arrangement of genes and regulatory elements along the chromosome is well organized at multiple scales [19, 20] (see Figure 2.14). One or more genes under the control of a single promoter is called a Transcription Unit (TU) [62, 63]. Group of genes along with their associated regulatory elements is called an operon [62, 63], which can be transcribed as a single polycistronic mRNA carrying multiple Open Reading Frames (ORFs) for coding different proteins at the same time. Polycistronic mRNAs are most frequently observed in prokaryotes due to their overlapping genome structure. An operator is a sequence with which transcription factors can combine. In the case of a repressor regulator, physical interaction between the repressor protein and the operator will block the access to the promoter by RNA polymerase and thus inhibit gene expression. On the

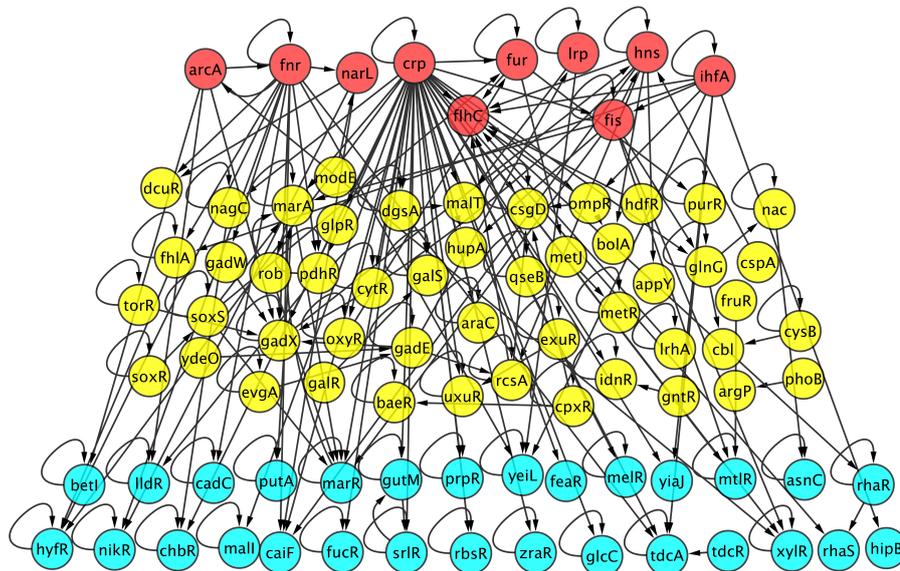
contrary, the combination of an enhancer protein will initiate gene expression by looping out the chromatin fiber in between the operator and the promoter. For a complex operon, several promoter regions might coexist, which can selectively transcribe partial set of genes. And for a complex regulon, a set of genes are exactly subjected to the same two or more transcription factors. In some extreme case, all transcription factors involved have the same effect on the genes included in the regulon, which is called a strict complex regulon [62, 63].



**Figure 2.14: Schematic representation of diversified transcription units.** (a) indicates the structure for a transcription unit (TU), (b) for an operon, and (c) for a regulon.

Numerous efforts have been made in elucidating potential transcription factors (TF) and target genes. TFs are themselves encoded by TF genes. Therefore protein products of some kinds of genes can influence the expression of others. A single TF can regulate the expression of multiple genes, and a gene is usually subject to the regulation of multiple TFs. A genome-wide interplay between

TFs and target genes will reveal a network structure, called gene transcriptional regulatory network (TRN) [64]. Figure 2.15 shows a transcriptional regulatory network involving only TF genes in *E. coli*. TFs can regulate the activities of genes globally as well as locally. Global TFs, which are in charge of a great number of genes, can change the expression status of a cell globally. As a result, global TFs are distantly located on the genome with respect to their target sites, and expressed in large amounts. By contrast, local regulators are genomically close to their regulated genes on the chromosome, and have a much lower copy number. In light of these observations, an alternative three-dimensional jumping between DNA-strands and one-dimensional sliding along the DNA chain mechanism has been postulated to explain the way of TFs in searching and binding their target sequences [10, 65, 66].



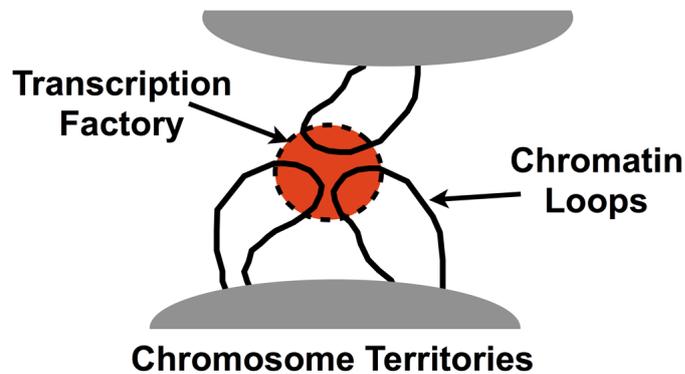
**Figure 2.15: Diagram representation of a gene transcriptional regulatory network identified from *E. coli*.** In this network, only the transcriptional interplay involving TF genes of *E. coli* is presented. Regulatory relationships are indicated by arrowed lines extending from TF genes to their regulated genes.

## 2.2.2 Transcription Factory as a Structure Organizer

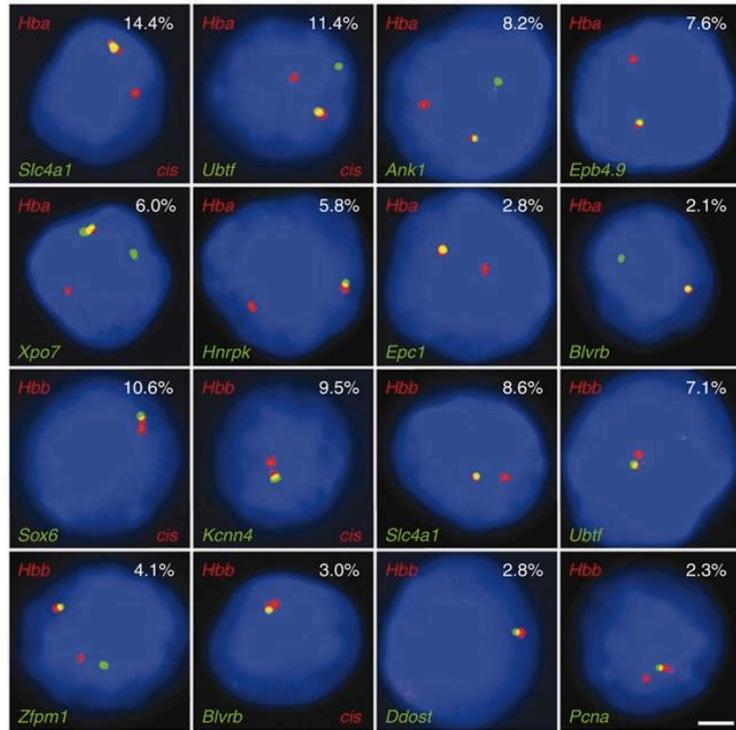
Gene transcription sits on top of the flow of genetic information, and receives most intensive controls. When genes are expressed, RNA polymerases are recruited to the promoter regions of the target genes. Since the presence of

RNA polymerases can indicate an active state of gene expression, Jackson *et al.* and Wangsink *et al.* scrutinize the distribution of RNA polymerases (type II for specific) in the cell nucleus [67, 68]. They find that RNA polymerases II are concentrated at a very limited number of discrete foci, instead of being ubiquitously distributed. The compartments enriched with RNA polymerases are named transcription factories [69, 70]. Genes that are concurrently transcribed share these sites for expression. Several studies have observed that once activated, genes are able to extrude from the chromosome territories and relocate themselves to the transcription factories [71, 72, 73]. Some genes from the same chromosome or from different ones are more likely to visit the same transcription factory [70], as shown in Figure 2.16. For co-localized genes from the same chromosome, they are not necessary to be genomically proximate to each other, such as genes *Hbb*, *Eraf*, and *Uros*, which are separated by tens of Mb on the human genome [70]. From mouse erythroid cells, Schoenfelder *et al.* [74] discover that active globin genes prefer to associate with hundreds of other co-transcribed genes, resulting in both intra- and inter-chromosomal transcription interactoms (see Figure 2.17a). Genes regulated by the transcription factor *Klf1* are organized inside a limited number of specialized transcription factories (see Figure 2.17b). According to this observation, one can infer that co-regulated genes coupled with specific transcription factors might be assembled at some specialized nuclear compartments to guarantee an efficient and coordinated transcriptional control. Later, Xu *et al.* prove that transcription factories are transcript-specific, which depends on the promoter type and whether or not the gene contains an intron [75]. Transcripts transcribed by different types of RNA polymerases are allocated at distinct transcription factories. Evidence shows that specific transcription factories are enriched with particular regulatory factors [76], which might be the reason why transcription factories have their own appetites on the transcripts being transcribed. Binding of transcripts to a factory with suitable factors will increase the possibility of transcription initiation. Although it is possible for a specialized transcription factory to transcribe other types of genes, the efficiency will be much lower. Based on these observations, a conceptional model with regard to the role of transcription factory in shaping the genome organization has been postulated by Cook [15]. Cook suggests that DNA or chromatin loops can be

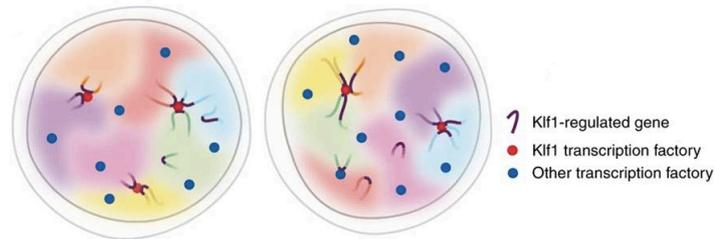
tethered to the transcription factories by means of transcription machineries. Transcription initiation and termination, or transcription factors' binding and dissociating could establish and disrupt loops, which makes the genomic architecture dynamic and self-sustaining. Transcription induced gene association at the same transcription factory could reflect the neighboring relationship of genes in the cell nucleus in a probabilistic way.



**Figure 2.16: Chromatin kissing mediated by the transcription factory.** Chromatin loops (black) extruding from the chromosome territories (gray) can migrate to the transcription factory (red) for gene expression. Due to a limited number of transcription factories exist in the cell nucleus, multiple transcribed genes from the same chromosome or from different chromosomes need to share the same transcription factory.



(a)



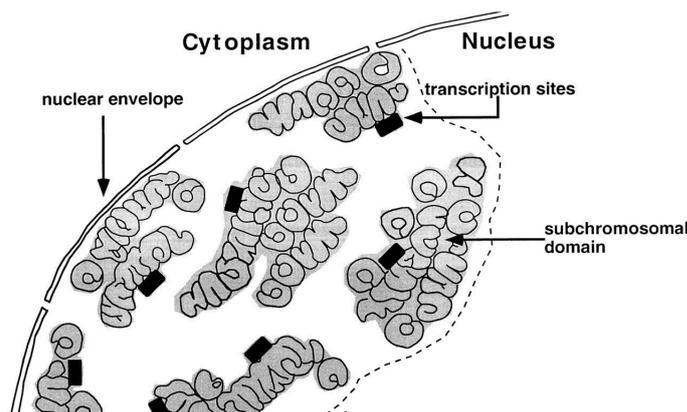
(b)

**Figure 2.17: A probabilistic and dynamic model of genes co-localization.**

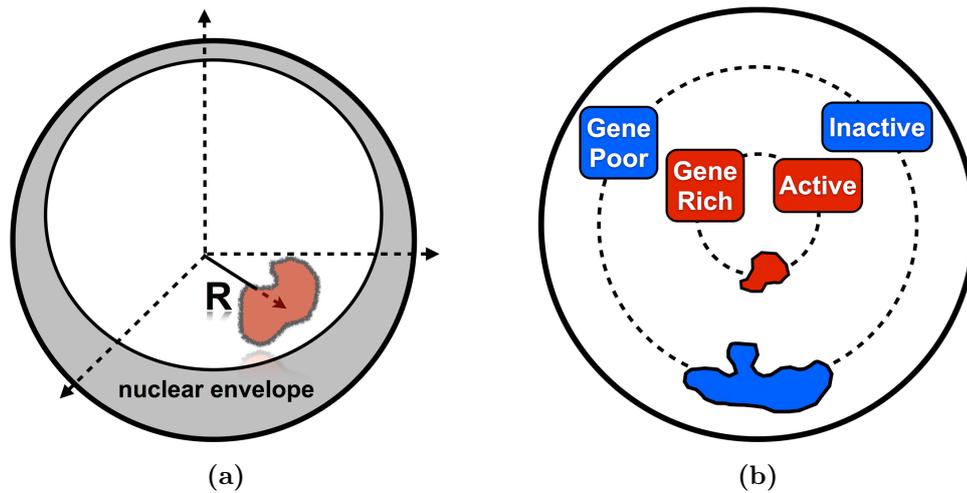
(a) shows the co-localization of erythroid-expressed gene in mouse erythroid cells. The frequencies of co-localization between *Hbb* gene (in red) and several other erythroid-expressed genes (in green) are given. Gene loci are displayed by using double-label RNA FISH. (b) Dynamic association between genes localized in specialized transcription factories shows a higher probability. Transcription factories enriched with the transcription factor Klf1 are labeled as red nodes, and other types of transcription factories are labeled as blue nodes. Chromatin loops containing Klf1-regulated genes are represented as strips in purple. Klf1-regulated genes will surround the Klf1-specific transcription factories when activated, and diffuse in the nucleoplasm when inactivated. Image adapted from [74].

### 2.2.3 Role of Chromosome Territories

In interphase cell nucleus, chromosomes are observed as separated entities, called ‘chromosome territories’ [58, 77, 78, 79], which occupy different volumes of the cell nucleus, as illustrated in Figure 2.18. Chromatin fibers from different territories or sub-territories can interact with each other, but they do not intermingle extensively. The underlying mechanism for the formation of chromosome territories, their positioning in the cell nucleus, and their function are not well understood. It’s still far from clear whether there exists a general pattern of organization in different cell types, or whether the organization pattern observed can be reproduced through the cell division. Some studies with certain type of cells show the lack of such an order at large [80, 81], while others do disclose a rigidly maintained order [82, 83, 84]. Multiple factors (i.e. DNA content [85], gene content [86, 87], and so on) that can influence the radical positioning of a chromosome in the cell nucleus, have been identified, as shown in Figure 2.19. But none of these factors can serve as a solo determinant. Controversies also relate to the question whether there is any preference on the homologous associations between different chromosomes (or to say chromosomal neighbors) [84, 88, 89, 90, 91, 92].

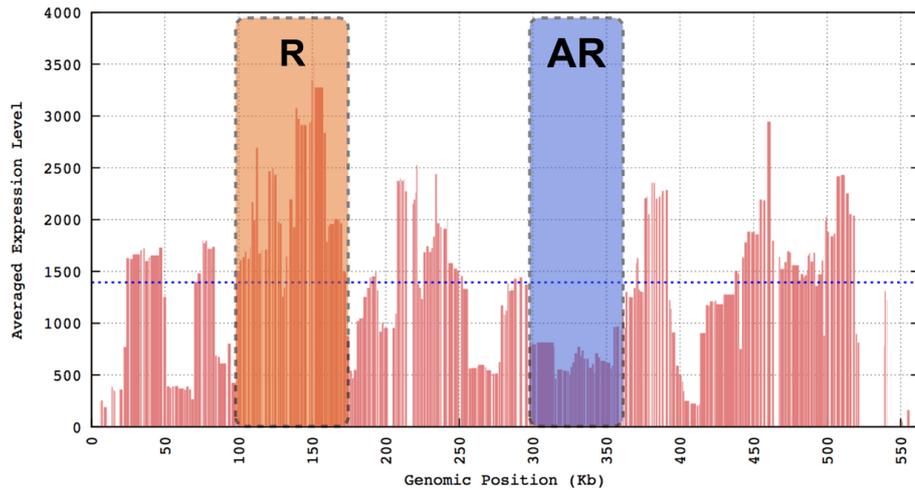


**Figure 2.18: Cartoon representation of a chromosome territory.** Chromatin fibers from the same chromosome or from different ones are folded into discrete chromosomal sub-domains, which are interconnected, and grouped as a larger chromosomal domain. Image adapted from [77].

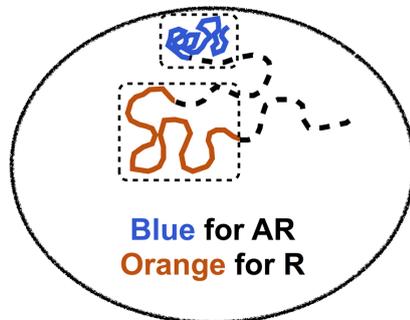


**Figure 2.19: Factors that can influence the radial positioning of a chromosome.** (a) gives a schematic representation of the radial positioning of a chromosome in a spherical nucleus. (b) Factors, like chromosome size, gene density, and expression activity, can influence the radial positioning of a chromosome.

A deeper insights into the functional role of the chromosome territories is given by Sandra *et al.* [93]. According to the transcriptome map of human, the author define two types of chromosomal domains on the chromosome: ridge and anti-ridge. The ridges are the chromosome regions where gene density and expression level are high. The anti-ridges, by contrast, are gene poor regions and low in expression. Analysis of the three-dimensional structures of the ridge and anti-ridge domains shows that the ridge domains, in contrast to the anti-ridge domains, are generally less compacted, more irregular in shape, and take a more central localization, as shown in Figure 2.20. All of six human cell lines involved in the study display a compatible organization motif at large, in spite of the differences in tissue types and differentiation states. Additionally, they observe that chromatin fibers included in the same large chromosome territory but from different parts of the same chromosome intermingle to a very limited degree.



(a)



(b)

**Figure 2.20: Illustration of the ridge and anti-ridge positioning motif.**

On the transcriptome map shown in (a), the ridge and anti-ridge domains can be delineated according to the expression level. The average expression level is indicated by a blue dashed line. The ridge and anti-ridge positioning motif is shown in (b). The ridge domain contains more loosely packaged chromatin fiber and has a more interior localization. By contrast, the anti-ridge domain is more condensed and peripherally localized.

Although individual chromosomes of human are folded into discrete territories occupying different volumes of the cell nucleus, their positions are not completely fixed. Instead, chromosome territories are dynamic to a certain degree. Fluorescently labeled chromatin locus can display a rapid movement with ease throughout a relatively small scope, but it can seldom migrate far away from its chromosomal territory. The movability of a chromatin locus depends both on its nucleus localization, and on its surrounding genome environ-

ment. Chromatin loci that are associated with nucleolus or nuclear periphery are more restricted in movement than those floating in the nucleoplasm [94]. When compared with human, budding yeast has a less dense nucleus, and contains few heterochromatin domains, which might be the reason for a higher diffusion constant observed in it [95]. Modification of DNA-binding proteins or association/dissociation with different DNA-binding proteins can also alter the chromatin structure, which in turn could result in the movement of chromosomes. Taddei *et al.* has revealed that inhibiting the activities of histone deacetylases can induce large-scale movements of centromeric and pericentric heterochromatin to the peripheral environment of nucleus [96], which can be reversed by reactivating histone deacetylases afterwards. During regular cell cycles, redistributions of chromosomes are also frequently observed. Ferguson *et al.* traces the movement of centromere and chromosomal arms of chromosome 8 in human T-lymphocyte from G1 to G2 cell cycles [91]. They find that chromosome 8 moves in a cell cycle dependent way. In G1 phase, the centromere of chromosome 8 positions at the nuclear periphery, while the chromosomal arms protrude toward the interior of nuclear. After going through the G1 phase to the G2 phase, the positions of the centromere and the chromosomal arms are interchanged.

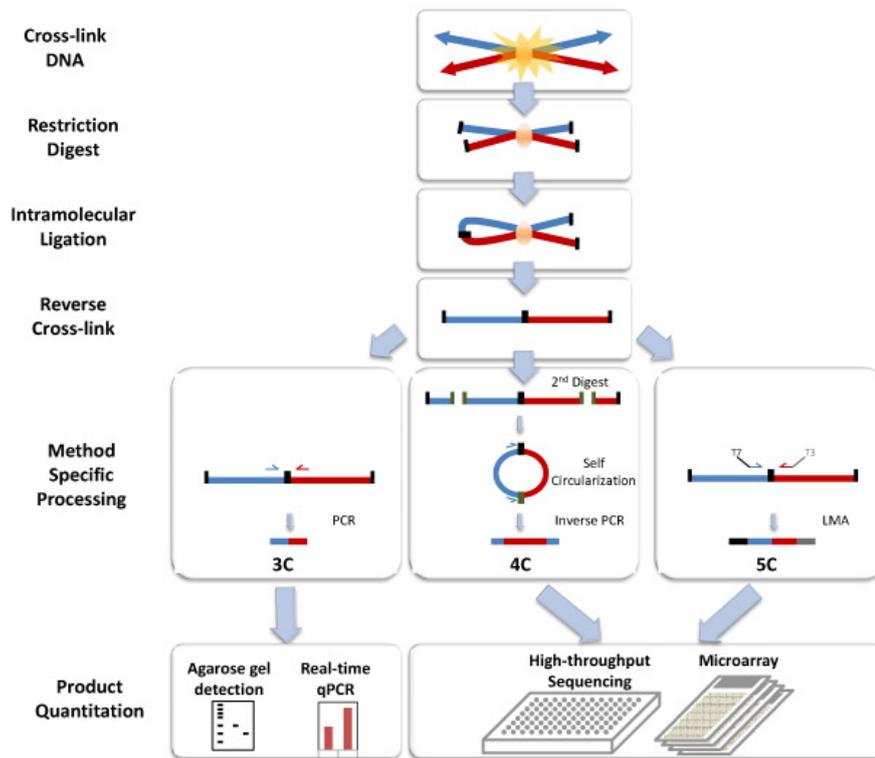
#### 2.2.4 Gene Expression Regulation Mediated by Specific Chromosomal Interactions

Unlike the gene association observed at the sites of on-going transcription due to a limited number of transcription factories above-mentioned, some physical interactions among specific gene loci are more unique in function, even though the possibility of specific gene association at the same transcription factory can not be eliminated, as reported in some cases [75, 97]. While *cis*-regulatory sequences are most frequently observed at short genomic distances from their target genes, ever-increasing evidence demonstrates that *cis*-regulatory elements can actually be really distant from their controlled genes (sometimes up to hundreds of kb) [55]. It's even more surprising that *trans* regulatory elements located on different chromosomes have been identified as well. In order to tune up the expression of target genes, corresponding regulatory sequences

must be involved to directly form physical interactions with their target genes. As a consequence, a genome-wide chromosome physical interactions network is established so as to coordinate the expression of distantly localized genes. With more and more distant regulatory sequences being identified, it's clear that we have underestimated the importance of long-range physical chromosome interactions, which might function as a common architecture motif in the genome. For *cis* and *trans* interactions between regulatory element and target gene, the spatial gap can be overcome by the formation of chromatin loops, or chromatin bridges [72, 98, 99, 100, 101]. Long-range gene regulation has been intensively investigated in the case of  $\alpha$ - and  $\beta$ -globin loci [101, 102, 103]. The  $\beta$ -globin gene cluster includes a group of developmentally specific genes encoding different  $\beta$  chains of hemoglobin. All these variant genes are under the control of the Locus Control Region (LCR), which is 25 kb upstream of the closest variant gene. To activate target genes with a genomic distance of 80 kb, a direct chromatin loop can be formed between the LCR and the promoter of target gene [100, 101, 104]. Similar to the  $\beta$ -globin locus, the  $\alpha$ -globin locus is also controlled by a set of remote regulatory elements positioned 40-60 kb upstream of the gene locus [105, 106], and activated by long-range looping [105, 107]. While *cis* chromosomal interactions are more frequently observed, *trans*-regulation is not scarce. *Cis* elements on one chromosome can regulate not only the target gene *in cis*, but also the homologous allele *in trans* [108], as reported in the case of X-chromosome inactivation [109, 110]. *trans*-regulations are not limited to the invertebrate genomes, or to the homologous chromosomes. Non-homologous *trans* chromosomal interactions have also been observed from the mouse  $T_H2$  gene locus [111] and H-enhancer locus [112]. Moreover, gene silencing or heterochromatin formation can also be mediated by the long-range chromosomal interactions, as observed from the telomere clustering in *S. cerevisiae* [113] and the polycomb complex dependent silencing in *Drosophila melanogaster* [114]. All of these findings emphasize the importance of loci-specific chromosomal interactions over a large genomic distance or from different chromosomes.

## 2.2.5 3D Genome Organization Revealed by 3C-based Methods

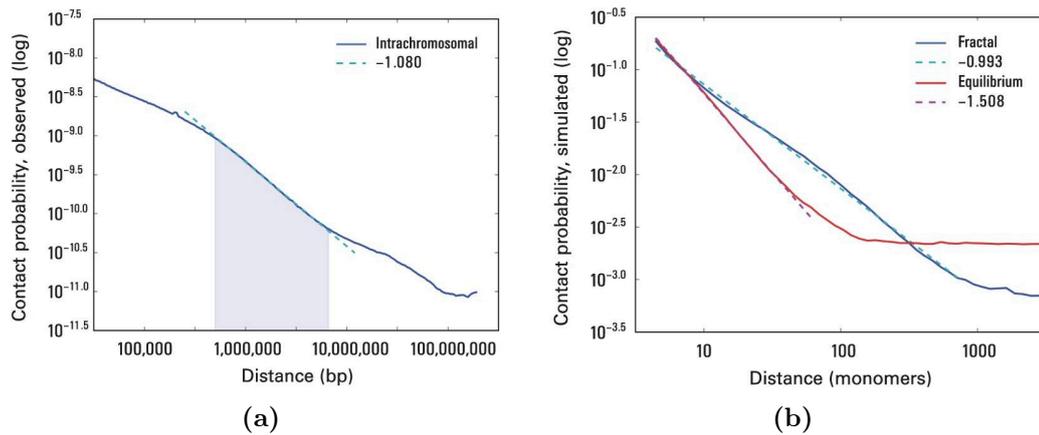
Traditionally optical technologies, such as fluorescent *in situ* hybridization (FISH) or chromosome painting, are only able to display the organization of chromosomes or chromosomal territories with a low resolution. Through these methods, chromosomes have been visualized as distinct territories. Multi-labeling method developed later makes it possible to display several chromosomes or gene loci at the same time, but it is still impossible to get a large picture of the genome organization. The chromosome conformation capture (3C) technology developed by Dekker *et al.* [115] provide us a novel way to discover the chromosomal regions in physical proximity. Chromosomal contact maps or gene loci contact maps of various organisms have been generated via varieties of 3C-based methods [8, 18, 104, 105, 116]. ChIP-loop assay [112, 117, 118, 119] has been used to investigate cell-to-cell differences in chromatin conformation. Circular chromosome conformation capture (4C) method [4, 120, 121] makes it possible to screen the entire genome in an unbiased fashion for DNA segments physically interacting with a DNA fragment of interest. 3C-carbon copy (5C) method [5, 122] is a combination between standard 3C protocol and large-scale mapping via microarray detection or high-throughput sequencing. Hi-C method developed by Lieberman-Aiden *et al.* [18] has realized the genome-wide screening by introducing biotin-based ligation protocol. With a novel method named genome conformation capture (GCC), the network of chromosome interactions for the yeast *S. cerevisiae* is first studied [116]. After that, a Hi-C like method has been employed to investigate the genome organization in budding yeast [8]. In addition to these most widely used methods, other derivatives, like combined 3C-ChIP cloning (or 6C) [117, 123], enhanced ChIP 4C (e4C) [74], CHIP-PET [124] have also been worked out. A comparison between the 3C technology and its derivatives is presented in Figure 2.21. All these methods provide us significant implications for understanding eukaryotic genome organization.



**Figure 2.21: Experimental pipelines of the chromosome conformation capture (3C) technology and its derivatives.** These methods differ in the step for ligation products quantification. Image adapted from [125].

One telling example involves the study conducted by Lieberman-Aiden [18]. In 2009, Lieberman-Aiden *et al.* construct a genome-wide spatial proximity map of human chromosomes with the Hi-C method. Under large genomic scale, they confirm that chromosomes are organized into individual domains or territories. Chromosomes that are small in size and high in gene density are more likely to cluster with each other. Besides, two types of distinguishable genomic compartments have been identified. In contrast to inactive genomic compartments, those active ones show higher gene expression activities, and more open and accessible chromatin structures. They also find that when averaging across the whole genome, chromatin contact probability as a function of genomic distance shows a power law scaling behavior extending from 500 kb to 7 Mb with slope around -1 (Figure 2.22). Different from the equilibrium globule model, where the scale is  $-3/2$ , this study indicates a fractal globule organization. Take all these findings together, they draw the conclusion that

in human at the scale of several mega bases, the chromatin is organized in the form of ‘fractal globule’, which is self-organized, long-lived, non-equilibrium and free of knots. To achieve a conformation of this type, an unentangled polymer is first packed into a series of small globules into a beads-on-a-string conformation, which serves as packaging units in succeeding rounds of compaction. Eventually, a single globule-of-globules conformation will be reached.



**Figure 2.22: The fractal globule organization revealed by the Hi-C method.** (a) On a double-logarithmical plot, the contact probability as a function of genomic distance shows a power law scaling behavior with a slope around  $-1$  between the genomic distance from 500 kb to 7 Mb. (b) The equilibrium globule model predicts that the contact probability will scale as  $s^{-3/2}$ , while the fractal globule model predicts the scaling as  $s^{-1}$ . Image adapted from [18].

## 2.2.6 Structural Alteration and Genome Malfunction

Increasingly evidence points out that the transformation of normal cells might be resulted from the aberrant genome organization. In addition to copy number change, another remarkable alternation observed in varieties of tumor cells is the chromosome rearrangement, such as chromosomal deletion, insertion, reversion, and translocation. Among these changes, chromosomal translocation is of special importance. As a consequence of chromosomal translocation, oncogenic genes might be misregulated after being inserted into different genomic environments. It’s also possible that some genetic fusion events will result in oncogenetic fusion proteins. One of telling case involves human *BCR*

gene from chromosome 9 and *ABL* gene from chromosome 22. A fusion gene composed of these two genetic loci can serve a constitutively active kinase, prevalently observed in chronic myelogenous leukemia [126]. In higher eukaryotic organisms, the spatial proximity between genetic loci is important to the occurrence of chromosomal translocation. Due to the position variation of individual cells, some kinds of organizations have higher probability than others. For instance, some chromosome territories or gene loci are more likely to stay side-by-side as neighbors. This explains why certain type of chromosomal translocation events are more frequently observed than others. Intriguingly, several chromosomal translocations have been found to be tumor type specific, especially clear in leukemia and lymphomas [126, 127, 128]. Obviously, the frequency of chromosomal translocation depends also on the size of chromosome involved [129]. Evidence also shows that the spatial proximity between chromosome territories can be reflected by the translocation frequencies [130, 131, 132, 133, 134]. Likewise, such a correlation is also conserved for the proximity of genes and their translocation probabilities. As a consequence of the chromosomal translocation, the position of fusion chromosome will largely not be affected. By contrast, the translocated gene loci show a tendency to relocate to the place where they usually stay [135].

# Chapter 3

## Introduction to Polymer Physics

In this chapter, some fundamental conceptions which are able to describe the properties of polymers are introduced. After that, basic and improved polymer models developed in last decades are presented and compared. For a detailed review on polymer physics, reader can refer to Refs. [136, 137, 138, 139, 140].

Various models have been postulated from different perspectives to explain the complicated behaviors of DNA folding as observed in experiments. Therefore, in the second part of this chapter, several genome architecture models taking the role of genome function into consideration are introduced, and some enlightening findings are emphasized. For the reader, who is already familiar with these topics, can skip this chapter.

### 3.1 Describing DNA as a Polymer

Fitting well with the definition of a polymer, DNA is a biological macromolecule composed of the repetition of four types of monomers bound together via covalent bonds. Base pairs are connected in the form of a one-dimensional string. In most cases, information essential for life is maintained on a double-stranded DNA. Due to the chemical nature of the double-stranded DNA, it is stiff at short scale, and flexible when the scale is large enough. To measure the stiffness associated with the DNA, a notion termed the persistence length  $l_p$  is introduced, which relates to the decay length of correlations of directionality

along the chain. The correlation function is defined as

$$C(x, y) = \langle \vec{e}(x) * \vec{e}(y) \rangle, \quad (3.1)$$

where  $\vec{e}(x)$  and  $\vec{e}(y)$  are the unit vectors tangent to the chain at the respective positions  $x$  and  $y$  on the chain, and the symbol  $\langle \rangle$  refers to an ensemble averaging over time with the same polymer or over many conformations of different polymers. The correlation function decays exponentially with distance for a long enough homopolymer, and has a characteristic scale  $l_p$

$$C(|x - y|) = \langle \vec{e}(x) * \vec{e}(y) \rangle \propto \exp(-|x - y|/l_p). \quad (3.2)$$

$l_p$  is the scale over which the directions of unit vectors become uncorrelated. The persistence length of a double-stranded DNA is 10 nm (with a separation of about 200 base pairs). The DNA molecule becomes stiff at length scales  $L \ll l_p$ , and flexible at length scales  $L \gg l_p$ .

A polymer's end-to-end distance is a measure of its spatial extension. Most frequently, the mean square end-to-end distance (eed)  $\langle R_N^2 \rangle$  is used as one of the characteristic features of polymer models. For the models to be introduced later (the random walk model, the worm-like chain model, the self-avoiding walk model, the equilibrium globule model, and the fractal globule model, excluding the random loop model), their end-to-end distances all scale as

$$\langle R_{eed}^2 \rangle = l^2 N^{2\alpha}, \quad (3.3)$$

where  $l$  is the linker length,  $N$  the chain length, and  $\alpha$  a model specific constant.

In polymer physics, the radius of gyration can also characterize the dimensions of a polymer. The radius of gyration is defined for a particular polymer as

$$\langle R_{gyr}^2 \rangle = \frac{1}{N} \sum_{i=1}^N (r_i - r_{mean})^2, \quad (3.4)$$

where  $r_{mean}$  is the mean position of the monomers.

### 3.1.1 The Random Walk (RW) Model

The random walk (RW) model, also known as the freely-jointed chain model, is the simplest model to describe a polymer, which takes no interaction among

monomers and no excluded volume of monomers into consideration [141]. Monomers in this model are represented as rigid rods with a fixed length. Each monomer's orientation is independent of their neighboring monomers' positions and orientations, which allows the polymer visiting the same place twice or more. When the length of a DNA molecule is long enough compared with its  $l_p$ , the flexibility associated will allow it to have lots of different conformations. As mentioned above, we can characterize the properties of an ideal chain model by means of the average (root mean square) end-to-end distance or the average radius of gyration, given a chain with  $N$  monomers. Let's denote the end to end vector by  $\vec{R}$ , and individual monomer's vector by  $\vec{r}_i$ , where  $i$  is from 1 to  $N$  ( $N$  is large enough to obey the central limit theorem). Due to the symmetry of system, we can get

$$\langle \vec{R} \rangle = \sum_{i=1}^N \langle \vec{r}_i \rangle = \vec{0}. \quad (3.5)$$

The symbol  $\langle \rangle$  means the mean of a random vector throughout the article. By applying the central limit theorem, the distribution of  $\vec{R}$  follows a normal distribution, which gives us

$$\sigma^2 = \langle R_j^2 \rangle - \langle R_j \rangle^2 = \langle R_j^2 \rangle - 0, \quad (3.6)$$

where  $j = x, y, \text{ or } z$ .

Seeing that

$$\langle R_x^2 \rangle = \langle R_y^2 \rangle = \langle R_z^2 \rangle = N \frac{l^2}{3}, \quad (3.7)$$

we will get

$$\langle R_{eed}^2 \rangle = N l^2 = l l. \quad (3.8)$$

The probability density function of end to end vector of the chain is calculated as

$$P(\vec{R}_{eed}) = \left( \frac{3}{2\pi \langle R_x^2 \rangle} \right)^{\frac{3}{2}} \exp - \frac{3\vec{R}^2}{2 \langle R_x^2 \rangle}. \quad (3.9)$$

Thus the average end-to-end distance of the chain is calculated as

$$\sqrt{\langle R_{eed}^2 \rangle} = \sqrt{N} l = \sqrt{L} l. \quad (3.10)$$

And the radius of gyration  $R_{gry}$  is calculated as

$$R_{gry} = \left\langle \left( \frac{1}{N} \right) \sum_i (\vec{r}_i - \langle \vec{r} \rangle)^2 \right\rangle^{\frac{1}{2}} = \sqrt{\frac{\langle \vec{R}^2 \rangle}{6}} = \frac{\sqrt{N} l}{\sqrt{6}}. \quad (3.11)$$

Now, let's see another property of the ideal chain model, the scaling.

$\langle R_{eed}^2 \rangle$  can also be calculated as

$$\langle R_{eed}^2 \rangle = \langle \left( \sum_{i=1}^N \vec{r}_i \right)^2 \rangle = \langle \sum_{i,j=1}^N \vec{r}_i \cdot \vec{r}_j \rangle. \quad (3.12)$$

Since

$$\langle \vec{r}_i \cdot \vec{r}_j \rangle = \langle |\vec{r}_i|^2 \rangle \delta_{i,j}, \quad (3.13)$$

we can see that

$$\langle R_{eed}^2 \rangle = N \langle |\vec{r}_i|^2 \rangle. \quad (3.14)$$

If we define

$$R_0 = \langle R_{eed}^2 \rangle^{\frac{1}{2}}, \quad (3.15)$$

and

$$a_s = \langle |\vec{r}_i|^2 \rangle^{\frac{1}{2}}, \quad (3.16)$$

we will get

$$R_0 = a_s N_s^{\frac{1}{2}}. \quad (3.17)$$

### 3.1.2 The Worm-like Chain Model

The worm-like chain (WLC) model, or called the Kratky-Porod model, was developed to describe the behavior of chains with a high backbone rigidity [140]. To characterize the properties of a worm-like chain, one can use the orientational correlation function:

$$K_{or}(\Delta l) = \langle e(l) e(l + \Delta l) \rangle = \exp - \frac{\Delta l}{l_{ps}} \quad (3.18)$$

It's clear that

$$K_{or}(\Delta l_1 + \Delta l_2) = K_{or}(\Delta l_1) K_{or}(\Delta l_2). \quad (3.19)$$

The mean squared end-to-end distance  $\langle R_{eed}^2 \rangle$  is calculated as

$$\langle R_{eed}^2 \rangle = \int_{l'=0}^{l_{ct}} \int_{l''=0}^{l_{ct}} \langle e(l') e(l'') \rangle dl' dl'' \quad (3.20)$$

$$= 2 \int_{\Delta l=0}^{l_{ct}} \exp\left(-\frac{\Delta l}{l_{ps}}\right) (l_{ct} - \Delta l) d \Delta l \quad (3.21)$$

$$= 2l_{ps}l_{ct} - 2l_p^2(1 - \exp(-\frac{l_{ct}}{l_{ps}})) \quad (3.22)$$

In the case  $l_{ct} \gg l_p$ , we can find that the scaling law of an ideal chain as

$$\langle R_{eed}^2 \rangle = 2l_{ps}l_{ct}. \quad (3.23)$$

Since the Kuhn segment length of an ideal chain is defined as

$$\langle R_{eed}^2 \rangle = l_K l_{ct}, \quad (3.24)$$

we can get the relationship between the Kuhn segment length and the persistence length as

$$2l_{ps} = l_K. \quad (3.25)$$

On the other hand, if  $l_{ct} \ll l_{ps}$ , we will find

$$\langle R^2 \rangle = l_{ct}^2. \quad (3.26)$$

The equation 3.26 show the transition process from a rod-like structure to a coil structure.

### 3.1.3 The Self-avoiding Walk (SAW) Model

While the ideal chain model and the worm-like chain (WLC) model are inspiring in some aspects, they are still not realistic enough to describe the behavior of a real polymer in solvent. In the ideal chain model, the same position in space can be visited by different monomers several time, which is actually not possible. To devise out a more realistic model, the constraint that no point can be revisited is added. Accordingly, a self-avoiding walk model is developed [136].

By definition, a self-avoiding walk of a chain with  $N$  monomers in the  $d$ -dimensional lattice  $Z^d$  starting at  $x$ , is described as a path  $w = (w_0, w_1, \dots, w_n)$  with  $w_j \in Z^d, w_0 = x, |w_j - w_{j-1}| = 1, j = 1, 2, \dots, n$ ; and  $w_i \neq w_j$  for  $i \neq j, 0 \leq i \leq j \leq n$ . We use  $|w| = N$  to represent the length of  $w$ . When  $N$  is enlarged, the number of paths is increased exponentially, which makes it difficult to get a general result. Also, it's difficult to answer the question in high-dimensional case. To understand the behavior of a self-avoiding walk, following questions need to be answered: 1) How many possible paths exist for

an  $N$ -step self-avoiding walk? 2) What's the average end-to-end distance for an  $N$ -step self-avoiding walk assuming that all conformations have the same probability? 3) What's the asymptotic behavior of an  $N$ -step self-avoiding walk, when  $N$  goes to infinity? For the simplest case, where dimension  $d = 1$ , a self-avoiding walk will go towards the same initially chosen direction. Hence, for each value of  $N$ , there are only two paths available. And the maximum distance from the end to the origin point is exactly  $N$ . Above dimension  $d = 4$ , the critical exponent of a self-avoiding walk is dimension independent. When the dimension is high, the SAW will be closer to a simple random walk. Assuming that the scaling limit is the law of the path  $n^{-\nu}w$  when  $n \rightarrow \infty$ , where  $w$  is an  $N$ -step self-avoiding walk, and that the limit exists and is conformally invariant, for  $d = 2$ , it is conjectured to be  $SLE_{\frac{8}{3}}$  (stochastic Loewner evolution, SLE for short), and for  $d = 4$ , the scaling limit is believed to be Brownian motion by using the logarithmic correction factor  $[\log N]^{\frac{1}{4}}$ . And for dimension  $d \geq 5$ , the corresponding scaling limit has been proved as a Brownian motion. Unfortunately, for  $d = 3$ , it's still not well understood. To evaluate the average end-to-end distance from the origin to  $x$  after  $N$ -steps, we can define  $P_n$  as the uniform probability measure on  $\Gamma_N$ , which is the ensemble of  $N$ -step self-avoiding walks initiating at 0. Let  $C_N$  be the cardinality of the set  $\Gamma_N$ , then  $P_n$  is given by

$$P_n(w) = \frac{1}{C_N}, \forall w \in \Gamma_N. \quad (3.27)$$

According to Madras and Slade [142], when dimension is larger than four ( $d \geq 5$ ), the conjectured behavior of  $C_N$  is

$$C_N \propto A\mu^N N^{\gamma-1}. \quad (3.28)$$

And the mean-squared displacement  $E^{P_n}[|w(N)|^2]$  behaves as

$$E^{P_n}[|w(N)|^2] \propto D N^{2\nu}, \quad (3.29)$$

where  $E^{P_n}[\cdot]$  is the expectation value of the uniform measure  $P_n$ , and the values  $A, \mu, D, \gamma$  are positive constants that depend on the dimension.  $\mu$  is named as the connective constant, and  $\gamma$  as well as  $\nu$  are called critical exponents. For dimension  $d = 4$ , the relationships are given by

$$C_N \propto A\mu^N [\log N]^{\frac{1}{4}}, \quad (3.30)$$

and

$$E^{P_n} [|w(N)|^2] \propto D [\log N]^{\frac{1}{4}}. \quad (3.31)$$

In summary, the critical components  $\nu$  and  $\gamma$  take following values under different dimensions:

$$\nu = \left\{ \begin{array}{l} 1, \quad d = 1, \\ 3/4, \quad d = 2, \\ 0.588, \quad d = 3, \\ 1/2, \quad d = 4, \\ 1/2, \quad d = 5. \end{array} \right\}, \quad (3.32)$$

and

$$\gamma = \left\{ \begin{array}{l} 43/32, \quad d = 2, \\ 1.162, \quad d = 3, \\ 1, \quad d = 4, \\ 1, \quad d = 5. \end{array} \right\}. \quad (3.33)$$

### 3.1.4 The Equilibrium Globule Model

In this model, the repulsive interactions between monomers are assumed to be overwhelmed by the attractive interactions, or the polymer is included inside a small enough confining volume. Under either scenario, the polymer will self-organize into an equilibrium globule conformation. And we will see that the root mean squared end-to-end distance scales with the polymer length ( $N$ ) as

$$R \sim N^{\frac{1}{3}}, \quad (3.34)$$

and the occupied volume of the polymer scales linearly with the polymer length as

$$V \sim R^3 \sim N. \quad (3.35)$$

The monomer density is calculated as

$$\rho = \frac{N}{V}, \quad (3.36)$$

It is a constant, which means a uniform distribution of monomers. Grosberg *et al.* demonstrate that inside single globule, sub-chain is organized in a different

way [139], which behaves roughly like a random walk included inside a confining volume as

$$R \sim N^{\frac{1}{2}}. \quad (3.37)$$

For a random walk, the occupied volume of the polymer scales with the polymer length as

$$V \sim R^3 \sim N^{\frac{3}{2}}. \quad (3.38)$$

Accordingly, the density depends on the polymer length as

$$\rho \sim \frac{N}{V} \sim N^{-\frac{1}{2}}. \quad (3.39)$$

Therefore, the root mean squared end-to-end distance for a subchain scales as

$$R(s) \sim \begin{cases} s^{\frac{1}{2}} & \text{for } s \leq N^{\frac{2}{3}} \\ \text{const} & \text{for } s > N^{\frac{2}{3}}. \end{cases} \quad (3.40)$$

Furthermore, Lua *et al.* [143] show that the contact probability for a subchain included in an equilibrium globule scales as

$$P_c(s) \sim \begin{cases} s^{-\frac{3}{2}} & \text{for } s \leq N^{\frac{2}{3}} \\ \text{const} & \text{for } s > N^{\frac{2}{3}}. \end{cases} \quad (3.41)$$

Since an equilibrium globule polymer is featured by intensive entanglements verified by numerical simulations and theoretical calculations [18, 144, 145, 146], it is less possible to be a candidate model for chromatin architecture.

### 3.1.5 The Fractal Globule Model

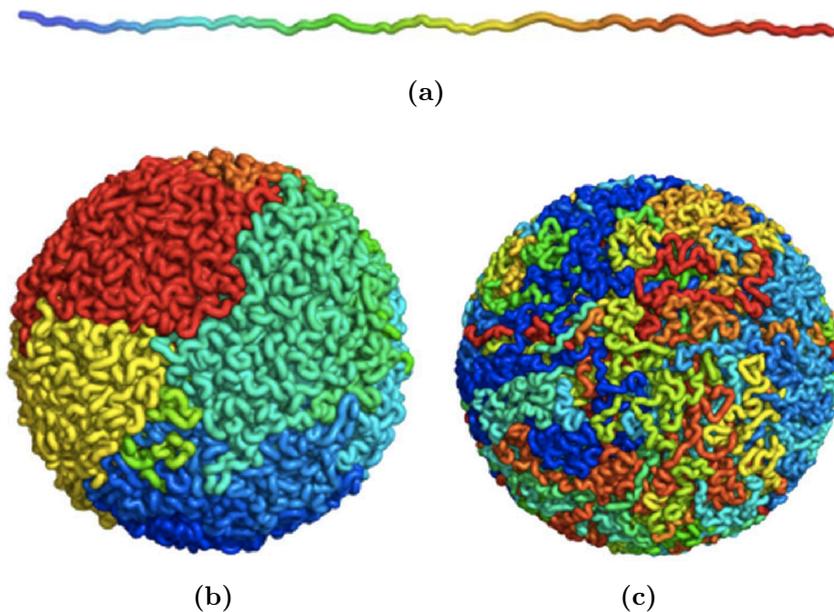
The fractal globule model (or called the crumpled globule model) has been proposed by Grosberg *et al.* [147]. According to this model, a polymer is compacted into a series of connected crumples at first; crumples formed serve as monomers for further rounds of compaction until a single large crumple globule is formed. The fractal globule is actually a hierarchical structure composed of crumples at different scales with a self-similar structure. Like the equilibrium globule, the volume of the polymer scales with the polymer length ( $N$ ) as

$$R \sim N^{\frac{1}{3}}. \quad (3.42)$$

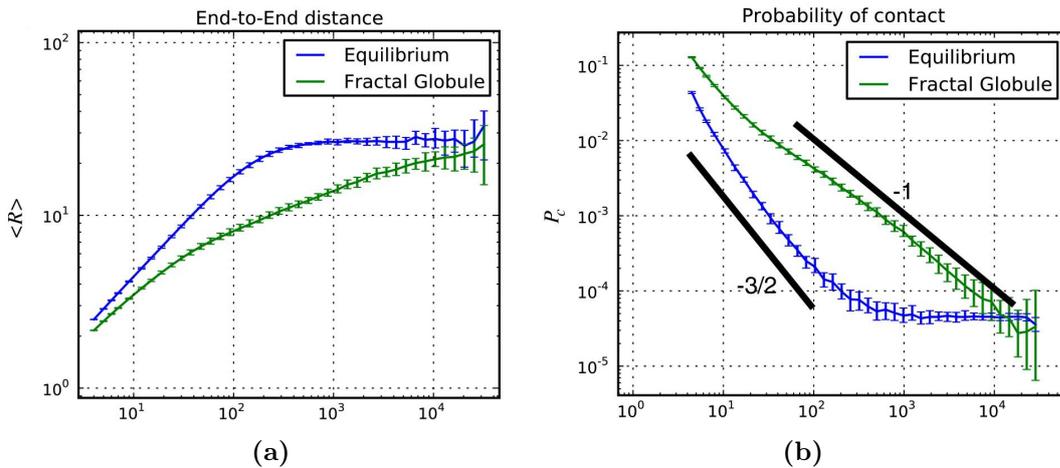
Due to the formation of the fractal globule on all scales, the sub-chain also scales linearly with the sub-chain length ( $s$ ) as

$$R(s) \sim s^{\frac{1}{3}}. \quad (3.43)$$

Numerical simulations with a 500,000 monomers long polymer for the fractal globules and the equilibrium globules show that the contact probability of the former have a robust scaling over a wide range of polymer lengths as  $P_c(s) \sim s^{-1}$ , and the latter displays a different exponent as  $-3/2$  [148]. A visual comparison is presented in Figure 3.1. We can see the differences between the equilibrium globule and the fractal globule in that: 1) the scaling of the end-to-end distance for the fractal globule has a power of  $1/3$ , while for the equilibrium globule the power is  $1/2$ ; and 2) the equilibrium globule shows a leveling-off behavior in the plot of root mean squared end-to-end distance  $R(s)$  (see Figure 3.2).



**Figure 3.1: Folding behaviors of the equilibrium globule and the fractal globule.** (a) shows a polymer continuously labeled with different colors. (b) represents a conformation of the fractal globule, and (c) for the equilibrium globule. Both conformations are folded from the fully extended chain displayed in (a). In the conformation of the fractal globule, sub-chains that are colored differently are well separated, and the boundaries of polymer territories can be clearly seen. By contrast, chains are more mixed in the equilibrium globule. Image adapted from [148].



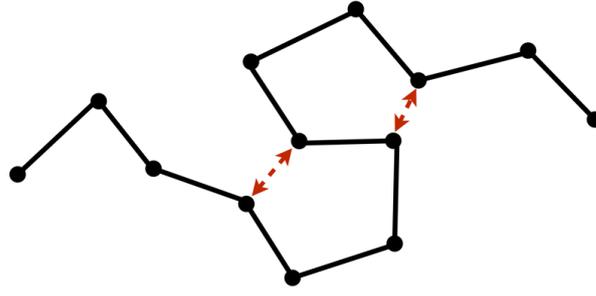
**Figure 3.2: Comparison of the equilibrium globule and the fractal globule models.** (a) shows different scaling behaviors of the end-to-end distance for the fractal globule and the equilibrium globule. For the equilibrium globule, a leveling-off behavior is visible. As shown in (b), the contact probability of the polymer scales as the genomic distance differently in the fractal globule and the equilibrium globule. Image adapted from [148].

A genome-wide chromosomal contacts screening in human performed by Lieberman-Aiden *et al.* has substantiated the fractal globule organization of chromatin in the range from 500 kb to roughly 7 Mb [18].

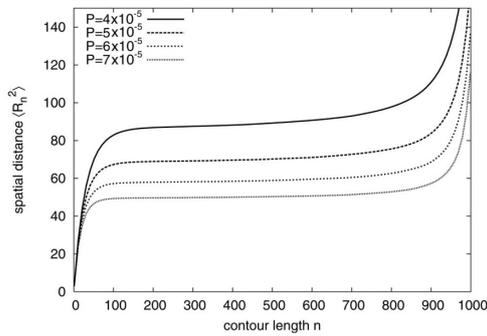
### 3.1.6 The Random Loop Model

The random loop model postulated by Bohn and Heermann [56] states that the backbone of a polymer chain is simplified by a random walk chain, and no excluded volume interaction is considered. Loops are allowed to be formed between any two attachment points of the chain with random polymer lengths. Polymer is coarse-grained by dividing the chromatin fiber into  $N$  equal subunits of equal length  $b$ . The subunits are assumed to be uncorrelated so that they can freely rotate around each other. On the scale, where the number of subunits  $N$  is sufficient small, and the subunit length  $b$  is large than the persistence length of chromatin, the bending energy can be neglected. In this model, two parameters are involved, namely the chain length  $N$  and the looping probability  $P$ . With this model, it turns out that loops on all scales are needed to achieve the leveling off behavior observed in the experiment [54].

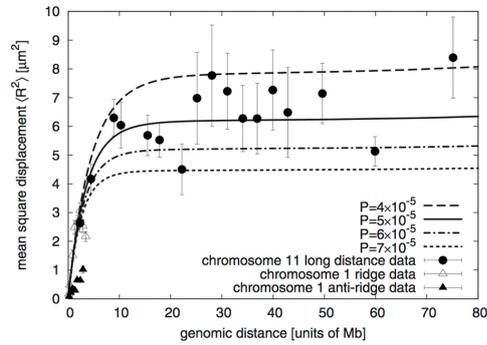
Under short contour lengths, the pattern of mean square displacement as a function of genomic distance behaves like a random walk. Large loops formed by long-range interactions, which are responsible for the collapse of the chain, accelerate the leveling off of mean square displacement. Long-range loops are the key that makes the random loop model having different traits from a simple random walk or a self-avoiding walk model. Actually, only a few loops of large size are enough to cause the leveling off. After approaching the contour length  $N$ , the mean square displacement returns to a random walk-like behavior. With this model, the leveling-off behavior observed over several mega-bases in the experiment [54], where the mean squared end-to-end distance scales as  $\langle R^2 \rangle \sim O(1)$ , can be explained, as shown in Figure 3.3.



(a)



(b)



(c)

**Figure 3.3: The random loop model can explain the leveling-off behavior experimentally observed.** (a) displays a schematic representation of the random loop model, in which the backbone of a polymer chain is represented as a random-walk, and any two monomers on the chain can interact with each other via a harmonic bond under certain probability. (b) shows the relationship between the mean square displacement of two chain segments and the contour lengths under different looping probabilities. The chain is sized as  $N = 1,000$ . Loops of all sizes are allowed. In Figure (c), the experimental data from the study [93] is compared with the random loop model with different values of  $P$ . The experiment measured the mean square displacements as a function of genomic distance for different chromosomes (chromosome 1 and 11), as well as for different compartments on the same chromosome. Image (b) and (c) adapted from [56].

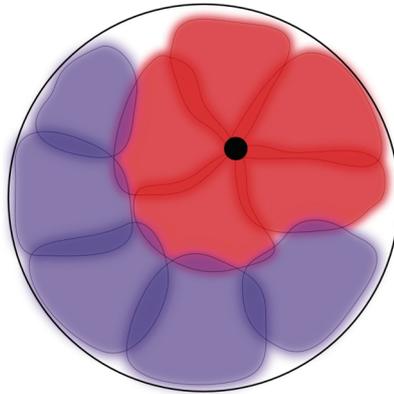
## 3.2 Application of Polymer Models in Elucidating Genome Function

### 3.2.1 Role of Transcription Factory

A simple thermodynamic framework for the role of transcription factory in forming the topological order of the genome has been proposed by Junier *et al.* [149]. They focus mainly on the mechanism by which highly transcribed genes, RNA polymerases and transcription factors could gather into discrete foci, named transcription factories [69, 70]. To this end, they simulate a chromosome with the worm-like chain model, and design a limited number of potential interaction sites along the chain. These sparse sites can interact with each other whenever they are spatially close enough. Based on the results obtained from the numerical simulations, they postulate that clumps formed along the chain mediated by interacting sites can be regarded as a micro-phase separation of these sites from the rest of non-interactive chain. And they predict that the transcription factors assembled inside the transcription factories are responsible for the congregation of genes and mediating physical interactions between them. Moreover, they find that the clustering way of those interacting sites are versatile, for the purpose of enhancing local concentrations of the interacting sites. Linear clustering of members belonging to the same gene family has been widely observed in varieties of organisms. They demonstrate that long-range interactions mediated by distal binding sites could serve as an effective way to group multiply genes from different families, which is increasingly obvious from different experiments, in which genomically distant genes can be identified from the same transcription factory [70, 74].

From a different perspective, Dorier *et al.* has addressed the role of transcription factories-induced inter-chromosomal interactions, and the role of chromosome territories organization in shaping the nuclear architecture [150], as represented in Figure 3.4. With the help of a polymer model, they postulate that gene transcription induced interactions between different chromosome territories are sufficient to explain genomic architecture features experimentally disclosed. With this model, Dorier *et al.* are able to answer the following questions: 1) why chromosomal territories with higher overall expression level

are positioned more frequently in the interior part of the cell nucleus; 2) why chromatins including actively expressed genes appear more frequently on the surface of associated chromosome territories; and 3) why genes are more frequently associated with other genes that have a comparable expression level.



**Figure 3.4: Schematic representation of a genome organization model on account of the transcription factories-induced inter-chromosomal interactions.** Chromatins are self-organized into discrete entities, known as the chromosome territories. Red entities indicate actively expressed domains, while the blue ones are transcriptionally suppressed. Active and inactive chromosomal territories tend to aggregate separately. A transcription factory shown as a black node is surrounded and shared by multiple active domains, which stay more centrally in the cell nucleus. By contrast, inactive domains prefer periphery localizations. Chromatins clustered at on-going transcription sites are more likely to be positioned on the surface of their chromosomal domains. Image adapted from [150].

### 3.2.2 Non-specific Chromatin Interactions

Nooijer *et al.* highlight the importance of non-specific genomic contacts, in contrast to those specific ones mediated by protein-protein interactions or protein-DNA interactions [151]. They postulate that in eukaryotic nuclei both specific and non-specific chromatin interactions contribute to the formation of functional compartment, like chromosome territories [152, 153], nucleoli, and so on. They construct a coarse-grained molecular dynamics simulation to show the chromatin organization of *Arabidopsis thaliana*. They present that non-specific interactions are sufficient to reconstruct already clear *in vivo* localization of nucleoli and chromocenters in *A. thaliana*. The general chromatin

organization generated by this model is in agreement with previously identified Rosette model of *A. thaliana*. They also find that loops that surround chromocenters suppress the aggregation of them.



# Chapter 4

## Gene Transcription-Mediated Chromosomal Interactions

### Chapter Summary

Chimeric transcripts, the potential post-transcriptional processing products, might reflect the spatial proximity of actively transcribed genes co-localized in transcription factories. A growing number of expression data deposited in databases provides us with the raw material for screening such chimeric transcripts and using them as the probes to identify the patterns of inter- and intra-chromosomal gene-gene interactions. On top of the contact pattern from inter-chromosomal interactions, we also observe an exponential behavior for the intra-chromosomal interactions within a certain length scale, which is consistent with the independent experimental results from Hi-C screening and with the Random Loop Model.

Transcription induced chimeric transcripts sheds light on the spatial organization of chromosomes. These inter- and intra-chromosomal interactions might contribute to the compaction of chromosomes, their segregation and formation of the chromosome territories and their spatial distribution within the nucleus.

### 4.1 Introduction

Numerous studies, focused on the transcriptomes of model organisms, have revealed novel ways, by which the genomic information is stored and organized. One of the striking examples is the chimeric transcripts composed of

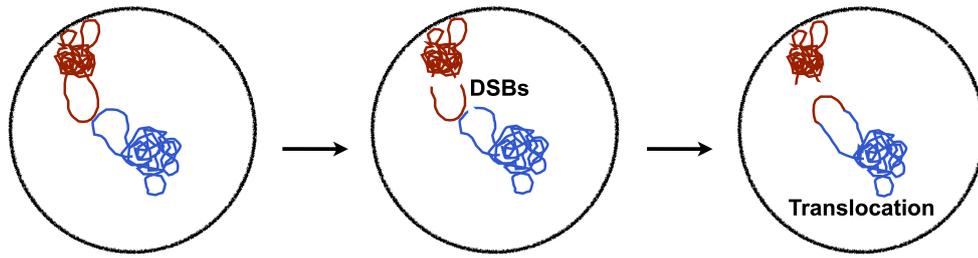
sequences from different sources. So far, three categories of chimeric transcripts have been identified: those that are encoded by two or more gene loci, those that are encoded by two different strands of the same locus, and those that have the shuffled exon order compared to their sequences [154]. Among them, the chimeric transcripts consisting of heterologous sequence segments originating from distant regions on the same chromosome, or from two or more distinct chromosomes, which is known as the transcription induced chimerism (TIC) [16, 17], are of especial interest. Such transcriptional behavior disrupts the conventional belief that the information stored in the genomic sequences is transferred to RNAs in a co-linear fashion. This attribute might contribute to the complexity and diversity of the genome of organisms, but its significance remains to be assessed. Chimeric transcripts of this type have been reported for a variety of organisms every now and then. Transcripts from large-scale transcription databases might be the contaminants as a consequence for various reasons [155, 156], such as the defects in experimental design, or the contamination from the host genomic sequences. Some hybrid transcripts have indeed been verified via *in vivo* experiments and their number is continually increasing [154]. Thus, there is good reason to take chimeric transcripts serious.

To decipher the underlying mechanism of the generation of chimeric transcripts, several hypotheses have been postulated. One plausible explanation is the trans-splicing, first discovered in trypanosomes [157], a process in which the splicing machinery recognizes and cleaves nascent transcripts at their consensus splice sites (GU-AG), and then heterologous exons are jointed together [158, 159]. Subsequent studies confirmed the ubiquitous nature of such chimera in a variety of organisms, including protist [160, 161, 162, 163], plants organelles [164], and higher eukaryotes [165]. But its occurrence is rare [166]. Cases from lower eukaryotes are better understood than those from higher eukaryotes; and some chimeric transcripts were shown to be indispensable to normal cellular functions [167, 168]. However, with the continually growing number of chimeric transcripts reported, only a small portion of chimera was found to fit well with this model [169]. Later, in early 2009, Li *et al.* showed that a high percentage of chimeric transcripts have short homologous sequences (SHSs) at the junction sites in between the fusion sequences. Based on this observation, they proposed a transcriptional-slippage model [170], and tested it

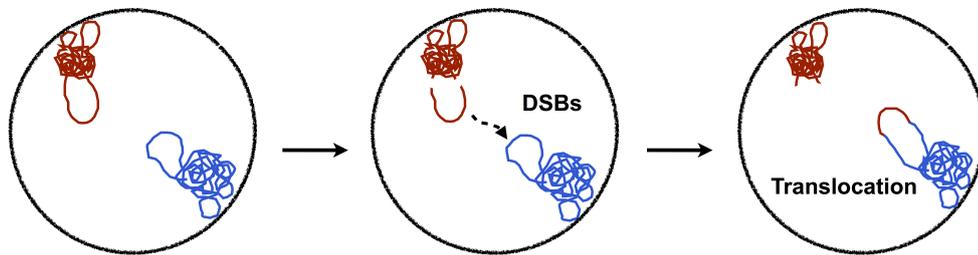
via in vitro and in vivo experiments. According to this alternative mechanism, SHS site on one pre-mRNA can misalign with SHS site on another pre-mRNA, which makes the template switching during transcription and the consequent transcriptional slippage possible. With this model, now a lot of chimera have clear origin, but still some others have their source unknown.

It is generally acknowledged that chimeric transcripts derive from the genomic rearrangement of the transformed cells. However, recent studies suggested that chimeric transcripts are more common in normal tissue than previously appreciated. Some chimeric transcripts expressed in genetically rearranged neoplastic cells were shown to exist in normal cells as well [16, 17], like human fusion genes JAZF1-JJAZ1 [171], SLC45A3-ELK4 [172], and IGH-BLC2 [173]. Except for a few cases reported, most of the chimeric transcripts identified in higher eukaryotes show no function yet. But most recent study indicates that chimera, which is an oncogenic transcription factor, can lead to mistarget on the chromosome and has the ability to alter chromatin structure, as reported in the case of EWS-FLI fusion protein, which causes a malignant bone and soft tissue tumor [174]. It is assumed that chimeric transcripts expressed in normal and transformed cell might be mediated by two distinct mechanisms (transcription induced or chromosomal translocation induced). The production of chimeric transcripts in normal cells mediated by pre-RNA joining mechanisms could be the prelude to the chromosomal translocation events, since chromosome translocation needs physical proximity between loci as well. Increasing evidence shows that the spatial proximity of chromosomes or genetic loci in interphase cell nuclei is strongly correlated with the translocation frequency [175, 176]. Based on these findings, the ‘contact-first’ model has been postulated [177, 178], as shown in Figure 4.1a. According to this model, broken chromosomes can only be translocated to other chromosomal loci when chromatin fibers are co-localized, and at the time double-strand breaks are created. Controversial findings, however, have also been demonstrated from *S. cerevisiae*, in which intra-chromosomal single strand annealing of homologues sequence occurs with a comparable frequency as inter-chromosomal translocation [179]. This observation has given rise to the ‘break-first’ model, as shown in Figure 4.1b, which indicates that broken ends of chromosomes are able to travel throughout nucleus without many constraints while seeking their

homologous loci. However, this ‘break-first’ model might only be true for those budding yeast-like simple eukaryotes, whose nucleus densities are relatively low. For mammalian cells, the ‘contact-first’ model is more probable, since mammalian nuclei are more crowded, and chromosomes are organized into individual territories, which makes the diffusion of broken chromosomal segments not easy. More recent evidence stems from fluorescent in situ hybridization (FISH) experiments performed by Osborne *et al.*, in which they show that the human gene MYC can co-localize to the same transcription factory with another gene IGH located on the different chromosomes [70]. In human, these two loci are the recombination hotspots, where the most frequent translocation events were observed in plasmacytoma and Burkitt’s lymphoma. It is reasonable to assume that since different sequence modules of chimeric transcripts need to be conjugated, their expression should be concomitant, and that distinct pre-RNA products should be confined in a specific cell sub-compartment with close three-dimensional proximity. The candidate of such structure might be the transcription factory [180], a long-lived discrete structure in the cell nucleus. The fact that fewer transcription factories exist than the number of expressed genes is revealing [180], which indicates that multiple actively expressed genes need to share the same transcriptional machinery [70], and that genes separated by a long genomic distance can migrate to each other closely [181, 182], which might expedite the potential long-distance interaction between genes. Thus, spatial proximity of distinct transcripts in the transcription factories might facilitate the production of chimeric transcripts due to the expression activity, which in turn can be used as the signal of chromosome loci contact [183], as represented in Figure 4.2.

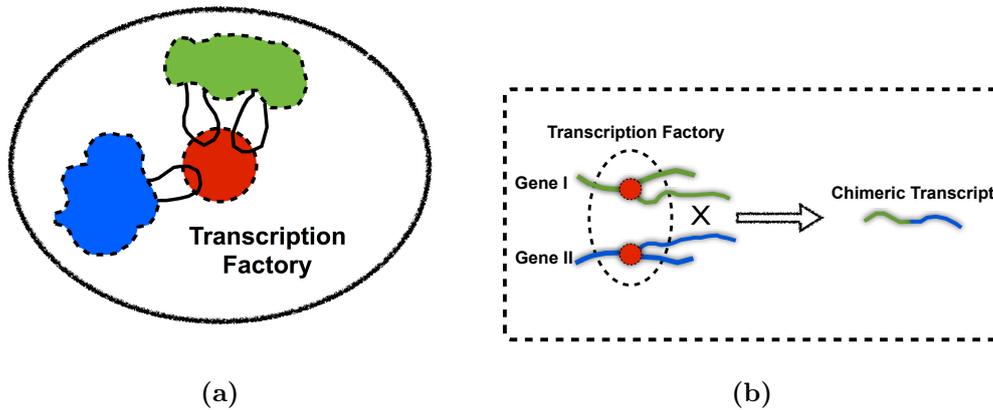


(a)



(b)

**Figure 4.1: The ‘contact-first’ and ‘break-first’ models for the formation of chromosome translocation.** According to the ‘contact-first’ model, physical contact between chromatin fibers is the prelude to the join of double-strand breaks (DSBs). And for the ‘break-first’ model, broken chromosome segment can travel within a large space till the occurrence of chromosomal recombination event.



**Figure 4.2: Schematic representation of the ‘Transcription Induced Chimerism’ (TIC).** As shown in Figure (a), chromatin fibers from the same chromosome or from different chromosomes (one in blue, and the other in green) that are loosely compacted and loop away from their respective chromosome territories for expression, could visit at the same transcription factory (colored in red). (b) Error ligation products mediated by those unclear mechanisms can be utilized as the probes to identify physical interactions between genetic loci *in vivo*.

Due to the rapid development of screening methodologies, like chromosome conformation capture technology (3C) [184] and the derivative 4C [4] or 5C [122] technology, we now have a deeper insight into the three-dimensional organization of the chromosome. In the interphase cell nucleus of higher eukaryotes, chromosomes are folded and condensed into a compact but flexible structure, which is critical to store the extremely long chromatin fiber in such a tiny volume, and, at the same time, keep the structure highly mobile to exert multiple functions with different expression profiles [133]. Hence, the architecture of chromatin compression is closely related with gene expression activity. Beyond the scale of nucleosomes, however, the higher-order chromatin structure is still far away from being clear. Various polymer models have been postulated to describe chromatin fiber folding. Among these models, the fractal globule model [143] proposes that chromatin presents an unentangled non-equilibrium globule organization achieved by progressively crumpling chromatin fiber into a series of small globules which function as the element for crumpling in subsequent round of folding. Experimental support for this model came from the *in vivo* screening on the chromatin contact. In 2009,

Lieberman-Aiden *et al.* developed a Hi-C method, with which genome-wide chromatin contact loci can be screened with the resolution of 1 Mb, and the contact partners can be confirmed by high-throughput sequencing method. According to their results, chromatin folding folds as ( $f(x) \propto x^\alpha$ , where  $f(x)$  is the number probability of contact between two loci that are genomically separated by  $x$  base pairs) with a span from 500 kb to 7 Mb with an exponent  $\alpha = -1$  [18]. An alternative view comes from [54, 56, 57] where a chromosome is viewed as an on all scales looped polymer. Not only can this model explain FISH experiments [54] and the experimental results of Lieberman but can also explain the segregation of chromosomes into territories due to entropic repulsion between the loops [59] as well as the shape of the chromosome territories. Thus loops provide a unified model for the organization of the nucleus.

At present the underlying mechanism of transcript fusion is still obscure and a model for transcripts fusion process is needed. However, what seems to be tenable from the evidence presently available is that such a fusion process is closely related to the spatial proximity of actively transcribed genes. In 2007, Unneberg *et al.* published the mapping results of transcription-induced inter-chromosomal interactions by using the chimeric EST as the probe [185]. After that, the high-throughput sequencing methods developed have made feasible the large scale screening on the transcriptomes in higher organisms (i.e. human and mouse). Thus the increased number of data deposited in the databases have made possible a comprehensive study of chimera and its relation to the 3D organization of chromosomes. Here, with chimeric transcripts gleaned from human and mouse expression sequence tag (EST) database, we test the idea of transcription induced chimerism (TIC), and find the functional relevance of chimeric genes in terms of their co-expression pattern and of their GO terms' semantic similarities. Comparing with normalized Hi-C contact map [186] also confirm that these chimeric genes are more likely to be physically proximal. Moreover, besides the inter-chromosomal interaction analysis, we present also the intra-chromosomal gene-gene interactions pattern of human and mouse. Based on interacting gene pairs from distinct chromosomes, we obtain the pairwise contact frequencies of different chromosomes. Unlike the previous study result [185], we show that the arrangement of chromosomes within the cell nucleus is non-random, by comparing with the interaction pat-

tern of chromosomes that are arbitrarily arranged. No evident changes in the inter-chromosomal interaction pattern can be observed based on the chimeric ESTs collected from the normal and the tumor cells. Based on the gene pairs located at every single chromosome, we obtain the relationship between the genomic distance and the gene contact frequencies. A functional behavior ( $f(x) \propto x^\alpha$ ) is observed from our data with the genomic distance ranging from 500 kb to 7-8 Mb, and an exponent  $\alpha = -1$  on a double logarithmic plot, which agrees with the experimental results from Hi-C screening.

## 4.2 Material and Methods

### 4.2.1 Data Collecting and Processing

mRNA data was gathered from gbEST, gbHTC, and gbPRI divisions of NCBI Genbank database (Release 186) [187]. We neglected the transcripts containing mitochondrion genomic elements. From the gbEST division, we collected 8,315,246 EST sequences of human, which had highest expressed sequence tag (EST) coverage, and 4,853,475 sequences from mouse. From the gbHTC division, we obtained 59,897 mRNA sequences of human and 149,903 sequences of mouse. And from the gbPRI division, 470,261 sequences of human were collected. For mouse, no record existed in gbPRI databases. EST library report and annotation [188, 189] were downloaded for restriction enzyme information and library ID number.

To conduct a genome-wide blast, genome sequences were downloaded from NCBI Genome database. Human genome reference build (hs\_ref\_GRCh37) was downloaded [190]. Mouse genome reference build (mm\_ref\_37) was downloaded [191].

Ensembl human genome database (homo\_sapiens\_core\_64\_37) in mysql format was downloaded [192], and stored by using MySQL for gene mapping. For mouse, mus\_musculus\_core\_64\_37 database was downloaded [193]. Ensembl ontology information (Version 64) was download from Ensembl genome database [194] and installed, in order to perform an analysis calculating semantic similarities of GO terms associated with gene pairs.

Recognition sequences of restriction enzymes were download from REBASE [195].

## 4.2.2 Sequence Alignment

We used BLAT [196] for sequence alignment. Different from BLAST, Blat works by constructing an index of an entire genome in memory. Therefore, the target database is an index derived from the assembly of the entire genome, instead of a set of GenBank sequences. Due to this unique structure, Blast runs with a higher efficiency. We set up a local BLAT aligning server for performance consideration, with which we only need to build an index of the whole genome once, and then we can align sequences in a batch at very low cost on the memory. We set the minimal identity threshold as 95% and the rest of the parameters by their default values. After aligning, identity and score were calculated for each alignment, and alignments in each query were sorted according to the score and then the identity.

## 4.2.3 Chimeric Transcripts Identification

Similar to previous studies [170, 185], we adopted their method with some improvements. Based on sorting results, we picked out the queries with top two ranked alignments having the sequences identity not less than 95%, and minimal alignment length not less than 50 bp for gbEST sequences. Considering the probability that the chimeric transcripts might be the noisy products from the cell transcriptional process and the accompanying uncertainty on the conjugation position, we set the shortest alignment length to 50 bp, instead of the length 100 bp commonly used for the potentially functional transcripts. For gbHTC and gbPRI sequences, which are much longer, a threshold of 100 bp was used instead. A 10 bp overlapping at the binding site of the query was allowed due to the uncertainty in aligning.

Further, several validation steps were performed to remove potential artifacts. In order to make sure that the two best alignments in a query were unambiguously aligned, a uniqueness criterion was imposed. If one query had more than two alignments, we compared the lower ranked alignments with the best two, respectively. If the two compared alignments had more than 80% similarity and their target regions were not overlapping, this query was treated as an ambiguous aligning, and eliminated from further consideration. After ambiguity removal, accounting for the artifacts stemming from the liga-

tion procedure in the EST library construction, we inspected the recognition sequence of the restriction enzymes used at the boundary site of the chimeric sequence. If the restriction sites were identified, the corresponding chimeric transcripts were deemed as contamination and discarded. For transcripts from the libraries without annotation on the restriction enzymes we checked them with the most commonly used enzymes for EST library construction. Finally, we reconstructed the chimeric sequences by fusing the two best alignments from the remaining queries that passed all the validation steps.

#### 4.2.4 Gene Mapping

We mapped the fusion sequences onto the chromosomes according to the Ensembl Human Genome database, and selected transcripts whose partners were both mapped onto the gene regions with the same directions as the annotation. Since the ESTs were sequenced with single reading, some sequences were deposited into the database with 3'-direction reading. Thus, for sequences with both partners mapped with the opposite directions of the corresponding genes, we re-evaluated their directions according to their GT/AG pairs ratio, the so-called GT/AG rule [197]. We made statistics on the number of GT/AG pairs from one direction, and the number of complementary CT/AC pairs from the reversed direction. If the GT/AG to CT/AC ratio is smaller than 1, then we reversed its orientation. We retained the transcripts with both partners re-mapped to the same directions as the annotation. Transcripts with their directions difficult to assess were removed. Then, we separated the remaining chimeric sequences into two groups, according to gene partners' location (whether they originated from the same chromosome, or from distinct chromosomes). For intra-chromosomal interactions, considering the possibility of tandem transcription or intergenic splicing, we removed the chimeric transcripts with genomic distance between gene loci smaller than 500 kb.

Similar to [170], we made a scrutiny on the boundary sites of these retained transcripts to check whether Short Homologous Sequences (SHSs) exist at the junctions of their source genomic sequences. We took the SHSs with length from 3 to 10 bp, and calculated their abundance.

### 4.2.5 Co-expression Pattern of the Chimeric Gene Partners

Co-expression pattern of 19,777 human genes and 21,036 mouse genes was obtained from COXPRESdb Version 4.1 [198, 199]. In this study, the Mutual Rank (MR) score was used as a measure for gene co-expression correlation. For a given gene pair, like gene A and gene B, the MR is calculated by averaging the rank of gene B in the co-expressed genes to gene A and the average of the rank of gene A to gene B. The smaller the MR value is, the stronger positive correlation the gene pair has. Co-expression correlation data was processed and stored in a MySQL database. Genes were indexed by their Entrez Gene IDs. We grouped the inter- and intra-chromosomal interactions together. Ensembl Gene IDs of gene pairs from those validated chimeric transcripts were obtained after Gene Mapping procedure. R package ‘biomaRt’ (Version 2.10.0) was used for mapping the Ensembl Gene ID to the Entrez Gene ID. For each chimeric transcript, the MR value was extracted by querying gene pair’s Entrez Gene IDs. To check whether the co-expression correlation of gene pairs in the chimeric transcripts are non-preferential, MR scores of all possible gene pairs were used as a background. Plots of the MR value’s distribution were plotted for comparison.

### 4.2.6 Semantic Similarities of GO Terms Associated with Chimeric Gene Partners

In order to have a deeper insight into the functional similarities of chimeric gene pairs, we employed the algorithm as described in [200]. With this method, functional similarities between two genes were determined based on their annotated Gene Ontology (GO) information. For a given gene pair, their GO term’s semantics were measured by their semantic contributions of their ancestor terms in the GO graph. We analyzed similarities of GO molecular function terms, GO cellular component terms, and GO biological process terms of chimeric gene pairs, and compared them with randomly selected gene pairs. A python module named Networkx [201] was used for graph manipulation and calculation.

## 4.2.7 Comparison with Normalized Hi-C Contact Pattern

We used normalized Hi-C contact pattern of human lymphoblast generated by Yaffe *et al.* [186]. From the raw Hi-C contact data, they identified several systematic biases, and removed them by use of an integrated probabilistic background model. After renormalization, the corrected contact maps constructed by two different restriction enzymes are more comparable. The pre-prepared data was download from the author’s website [202]. For each kind of enzyme restricted map (HindIII and NciI), fragments were aggregated by a 1 Mb bin. The ratio of the observed to the expected contact number in each bin was transferred by the logarithm to base 2 as a measurement of contact enhancement. The distribution of all possible inter- and intra-contacts were used as the backgrounds for comparison. Gene pairs of the chimeric transcripts were mapped onto the corresponding bins of the chromosomes, and the enhancement was screened. For genes spanning over more than one bin, the contact enhancement was calculated by averaging weighted contact enhancement for each 1 Mb bin pair covered. The distribution of contact enhancement observed from the inter- and intra-chromosomal chimeric transcripts were compared to the background, respectively. The Kolmogorov-Smirnov test was applied to check the difference of the contact enhancement between chimeric gene pairs and the background.

## 4.2.8 Inter- and Intra-chromosomal Interactions Pattern Analysis

A parallel analysis was performed on sequences from three different expression databases. Due to the limited amount of sequences from gbHTC and gbPRI databases, we only used data from gbEST for interaction pattern analysis.

From the inter-chromosomal interactions identified, some chimera overlapped and had the same gene partners. Therefore, to remove the over-counts of these interactions between the same gene partners, we grouped chimera on the basis of gene partners participating in the interactions. From these non-overlapping inter-chromosomal interactions, we picked out transcripts composed of two genes labeled by ‘protein coding’ biotype in the Ensembl database,

which were far more abundant than others, as the observed interaction pattern. We collected the frequencies of contact between two given chromosomes to obtain the interaction matrix, and plotted the data as a heat-map. Also, a random interaction matrix was constructed to screen if interaction preference existed on the chromosome partners. From the Ensembl database, we obtained the frequencies of protein coding genes on 23 chromosomes except chromosome Y. With these expected gene frequencies, we constructed a chromosome pair-wise interaction matrix by multiplying the gene frequencies from any two given chromosomes. The diagonal elements of this matrix were all zero. For this symmetric matrix, each element in the upper half of the matrix was multiplied by a random number, the corresponding element in the lower half was assigned the same value. This finished a shuffle step on the interaction pattern, and the eigenvalue of the matrix was calculated by using the software R. Totally,  $10^5$  steps were carried out. After that, the mean eigenvalue was obtained and plotted against the eigenvalue from the observation. Also, 95% percent confidence interval of the eigenvalue from the simulation was calculated by using software R. Furthermore, based on the cell source annotation from the EST library report, we separated the interactions from normal cells with those from tumor cells, and plotted the eigenvalue of interaction, respectively to check if some difference existed.

For intra-chromosomal interactions, transcripts with partners mapped to the identical gene partners but to different regions of respective gene were grouped. Further, we made statistics on the frequencies of the gene-gene interaction under given genomic distance within every chromosome, and plotted the frequencies of the interactions versus the genomic distance in a double logarithmic plot. Due to the large fluctuation of data points for the long genomic distance, logarithmic bins were used to minimize the fluctuation. At the short length scale, genomic distances between interacting genes were grouped by a fixed bin size. The contact probabilities were divided by their corresponding genomic distance for normalization. Next, we replotted the contact frequencies versus the normalized genomic distance.

As mentioned before, steps done on human expression data were also done on the mouse expression data.

## 4.3 Results and Discussion

### 4.3.1 Chimeric Transcripts Identification and Validation

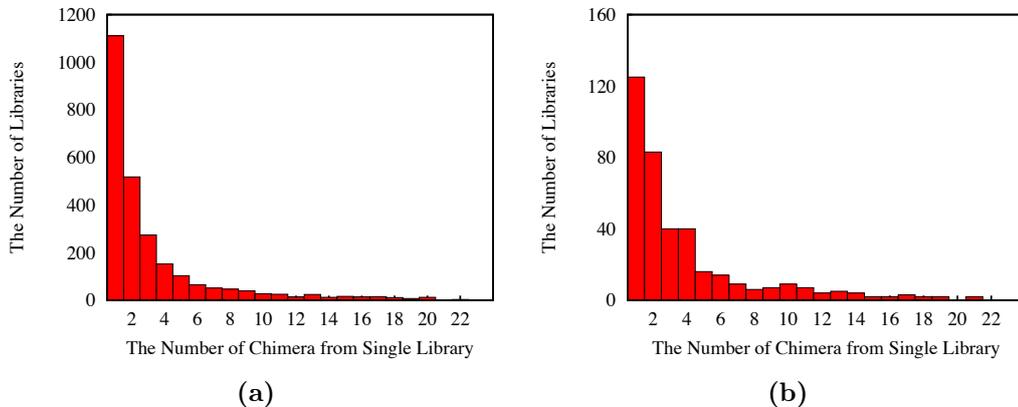
After aligning EST sequences with BLAT, we filtered chimeric transcripts with the selection criterion described in the Material and Methods section. Details on the number of transcripts that remained after each validation step are listed in Table 4.1. Transcripts downloaded from the gbPRI and gbHTC databases are the products from direct sequencing of intact mRNA sequences. Thus, they have higher quality than the ESTs data. Due to the small size of gbPRI and gbHTC databases, and accordingly the limited number of chimeric transcripts identified, they were not suitable to be used as the probes for gene loci interaction analysis. But the fact that still some chimera can be found from gbPRI and gbHTC databases suggests that at least some chimeric sequences were not results of artifacts or contaminations. Information about the chimeric transcripts identified from three different databases can be found at the additional file 1, 2, and 3, respectively.

| Interaction | Human      |             | Mouse      |             |
|-------------|------------|-------------|------------|-------------|
|             | Uniqueness | Restriction | Uniqueness | Restriction |
| Inter       | 35,676     | 33,958      | 6,141      | 5,729       |
| Intra       | 16,255     | 15,885      | 4,890      | 4,782       |

**Table 4.1: Validation results of chimeric transcripts.** The **Uniqueness** column corresponds to the number of transcripts remained after removing those ambiguous ones. And the **Restriction** column shows the number of transcripts with no recognition sequence at their boundary sites.

Next, we analyzed the distribution of human chimeric transcripts among the EST libraries. Of 8,995 libraries screened, 2,617 were involved in the production of chimera. As shown in Figure 4.3, most libraries contributed very few chimera. 1,111 and 518 libraries from a total of 2,617 libraries provided only one and two chimera, respectively. Such a result indicates that chimera formation is an ubiquitous event with low frequency. A previous study performed by Sorek and Safer [155] identified those highly contaminated EST libraries. We checked the libraries contributing more chimeric sequences in our study,

and found that none of our libraries overlapped with those problematic ones.



**Figure 4.3: The distribution of human and mouse EST libraries according to the number of chimera they produced.** (a) for human, (b) for mouse. The x-axis represents the number of chimera that can be identified from single library. The y-axis indicates the number of libraries producing the corresponding amount of chimeric ESTs.

### 4.3.2 Gene Mapping Results

After validation, we mapped the chimera to the human genome based on current annotation from the Ensembl database. Chimeric transcripts having sequence modules mapped to intergenic regions were disregarded, since these intergenic components might be derived from the contamination of host genomic sequences [155], or pertain to the sequences with no annotation in the genome. Curiously, some chimeric transcripts were composed of one segment mapped to a gene with the correct orientation and the other segment mapped to the gene with the opposite direction indicated in the annotation. In other words, one segment of the chimeric transcript was mapped to the sense strand of one gene, while the other was mapped to the anti-sense strand of another gene. Similar observations were found from the pilot ENCODE studies, in which chimeric RNAs were composed of sequences of genes that mapped to the opposite strand of the index gene [2], but its significance remains unclear. Thus, we only retained transcripts with both partners mapped to the gene and oriented to the same direction as that indexed in the annotation. The transcripts passing all the validation steps were the potential candidates for

gene loci interaction. Most interactions were observed only once, while some gene partners associated with each other more frequently. In addition, some of these re-occurring chimera can be identified from multiple different EST libraries, while other repeats originated only from a single library. Due to the redundancy of chimera, we categorized them according to the gene pairs associated to obtain unique gene loci interaction pattern. Intra-chromosomal interactions with genomic distance larger than 500 kb were retained. Detailed mapping results are shown in Table 4.2. Results of the chimera selection with high stringency (100 bp) from gbPRI and gbHTC database are given in Table 4.3. While the number of human and mouse chimera gleaned from the mRNA high-throughput sequencing database is not quite large, there still existed some overlapping gene pairs when compared with chimera picked out from the EST data (see Table 4.3).

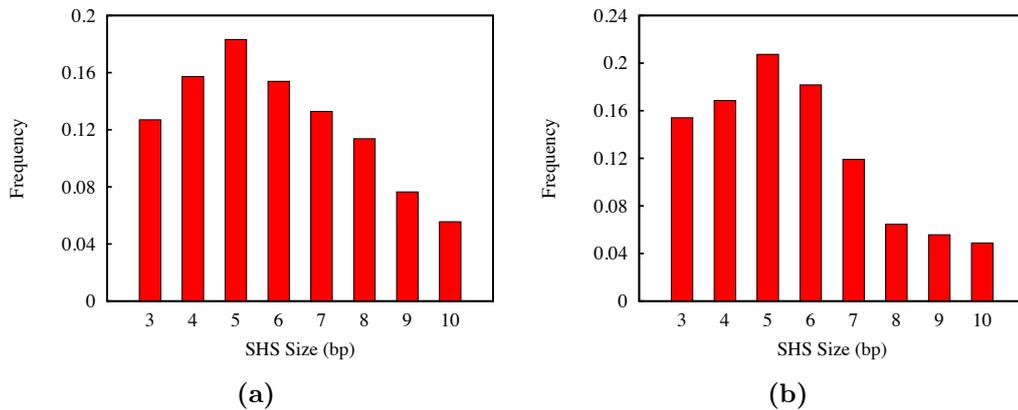
| Interaction Type | Human  |       |       |      | Mouse |       |       |      |
|------------------|--------|-------|-------|------|-------|-------|-------|------|
|                  | GG     | UG    | Genes | Size | GG    | UG    | Genes | Size |
| Inter            | 11,262 | 8,261 | 7,871 | 516  | 2,493 | 2,230 | 3,316 | 580  |
| Intra            | 773    | 673   | 1,165 | 540  | 269   | 223   | 426   | 602  |

**Table 4.2: Summary of the chimeric transcripts.** The **GG** column represents the number of chimeric transcripts whose partners were both mapped onto the gene regions with the same directions as the annotation. The **UG** column indicates the unique gene pairs identified from chimeric transcripts of gene-gene interaction. The **Genes** column corresponds to the number of genes participating in the chimera production. And the **Size** column shows the averaged length (unit in bp) of validated chimeric transcripts.

|          | Human |       |        |         | Mouse |       |       |         |
|----------|-------|-------|--------|---------|-------|-------|-------|---------|
| Database | Inter | Intra | Size   | Overlap | Inter | Intra | Size  | Overlap |
| gbHTC    | 266   | 29    | 2,634  | 51      | 322   | 27    | 1,846 | 71      |
| gbPRI    | 657   | 109   | 15,079 | 72      |       |       |       |         |

**Table 4.3: Chimeric transcripts identified from gbHTC and gbPRI database.** The **Size** column indicates the mean length (unit in bp) of chimeric transcripts identified from the corresponding database and organism. The **Overlap** column indicates the number of gene-gene partners which can be screened from gbEST database and from gbHTC/gbPRI database. For mouse, no record existed in the gbPRI database.

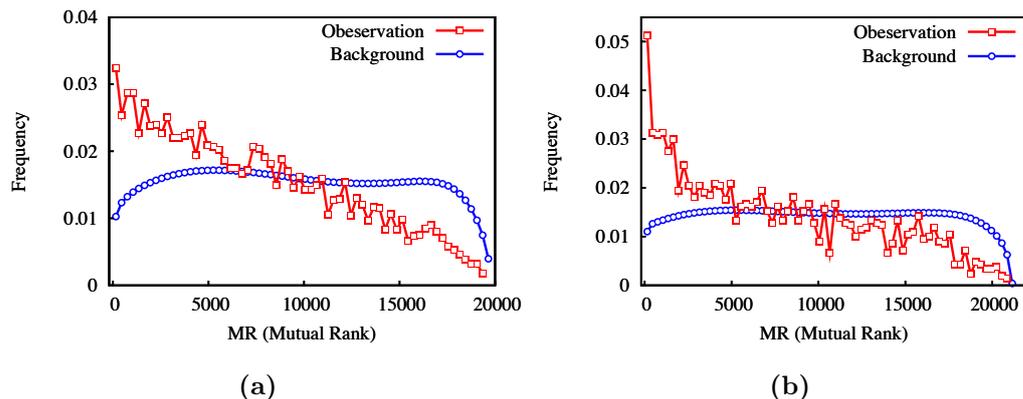
In addition, we checked the Short Homologous Sequences (SHSs) at the boundary sites of all examined homo and mouse chimeric ESTs. Human chimeric sequences with SHSs (size from 3 to 10 bp) accounted for about 42.6% (5,126/12,035). For mouse, the ratio was 57.8% (1,596/2,762). Under either scenario, this ratio is close to the previous study result [170]. The abundance of SHSs with different size was shown at Figure 4.4.



**Figure 4.4: SHSs abundance in validated chimeric transcripts.** SHS length was taken from 3 bp to 10 bp.

### 4.3.3 Co-expression Pattern of the Chimeric Gene Partners

After transforming Ensembl Gene IDs to Entrez Gene IDs, chimeric transcripts with both genes Entrez Gene IDs known were retained for further analysis. Of the total of 8,933 validated human chimeric transcripts, 7,412 gene partners were collected. For mouse, 2,229 gene pairs were obtained from 2,453. As shown in Figure 4.5, gene pairs from the chimeric transcripts associated more frequently with small Mutual Rank (MR) score than with a large one, when compared with the co-expression pattern of all possible gene pairs. This observation revealed that gene pairs producing chimeric transcripts showed stronger positive correlation in terms of their expression activities. This observation further supported the assumption that such chimeric transcripts were transcription-induced [16, 17].

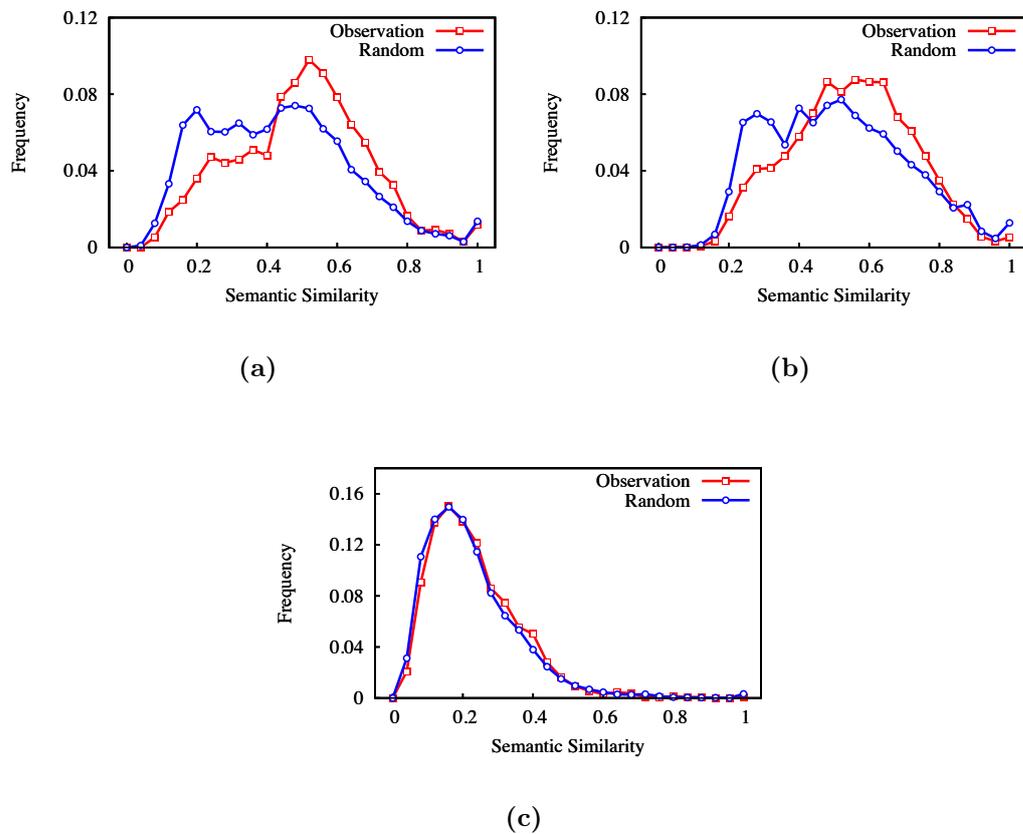


**Figure 4.5: Co-expression pattern of genes from the chimeric transcripts.** (a) for human, and (b) for mouse. The smaller the Mutual Rank (MR) score is, the higher the correlation between gene pairs' expression is.

### 4.3.4 Semantic Similarities of GO Terms Associated with Chimeric Gene Partners

Chimeric gene pairs of inter- and intra-chromosomal interaction were grouped together. For the chimeric transcripts with both genes having GO information annotated, their semantic similarities of GO molecular function terms, GO cellular component terms, and GO biological process terms were calculated,

respectively. Then the distributions of semantic similarities under different GO ontologies were screened. For comparison, the semantic similarities of randomly selected gene pairs with the same sample size were calculated in the same way. As shown in Figure 4.6, similarities of chimeric gene pairs displayed a tendency with larger similarity values, especially obvious in the cases of GO molecular function terms and GO cellular component terms. For GO biological process terms, however, the difference was not so observable. GO terms information for gene pairs of chimeric products can be found in the Supplementary Material. GO molecular function terms and GO cellular component terms have 2858 and 9026 items, respectively, while GO biological process terms have 21010 items, which might be too diverse to congregate genes.

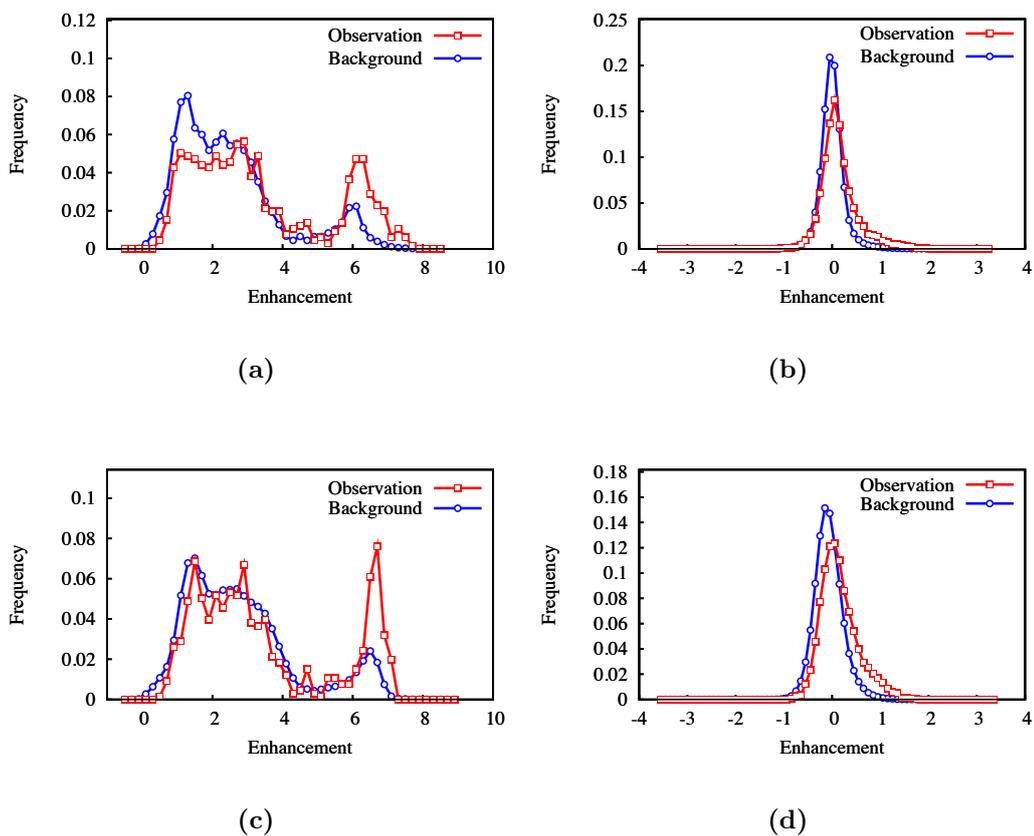


**Figure 4.6: Semantic Similarities of GO Terms Associated with Chimeric Gene Partners.** (a) for GO molecular function terms; (b) for GO cellular component terms; and (c) for GO biological process terms. Semantic similarity ranges from 0 to 1. The larger the similarity value, the more similar the gene pairs are for specific GO ontologies.

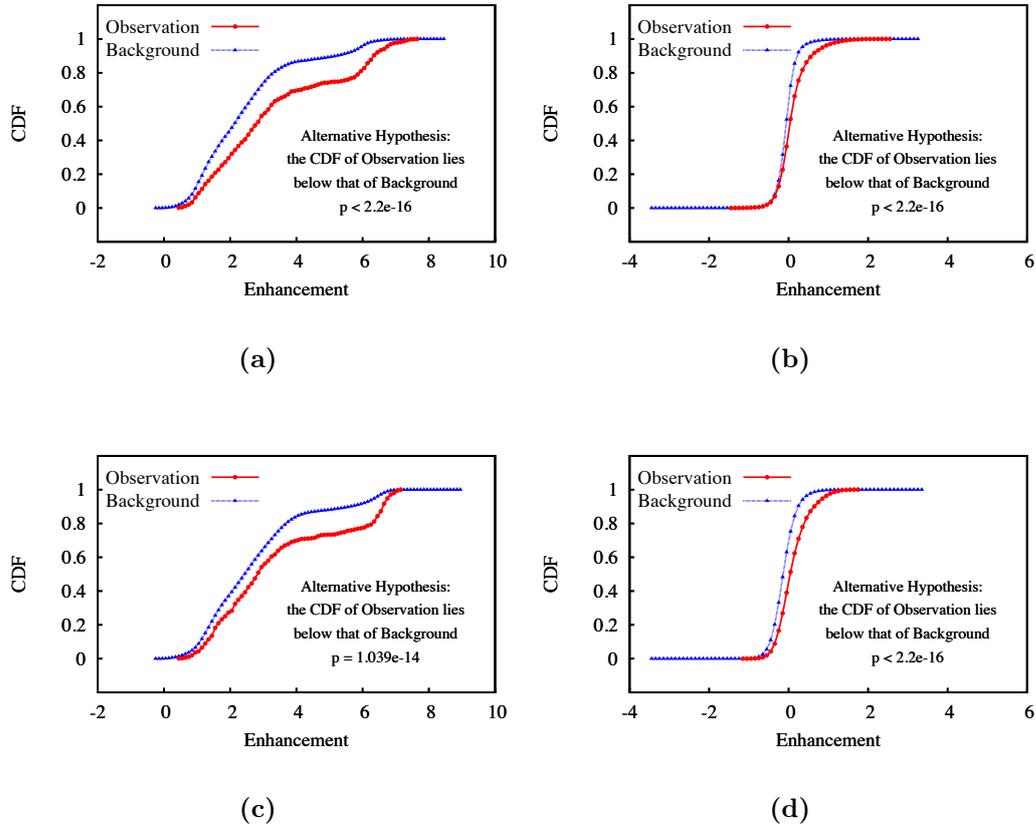
### 4.3.5 Comparison with Normalized Hi-C Contact Pattern

After mapping chimeric genes onto the corresponding chromosome bins, contact enhancement between gene pairs were weighted and averaged for a 1 Mb bin. From the enhancement distributions shown in Figure 4.7, both for inter- and intra-chromosomal chimeric transcripts, a shift towards larger enhancement can be seen with either kind of restriction enzyme. For intra-chromosomal contacts a large enhancement can be seen for the tail, i.e. for gene pairs, which are several Mb far away. These long-range interactions have

a relatively low frequency. Kolmogorov-Smirnov tests show that the p-values are small enough to reject the null-hypothesis that observation and background abide by the same kind of distribution. From the cumulative distribution function (CDF) curves shown in Figure 4.8, one can clearly see that the CDF of observation lies below that of background, which indicates a higher contact probability between chimeric genes is given over that of the random case.



**Figure 4.7: Comparison with Normalized Hi-C Contact Pattern.** Pattern (a) and (b) were extracted from the contact map restricted by the enzyme HindIII, (a) for intra- and (b) for inter-chromosomal contact; Pattern (b) and (d) with restriction enzyme NcoI, (c) for intra- and (d) for inter-chromosomal contact.

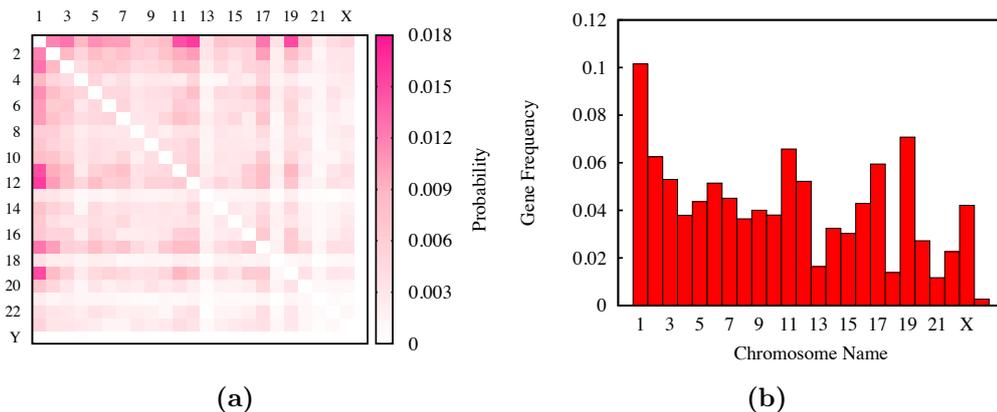


**Figure 4.8: Kolmogorov-Smirnov Tests on the Enhancement Distribution.** Cumulative Distribution Function (CDF) of corresponding distribution pattern shown in Figure 4.7 was plotted. Pattern (a) and (b) were extracted from the contact map restricted by the enzyme HindIII, (a) for intra- and (b) for inter-chromosomal contact; Pattern (c) and (d) with restriction enzyme NcoI, (c) for intra- and (d) for inter-chromosomal contact. P-values from the Kolmogorov-Smirnov tests and alternative hypotheses were shown.

### 4.3.6 Inter-chromosomal Interaction Pattern

Based on the gene mapping results, we find that far more abundant chimeric transcripts consist of two genes belonging to the gene type ‘protein coding’ annotated in the Ensembl database (for human, 6,999 out of 8,261 transcripts). Thus we picked out these unique ‘protein coding’ cases of inter-chromosomal gene interactions and plotted these as a heat-map according to where the chromosomes gene partner was localized. The density of the pink color in

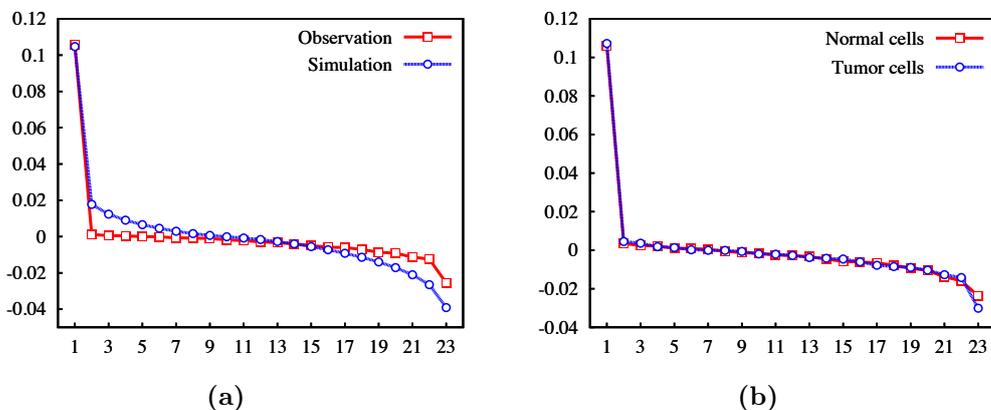
each square of the heat-map (Figure 4.9a) is proportional to the frequencies of chimera composed of gene partners derived from two given chromosomes. As illustrated by the ‘protein coding’ gene frequencies on human chromosomes (as shown in Figure 4.9b), the interactions between chromosomes with higher gene content show a correlation as well. Since the expression data was collected from a variety of cells with different types and developmental stages, it cannot easily be determined whether or not there is a preference on the interaction of given chromosome pairs. But what is evident is that the observed pattern is different from the pattern achieved from random interaction. This comparison can mathematically be done by comparing the eigenvalues of the heat-maps viewed as a matrices (Figure 4.10a). A 95% confidence interval for the eigenvalues based on the simulation pattern was calculated, which however is too small to be shown. None of the eigenvalues from the observed pattern fell into the corresponding confidence interval.



**Figure 4.9: Inter-chromosomal interaction pattern for human chromosomes.** (a) Contact as heat-map of human inter-chromosomal gene interactions. Protein-coding genes were used as a measure of interactions; (b) Protein-coding gene content of human 24 chromosomes.

Further, we separated the chimera according to the cell sources, which they originate from (normal or tumor cell lines). Then, we constructed the interaction matrix and tried to compare the patterns from normal and tumor cell lines. For both cases, we acquired about 3,500 chimera. Indeed, a slight difference is observed for normal and tumor cells, but it is not clear whether such differences are due to the statistics or if these are real (as shown in Fig-

ure 4.10b). Previously, several studies assessing the chromosome positioning in transformed or immortalized cells found similar or comparable pattern in normal cells as well [176]. From murine lymphoma, Parada *et al.* found that pairs of chromosomes that were spatially positioned in close proximity were conserved in normal cell as well [134]. In another study [135], however, alternation in chromosome's radial positioning was prominent from certain tumor cell lines. These findings indicate that cell line or disease specific factors should be taken into consideration. But due to the limited amount of chimeric transcripts picked out from every single library, it is difficult to perform such an analysis in a cell line specific way.

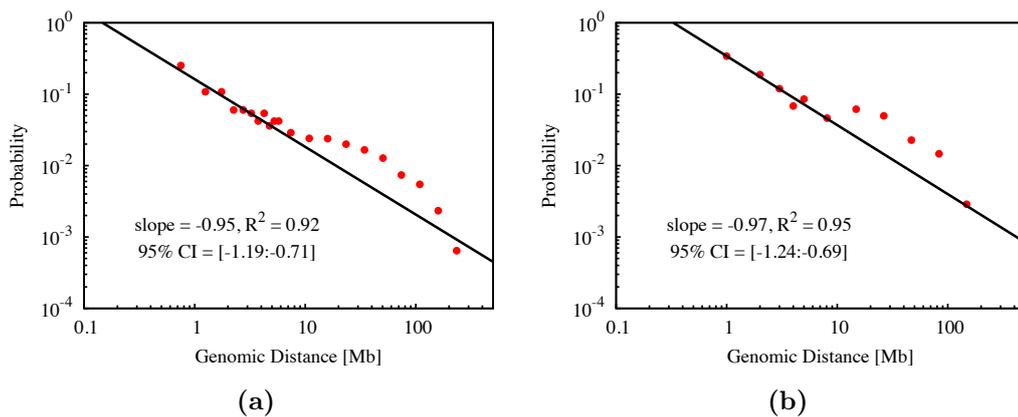


**Figure 4.10: Comparison of inter-chromosomal interaction patterns via eigenvalue.** (a) Patterns from human EST chimeric transcripts and random distribution simulation (excluding the chromosome Y); (b) Patterns from human normal cells and tumor cells (excluding the chromosome Y).

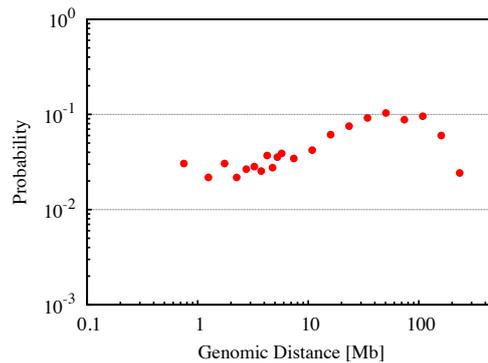
### 4.3.7 Intra-chromosomal Interaction Pattern

For the intra-chromosomal gene loci interaction data, we calculated the frequencies of gene-gene interactions for given genomic distances. The genomic distances between gene partners ranged from 509 Kb to 244.4 Mb. We plotted the probability of contact versus the genomic distance between gene partners on a double logarithmic plot. Two different types of behavior can be observed. Below the distance of 7-8 Mb, a functional  $f(x) \propto x^\alpha$  exists with an exponent  $\alpha = -1$  (see the double logarithmic plot shown in Figure 4.11). Above this distance, a quite different pattern is observed. The rapid drop for large

genomic distances might result from the limitation of interaction data when the genomic distance is increased. Under different threshold values for the chimera selection, the general patterns are very similar. To our surprise, a comparable pattern was observed for mouse as well, even though its expressed sequence tag coverage is not as high as that of human. Due to relative depletion of EST sequences from other model organisms compared to that from human or from mouse, the amount of chimeric sequences detected was too limited to be informative. With caution we speculate that a general strategy for chromatin compaction might be utilized in mammalian cells. On account of the difference in the contact probability of chromatin with different length, we normalized the contact frequencies by the genomic distance and re-plotted the pattern. After normalization, as shown in Figure 4.12, under relative short genomic distance, the contact probability remains stable. When the genomic distance is increased, the contact probability rises until a plateau reached. The dropping at the end might be due to the data depletion.



**Figure 4.11: Power-Law fit of intra-chromosomal interactions from human and mouse.** Intra-chromosomal gene pairs with genomic distance larger than 500 kb were considered. (a) Interaction patterns of human; (b) Interaction patterns of mouse. 95% Confidence Interval (CI) of the slopes are shown.



**Figure 4.12: Human intra-chromosomal interactions with normalized genomic distance.** Pattern for chimeric transcripts collected under the threshold value 50 bp.

## 4.4 Conclusion

From the human and mouse expression sequence tag databases, we identified chimeric transcripts. These gene partners might locate on the same chromosomes or originate from distinct chromosomes. The EST database encompassed artifacts and contaminants owing to its experimental tactics. By applying validation procedures, such artifacts can be eliminated from the raw data of chimera. Based on what we currently know about transcription induced chimerism (TIC) and transcription factories, ‘genuine’ chimera, functional or not, might originate from product interactions while being transcribed. Despite their unknown underlying mechanism of production, several distinct molecular mechanisms have been discovered that are able to mediate some sorts of chimera formation (i.e. trans-splicing or transcriptional-slippage model). The chimeric transcripts, whose occurrence is rare, are thought to be the inevitable burden the cells need to take on for the crowded environment in the cell nucleus. Transcription of pre-mRNA molecules in the same transcription factory might bring them close to each other and facilitate their aberrant interaction. Due to the increasingly accumulated expression data, we can perform an extensive screening on chimera production. We found that 29% of human libraries investigated yield chimeric sequences and among these libraries, most of them contributed a limited number of chimera. This fact might reflect the biological nature of chimeric transcripts, which in general

are the unavoidable but rare erroneous ligation products from transcriptional process. Some EST libraries, however, contributed a much larger amount of chimera. This observation called into question their authenticity. But the size of the library should also be taken into account. In addition, some chimera, with the same gene partners, were revealed multiple times. An analysis of their distribution among EST libraries shows that some of these chimera can be observed from different independent libraries, while others can only be seen from a single library with high repetition. Lacking independent support for the latter case made its reliability hard to assess. Further, from the entirely independent gbHTC and gbPRI databases, we discovered some chimera having the same gene pairs as those from the EST libraries, which might consolidate the authenticity of these fusion events.

A further analysis on the co-expression pattern of gene pairs in the examined chimeric products revealed a strong correlation in their expression. Unlike the correlation pattern seen from the background, gene partners from the chimeric transcripts showed an enrichment in positive correlation, and a depletion in negative correlation. To make the expression of chimeric gene partners synchronous, two gene loci might be recruited to the same transcription factory, which, in turn, would increase the possibility of their aberrant ligation. The same should be true for chimeric gene pairs localized on the same chromosome. By looping out the chromatin between two gene loci, they can decrease their spatial distance and share the same transcription factory. This observation further supports the assumption of transcription induced chimerism (TIC) [16, 17] and the role of the transcription factories in gene transcription events [180] as well. To validate spatial association of these chimeric gene pairs, we checked the contact probability of chromosomal regions where chimeric genes localize. Although only a normalized contact map of human lymphoblast is available so far, we still can see contact enhancement between some chimeric genes. Considering the difference in expression profile of different cell lines, chimeric gene pairs collected from other cell lines might not be actively expressed in the lymphoblast, which meets no requirement of the transcription induced chimerism (TIC). Thus more chromosome contact maps from diverse cell lines would be helpful to confirm the spatial proximity of these chimeric gene pairs. On the other hand, however, it is not necessary for each chimeric gene pairs

to be functionally related. Considering the fact that a limited amount of the transcriptional factories exists in the cell nucleus, genes sharing the same transcriptional apparatus might just be temporal related. This might be the reason why biological significance of interactions between some gene partners is obscure. However, functional analysis did tell something. GO terms associated with chimeric gene partners showed higher semantic similarities.

We used the chimeric transcripts that passed the validation as the probes for gene loci interactions mapping. For the time being, it is still too early to draw definitive conclusions about inter-chromosomal interaction pattern, but a probabilistic non-random arrangement of chromosomes seems plausible in mammalian cell nucleus [203]. From our observation, what can be confirmed at present is that the chromosomes pair-wise association pattern is not random, compared to the pattern with chromosomes distributed arbitrarily. Some researches have shown that relative arrangement of some chromosomes are conserved and independent of cell types, while others seem to be cell-type specific [203]. The same is true with regard to the chromosomal re-arrangement in tumor cell [176, 204], a cell line and disease related fashion was observed. The observation that some chimeric transcripts can be expressed in genetically rearranged cells as well as in normal cells raises the possibility that such physical proximity between gene loci might be preserved after the chromosomal translocation. This might be an explanation of the slight difference detected from the inter-chromosomal interaction patterns of normal and tumor cell lines.

The pattern reflected by the intra-chromosomal loci interaction is very informative. From 500 kb to 7-8 Mb, the interaction probability shows a specific mathematical form (i.e., an exponential form) which is consistent with the data from chromosome conformation capture experiments. While such a relationship can be extended to shorter length scales, to exclude the possibility of tandem transcription or intergenic splicing induced chimerism, we only considered the gene partners with genomic distance over 500 kb. Thus, at this scale, the architecture of chromatin folding fits well with the experimental results [18] and with a model recently proposed [56, 57]. In mouse, while it has fewer data, a coarse pattern compatible with that from human is seen as well. For the complexity of higher eukaryotic nucleus, the genome must be well organized covering several different length scales to keep an ordered but flexible

architecture and to exert their functions properly. The strategy used by mammals for chromosome organization might be conserved across the species. Further, the pattern shown after distance normalization indicated that a basal folding schema might exist. This basal form of condensation might organize the chromatin fiber into a series of units for subsequent folding. Due to the increase in the stiffness and the bulk size of the chromatin fiber after folding, the contact probability rises along with the extended genomic distance.



## Chapter 5

# Transcriptional Regulatory Network Based Chromosomal Domain Formation in *Escherichia coli*

### Chapter Summary

In this chapter, we investigate the role of gene transcriptional regulatory networks (TRNs) in shaping the genome physical structure in *E. coli*. We suggest a prospective role of TF-gene co-localization in the formation of chromosomal domains. Based on the assumption of TF-gene interactions as drivers of domain formation, we construct a framework model and prove that the physical constraints originated from the co-localization between TF genes and controlled genes facilitate the self-organization of the *E. coli* chromosome into topologically distinguishable domains. This model well explains the experimentally observed precise subnuclear positioning of single genetic loci, and their ordering as well. Furthermore, the domain sizes estimated from this model are in agreement with the size of topological domains identified in the context of DNA supercoiling.

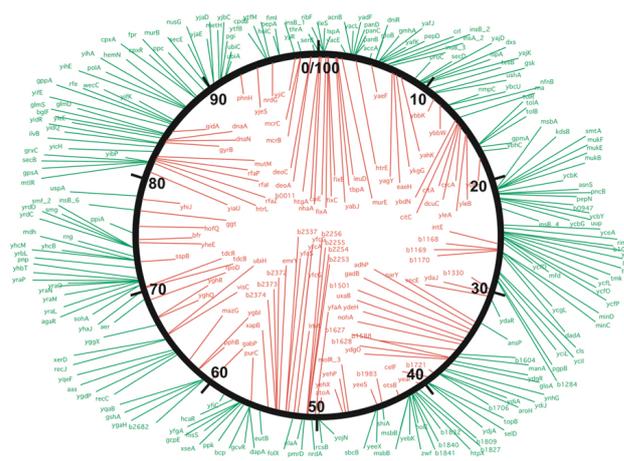
## 5.1 Introduction

*E. coli* is one of the most intensively studied prokaryotic organisms. It has been employed in the fields of biotechnology and bioengineering for decades on the basis of our comprehensive understanding of its genome. It's a kind of rod-shaped bacterium, with a typical dimension of 2.0  $\mu\text{m}$  long and 0.5  $\mu\text{m}$  in diameter. *E. coli* has a single circular chromosome. In eukaryotes, chromosomes are contained in a compartment called nucleus, which is surrounded by a double-layered lipid membrane. Because of this organization, eukaryotic chromosomes are separated from the cytoplasm. Genomic material in *E. coli*, however, is not well compartmentalized. Similar to other prokaryotic organisms, its genetic material is concentrated inside an irregularly-shaped region, termed nucleoid. The genome of *E. coli K-12* has been completely sequenced and published in 1997 [205]. The entire chromosome is around 4.6 million base pairs in length, and encodes more than 4,000 genes. After studying the *E. coli* genome for several decades, we have made most of its genes' functions clear. Compared to its genome context, much less is known about its genome physical structure. For an entirely unwrapped and relaxed *E. coli* chromosome, its dimension could be thousands of times larger than that of the cell. After being compacted into the nucleoid with the attendance of other DNA-binding proteins, the chromosome of *E. coli* occupies around one fifth of the cell volume [206]. DNA supercoiling has been postulated to explain how could the chromosome of *E. coli* be packaged with such a high compaction ratio in order to fit into the limited volume of the cell [23]. Cooperation of various topoisomerases and nucleoid-associated proteins (NPAPs) renders a steady-state level of negative supercoiling DNA in *E. coli*. DNA supercoiling, which packages the *E. coli* chromosome with an astonishing compaction ratio, is just one face of the chromosome organization in *E. coli*. On the other hand, the chromosome of *E. coli* is suggested to be organized as separated domains or loops with heterogenous sizes. Compaction and relaxation of individual domains are independent of the supercoiling status of neighboring domains [31]. Sinden *et al.* have estimated the number of total independent domains as approximately 50, with an average domain size of 100 kb [207]. Postow *et al.*, however, shows that as many as 400 topologically distinguishable domains might co-exist, with an

average domain size of 10 kb [30]. Jeong *et al.* confirms this result when finding that the number of genes exhibiting coherent transcriptional activities can be up to 16, which means that the largest confined supercoiled domains can be up to 16 kb [208] (the average gene size in *E. coli* is around 1 kb). Other mechanisms, like entropy force-driven [209], protein-DNA interactions [24, 29, 210], as well as chromosome tethering [211, 212], have also been put forward to explain the higher-ordered genome structure observed from the *E. coli*.

Decondensation of the supercoiled DNA molecule in *E. coli* is as important as the reversed progress, since it's indispensable to the proper function of the genome, especially for gene expression control, in which the topological structure of the chromatin fiber is determinant. Threads of evidence has been identified. Sousa *et al.* have discovered a chromosomal positioning dependent fashion of gene expression in *E. coli* [213]. In the beginning, they notice differential expression of a reporting gene cassette when being randomly inserted into different positions on the *E. coli* chromosome. The possibility of read-through transcription from nearby external promoters is excluded by combining strong terminators flanking the reporter system to avoid a tandem transcription. The variation of gene expression shows a roughly linear pattern, which is high when close to the replication origins, and low when approaching to the replication terminus. The expression level is propositional to the distance to the origin of replication. They infer that this observation is due to the increased gene dosage associated with regions close to the origin of replication, rather than to the local differences of chromosome organization. However, study from Peter *et al.* shows something different [214]. In this experiment, they use microarrays representing nearly the entire genome of *E. coli K-12* to check the correlation between gene expression and chromosome dynamics structure. By impairing topoisomerases genetically or by applying specific topoisomerase inhibitors, they are able to change the topological structure of DNA. In the meantime, they scrutinize genes, whose expressions are significantly changed (either up-regulated or down-regulated). By doing so, they systematically identify totally 306 genes (accounting for 7% of the *E. coli* genome), which can rapidly and reproducibly response to changes on the level of supercoiling during log-phase growth. Among these genes, expressions of 106 genes are increased, and the remaining 200 genes are declined in expression. Genome mapping result of

these inducible genes is shown in Figure 5.1. A kinetic analysis tells that the variations of genes in expression are possibly a direct consequence of the alternation on DNA supercoiling status. These supercoiling-sensitive genes (relaxation-induced and relaxation-repressed genes) are dispersed throughout the chromosome, and diverse in function. Footed on these findings, they propose that DNA supercoiling could function as a control means and affect gene expression genome-widely.



**Figure 5.1: Supercoiling-sensitive genes mapped across the *E. coli* genome.** Relaxation-induced genes are in red, and green for relaxation-repressed genes. Image adapted from [214].

Genes are organized sequentially on the chromosome, therefore transcription operons or regulons are usually defined linearly along the genome. However, chromosome folding makes the physical interactions between distant genes possible, which has been suggested to be functionally important [55, 105, 111]. After taking a closer look at the properties of global transcriptional profile of *E. coli K-12* as a function of genes' positions on the chromosome by Jeong *et al.* [208], some spatial correlations between the expression of distant genes are discovered. They cluster genes according to their transcriptional similarities, and check their linear distances. Three categories of transcriptional similarities are classified: short-range class, below 16 kb; medium-range one, from 100-125 kb; and long-range one, over 600-800 kb. For the short-range correlation, they find that it's the local DNA supercoiling states that governs the similarity pattern. By contrast, for the medium- and long-range correlation, the similarity

depends not too much on the DNA supercoiling states, but on the distribution of DNA gyrase along the chromosome.

The topological organization of the *E. coli* chromosome is found to be important to gene translation as well. In prokaryotic cells, gene transcription and gene translation are coupled and occur in the same cellular milieu, unlike what happens in eukaryotic cells, in which matured mRNAs are transported from the nucleus to the cytoplasm for translation. However, it remains poorly characterized how mature mRNAs are distributed inside the prokaryotic cells due to the deficiency of labeling and microscopic technologies. The diffusion coefficient obtained by using extraneous plasmids tells that mRNA molecules disperse fast enough to travel throughout the cell before being degraded, which naturally implies that translation of mRNAs could occur at any cellular space. To confirm whether this is the case, with a far more enhanced quantitative FISH technology, Llopis *et al.* trace the distribution of chromosomally expressed endogenous mRNAs from their coding sites till they are degenerated [13]. They find that mRNA molecules are only able to diffuse within a small scope after being generated. Most evident in *C. crescents*, positions of its ribosomes depend largely on its mRNAs' localizations. All these evidence supports the idea that prokaryotes possibly use their chromosomes as templates to anchor their translation sites. Accordingly, gene translation in prokaryotes might be well-compartmentalized and occur in discrete subnuclear domains with a chromosome-centric organization. In other words, functionally specific proteins are generated within individual sub-cellular domains, which is determined by the genetic map and the chromosomal organization. We might be able to go a step further to infer that physically related genes (genes that encode interacting proteins), or transcriptionally related genes (transcription factor genes and their target genes) might be clustered more frequently in order to facilitate their interactions. On the contrary, spatially clustered genes (unlike those linearly clustered genes along the chromosome) might shed light on their functional relevance. In split of the lack of membrane-based organelles, bacteria have developed a unique strategy to use their chromosomes as a framework to spatially organize the gene transcription and translation. At bottom, such an organization can still be attributed to a compartmentalization motif.

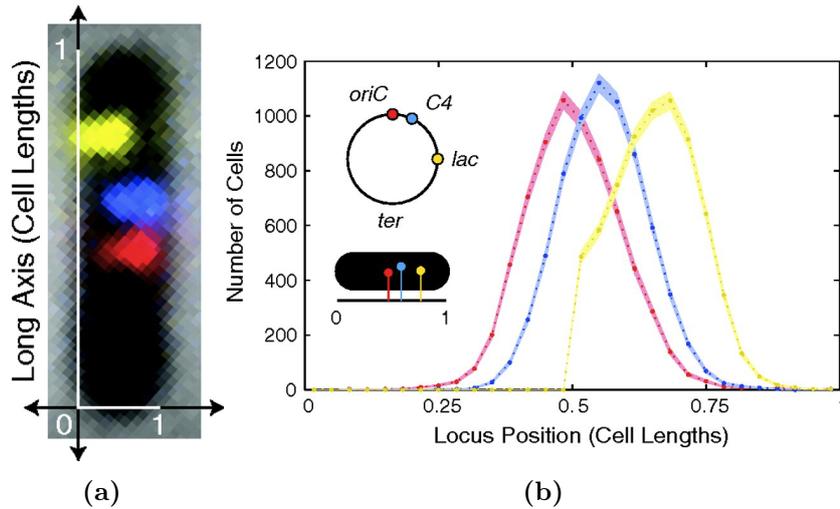
It's possible that the spatial co-localization of genes in the cell nucleus can

not only facilitate the physical interactions between their protein products, but also regulate gene expression more efficiently. In bacteria, the expression induction or repression of genes are controlled by transcription factors (TFs), which are proteins that can combine with specific DNA sequences to promote or block the recruitment of RNA polymerases. Therefore TF genes can indirectly interact with their target genes through the TFs they encode. As a result, a collection of TF genes and corresponding target genes can be connected together in the form of a network, named transcriptional regulatory networks (TRNs). Usually, a TFN can be depicted with a hierarchical structure, like a cascade. Nodes on the top of the hierarchical network are more crucial and deterministic, which can genome-widely alter the regulatory flow in the network. TF genes on the middle level of the network receive transcriptional controls from their upstream TFs, while their protein products are responsible for delivering signals to downstream genes. Gene nodes on the bottom of the TRN generally play very specific roles in generating structural or catalytic proteins.

By using genome-wide approaches, the transcriptional regulatory networks of some model organisms (i.e. *E. coli*) have been well characterized. But it's still poorly understood how could the TRNs be linked with the genome physical structure. A conceptual model has been postulated by Janga *et al.* [10], who suggests that the dynamic nature of biophysical requirements for targeting DNA-binding sites by TFs could guide the way how genes are spatially organized in bacterial nucleoid. In their view, the biophysical nature of TFs for reaching their DNA-binding sites might be an important driving force in shaping the genome structure. From the angle of gene regulon, they investigate the relationship between the size of a regulon and the average distance of a controlling TF to all of its target genes. The analysis on all of the *E. coli* regulons indicate that all its TFs can be roughly grouped into three categories: 1) top TFs that regulate large regulons, 2) intermediate TFs that control some medium-sized regulons. TFs of this category are heterogeneous in their regulon sizes and chromosomal distances, and 3) bottom TFs that regulate lots of local regulons with smaller sizes and shorter genomic distances. Those global effectors can be either Nucleoid Associated Proteins (NAPs) [24], or growth condition associated global regulators that are crucial in response to different

growth conditions. Except for a handful of global TFs, vast majority of TFs work more locally and regulate smaller regulons. The expression level of TFs is correlated with their hierarchical positions in the network. Global TFs are far more abundantly expressed than those local ones. That is to say, TFs with higher intracellular concentration correspond to the nodes with more out degrees in the regulatory network. Seeing the fact that local TFs that are in charge of smaller regulons are proximate to their regulated genes, and the fact that gene transcription and translation occur in the same compartment in prokaryotic cells, local TFs after being generated could efficiently mount onto their target sequences by searching locally. While for those global TFs, whose target genes are much larger in number, and more distant to the regulator genes, they might use a large amount of protein products to compensate for the differences in the number of controlled genes and the genomic distances to target genes.

More significant findings about the *E. coli* genome organization has been achieved by Wiggins *et al.*, who observe a domain-centric organization motif in the *E. coli* genome. They reveal that the positioning and the ordering of genomic loci within slow growing *E. coli* cells in G1 phase are precisely organized inside the nucleoid [7], as shown in Figure 5.2. Qualitative measurements of the *E. coli* chromosome discover that *E. coli* nucleoid has a linear organization along the long-axis of the cell, with the origin of replication at the middle between replication cycles [215, 216, 217, 218]. This organization has a roughly constant linear compaction ratio, except for the region approaching to the *ter* site, which connects two arms of the chromosomes [7, 218]. To assess how precisely gene loci are positioned inside the nucleoid, Wiggins *et al.* measure the position fluctuations of three gene loci (namely *oriC*, *C4*, and *lac*) and evaluate the correlation between the fluctuations of these loci pairs (see Figure 5.2b).



**Figure 5.2: The *E. coli* chromosome is organized into separated domains.**

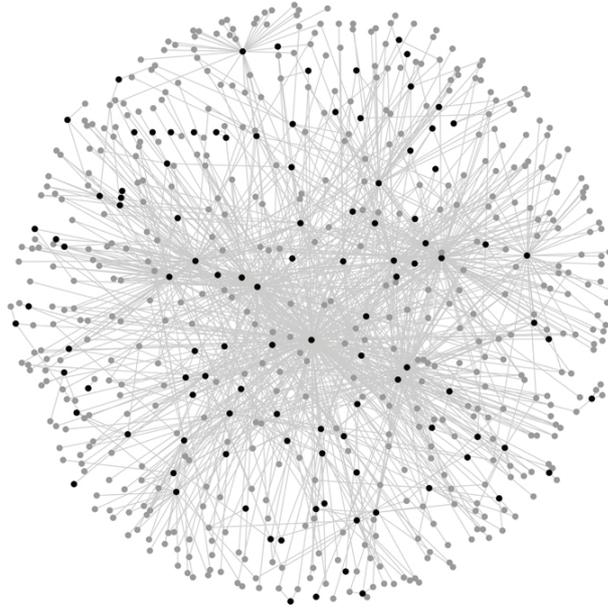
In the experiment conducted by [7], three genomic loci along the *E. coli* genome, namely *oriC*, *C4* and *lac*, are labeled with different tags and their projections along the long-axis of the cell are determined in a population-averaged way. (a) shows a typical image of microscopic screening. Loci's schematic genome positions are represented in the inset of Figure (b). The position distributions of three genomic loci along the long-axis of the cell are summarized as a histogram shown in (b). It is clear that three labeled genomic loci do not intermingle with each other at large. Image adapted from [7].

Inspired by all the findings above-mentioned, we propose that the gene regulatory network of *E. coli* might serve as a driving force to organize the DNA chain into discrete domains. It's the long- and short-range interactions mediated by the co-localization of TF genes and their controlled genes that constrain the 3D organization of the *E. coli* genome. We investigate the consequences of this assumption and prove it as one possible mechanism for the formation of chromosomal domains in *E. coli*. With this model, we manage to reproduce the high precision of subnuclear positioning of genetic loci as observed in the experiment [7]. The observed precise ordering of the chromosome is further consolidated by investigating the domain sizes that are distributed between 10 and 700 kb. Therefore, this model is able to re-construct the size distribution of topological domains found in the context of DNA supercoiling.

## 5.2 Material and Methods

### 5.2.1 The Gene Transcriptional Regulatory Network of *E. coli K-12*

Numerous efforts have been made in understanding how can activities of group of genes interactively influence each other in order to undergo programmed cellular events, or to response to the stimulus from the environment. Meanwhile, increasing number of unexplored transcription factors together with their regulated genes have been identified. With the application of microarray technology, more detailed expression regulation relationships have been elucidated. The genome of *E. coli K-12* has been thoroughly studied. It's 4,639,676 bp long and encodes over 4,000 genes. Among these genes, products of around 180 genes are capable to serve as transcription factors. With the genome component of *E. coli K-12* clear for the most part, we aim to construct a model to evaluate the role of transcriptional regulatory relationships in shaping the 3D *E. coli* genome. Updated genome information about the *E. coli K-12* can be found at EcoGene database (<http://maxd.cs.purdue.edu:9455/databasetable.php>) [219]. EcoGene database focuses on collecting and revising genome and proteome information relating to the *E. coli K-12*. The regulatory network of transcription factors and target genes (TF-gene interactions) can be obtained from RegulonDB (Version 7.3) [220], which is the primary reference database of the best-known regulatory network of *E. coli K-12*. A graphical representation of the gene transcriptional regulatory network applied in this model is given in Figure 5.3. The regulatory network applied includes 177 transcription factors and 1,519 target genes involving 3,984 regulatory interactions.

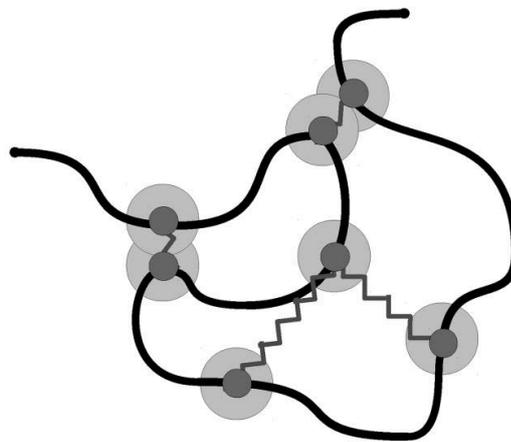


**Figure 5.3: Graphical representation of the transcriptional regulatory network as applied in our polymer model.** Black nodes indicate the TF genes, dark gray nodes indicate the target genes, and links represent regulatory interactions between them.

## 5.2.2 Modeling and Simulation

Based on the genome localization of TFs and target genes involved, we set up a coarse-grained polymer for simplicity. A chromosome is simplified as a fragment of a polymer composed of homogenous monomers with excluded volume and bonded via a FENE potential. A monomer is set at every 1 kb genome distance (about 4640 kb for *E. coli K-12* genomic size) leading to  $N = 4641$  monomers building up the polymer. Then the TFs' and target genes' positions are assigned by taking the middle points of genes' genomic positions. Since the chromosome is coarse-grained, genes that are genomically proximate could be allocated into the same monomer. Each monomer is a hard sphere with a bead diameter. Initially, the conformation of polymer is randomly assigned inside a large enough cubic box without any confinement. After that, in the light of gene regulatory network, those monomers corresponding to mapped interacting genes are linked via build up elastic harmonic bonds in between so as to mimic presumed physical proximities. By carefully setting the balanced distances between gene loci with transcriptional interactions, harmonic

bonds in between will recruit gene pairs closely, as illustrated in Figure 5.4. In order to make the density and the geometry of simulated polymer compatible with experimental observation, an elongated rectangular cuboid are used to confining the geometry of the polymer included. The final aspect ratio of the rectangular cuboid is around 6:1, which is similar to the aspect ration of the nucleoid. And the eventual volume fraction of the polymer is about 10% [221, 222]. To make sure that the polymer has enough time to relax itself after each time resizing the constraint box, the box geometry is slowly shrank at each integration step during the run of simulation.



**Figure 5.4: Mimic the regulatory control between TF genes and target genes.** On this 3D self-avoiding DNA chain, presumed interacting TF genes and target gene sites are represented by small gray filled circles connected by springs. The outer gray circles define the strength of the harmonic interaction potential. Image adapted from [149].

Molecular dynamics simulations are performed by applying the software package ESPResSO (<http://espressomd.org>) [223]. The parameters used in the simulations are summarized in Table 5.1 and explained in the following.

**Table 5.1: List of parameters.**

|  |
|--|
| 1. Parameters for system   |
| density = 1E-6<br>temperature $T = 1.0$<br>friction $\Gamma = 1.0$<br>time_step $t = 0.01$   |
| 2. Parameters for L-J potential  |
| $\epsilon = 1.0$<br>$\sigma = 1.0$<br>$r_{cut} = 2^{1/6}$  |
| 3. Parameters for FENE potential   |
| $K_{FENE} = 10.0$<br>$r_0 = 1.5$<br>$\Delta r_{max} = 1.5$   |
| 4. Parameters for HARMONIC potential   |
| $K_{HARMONIC} = 0.005$<br>$R_{HARMONIC} = 1.5$   |
| 5. Parameters for NPT isotropic thermostat   |
| nptiso_gamma0 = 1.0<br>nptiso_gammaV = 1.0<br>initial npt_p_ext = 1.0<br>initial piston = 1.0<br>npt_p_ext step = 0.08<br>piston step = 0.08 |

Several types of interactions involved in the simulation are listed as follows.

### **I) Backbone Potential**

The FENE (finite extension nonlinear expander model of a long-chained polymer) interaction is used for the backbone of the chain. It simplifies the linear polymer by a sequence of connected beads with nonlinear spring force between linked monomers. This type of bond is a rubber-band-like, symmetric interaction between two particles and is defined by three parameters: prefactor  $K$ ,

equilibrium bond length  $r_0$ , and maximal stretching  $\Delta r_{max}$ . The bond potential diverges at a particle distance  $r = r_0 - \Delta r_{max}$  and  $r = r_0 + \Delta r_{max}$ .

$$U_{FENE}(r) = \begin{cases} -\frac{1}{2}K\Delta r_{max}^2 \ln[1 - (\frac{r-r_0}{\Delta r_{max}})^2] & , r < \Delta r_{max} \\ +\infty & , r \geq \Delta r_{max} \end{cases} \quad (5.1)$$

## II) Excluded Volumes

To avoid monomers sitting on the top of each other, excluded volume interactions were included by using Lennard-Jones potential, which is a ‘work-horse’ potential of interactions between particles in coarse-grained simulations. Lennard-Jone (L-J) potential is defined as

$$U_{L-J}(r) = \begin{cases} 4\epsilon((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6 + C_{shift}) & , r < r_{cut} \\ 0 & , r \geq r_{cut} \end{cases} \quad (5.2)$$

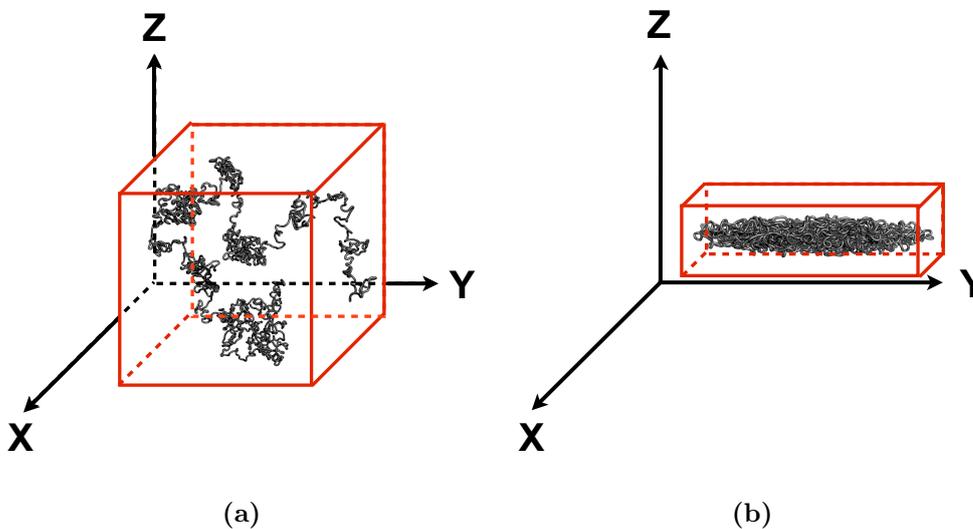
It’s a simplified model of the van-der-Waals interaction, which is attractive at large distance, but strongly repulsive at short distances. The minimum potential is achieved at the radius  $r = r_{off} + 2^{\frac{1}{6}}\sigma$ . Above this radius, the attractive part of the potential starts to affect. Weeks-Chandler-Andersen (WCA) potential, a special case of Lennard-Jones interaction, is applied by setting the radius cutoff at the site of the minimum potential ( $r_{cut} = 2^{\frac{1}{6}}\sigma$ ). The WCA potential turns off the attractive part of the interaction between particles, while it is purely repulsive at short ranges when the distance is smaller than a cut-off  $r_{cut}$ .

## III) TFs Interaction Potential

To imitate the spatial proximities between transcription factors and their regulated genes, the classic harmonic potential is chosen for the regulatory interactions. This potential is determined by the prefactor  $K$  and particle distance  $R$ , where it takes the minimal values at distance  $r = R$ ,

$$U_{HARMONIC}(r) = \frac{1}{2}K(r - R)^2. \quad (5.3)$$

The NPT ensemble and the `npt_isotropic` thermostat modules available in ESPResSO are applied in order to perform the isotropic changes of the box geometry (from cubic to rectangular) during equilibration, as shown in Figure 5.5. To make sure that the polymer has enough time to relax itself after each time the box is resized, the box geometry is slowly shrunk at each integration step during the run of simulation. During each integration step in the simulation, the system will perform an isotropic change of the box geometry. Two sides of the confining box are evenly shrunk a little bit for each cycle, while keeping the third side unchanged. By doing so, the aspect ration of the confining box is increased gradually until the final aspect ratio 6 is reached. After the final aspect ratio is reached, the simulation is continued for a while in order to relax the polymer further. In order to speed up the equilibration, the parameters of FENE potential are selected so as to make the back-bone extensible and allow chains to pass through each other. Also, the Lennard-Jones potential is capped and increased gradually to get ride of the overlapping of particles and to facilitate the equilibrium.



**Figure 5.5: The isotropic changes of the box geometry during equilibration.** (a) At the beginning, the conformation of the polymer is randomly assigned. The polymer is included inside a cubic confining box. (b) During equilibration, the confining box is evenly and slowly shrunk along the X- and Z-axes until the geometrical aspect ratio of the long side of the chain to the short one is approximately reached to 6, which gives us a geometrical transition from a cubic box to a rectangular one.

## 5.3 Results and Discussion

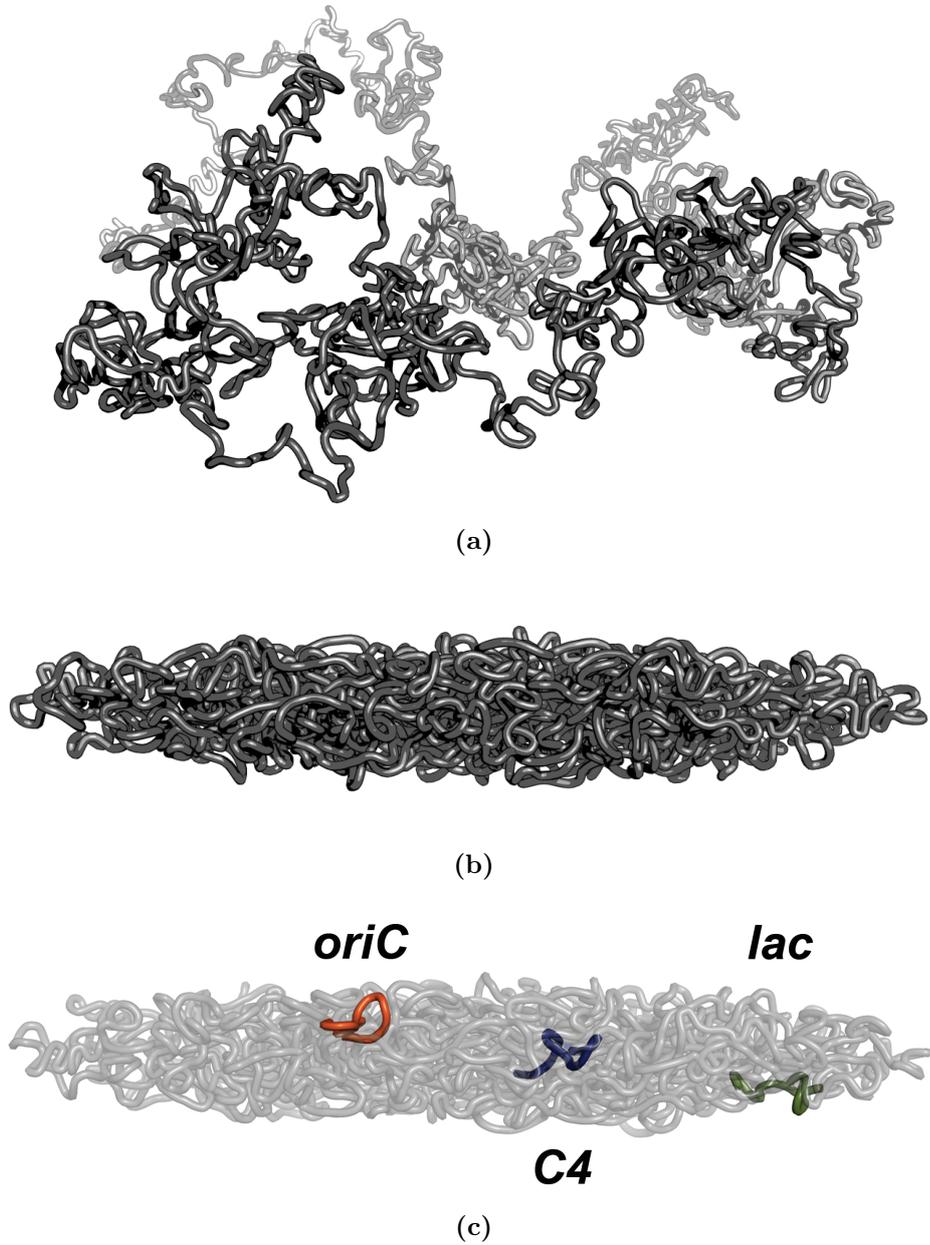
### 5.3.1 The Gene Regulatory Network as a Mechanism for Domain Formation

We propose the co-localization between the TF genes and their target genes as a possible mechanism for the formation of chromosomal domains in *E. coli*. On a 3D self-avoiding DNA chain with confinement, the regulatory interaction between TF genes and target genes in the transcriptional regulatory network is mimicked by assuming a harmonic interaction between these sites, as illustrated in Figure 5.4. As a consequence of this construction, we find that the DNA chain indeed self-organizes into topologically distinguishable domains. Figure 5.6c shows a snapshot of the genome organization as obtained after the equilibration of the self-interacting DNA chain. Consequently, we get the same linear ordering of the three gene loci as observed in the experiment [7].

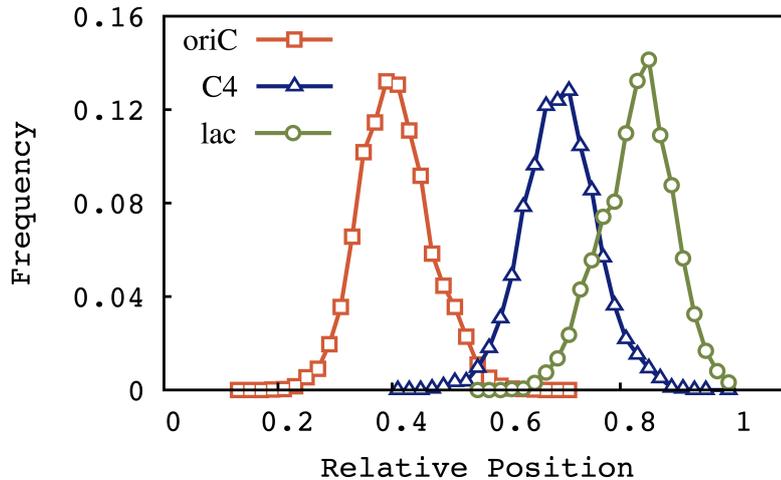
More analyses indicate that the three genetic loci are separated into topologically distinguishable domains. Figure 5.7 shows the position distribution of the three genetic loci *oriC*, *C4* and *lac* along the long axis of the confining cavity as obtained by our model of the *E. coli* nucleoid. We can see that the three genetic loci are located on different chromosomal domains considering the limited overlap between any two of them. On the other hand, the genomic distances between the loci *oriC* and *C4* as well as between the loci *C4* and *lac* are about 300 and 600 kb respectively, which also indicate that they are positioned on different chromosomal domains when comparing with the domain size distribution experimental estimated [30, 207, 208]. We also find that the terminus region in our model is localized at the mid-cell position. All these findings are compatible with [7]. Consequently, our model is capable of reproducing the linear correlation between the position of a gene on the chromosome and its sub-cellular position inside the nucleoid.

One should notice that the conformation shown in Figure 5.6c just represents one of numerous possible conformations that the *E. coli* genome could have. Due to thermal dynamics and cell-to-cell variance, the conformation of *E. coli* genome can never be underpinned. Nevertheless, we can use a probabilistic model to characterize the genome conformation. We can see from Figure 5.7 that the relative position of genomic locus along the long cell axis

is not uniformly distributed, but rather in a probabilistic way. Scrutinized genomic loci in the body of the nucleoid show a precision of positioning of better than 10% of the cell length. Also, the precision of interlocus distance of genomically proximate loci is found to be high.



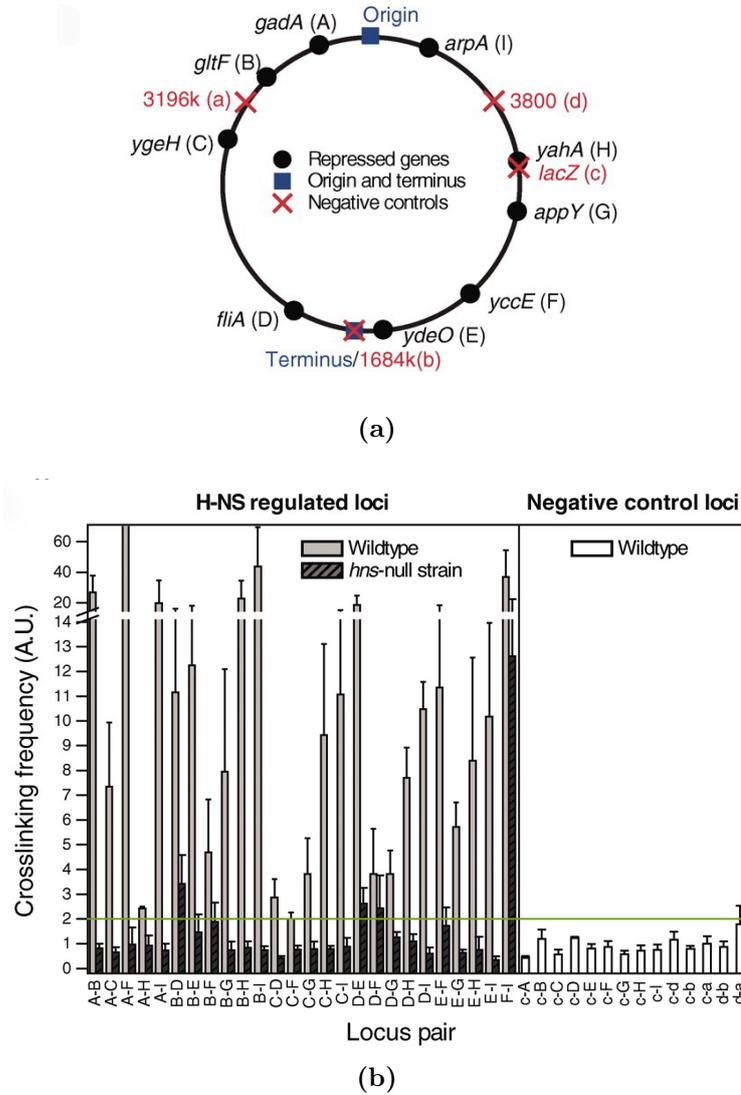
**Figure 5.6: Snapshots of the initial and equilibrated conformations of the same polymer.** The polymer shown in (a) is created randomly according to the method above-mentioned. By introducing physical proximities between TF genes and target genes, the polymer within a confining space is self-organized into a rod-shaped geometry after the equilibration, as shown in (b). The three gene loci screened in the experiment are roughly mapped onto the polymer, and highlighted in (c). As we can see, the linear ordering of these three gene loci is in agreement with the experimental observation, as shown in Figure 5.2.



**Figure 5.7: TF-gene interactions as drivers of chromosomal domain formation.** The distribution of three gene loci along the long axis of confining box achieved by our model is shown. It is clear that these three genomic loci are well separated with respect to projections on the long axis of the confining envelope and have the same ordering as observed in the experiment [7].

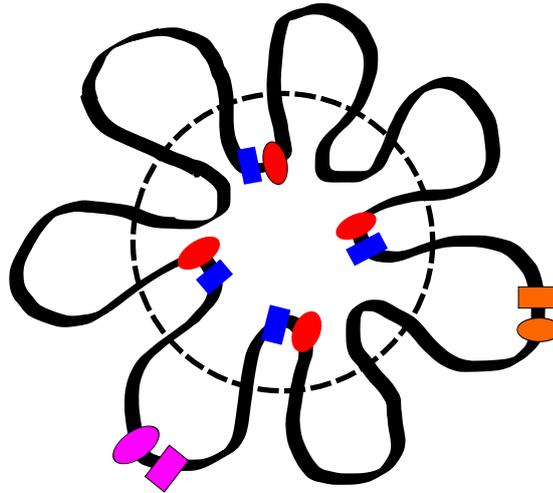
A work published later consolidates our assumption of TF-gene interactions as drivers of domain formation [11]. In this work, Wang *et al.* find that H-NS, a global transcriptional silencer regulating  $\approx 5\%$  of all *E. coli* genes [224], is largely aggregated into two compact clusters for each chromosome. Two H-NS enriched compartments sequester their regulated operons into the clusters and recruit heterogeneous DNA segments closely, which are distributed throughout the chromosome. While this is not always the case, which reflects a dynamic genome organization, but genes that are targeted by H-NS indeed show a higher probability to move towards H-NS clusters. As shown in Figure 5.8, frequencies of pair-wise cross-linking of H-NS regulated genes as measured via chromosome conformation capture (3C) technology are much higher than those of negatively controlled gene loci. Deleterious mutation on H-NS can drastically decrease the pair-wise interaction frequencies of H-NS regulated genes, and result in a chromosome reorganization. In *hns*-null cells, the positions of genes that are regulated by H-NS can be shifted up to  $\approx 300\text{nm}$ , which is comparable to the radius of the nucleoid. All these observations demonstrate

that the transcription factor H-NS might play a key role in shaping the genome organization globally.

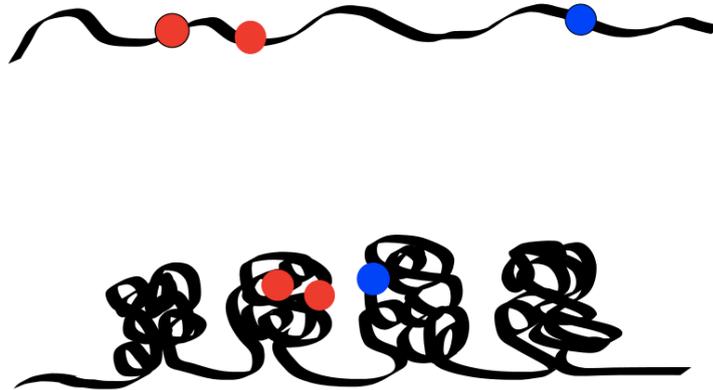


**Figure 5.8: Co-localization of genes regulated by the transcription factor H-NS.** (a) shows the genomic positions of genes that are repressed by the transcription factor H-NS (labeled as circles in black from ‘A’ to ‘I’), and genes used as negative controls (labeled as red crosses from ‘a’ to ‘d’). In Figure (b), the pair-wise cross-linking frequencies of genes labeled in (a) are compared. As we can see, genes that are both regulated by H-NS have much higher interaction frequencies than the gene pairs including at least one negative control gene. In *hns*-null cells, however, predominant interactions found from wild type cells are largely abolished. Image adapted from [11].

Seeing the compatible results obtained from our model and from other biological experiments, domain-based organization motif disclosed from the *E. coli* genome is possibly driven by the topological constraints derived from gene transcriptional regulation. Spatially specific distribution of TF genes, and physical proximities between TF genes and regulated genes might turn the *E. coli* genome into functionally separated territories, as illustrated in Figure 5.9. TF genes that are proximate to their controlled genes on the chromosome has been repeatedly found in *E. coli*. Most of these TF genes work locally and target more specific functions. TFs of this kind are classified as the local TFs. Such an organization will expedite the combination of newly synthesized TFs with their target genes through a sliding mechanism, since in prokaryotes gene transcription and translation are spatially coupled [13]. This linear arrangement could guarantee an efficient control of genes that are targeted by local TFs. Global TFs, by contrast, usually regulate far more abundant genes that are spread all through the genome and have heterogeneous functions. In this scenario, global TF genes have been postulated to assume a more central localization in the nucleoid, which makes it easy for them to undergo 3D diffusion or jump between DNA strands in order to get access to a large number of target DNA sequences. These global TF genes might determine the genome organization globally in order to maximize the binding efficiency of their protein products. This conjecture has been supported by the experimental observation, in which global TFs have a much higher concentration than local TFs. In addition, the expression of global TF genes is more roughly controlled. Local TFs, on the contrary, are more dedicated in function and delicate in control. Genes that are included within the same chromosomal domain are more functionally related and compatible in expression, which might be responsible for shaping the internal structures of individual domains. Genes within the same chromosomal domains can also explain the short-range correlation in gene expression as observed in the experiment [208]. Via looping and compartmentalization, genomic loci that are genomically distant could be closely recruited, which could explain the medium- and long-range correlation in gene expression as observed in the experiment [208] (see Figure 5.10). Also, the biological significance of the interaction between two spatially neighboring genes that are positioned at different chromosomal domains has been elucidated [225].



**Figure 5.9: The role of global and local TFs in forming chromosomal domains.** A circular genome of *E. coli* can be split into several large chromosomal loops (domains), each of which can be further self-organized into discrete structures. Red oval and blue rectangle pairs indicate some kind of global TF and its target genes. The rest oval and rectangle pairs with a same color outside the dashed circle denote some kind of local TF and its controlled gene. Global TF genes, whose protein products regulate a large numbers of targets genes, are more likely to occupy a central location in the nucleoid. Compared with the global TFs, the movability of local TFs is more restricted. Local TFs are largely domain-specified, and responsible for the regulation of genes clustered within the same domain.

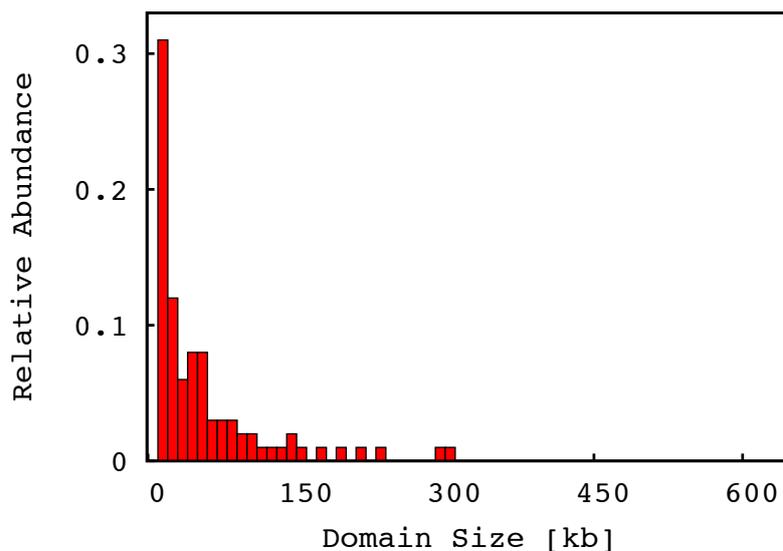


**Figure 5.10: Domain organization motif facilitates the communication between distant genes.** Experiments have already indicated that genes that are clustered within the same chromosomal domain are more related in function and prone to be expressed concurrently (genes in red). The formation of chromosomal domains could largely diminish the distance between genomically distant genes. Genomically distant but spatially close genes (one in red, the other in blue) might be the reason for the experimentally observed long-range correlation in gene expression [208].

### 5.3.2 The Distribution of Chromosomal Loop Sizes

Footed on our assumption, transcriptional regulation activities will give rise to chromosomal loops by inducing spatial proximity between specific genomic loci, which would otherwise be distant on the chromosome. These loci-specific loops might have a significant impact on the overall genome organization. Beside these specific contacts, other non-specific loops possibly mediated by transient chromatin contacts might also be of importance. Therefore we check the loop size distribution based on the numerous genome conformations generated by our model. To do so, we set a distance cutoff. If the distance between any two given particles on the polymer is below this value, a loop is formed. We set the minimal loop size as 10 kb, corresponding to 11 monomers in a row. After sampling numerous possible conformations, it's clearly seen from Figure 5.11 that most loops formed are relatively small, and the abundance of loops drops exponentially when the size is increased. The loops gleaned from our model

are sized from 10 to 700 kb, which fits well with the size of topological domains identified in the context of DNA supercoiling [226]. The average domain size obtained from our model is around 86 kb. While small loops are predominantly observed, large loops (sizing over 150 kb) can still be found. These larger domains can be called higher-order ‘super’ structures, which are built up from smaller subdomains. The aggregation of smaller domains into larger ones could facilitate the communication between sub-territories, which could explain the observed long-range correlation in gene expression as well [208, 226]. Long-range interactions involving chromosome regions that are genomically distant, but spatial proximate, might result from the higher-order genome organization [225]. From *Bacillus subtilis*, Berlatzky *et al.* have already observed some specific co-localization of distant genes within the same sub-cellular compartment for the purpose of co-regulation [227].



**Figure 5.11: The distribution of chromosomal domain sizes as obtained by our model of the *E. coli* nucleoid.** We find the domain sizes to be distributed between 10 and 700 kb, with the average domain size of 86 kb.

## 5.4 Conclusion

In this chapter, we have proposed a transcription regulatory network (TRN) based mechanism by which chromosome can self-organize into distinguishable

chromosomal domains. We quantitatively investigate this model by applying numerical simulations. Multiple threads of evidences imply a strong linkage between gene transcription and chromosome organization. It has been demonstrated that effector genes of *E. coli* are expressed in heterogeneous concentrations depending on their genomic distances from their target genes and the number of target genes that they regulate [10]. For a effector gene, the larger the distance from its target genes, the higher the concentration of effector products is needed. Such a positive correlation is also observed between the number of target genes and the concentration of effector products needed. From the perspective of mRNA translation, nascent mRNA molecules display a constrained dispersion from their transcription sites, which favors a compartmentalization strategy of translation by using the chromosome layout as a template [13]. Based on these recent discoveries, we propose that regulatory interactions between the TF genes and their target genes will induce physical proximity between them. We hold that the transcriptional regulatory network of *E. coli* plays a crucial role in spatially aggregating the TF genes and their target genes for the purpose of a more efficient transcriptional regulation. After testing this idea via numerical simulations, we find that the DNA chain can indeed self-organize into several topologically distinguishable domains. While neighboring domains can overlap to a certain degree, but they are not intermingled. The domain organization reflected by our model shows the same ordering and the distribution pattern as those observed from the experiment [7]. The loop sizes estimated from our model are distributed from 10 to 700 kb, which is in agreement with the size of topological domains as observed in experiments. With this conceptual and functional model, we shed light on the underlying mechanism that governs the genome organization in *E. coli*.

# Chapter 6

## Transcriptional Regulatory Network Shapes the Genome Structure of *Saccharomyces cerevisiae*

### Chapter Summary

More and more evidence is mounting that gene transcription is tightly connected with the 3D genome organization not only in prokaryotes but in eukaryotes as well. Spatial proximity of genes sharing transcriptional machinery is one of the consequences of this organization. Motivated by information on the physical relationship among genes identified via chromosomal conformation capture methods, in this chapter we complement the spatial organization with the idea that genes under similar transcription factor control, but possibly scattered throughout the genome, might be in physical proximity to facilitate the access of their commonly used transcription factors. Unlike the transcription factory model, ‘interacting’ genes in our model are not necessarily immediate physical neighbors but are in spatial proximity. Considering the stochastic nature of TF-promoter binding, this local condensation mechanism could serve as a tie to recruit co-regulated genes to guarantee the swiftness of biological reactions. We test this idea with a simple eukaryotic organism, *Saccharomyces cerevisiae*. Chromosomal interaction patterns and folding behavior generated by our model re-construct those obtained from experiments. We show that

the transcriptional regulatory network has a close linkage with the genome organization in budding yeast, which is fundamental and instrumental to later studies on other more complex eukaryotes.

## 6.1 Introduction

Gene transcription is central among all biological processes. Studies so far have indicated that transcription processes could play an architectural role and drive the genome organization [228]. Transcription of a single gene is always affected by the status of other genes, directly or indirectly. All these inter-correlated relationships are build into the Transcriptional Regulatory Network (TRN) [64]. The network consists of transcription factor (TF) coding genes and their target genes. To regulate the expression of a single gene, multiple TFs are usually recruited to form a transcription factor complex to initiate or to inhibit the transcription, while activities of multiple genes can be coordinated by the same group of TFs. Additionally, gene expressions can be controlled by internal or external stimuli. All these features make the transcriptional regulatory relationship a highly intermingled and complex network. This network can tune up gene activities with spatial and temporal difference in order to respond to the changing environment and to confer the organism viability and adaptability.

In recent years, molecular biological technologies, such as ChIP-ChIP and ChIP-seq, have been increasingly productive in finding unknown TFs and their prospective targets. Among all the model organisms, *Saccharomyces cerevisiae* (budding yeast) is one of the most intensively studied and evolution fundamental one. Numerous microarray data published by now have gradually augmented the network [229] topology. Little is know about what kind of constrains the transcriptional regulatory relationship might impose on the genome physical organization.

For the correlation between the genomic function and the spatial organization of the genome under different scales, several possibilities have been put forward [149, 230, 231]. One of ideas is the transcription induced gene co-localization. Transcription factories are discrete structures within the cell nucleus enriched with a transcription apparatus, like polymerases, TFs and

nascent RNA splicing apparatus. They are thought to play an architectural role in genome organization by tying up the actively expressed genes. Since the number of synchronously expressed genes is far beyond that of transcription factories, multiple activated genes have to co-localize at the same expression foci. Genes occupying the same transcription factories might come from distant parts of the same chromosome, or even from different chromosomes, which will induce the intra-chromosomal loops or inter-chromosomal bridges. Transcription factories with different transcription factor components have their selectivity on genes being transcribed, but transcription of other types of genes is still possible with lower frequency.

Activated genes involved in the same biological pathway are more frequently observed at the same transcriptional foci, like globin pathway genes in mouse [75]. To date, an unsolved problem why some genes are more likely to meet with each other, or even occupy the same transcription foci cannot be easily explained by diffusion and collision of gene loci onto transcription foci. In our view, in addition to the global TFs, gene expression regulation still needs more specific TFs for control. TFs of this kind are more stringently controlled in expression, and the number of expression products is low. Consequently, these local TFs are the key to assemble genes of specific types and could tether them closely. Intuitively, genes involved in the same pathway should be similar in their expression control profile, and tend to be transcribed in step or even at the same transcription foci. Even though exceptions cannot be excluded, we propose that specific TFs accumulated at the expression on-going loci might recruit regulated genes close to each other more or less, which in turn will reduce their physical distances.

In various organisms, adjacent genes along the chromosome tend to show higher positive correlation in expression, when compared with random gene pairs [232]. As seen in this study, genes on a chromosome are prone to be organized into separated globules, and genes within the same globule have comparable expression activities. However, such kind of expression correlation induced by the linear sequence proximity does not always show functional relationships among neighboring genes. This organization is not sufficient to explain significantly positive correlation found in other cases, where genes spanning large genomic distance or even on different chromosomes show strong

transcriptional correlation. These interacting genes are more frequently proximal in space. Interactions between gene loci of this kind have shown their biological significances [70].

As a general rule, genome-scale studies have found that target genes of many transcription factors usually spread throughout the genome, which raises the question how could transcription factors seek their binding sites on the targets spreading all over the genome. Studies on fission yeast (*Schizosaccharomyces pombe*) have shed light on this question. Tanizawa *et al.* have conducted a comprehensive chromosome contact screening on fission yeast [12] and observed that co-regulated genes display a tendency to move towards each other once activated concurrently. More interestingly, gene loci showing higher interaction probability share a consensus sequence on their promoter regions. Accordingly, operating TFs might not disperse all over the cell nucleus, but concentrate locally to attract their target genes. In this sense, TFs can be viewed as a tie, which could tether their target genes closely. Conversely, cluster of genes under similar control could serve to draw and elevate local TFs' concentration to make the transcription initiation more efficiently.

On the other hand, studies using chromosome conformation capture (3C) and its improved derivatives have been instrumental in investigating chromosomal contacts and chromosome higher order structures. With a novel method named genome conformation capture (GCC), the network of chromosome interactions for the yeast *S. cerevisiae* has been studied [116]. And more recently, in 2010 remarkable progress has been achieved by Duan *et al.*, who devised a method coupling 4C and massively parallel sequencing, and utilized it on budding yeast. With this method, they generate population-averaged intra-/inter-chromosomal interaction maps with a kilobase resolution [8]. This study consolidates some already known features about genome organization common to most organisms. They identify some global and local features specific to budding yeast. Some large segments of a chromosome on one arm show a tendency to interact more frequently with corresponding segment with a similar length on the other arm of the same chromosome, which forms a 'zippering' structure with two chromosomal segments. Some large segments of chromosomes also show higher frequencies of local contacts, on the contrary other given segment pairs show very limited or no interactions. Intra-chromosomal

interactions involving two telomere ends differ from one chromosome to another. Chromosome XII is spectacular in that it contains a rDNA repeats domain, which has 100-200 rDNA copies (1-2 Mb long) [233]. These rDNA gene copies aggregate at the nucleolus site, and function as a barrier by interrupting interactions between chromosomal segments flanking them.

Two models based on the transcription induced gene co-localization have been proposed by Junier *et al.* [149] and by Dorier *et al.* [150], respectively. A more generalized looping model postulated by Bohn *et al.* [231] well explains the chromosome as a looped polymer under all scales. Especially in prokaryotes, like *E. coli*, transcriptional regulation has been shown to modulate the genome structure, which is clearly seen from the study conducted by Fritsche *et al.* [234]. Based on these studies we address the problem how the gene transcriptional regulation shapes the genome organization. We test on *S. cerevisiae*. To this end, we propose a model based on the assumption that genes under highly similar transcription factors control are in spatial proximity. After analyzing the gene transcriptional regulatory network, we select genes under highly similar control and treat them as being in spatial proximity within the nucleus. The nuclear structure of yeast is also taken into account. With this model, we re-construct the genome structure of yeast and obtain compatible chromosomal interaction patterns as observed in experiments. Taken together, from one-dimensional gene organization along chromosomes and gene regulatory relationships, we manage to re-construct the spatial genome organization in budding yeast.

## 6.2 Material and Methods

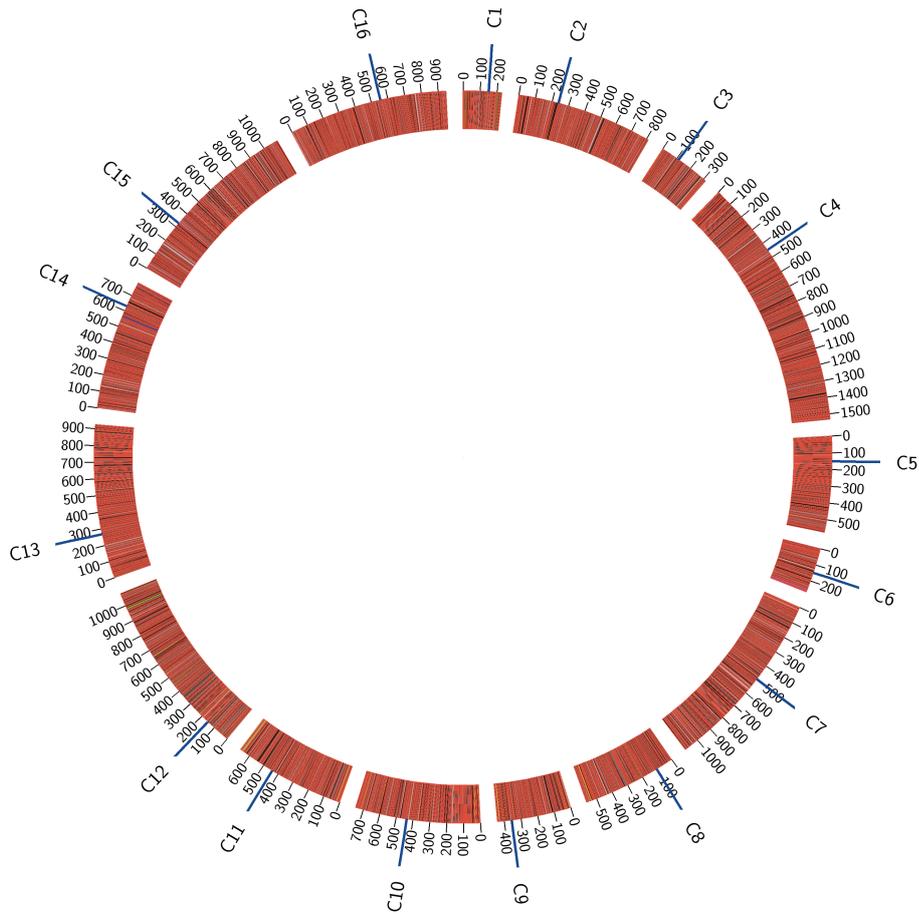
### 6.2.1 Neighboring Genes Selection

The transcriptional regulatory network of budding yeast is adapted from the study of Balaji *et al.* [229]. They reconstruct a regulatory network consisting of 157 transcription factors and 4,410 target genes involving 12,873 regulatory interactions in total. We determine transcription factors control similarities (TCSs), a notation proposed by Batada *et al.* [235] between all possible genes pairs involved in this network. For a given gene pair its TCS is defined as one

minus the ratio between the number of transcription factors regulating only one of two target genes and the total number of regulatory interactions. For instance, Gene I is regulated by TFs A, B, and C, while Gene II is regulated by TFs B, C, D, and F. Their TCS is calculated as:  $1.0 - (1 + 2)/(3 + 4) = 4/7$ . We assume gene pairs with high TCS as spatial neighbors in the nucleus. To make a small, robust and reliable neighboring list, we removed target genes that are regulated by less than two TFs, and set the TCS as 1.0. Thus selected neighboring genes should be regulated by at least two common TFs. By doing so, we cluster genes spatially according to their transcription regulatory control profiles.

## 6.2.2 Modeling and Numerical Simulation

To test our idea, we select *S. cerevisiae*, one of the most thoroughly studied species, either on its genome sequences, or more importantly on its transcriptional regulatory control profile. The genome information of budding yeast is shown in Figure 6.1. The nucleus of budding yeast is covered with a nuclear envelope and has a radius of approximately one micrometer. In the interphase cell, its genome takes a Rabl configuration [113, 236], in which centromeres are attached via protein fibers to the spindle pole body (SPB), a structure inserted into the nuclear envelope. Telomeres are positioned close to the nuclear envelope [95]. Such a conformation makes the chromatin within centromeres or telomeres less flexible than in other parts. Another distinguishable structural feature inside the budding yeast nucleus is the nucleolus, where amplified in tandem ribosomal RNA (rDNA) genes on chromosome XII form a cluster sitting opposite to the pole of the spindle pole body (SPB) next to the nuclear envelope [237, 238]. These intensively duplicated genes are vital house-keeping genes, and kept silent, but they are indispensable to homologous genes repairing or to nucleus structure maintenance [239].

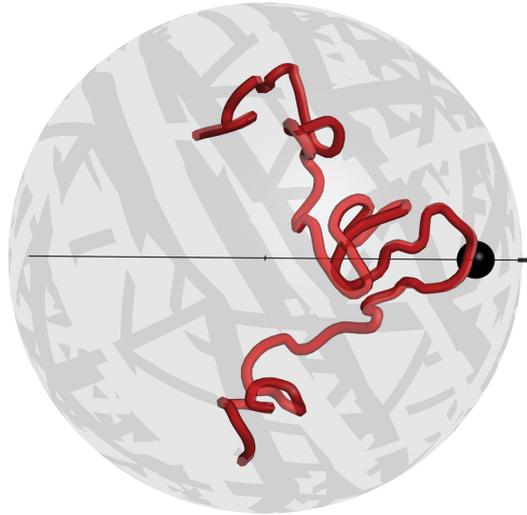


**Figure 6.1: Genome structure of *S. cerevisiae*.** The haploid genome of *S. cerevisiae* includes 16 euchromosomes and an extra mitochondria chromosome (not shown here). Scales on each chromosome indicate its length (unit in kb). On each chromosome ‘C’ indicates the position of its centromere in the form its affixed chromosome name.

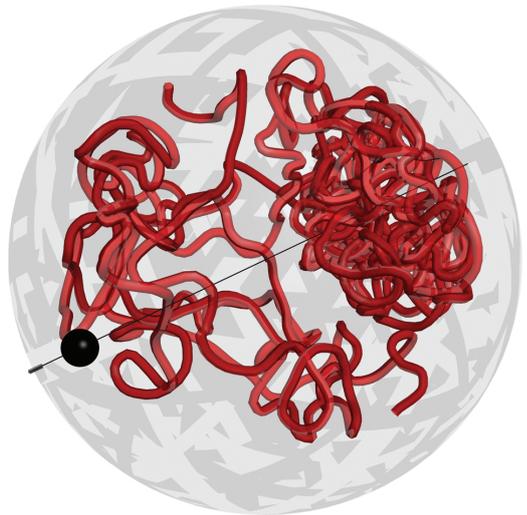
All molecular dynamics simulations are carried out with ESPResSo software package [223]. A chromosome is simplified as a fragment of a polymer composed of homogenous monomers with excluded volume and bonded via a FENE potential. The Weeks-Chandler-Andersen (WCA) potential (a special case of Lennard-Jones potential) is applied for the non-bonded interaction. The attractive part of the interaction between particles is turned off. Due to the large scale genomic studies conducted on the budding yeast its gene profile has largely been clarified. This makes simulations with high resolution possible. To make the simulation computationally affordable we con-

struct coarse-grained polymers. For all of 16 chromosomes in budding yeast, for every 2 kb a monomer is set. Assumed gene neighbors are mapped onto coarse-grained polymers by taking the gene's middle position for simplicity. The genome information is taken from the Ensembl genome database of budding yeast (*saccharomyces\_cerevisiae\_core\_58\_1j*) [240]. Also, to mimic the duplicated rDNA gene cluster on chromosome XII, we insert a patch of a 1500 kb long polymer (corresponding to 750 monomers) into the position 450 kb far away from the left end of chromosome XII. All these constructions give a system containing 6800 monomers. A typical monomer volume fraction of 0.15 is used [230] from which the radius of a simulated nucleus sphere is calculated. To construct the SPB structure, we set up an extra particle on the surface of sphere and deemed it as one pole. Genomic positions of centromeres are taken from the *Saccharomyces* Genome Database (SGD). Harmonic bonds between virtual centromere on each polymer and the particles that function as SPB are assigned. To simulate the rDNA cluster, we put another particle at the pole opposite to the SPB and on the surface of the confining sphere, which corresponds to the nucleolus site. Multiple harmonic bonds are used to connect the inserted rDNA polymer with the nucleolus particle. Telomeres have been shown to prefer localizations at the peripheral region of the nucleus. Thus, harmonic bonds are created between both ends of each polymer and the centre point of confining sphere, which will push polymer ends outwards to sit close to the virtual nuclear membrane.

Initially the 16 polymers are randomly distributed. A spherical confinement is defined by a Lennard-Jones potential between all the particles and the constraint surface to mimic the nuclear envelope. Depending on the particle density at the beginning of the simulation and the initial arrangement of the chains, the radius of the confining sphere is about 30 times larger than the final radius with monomer occupancy of 15%. At the first stage of the simulation, the large constraint sphere is slowly shrunk between every integration cycle. When the radius of the sphere reaches its expected value, the telomere interactions and rDNA interactions are created. To speed up the equilibration, the parameters of FENE potential is selected so as to make the back-bone extensible and allow chains to pass through each other. Also, the Lennard-Jones potential is capped and increased gradually to get ride of the overlapping of



(a)



(b)

**Figure 6.2: Modeling structural features of yeast nucleus.** In both panels, small black spheres represent the structure spindle pole body (SPB) where centromeres assemble. The surrounding transparent grey sphere defines a confining surface corresponding to the nucleus member. The black line indicates the axle extending from the SPB to the rDNA genes cluster. Harmonic bonds are employed to congregate centromeres and rDNA genes and to drive telomeres to the periphery of the confining sphere. Figure (a) shows chromosome 3 and figure (b) chromosome 12, which has bulk rDNA genes clustered on the pole opposite to the SPB.

particles and to facilitate the equilibrium. The simulation runs continually until various energies, the end-to-end distance, and the radius of gyration kept stabilized. To evaluate the role of specific gene-gene interaction, we construct specific control simulations where no gene-gene interaction is included. To get an ensemble of genome conformations, we run 10 independent simulations with totally different initial polymer arrangement both for the control simulations and for non-control simulations. The parameters used in the simulations are summarized in Table 6.1. For the explanation of different types of interactions involved in the simulation, the reader can refer to section 5.2.2 of the chapter 5.

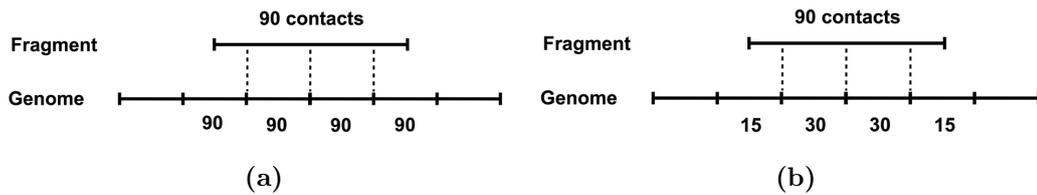
**Table 6.1: List of parameters.**

|   |
|---|
| 1. Parameters for system  |
| initial density = 1E-4<br>temperature T = 1.0<br>friction $\Gamma = 1.0$<br>time_step t = 0.01<br>final radius = 20.0 |
| 2. Parameters for L-J potential   |
| $\epsilon = 1.0$<br>$\sigma = 1.0$<br>$r_{cut} = 2^{1/6}$   |
| 3. Parameters for FENE potential  |
| $K_{FENE} = 1.0$<br>$r_0 = 1.0$<br>$\Delta r_{max} = 15.0$  |
| 4. Parameters for Gene HARMONIC potential   |
| $K_{Gene} = 0.00005$<br>$R_{Gene} = 1.0$  |
| 5. Parameters for Centromere HARMONIC potential   |
| $K_{Centromere} = 0.05$<br>$R_{Centromere} = 1.0$   |
| 6. Parameters for rDNA HARMONIC potential   |
| $K_{rDNA} = 0.5$<br>$R_{rDNA} = 10.0$   |
| 7. Parameters for telomere HARMONIC potential   |
| $K_{Telomere} = 0.5$<br>$R_{Telomere} = 20.0$   |

### 6.2.3 Experimental Data

Chromosomal contact data is kindly provided by Justin M O’Sullivan, who performs a genome-wide screening of chromosome physical interactions in *S.*

*cerevisiae* (data and technology details unpublished). Similar to Hi-C approach used to investigate the human genome organization, qualified chromosomal contacts identified from yeast genome are drawn as a genome-wide chromosomal contact map, which is used to compare with our simulation results. In brief, chromosomal contacts in yeast genome are fixed and digested with restriction enzyme. Ligated chromosome fragment pairs are sequenced before genome mapping. Those ambitiously mapped fragments are filtered out. The rest fragments, which can be uniquely mapped onto the genome, are used to construct chromosomal contact map. The resolution of the contact map is around 2 kb per bin. Since a fragment could possibly spanning several genomic bins, two different ways are used to assign interactions from a single fragment to covered genomic bins: 1) simple assignment, and 2) proportional assignment, as shown in Figure 6.3. In the case of simple assignment, if one fragment spanning several genomic bins, each bin involved is given the same number of interactions as this fragment has, no matter how many base pairs a bin is covered by the fragment. Alternatively, one could also do a proportional assignment by means of assigning interactions to a bin based on the ratio of its overlapped length with the fragment to the full length of that fragment. Still, there exists another assignment problem coming from the tail portions of chromosomes. Usually, the last bin on the chromosome is not intact and sometimes much smaller than complete segment size. To cope with this problem, Justin proposed three different strategies: 1) `fixed_length_short_last`, in which each segment is 2 kb long except for the last bin on the chromosome, no matter how small it might be; 2) `fixed_length_flexible_last`, similar to method 1, except that if last bin is smaller than 1 kb, it will be combined with the last second segment from the tail. Otherwise, it's kept as an individual segment; 3) `flexible_length`, in which the segment size is calculated individually for each chromosome such that all segments of one chromosome have the same length, and this length is as close to 2,000 bp as possible.



**Figure 6.3: A comparison between simple assignment and proportional assignment methods.** Numbers labeled on bins indicate the number of interactions assigned to each segment.

## 6.2.4 Chromosomal Interactions Analysis

The inter-chromosomal interaction maps of any two given chromosomes are obtained by determining those pairs of monomers having a distance below a given threshold. If the centromere sites are taken into consideration, a chromosome can be treated as two chains connected by a hinge (the centromere site), and an arm-based interaction map is generated.

For the intra-chromosomal interaction, genomic distances between interacting loci and their interaction frequencies are calculated. To determine their mathematical relationship the data is plotted on a double-logarithmic plot. Any straight line behavior indicates a power law relationship ( $f(x) \propto x^\alpha$ ). Interactions with a genomic distance smaller than 20 kb are eliminated, which is predominantly displayed on contact maps.

For the inter-chromosomal interactions we construct a random interaction matrix based on lengths of two given interacting chromosomes by assuming that they are arbitrarily arranged in the cell nucleus. To estimate the similarity between contact patterns, we use the cosine similarity measurement, which calculates the cosine of the angle between two vectors. Here a matrix is interpreted as an  $N \times N$  vector. The more the resulting cosine value approaches to 1, the higher the similarity is.

Another measure uses the eigenvalues of the contact matrix. These are invariant under unitary transformations and thus should provide information on the similarity irrespective of the absolute values of the matrix (a problem with the cosine similarity measure).

## 6.2.5 Territory Analysis

To create a locus probability map, we approximately tag the monomer corresponding to the gene loci screened in this study [241], and project this monomer onto a circular plane. The plane is divided into a grid. The frequency of gene loci falling into each box of grid is measured to generate a probability density map. The circular plane is divided into upper and lower halves, which are summed up and mirrored. A Gaussian smoother algorithm is utilized to blur the map in order to facilitate visual comparison. To define a territory of individual chromosome, a surface confining 70% of total volume occupied by the conformation ensemble of a single chromosome is constructed.

## 6.3 Results and Dissuasion

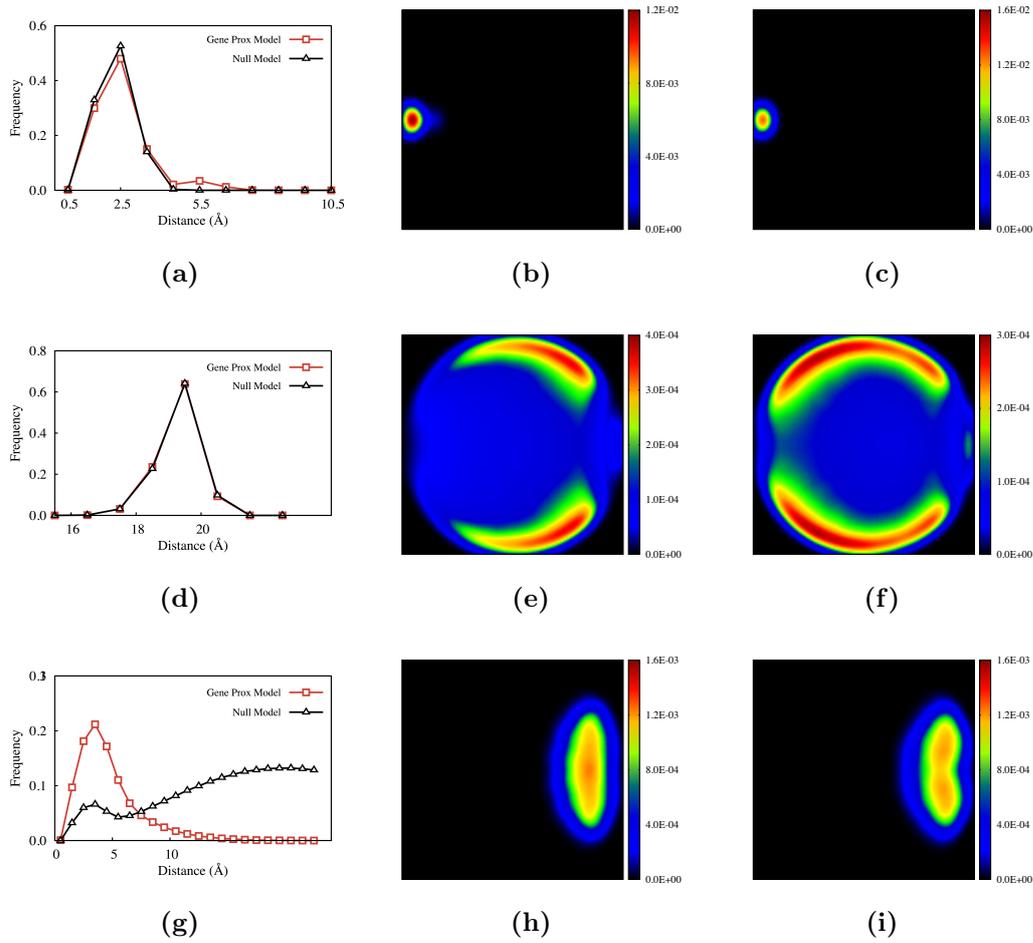
### 6.3.1 Neighboring Genes Selection

Using the method described in the ‘Modeling and Numerical Simulation’ section on the whole transcriptional regulatory network, a list of pair-wise genes’ TCSs is created. By setting the minimal number of TFs operating on the respective gene of gene pairs as 2, and the TCS as 1.0 (entirely identical), we extract a sub-network composed of 769 different genes, which gives us 1987 gene pairs with high TCSs. Among these genes, co-regulated ones are grouped as clusters. In total 485 clusters are obtained, i.e. genes that are under highly identical control. Most of the clusters consist of a small number of members, while few large clusters including up to 40 members are found.

### 6.3.2 Validation of the Model

To validate our model we check the position distributions of several genomic loci, including the centromere, the telomere, and the presumed interacting gene loci. As can be seen from Figure 6.4, the radical distributions of the centromere and the telomere loci are quite similar under either scenario (Gene Proximity Model and Null Model). While the telomeres in the Null Model are distributed peripherally, they occupy the space below the virtual nucleus membrane more uniformly compared with those in the Gene Proximity Model. Centromeres are aggregated around the SPB sites in both models. For interacting gene loci, as

we expected, by creating bonds between genes (see ‘Modeling and Numerical Simulation’) that are supposed to interact, their physical distances fluctuate around the mean distance set in the model with gene interaction. For the model without gene interaction, a small hump can still be observed at short scale. This might be due to short range intra-chromosomal interactions included in the interacting gene list, which are genomically close on chromosomes. By and large, in the control case, where no gene interaction is added, presumed interacting gene loci are much farther apart from each other.



**Figure 6.4: Distribution of genomic loci in Null Model and Gene Proximity Model.** Panel (a) shows the distribution of distances from the centromeres to the SPB site in the simulations with and without gene interaction. Panel (b) shows the two-dimensional projection of the position of centromeres in the simulation without gene interaction, and panel (c) with gene interaction. Panel (d) shows the distribution of distances from telomeres to the SPB sites for two kinds of simulations. Panel (e) shows the two-dimensional projection of telomeres in the simulation with gene interaction, and (f) shows the pattern without gene interaction. Panel (g) shows the distributions of distances between presumed interacting gene pairs for simulations with and without gene interaction. Finally panels (h) and (i) shows the two-dimensional projection of rDNA genes cluster in simulations with and without gene interaction, respectively.

### 6.3.3 Comparison of Single Chromosome Folding Pattern

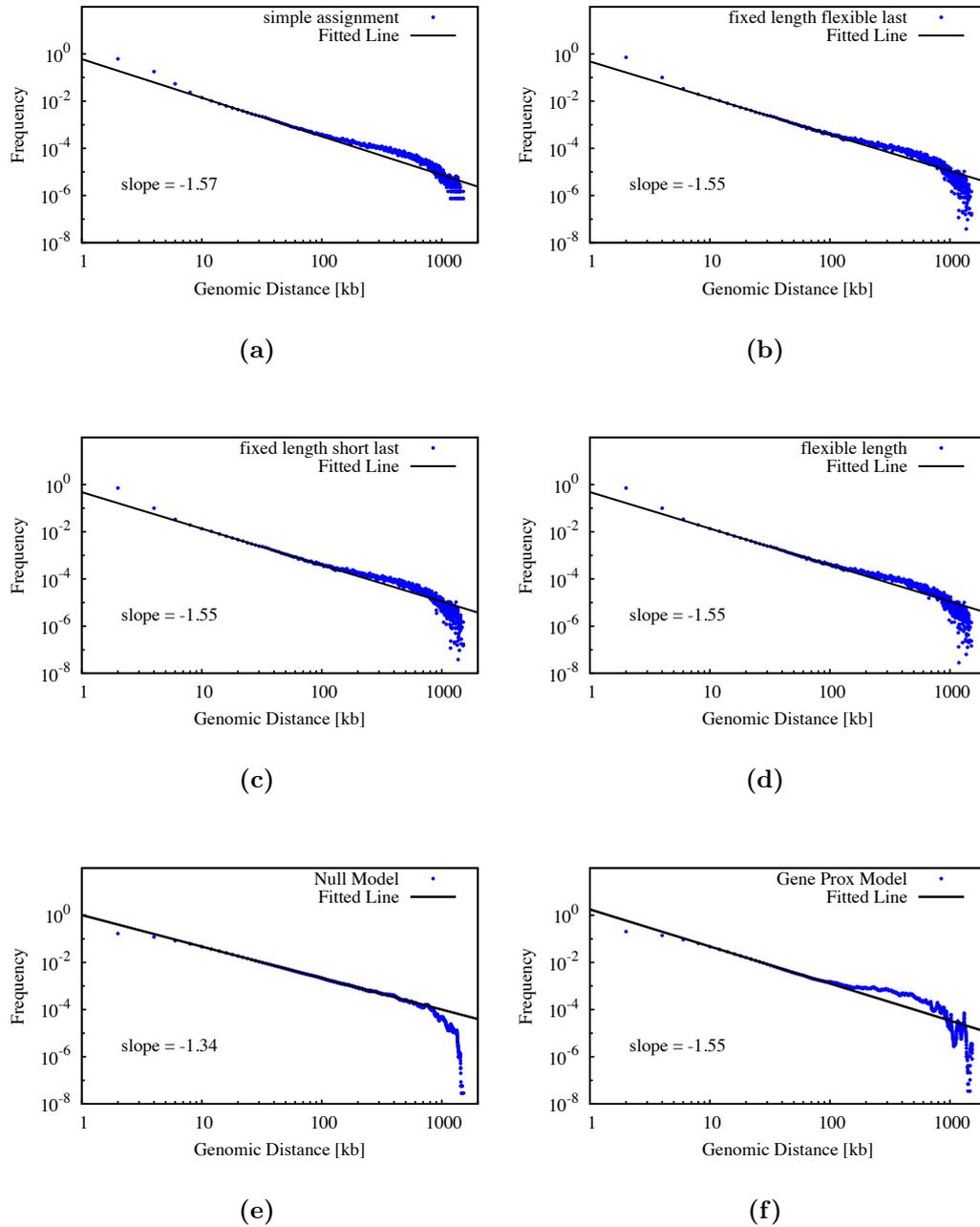
To study the folding of individual chromosomes, we check genomic distances between contacting chromosome loci and their contact frequencies. For two kinds of models (with and without gene interaction), we calculate genomic distances between all possible pairs on each chromosome and kept those with spatial distances below the contact threshold. From these the contact probabilities under given genomic distances are computed. For comparison the data from the experiment is calculated with four different contact assignment methods. We check their bin distances and bin pair contact frequencies.

The analysis of the data is based on ideas that come from polymer physics. If we assume that the chromosome is randomly folded like a random walk polymer (a polymer with no correlation along its contour), then the frequency with which parts come in contact is given by a power law ( $f(x) \propto x^\alpha$ ). The same holds true if we assume other polymer models, like self-avoiding walk, the globular state, fractal globular state or dynamic loop model. They differ in the exponent  $\alpha$ . Plotting the genomic distance versus the contact frequency on a double-logarithmical plot we can extract the exponent by fitting a straight line to the data.

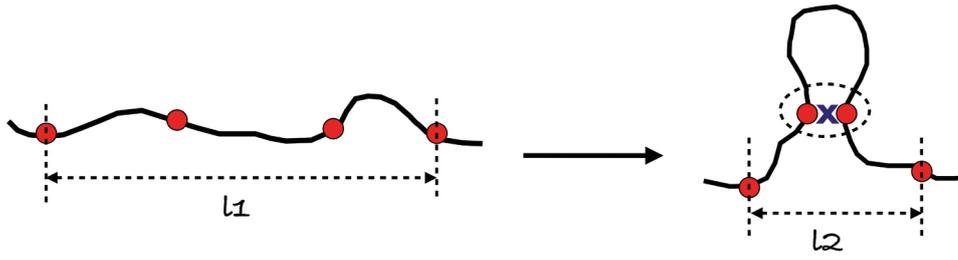
As shown in Figure 6.5, the values for the slope extracted from the experimental data with different assigning methods are generally compatible with  $\alpha$  around  $-1.5$ . For the Gene Proximity Model with gene pairs sharing high TCSs, we obtain a similar value for  $\alpha$ . In contrast, for the Null Model we obtain a value  $-1.34$ . Moreover, both in experimental pattern and in Gene Proximity, a hump of higher interaction frequency can be observed extending from several hundred kb to one Mb. However, such a hump is not visible in the control model. This evident hump at larger genomic scale implies an enhanced interaction probability and might result from gene interactions at shorter genomic scale. Interactions between genes with short to medium genomic distances could largely decrease spatial distance between distant gene loci, as an illustration shown in Figure 6.6, which, in turn, could enhance the contact probability between remote chromosomal loci. It should be pointed out that the control model, except for yeast nucleus structural features no other interaction between loci are used.

Another way of looking at the contacts within individual chromosomes is

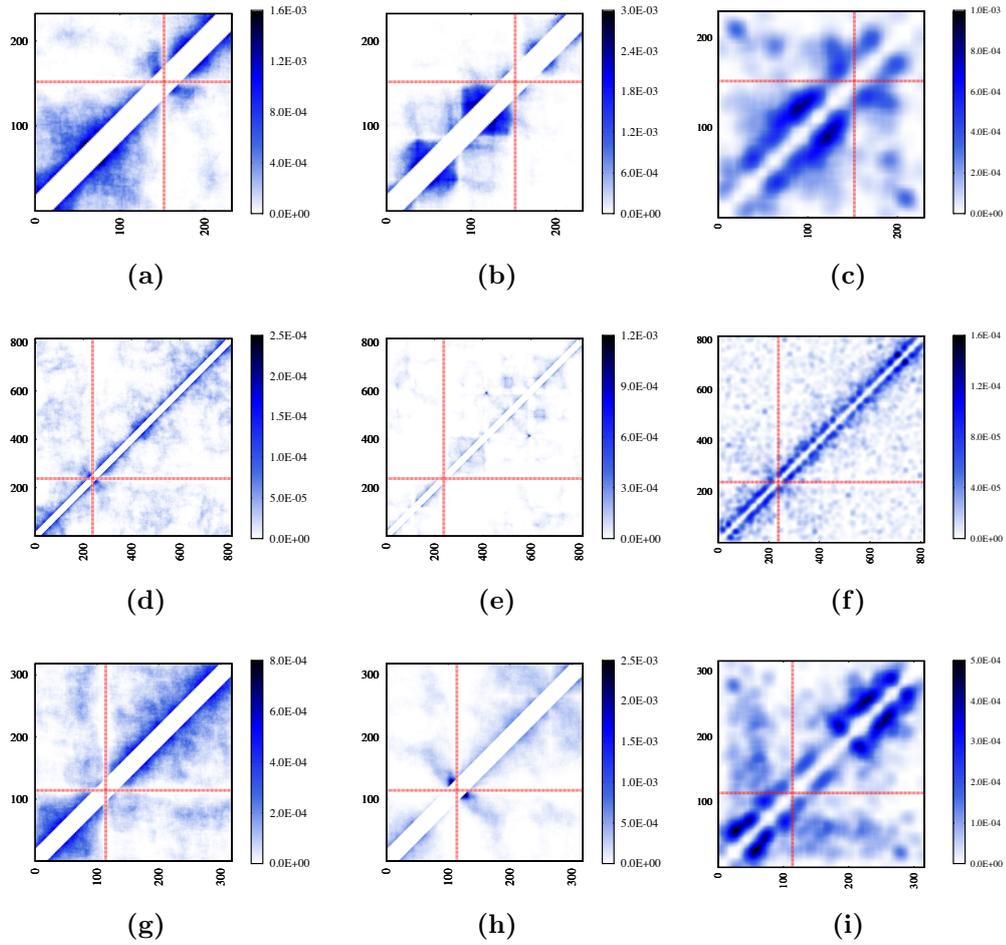
to use a heat map. In heat maps the contact frequencies among chromosome fragments are depicted color. No interaction is shown as with space. With increasing number of contact frequency the color shift from white towards one color. Figure 6.7 shows the results for the heat maps. Because interactions closed to the diagonal of contact map are extremely predominant, which means that intra-chromosomal interactions at short genomic scale are far more abundant, we only keep interaction signals between chromosome fragments with a genomic distance over 20 kb in order to focus more on the off-diagonal pattern. Since off-diagonal interaction signals on experimental contact maps are quite scattered, we use Gaussian blur to make the pattern more visible.



**Figure 6.5: Power-law behavior of intra-chromosomal organization.** The chromosomal contact probability is plotted double-logarithmically as a function of the genome distance. A (fitted) straight line indicates a genomic region where a power law is observed. Panels (a) to (d) show experimental data using four different fragment assigning methods [(a) for simple assigning, (b) for fixed length flexible last, (c) for fixed length short last, and (d) for flexible length, as explained in ‘Material and Methods’]. Panel (e) and (f) show patterns identified from the model without or with gene interaction, respectively.



**Figure 6.6: Gene interaction at short scale impact on genomically distant interactions.** Gene interactions at short scale can largely reduce the spatial distance between distant gene loci and hence increase the contact probability.

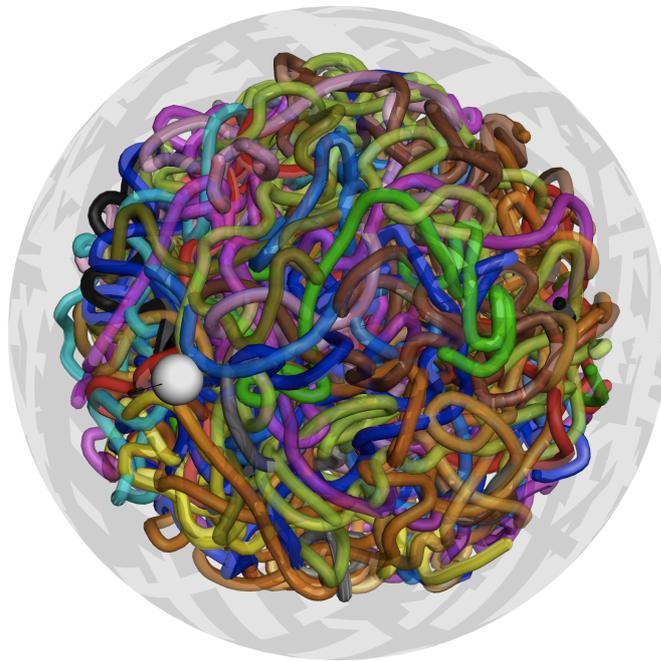


**Figure 6.7: Comparison of intra-chromosomal contact patterns.** Each row shows the contact patterns obtained from the Null model, model based on the transcription network and experiment (respectively from left to right). The three rows are for the chromosomes from 1 to 3. More patterns are available in the Appendix A.

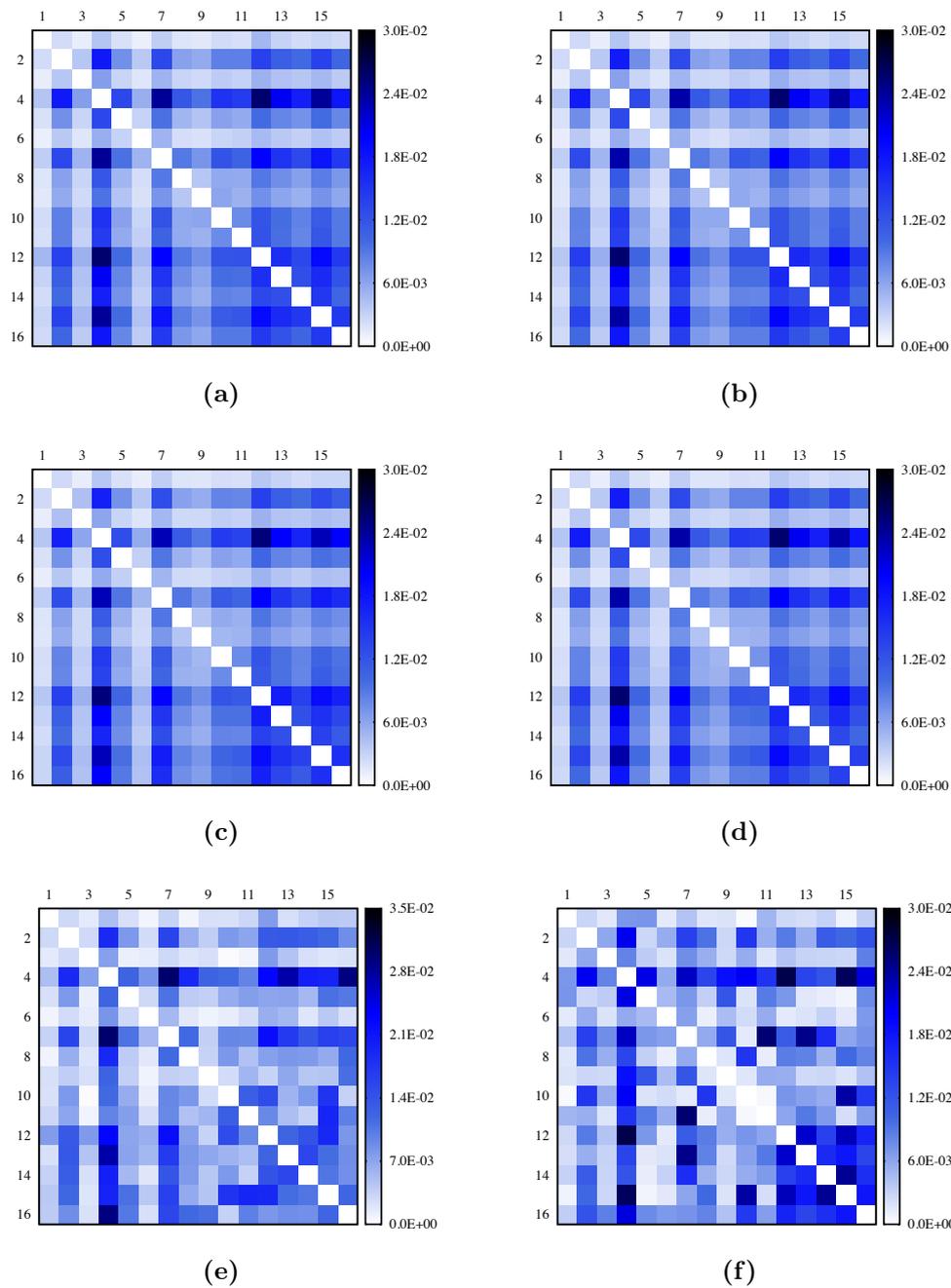
### 6.3.4 Comparison of Inter-chromosomal Contact Pattern

Shown in Figure 6.8 is a sample configuration using the model derived from the transcription network. By applying the above-mentioned assignment methods and calculating interaction numbers for any two given chromosomes, we can generate inter-chromosomal contact patterns (a 16 by 16 matrix) as heat maps (see Figure 6.9. Similarly, considering the existence of the centromere on every chromosome, we can view a single chromosome as two arms connected by the centromere site. By doing so, we can further generate heat maps for

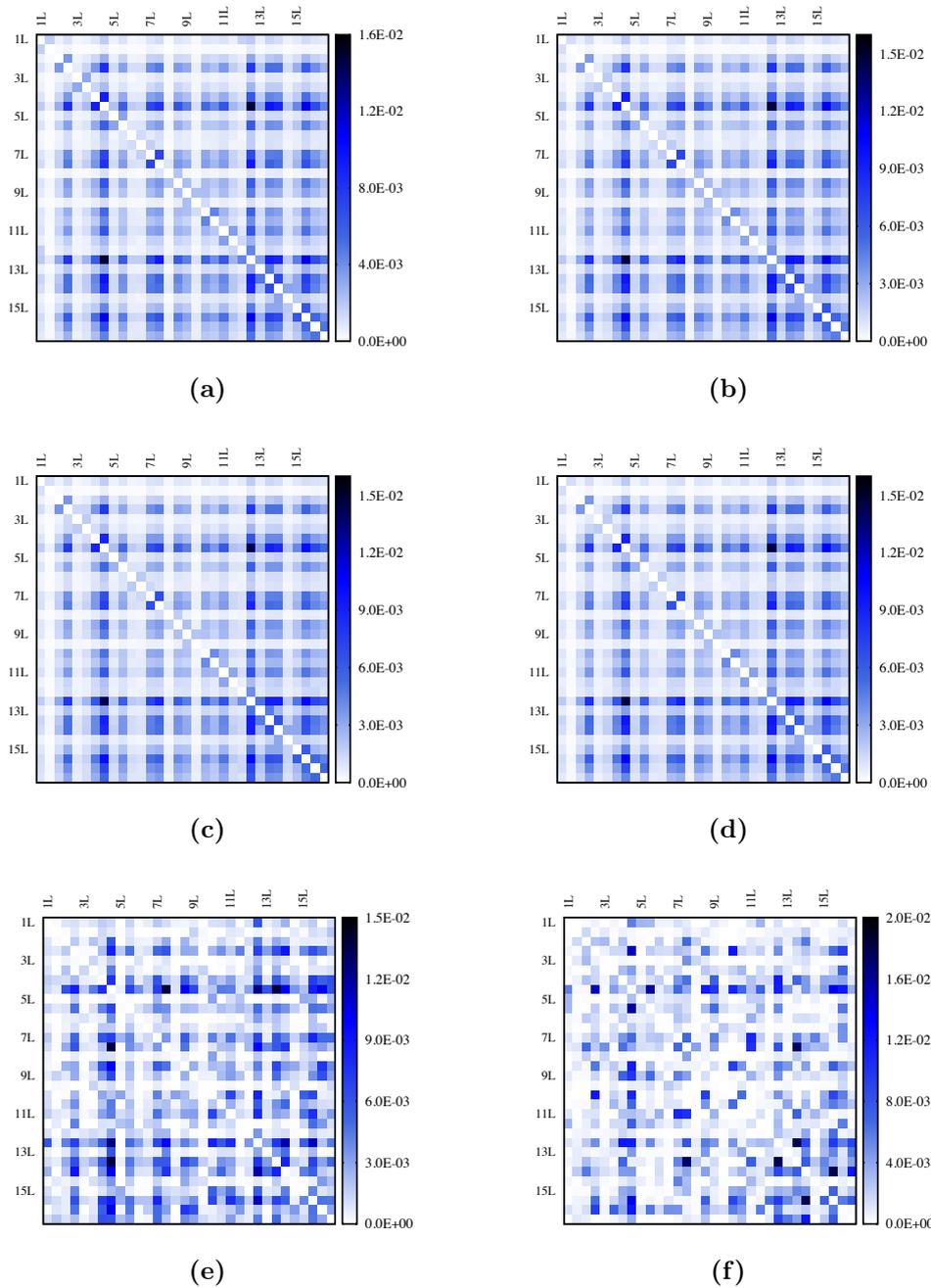
arm-interaction (matrix dimension 32 by 32), as shown in Figure 6.10. The interaction probability between two given chromosomes or arms are proportional to the color density.



**Figure 6.8: Illustration of genome architecture of budding yeast.** Here, one of conformations achieved from the Gene Proximity Model is shown. White sphere represents the SPB focus.



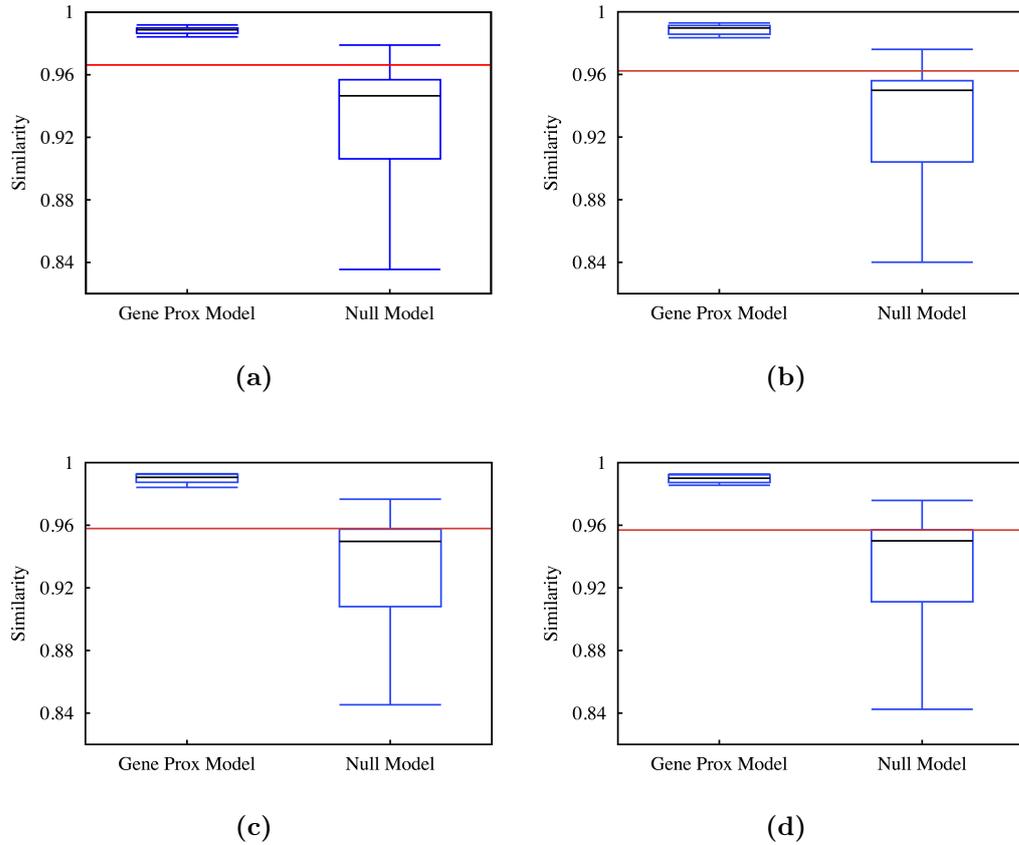
**Figure 6.9: Inter-chromosomal interaction patterns identified from the experiment and models.** Patterns from (a) to (d) correspond to: (a) for simple assigning, (b) for fixed\_length\_short\_last, (c) for fixed\_length\_flexible\_last, and (d) for flexible\_length. Pattern (e) and (f) correspond the model with interaction and without, respectively.



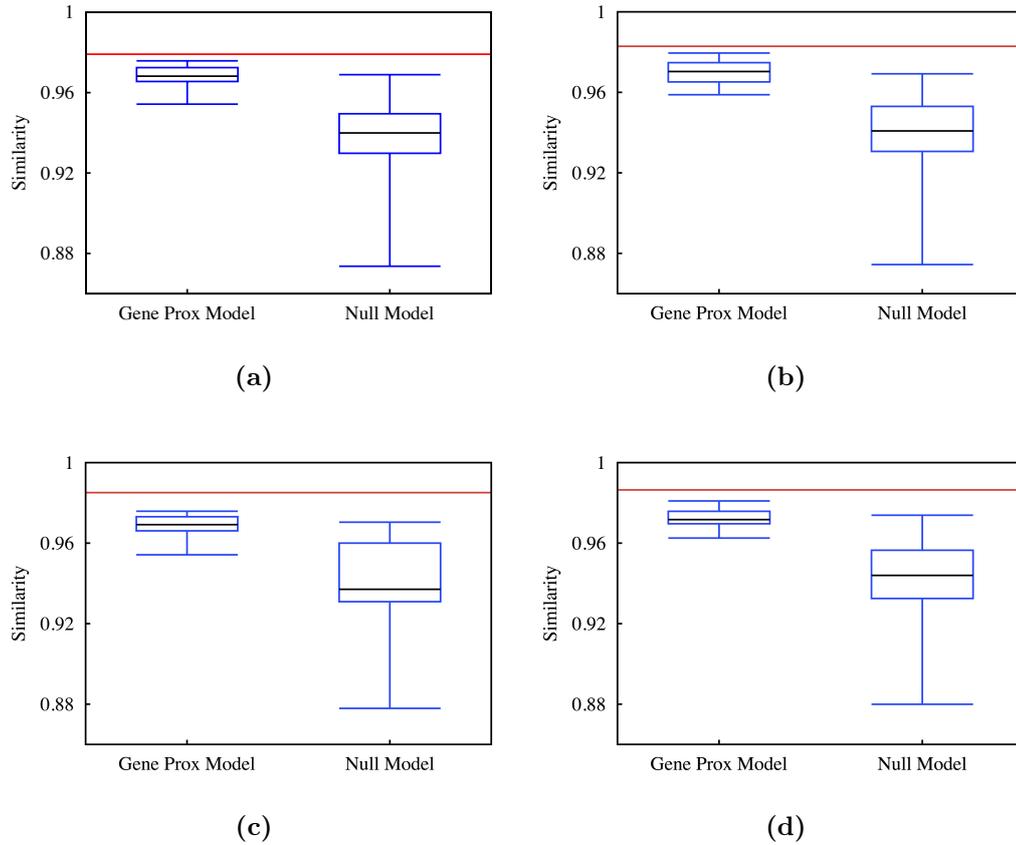
**Figure 6.10: Patterns of chromosomal arm interaction identified from experiment and simulations.** Patterns from (a) to (d) correspond to: (a) for simple assigning, (b) for fixed\_length\_short\_last, (c) for fixed\_length\_flexible\_last, and (d) for flexible\_length. Pattern (e) and (f) correspond the model with interaction and without, respectively.

Using the cosine similarity measurement, we evaluate the resemblance of

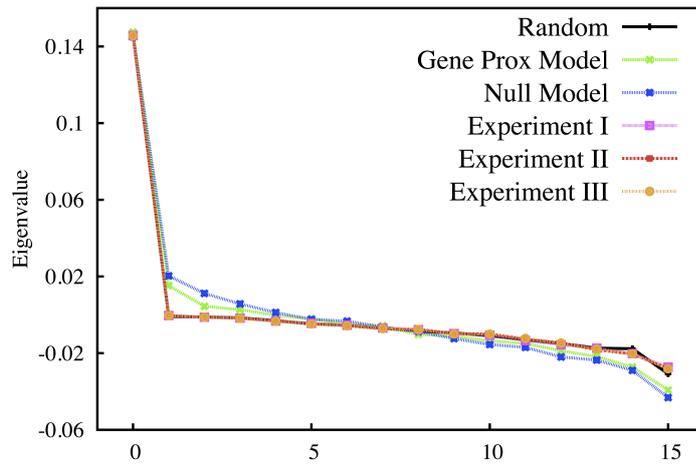
contact patterns achieved from simulation and from experiment, for inter-chromosomal interaction and arm interaction, respectively. As we can see from Figure 6.11 and Figure 6.12, by integrating gene-gene interactions in the model, we get a better resemblance, especially for inter-chromosomal interaction. Also, the similarity fluctuations of the Gene Proximity Model is much smaller than that of Null Model. This observation underpins the assumption that specific gene interactions play a role in spatial genome organization. In addition, the characteristics of the contact matrices can also be measured by their eigenvalues. Comparison of eigenvalues computed from contact matrices are shown in Figure 6.13. We find that Gene Proximity Model gives eigenvalues more close to the experiment than Null Model. We also calculate the average distances between eigenvalues contact matrices obtained from the model, random interactions, and experiments. The results are listed in Table 6.2 indicating that eigenvalues of contact matrix obtained from the interaction model are more close to the experimental pattern than the control model.



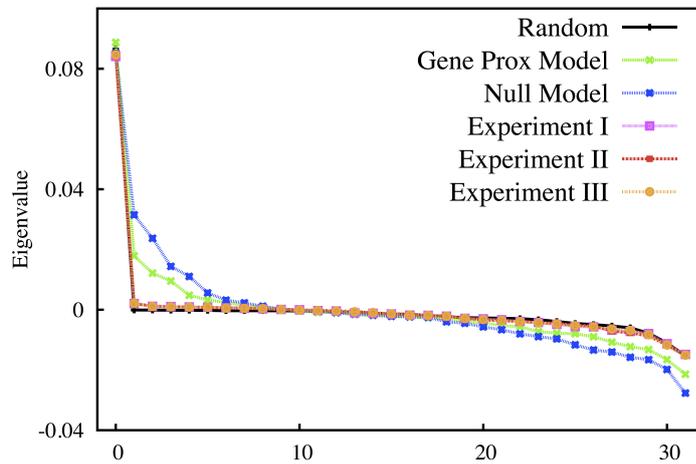
**Figure 6.11: Similarity measurement of inter-chromosomal interaction patterns identified from the models and experiment.** The red line indicates the similarity between random interaction pattern and the pattern extracted from the experimental data with different assigning methods. In each panel, the left candlestick shows the similarity between the Gene Proximity Model and the experiment. The right one indicates the results for the Null Model. (a) for simple assigning, (b) for `fixed_length_short_last`, (c) for `fixed_length_flexible_last`, and (d) for `flexible_length`.



**Figure 6.12: Similarity measurement of chromosomal arm interaction patterns identified from the models and experiment.** The red line indicates the similarity between random interaction pattern and the pattern extracted from the experimental data with different assignment methods. In each panel, the left candlestick shows the similarity between the Gene Proximity Model and the experiment. The right one indicates the results for the Null Model. (a) for simple assigning, (b) for fixed\_length\_short\_last, (c) for fixed\_length\_flexible\_last, and (d) for flexible\_length.



(a)



(b)

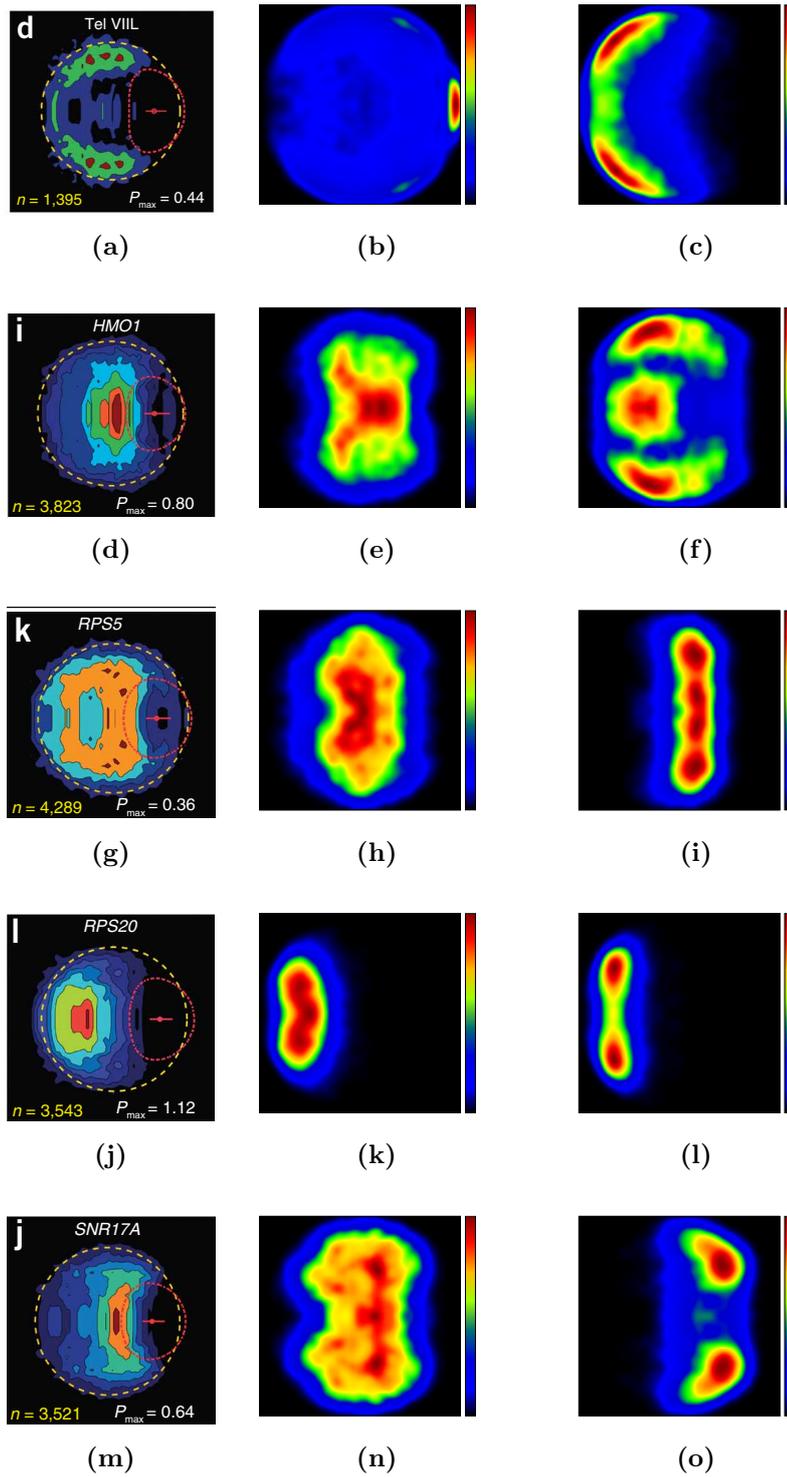
**Figure 6.13: Comparison of eigenvalues of contact matrices.** The characteristics of contact matrices seen from Figure 6.11 can be mathematically depicted by computing their eigenvalues. We used the eigenvalue comparison as a measurement of contact pattern similarity. The higher the similarity between two patterns, the closer two eigenvalue curves will stay. (a) for arm interaction pattern, and (b) for inter-chromosomal contact pattern.

| Simulation Type      | Experiment I | Experiment II | Experiment III |
|----------------------|--------------|---------------|----------------|
| Gene Proximity Model | 0.0041       | 0.0042        | 0.0042         |
| Null Model           | 0.0084       | 0.0084        | 0.0084         |
| Random               | 0.001        | 0.001         | 0.001          |
| Gene Proximity Model | 0.0043       | 0.0043        | 0.0043         |
| Null Model           | 0.0098       | 0.0098        | 0.0098         |
| Random               | 0.0011       | 0.0011        | 0.0011         |

**Table 6.2:** Average distances between eigenvalues computed for the **Gene Proximity Model**, **Null Model** and **experiment**. **Experiment I** for fixed\_length\_short\_last, **Experiment II** for fixed\_length\_flexible\_last, and **Experiment III** for flexible\_length.

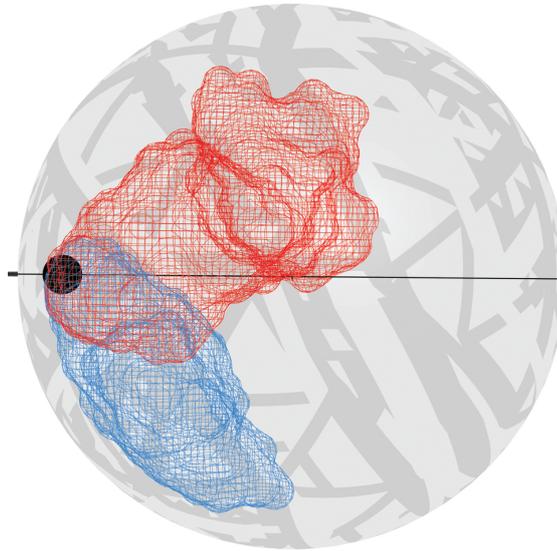
### 6.3.5 Territory Formation

We check the distribution of several genetic and structural loci (HMO1, RPS5, RPS20, SNR17A, and telomere of chromosome VII left arm) that have been screened experimentally [241], for comparison. While these loci can move and explore the nucleus to some extent, they stay more frequently within small region, which gives rise to an idea of ‘gene territories’. Shown in Figure 6.14 are these territories formed by genetic or structural loci for the Null Model, Gene Proximity Model and experiment. As we can see, some loci distribution patterns experimentally identified can be re-construct by the Gene Proximity Model (i.e. telomere of chromosome VII left arm in particular), while others are not compatible enough. Several factors might make the comparison difficult: 1) in the experiment [241], yeast cells can freely rotate around the central axis, which makes the alignment impossible; 2) in the Gene Proximity Model, all the genes are possible to be active once the requirement of spatial proximity is met, which is not the case in the experiment; and 3) it’s not clear if the conformation ensemble achieved from the Gene Proximity Model is comparable with that of the experiment.



**Figure 6.14: Comparison of ‘Gene Territories’ formed in the models and the experiment.** For each row, patterns are listed in the sequence of experiment, Null Model, and Gene Proximity Model, from left to right. Images of experimental patterns are adapted from [241].

We can, however, extend the idea of a territory for genetic loci to an individual chromosome. Although it's still doubtful whether in budding yeast chromosomes are organized in territories we can determine a three-dimensional space which specific chromosome occupy more frequently (Figure 6.15).



**Figure 6.15: Formation of chromosomal territory.** The meshes indicate the volume that a chromosome occupied during the course of simulation (here chromosome I). Red mesh represents the left arm, and blue for the right arm. The black sphere indicates the SPB loci.

## 6.4 Conclusion

The interplay between function and structure is becoming more and more apparent for chromosomes. Mechanisms, like transcription induced gene colocalization, have been postulated to explain how specific physical chromosome interactions can be induced. Improved chromosomal contact screening technology gives us more insight into the nature of the neighboring genes. Inspired by recently published findings that genes sharing highly homologous control sequences at their promoter regions will move towards each other and have higher interaction probabilities once activated [8], we propose here a genome organization model for budding yeast (Gene Proximity Model). The model emphasizes the role of gene transcriptional regulatory profile. We propose that the aggregation of specific TFs within the nucleus might function as a

recruiter to draw their target genes close in space, and probably to nearby transcription factories for coordinated expression. Our results show that our model does work to some extent. The model is able to explain some genome organization features of budding yeast. Admittedly, not all of the structural features from the experiment can be reconstructed, especially for interaction pattern at higher resolution.

At present, our model focuses only on the population-averaged chromosome behavior (which is also true for chromosomal conformation capture results available so far). In the scope of our model, interacting genes can be assumed active or inactive depending on the distance among them. In reality, activation of some genes are mutually exclusive, and unlike house-keeping genes, expression transducible genes can be periodically expressed along the cell-cycle [242]. Furthermore studies show that cell cycle transcriptional factors activated in one cell cycle stage could regulate those activated in the next cell cycle step [243] in a cyclic way. All of these factors make the genome conformation cell cycle and status specific. The model could take thus into account according to the time-resolved expression data to see the genome rearrangement along the cell cycle. However, so far we are lacking time-resolved conformation capture data.

# Chapter 7

## Conclusions and Extensions

### 7.1 A Summarization of the Results

Inspired by the evidence that genome function and three-dimensional genome organization are mutually dependent, we aim to develop precise models to explain the experimental evidence currently achieved, and to better understand the underlying mechanism of genome architecture from the perspective of structure-function relationships. To this end, we focus mainly on the role of gene transcription, which is one of most important processes in genome functions. With novel models developed, we confirm the interplay between functional and structural characterization of the genome organization both in prokaryotic and in eukaryotic cells. Genome structure modeling, which simultaneously takes structural and functional features into consideration, is a powerful approach that is instrumental to further experiments, and will give us a deeper insight into the genome-wide structure-function associations.

#### **Transcription Induced Chromosomal Contacts**

In chapter 4, we extract the inter- and intra-chromosomal contact patterns of human and mouse by making use of the chimeric transcripts as the probe. Chimeric transcripts induced by the gene transcription activity are believed to result from the erroneous ligation between the transcripts of concurrently transcribed genes within the same transcription apparatus. We confirm the authenticity of the chimeric transcripts discovered from the expressed sequence tags (ESTs) database in terms of their correlation in expression and functional

similarities. Gene loci that give rise to the chimeric transcripts are assumed to be spatial neighbors. Comparison with the chromosome contact pattern disclosed by the Hi-C chromosomal conformation screening method confirm that these chimeric genes are more likely to be physically proximal. And more important, we show that the arrangement of chromosomes within the cell nucleus is non-random. For the intra-chromosomal interactions, a functional behavior ( $f(x) \propto x^\alpha$ ) is observed from our data with the genomic distance ranging from 500 kb to 7-8 Mb, and an exponent  $\alpha = -1$  on a double logarithmic plot, which agrees well with the experimental results from Hi-C screening. Regardless of the fact that most chimeric transcripts are the inevitable burden the cells need to take on for the crowded environment in the cell nucleus, these chimeric transcripts do reflect the characterization of genome organization that is gene transcription relevant.

### **Transcriptional Regulatory Network Based Chromosomal Domain Formation in *E. coli***

In a subsequent step, we investigate the spatial packaging of the *E. coli* chromosome within the nucleoid. We suggest and verify a prospective role of the gene transcriptional regulatory network in shaping the *E. coli* chromosome as a domain-separated structure. Seeing the nature of transcription factors in searching and binding of their target DNA sequences, and the finding that mRNAs display a vary limited diffusion from their encoding sites, we postulate a transcriptional regulation based mechanism for the co-localization between effector genes and controlled genes. Taking into account the geometrical constraints imposed by the bacterial envelope and the spatial co-localization between the transcription factor genes and their target genes, we build up a genome architecture model and test it with *E. coli*. We find that the geometrical constraints stemming from the elongated nucleoid confinement and from the spatial proximity between specific genes are able to overcome the chromosome's propensity to mix and to organize the DNA chain into topologically distinguishable domains. This model well explains the experimentally observed precise subnuclear positioning of genetic loci, and their ordering [7]. Moreover, the domain size estimated from this model is in agreement with the size of topological domains identified in the context of DNA supercoiling. The

most recent study has also experimentally ascertained the role of transcription factors in aggregating their controlled genes [224]. Therefore loci-specific long- and short-range chromosomal interactions mediated by the gene transcriptional regulation could serve as a structure organizer in *E. coli*.

### **Genes under Similar Transcriptional Control are Spatial Neighbors in *S. cerevisiae***

Intrigued by the findings in *E. coli*, we further explore the interplay between gene transcriptional regulation and genome architecture in eukaryotic organisms by using *S. cerevisiae* as our model. Eukaryotic cells differ from prokaryotic ones in that their gene transcription and mRNA translation processes are dissociated and undertaken in different sub-cellular compartments. Protein products of effector genes are first generated outside the eukaryotic nucleus, and then imported back into the nucleus to execute their regulatory functions. In this context, considering the stochastic nature of TF-promoter binding, we postulate a local condensation mechanism, by which genes under similar transcription factor control, but possibly scattered throughout the genome, might be in physical proximity to facilitate the access of their commonly used transcription factors. Unlike the transcription factory model stated in chapter 5, ‘interacting’ genes in this model are not necessarily immediate physical neighbors but are in spatial proximity more or less. This local condensation mechanism could serve as a tie to recruit co-regulated genes to guarantee the swiftness of biological reactions. Chromosomal interaction patterns and folding behavior generated by this model re-construct those obtained from experiments. In the simple eukaryotic organism, budding yeast, we show again how could the transcriptional regulatory network be closely linked with the genome organization. This framework model is fundamental and instrumental to later studies on other more complex eukaryotes.

## 7.2 Extensions

### **Chimeric Transcripts as the Reference for Chromosomal Contacts**

The usage of chimeric transcripts in identifying chromosomal physical interactions works well in obtaining a general pattern of inter- and intra-chromosomal interactions. Because transcription-induced chimeras are very rare, from a single EST library or EST libraries for a specific cell line a very limited number of high quality chimeric transcripts can be found. Therefore we can hardly get enough chimeras data at the moment to analyze inter- and intra-chromosomal interaction patterns in a cell-line specific or development-stage specific fashion.

### **Other Factors that Facilitate Chromosome Packaging in *E. coli***

In this thesis, we have investigated the following factors that could contribute to the genome organization in *E. coli*: 1) excluded volume of particles, 2) geometrical confinement due to the bacterial envelop, and 3) transcriptional regulation-mediated specific gene co-localization. Although the results achieved are enlightening, there are still some other factors that are worthy to consider. Among them, the nucleoid-associated DNA-binding proteins (NAPs) are of special interest. NAPs are expressed in large amounts and serve as chromosome structural organizers to maintain the supercoiled structure of the nucleic acid [11, 24, 224]. Some NAPs, such as H-NS [11] and FIS, are also global transcription factors, which can regulate the expression of numerous target genes in a cell-cycle dependent way. Experiments show that each null-mutation of these NAPs can only impair the genome architecture to a limited degree. Some consequences resulted from different types of mutation of NAPs can be mutually compensated [244]. These observations imply the possibility that NAPs might work cooperatively in order to keep the whole DNA molecule in a dynamic equilibrium and resilient to perturbations. Consequently, a systematic approach will be fruitful in elucidating the structural roles of these NAPs, as seen from the studies in which chromosome conformation screening methods coupled with high-resolution imaging techniques are employed to track the distribution of chromosome structural organizers of interest [245, 246]. The observed distribution patterns are interpreted in relation to chromosome folding motifs [11]. Advanced fluorescent labeling methods, by means of which the

distribution of genetic loci can be scrutinized in a high throughput manner with a higher resolution, could also be helpful in characterizing the chromosome folding motifs in bacteria. Noninvasive genetic loci tracking approach with a spatial and temporal resolution is sorely needed as well, with which tracking of genetic loci during the whole cell cycle will be feasible. Then we can really appreciate the dynamic aspect of the genome organization.

### **3C-based Methods Reveal a Population Averaged Behavior of Genome Organization**

While 3C-based method makes it possible to measure the interaction frequencies between genomic loci, it's still suffering from several technical deficiencies at the moment: 1) cell samples for datasets construction include cell populations at various cell cycle states [5]. Since the cell samples are not well synchronized, it's hard to evaluate the percentage of cells under different cell cycle states. It's also evident that cells at various cell cycle states execute different gene expression profiles, which correspond to different and possibly mutually exclusive genome organizations. Consequently, some interactions included in the datasets can never simultaneously happen in any given cell. 2) 3C-based methods available so far can only reveal a population averaged behavior of chromosome organization. In order to increase the signal-to-noise ratio, a tremendous number of sample cells (usually millions of cells) are needed. After taking a close look at the number of total chromosomal contacts identified and the number of sampled cells, one can realize that a single sample cell only contributes to a very limited number of chromosomal contacts, which leads to the absolute number of interactions less informative and hard to compare, and further makes dubious the representability of the chromosomal contact pattern disclosed. For the above-mentioned reasons, a cycle-cycle specific or even a single-cell based chromosomal conformation capture approach is preferred, and will shed more light on the enigma of genome architecture.

### **Genome Architecture in Higher Eukaryotes**

In chapter 6, we test our idea of gene co-localization induced by transcriptional co-regulation in a simple eukaryote *S. cerevisiae* due to its relatively small genome size, and to its clearer genome architecture. In contrast to *S.*

*cerevisiae*, genome of a higher eukaryote generally associates a with much larger size, less gene density, and a more complicated gene regulatory network. Taking human as an example, the haploid genome contains about 23,000 protein-coding genes [247]. These gene coding regions represent, however, only 1.5% of the genome, and the rest of the genome includes non-coding RNA genes, regulatory sequences, introns, and noncoding DNA [248].

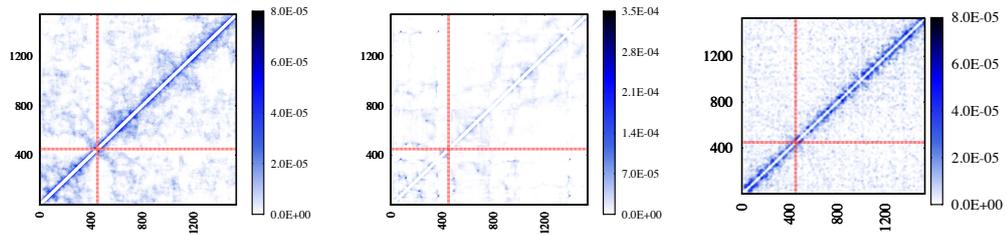
Two of most remarkable differences between higher and simple eukaryotic organisms reside in the euchromatin and heterochromatin, and the chromosome territory organization. Chromosomes in the nucleus of *S. cerevisiae* demonstrate a higher flexibility and the lack of chromosomal territories. Chromosomes of higher eukaryotes, by contrast, are delineated by heterochromatic and euchromatic domains. Euchromatin and heterochromatin show different preferences on their radial positioning inside the nucleus. The former occupies a more central radial position, while the later is more frequently peripherally distributed. Other factors, like chromosome size and gene density, also play a role in determining the radial position of a chromosome. Euchromatin represents genome regions with active gene expression. Heterochromatin, on the contrary, corresponds to silent genes or repeated sequences. In the interphase cell nuclei of higher eukaryotic organisms, only a subset of genes are actively expressed, which is cell-line specific. The organization of euchromatin and heterochromatin appeared in higher eukaryotic organisms provides them an extra controlling approach, through which they can keep those silent genes in condensed form.

Additionally, chromosomes in higher eukaryotic nucleus are folded and compacted as discrete territories. Chromosomal territories composed of heterogeneous parts from different chromosomes are also observable. Transcription-related segregation of chromosomes into distinguishable territories has been postulated [93]. Genes localized in the same chromosomal sub-compartment usually show a higher compatibility in their expression profiles. Unless we can have a deeper insight into the localization preference of euchromatin and heterochromatin, and the underlying mechanism for the formation of chromosomal territories, then a precise modeling of higher eukaryote genome will be possible. After that, we can really appreciate the genome organization strategy utilized by the higher eukaryotic organisms.

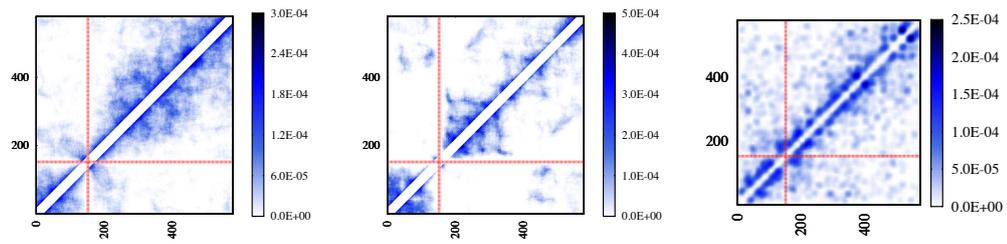
Due to the technological deficiencies of 3C-based methods with regard to their population-averaged nature as mentioned before, to make the results achieved from our model and from experiments comparable, varieties of gene expression profiles are used in our model. We assume co-regulated genes to be activated if the distance between them is small enough. Based on this assumption, we obtain an ensemble of genome conformations representing cells with heterogeneous expression profiles. For higher eukaryotic organisms, cells of which are more functionally specialized and consequently varied in expression, our model is not accurate enough at the moment. To overcome this problem, a model taking cell-line specific expression profile into consideration will be more powerful. Cell-line specific modeling will tell us more about the similarity and dissimilarity between different cell types with respect to their genome organization. And a cell-cycle dependent model, once devised, will demonstrate genome reorganization through the cell cycle.



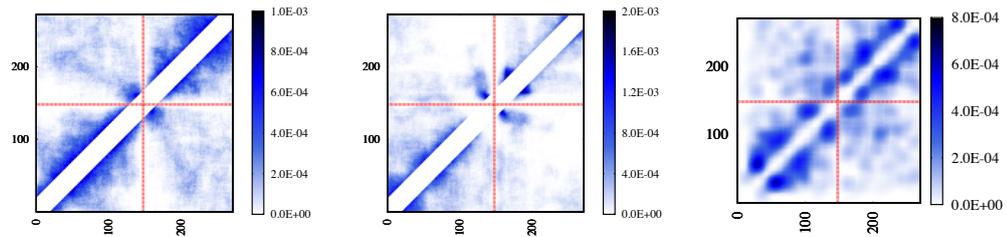
# Appendix



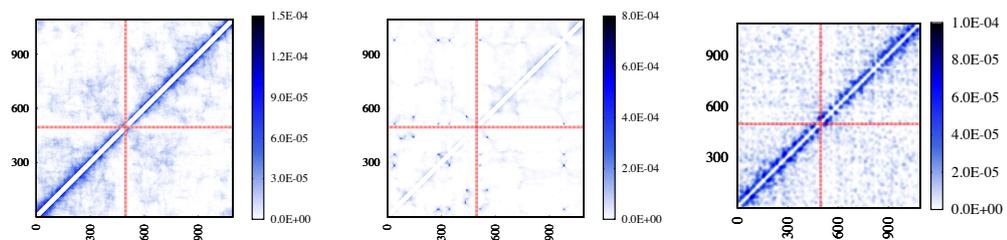
Chromosome 4



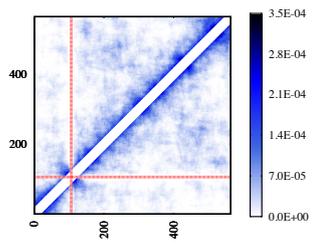
Chromosome 5



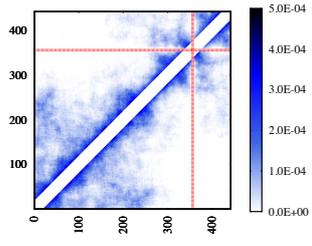
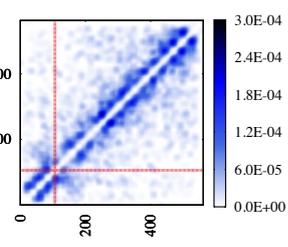
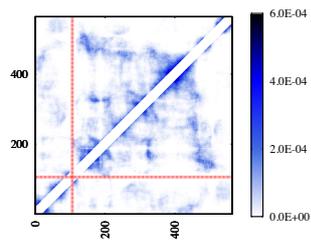
Chromosome 6



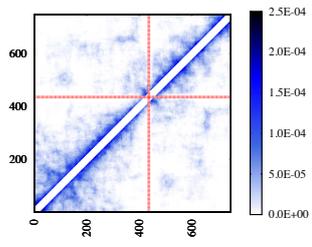
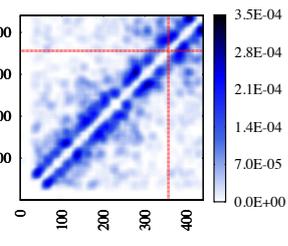
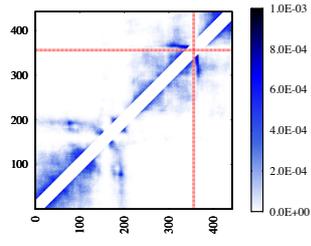
Chromosome 7



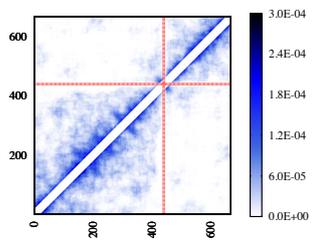
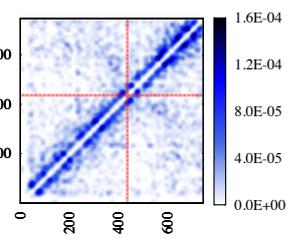
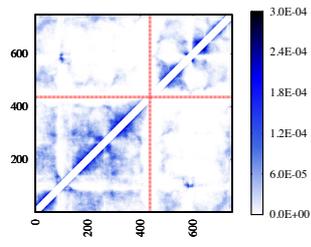
Chromosome 8



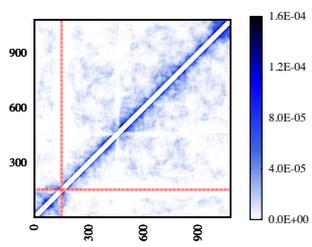
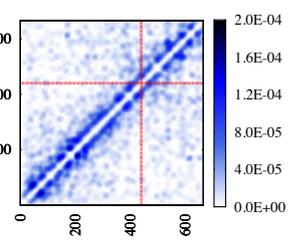
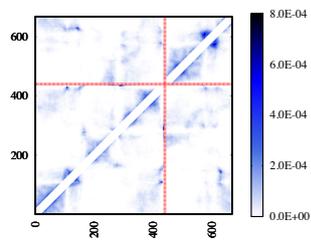
Chromosome 9



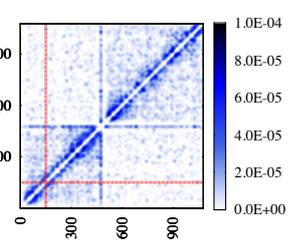
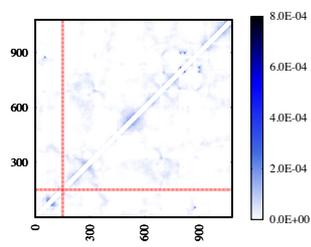
Chromosome 10

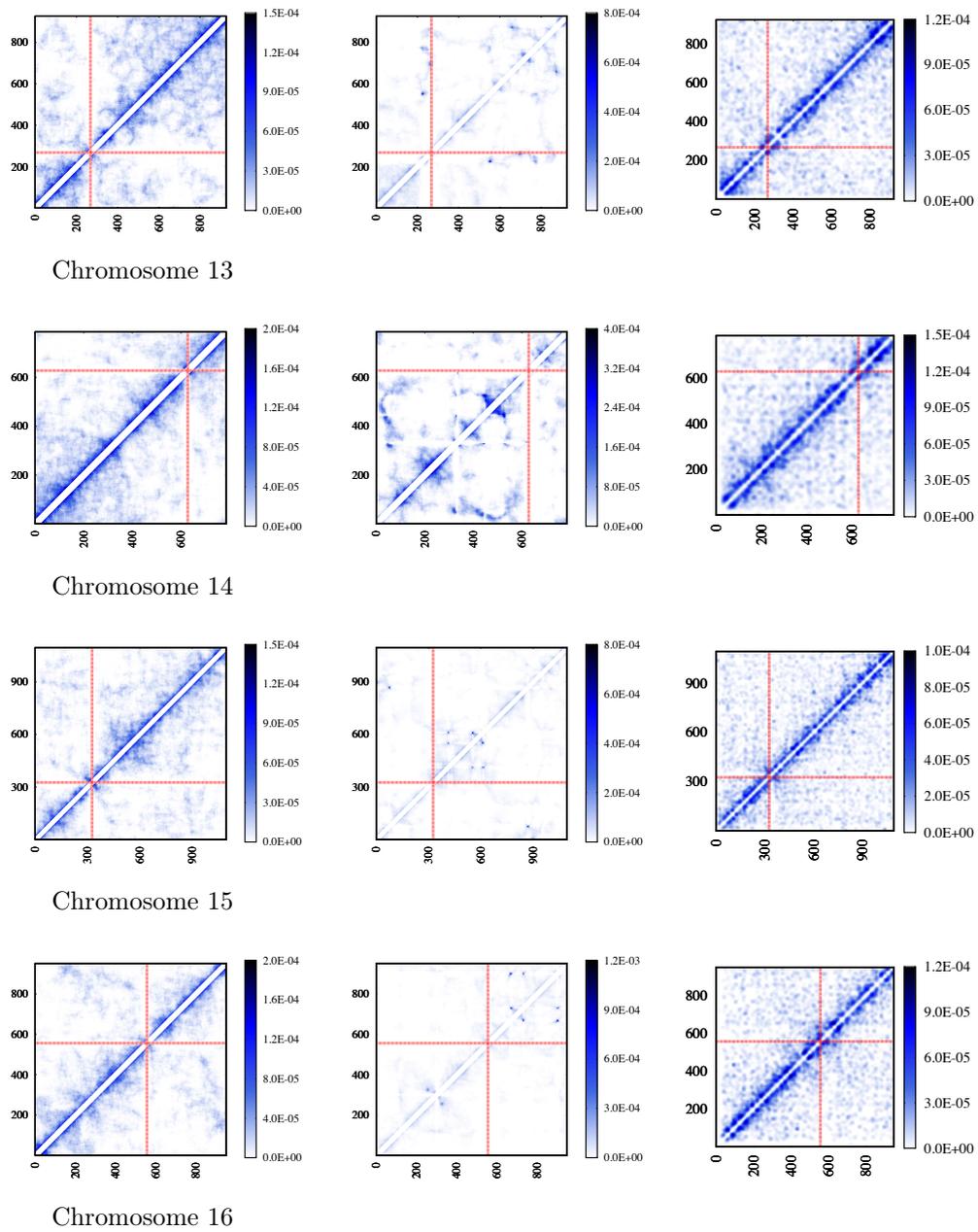


Chromosome 11



Chromosome 12





**Figure 1: Comparison of intra-chromosomal contact patterns.** Each row shows the contact patterns obtained from the Null model, model based on the transcription network and experiment (respectively from left to right). Each row for a single chromosome, from chromosome 4 to chromosome 16.



# Bibliography

- [1] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.
- [2] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.
- [3] J. Dostie, Y. Zhan, and J. Dekker. Chromosome conformation capture carbon copy technology. *Curr Protoc Mol Biol*, Chapter 21:Unit 21 14, 2007.
- [4] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet*, 38(11):1348–54, 2006.
- [5] J. Dostie and J. Dekker. Mapping networks of physical interactions between genomic elements using 5c technology. *Nat Protoc*, 2(4):988–1002, 2007.
- [6] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 3(5):e157, 2005.
- [7] P. A. Wiggins, K. C. Cheveralls, J. S. Martin, R. Lintner, and J. Kondev. Strong intranucleoid interactions organize the escherichia coli chromosome into a nucleoid filament. *Proc Natl Acad Sci U S A*, 107(11):4991–5, 2010.

- [8] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, et al. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–7, 2010.
- [9] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.
- [10] S. C. Janga, H. Salgado, and A. Martinez-Antonio. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Research*, 37(11):3680–8, 2009.
- [11] W. Wang, G. W. Li, C. Chen, X. S. Xie, and X. Zhuang. Chromosome organization by a nucleoid-associated protein in live bacteria. *Science*, 333(6048):1445–9, 2011.
- [12] H. Tanizawa, O. Iwasaki, A. Tanaka, J. R. Capizzi, P. Wickramasinghe, M. Lee, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*, 38(22):8164–77, 2010.
- [13] P. Montero Llopis, A. F. Jackson, O. Sliusarenko, I. Surovtsev, J. Heinrich, T. Emonet, et al. Spatial organization of the flow of genetic information in bacteria. *Nature*, 466(7302):77–81, 2010.
- [14] J. Fraser, M. Rousseau, S. Shenker, M. A. Ferraiuolo, Y. Hayashizaki, M. Blanchette, and J. Dostie. Chromatin conformation signatures of cellular differentiation. *Genome Biol*, 10(4):R37, 2009.
- [15] P. R. Cook. A model for all genomes: the role of transcription factories. *J Mol Biol*, 395(1):1–10, 2010.
- [16] P. Akiva, A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, et al. Transcription-mediated gene fusion in the human genome. *Genome Res*, 16(1):30–36, 2006.

- [17] G. Parra, A. Reymond, N. Dabbouseh, E. T. Dermitzakis, R. Castelo, T. M. Thomson, et al. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res*, 16(1):37–44, 2006.
- [18] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragozy, A. Telling, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.
- [19] Bruce Alberts, John H. Wilson, and Tim Hunt. *Molecular biology of the cell*. Garland Science, New York, 5th edition, 2008.
- [20] Harvey F. Lodish. *Molecular cell biology*. W.H. Freeman, New York, 6th edition, 2008.
- [21] T. D. Brock. The bacterial nucleus: a history. *Microbiol Rev*, 52(4):397–411, 1988.
- [22] Wikipedia, the free encyclopedia. celltypes.svg. <http://upload.wikimedia.org/wikipedia/commons/thumb/8/83/Celltypes.svg/2000px-Celltypes.svg.png>. [Online; accessed January-2012].
- [23] A. Martinez-Antonio, A. Medina-Rivera, and J. Collado-Vides. Structural and functional map of a bacterial nucleoid. *Genome Biol*, 10(12):247, 2009.
- [24] M. S. Luijsterburg, M. C. Noom, G. J. Wuite, and R. T. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *J Struct Biol*, 156(2):262–72, 2006.
- [25] K. Drlica and J. Rouviere-Yaniv. Histone-like proteins of bacteria. *Microbiol Rev*, 51(3):301–19, 1987.
- [26] L. D. Murphy and S. B. Zimmerman. Isolation and characterization of spermidine nucleoids from escherichia coli. *J Struct Biol*, 119(3):321–35, 1997.

- [27] A. Martinez-Antonio, A. Medina-Rivera, and J. Collado-Vides. Structural and functional map of a bacterial nucleoid. *Genome Biol*, 10(12):247, 2009.
- [28] C. Tendeng and P. N. Bertin. H-nfs in gram-negative bacteria: a family of multifaceted proteins. *Trends Microbiol*, 11(11):511–8, 2003.
- [29] R. T. Dame, O. J. Kalmykova, and D. C. Grainger. Chromosomal macrodomains and associated proteins: implications for dna organization and replication in gram negative bacteria. *PLoS Genet*, 7(6):e1002123, 2011.
- [30] L. Postow, C. D. Hardy, J. Arsuaga, and N. R. Cozzarelli. Topological domain structure of the escherichia coli chromosome. *Genes Dev*, 18(14):1766–79, 2004.
- [31] A. Worcel and E. Burgi. On the structure of the folded chromosome of escherichia coli. *J Mol Biol*, 71(2):127–47, 1972.
- [32] H. Willenbrock and D. W. Ussery. Chromatin architecture and gene expression in escherichia coli. *Genome Biol*, 5(12):252, 2004.
- [33] Joanne M. Willey, Linda Sherwood, Christopher J. Woolverton, and Lansing M. Prescott. *Prescott's principles of microbiology*. McGraw-Hill Higher Education, Boston, 1st edition, 2009.
- [34] A. G. West, M. Gaszner, and G. Felsenfeld. Insulators: many functions, many mechanisms. *Genes Dev*, 16(3):271–88, 2002.
- [35] B. V. Iyer, M. Kenward, and G. Arya. Hierarchies in eukaryotic genome organization: Insights from polymer theory and simulations. *BMC Biophys*, 4:8, 2011.
- [36] R. D. Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–71, 1974.
- [37] R. D. Kornberg. Structure of chromatin. *Annu Rev Biochem*, 46:931–54, 1977.

- [38] J. T. Finch, L. C. Lutter, D. Rhodes, R. S. Brown, B. Rushton, M. Levitt, et al. Structure of nucleosome core particles of chromatin. *Nature*, 269(5623):29–36, 1977.
- [39] T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, and A. Klug. Structure of the nucleosome core particle at 7 Å resolution. *Nature*, 311(5986):532–7, 1984.
- [40] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.
- [41] J. R. Paulson and U. K. Laemmli. The structure of histone-depleted metaphase chromosomes. *Cell*, 12(3):817–28, 1977.
- [42] K. van Holde and J. Zlatanova. Chromatin higher order structure: chasing a mirage? *J Biol Chem*, 270(15):8373–6, 1995.
- [43] K. Maeshima, S. Hihara, and M. Eltsov. Chromatin structure: does the 30-nm fibre exist in vivo? *Curr Opin Cell Biol*, 22(3):291–7, 2010.
- [44] S. J. McBryant, J. Klonoski, T. C. Sorensen, S. S. Norskog, S. Williams, M. G. Resch, et al. Determinants of histone h4 n-terminal domain function during nucleosomal array oligomerization: roles of amino acid sequence, domain length, and charge density. *J Biol Chem*, 284(25):16716–22, 2009.
- [45] C. David Allis, Thomas Jenuwein, and Danny Reinberg. *Epigenetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 2007.
- [46] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047):138–41, 2005.
- [47] J. Bednar and S. Dimitrov. Chromatin under mechanical stress: from single 30 nm fibers to single nucleosomes. *FEBS J*, 278(13):2231–43, 2011.
- [48] C. L. Woodcock, S. A. Grigoryev, R. A. Horowitz, and N. Whitaker. A chromatin folding model that incorporates linker variability gener-

- ates fibers resembling the native structures. *Proc Natl Acad Sci U S A*, 90(19):9021–5, 1993.
- [49] H. Schiessel, W. M. Gelbart, and R. Bruinsma. Dna folding: structural and mechanical properties of the two-angle model for chromatin. *Biophys J*, 80(4):1940–56, 2001.
- [50] B. Dorigo, T. Schalch, A. Kulangara, S. Duda, R. R. Schroeder, and T. J. Richmond. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, 306(5701):1571–3, 2004.
- [51] F. Thoma, T. Koller, and A. Klug. Involvement of histone h1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J Cell Biol*, 83(2 Pt 1):403–27, 1979.
- [52] J. T. Finch and A. Klug. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A*, 73(6):1897–901, 1976.
- [53] J. Widom, J. T. Finch, and J. O. Thomas. Higher-order structure of long repeat chromatin. *EMBO J*, 4(12):3189–94, 1985.
- [54] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. Manders, P. J. Verschure, et al. Spatially confined folding of chromatin in the interphase nucleus. *Proc Natl Acad Sci U S A*, 106(10):3812–3817, 2009.
- [55] S. Jhunjhunwala, M. C. van Zelm, M. M. Peak, S. Cutchin, R. Riblet, J. J. van Dongen, et al. The 3d structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell*, 133(2):265–79, 2008.
- [56] M. Bohn, D. W. Heermann, and R. van Driel. Random loop model for long polymers. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76(5 Pt 1):051805, 2007.
- [57] M. Bohn and D. W. Heermann. Diffusion-driven looping provides a consistent framework for chromatin organization. *PLoS One*, 5(8):e12218, 2010.

- [58] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2(4):292–301, 2001.
- [59] M. Bohn and D. W. Heermann. Topological interactions between ring polymers: Implications for chromatin loops. *J Chem Phys*, 132(4):044904, 2010.
- [60] D. L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336, 2003.
- [61] Wikipedia, the free encyclopedia. mrna-interaction.png. <http://upload.wikimedia.org/wikipedia/commons/f/fb/MRNA-interaction.png>. [Online; accessed January-2012].
- [62] C. Niehrs and N. Pollet. Synexpression groups in eukaryotes. *Nature*, 402(6761):483–7, 1999.
- [63] M. Kozak. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev*, 47(1):1–45, 1983.
- [64] E. Davidson and M. Levin. Gene regulatory networks. *Proc Natl Acad Sci U S A*, 102(14):4935, 2005.
- [65] R. B. Winter, O. G. Berg, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. the escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–77, 1981.
- [66] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.
- [67] D. A. Jackson, A. B. Hassan, R. J. Errington, and P. R. Cook. Visualization of focal sites of transcription within human nuclei. *EMBO J*, 12(3):1059–65, 1993.
- [68] D. G. Wansink, W. Schul, I. van der Kraan, B. van Steensel, R. van Driel, and L. de Jong. Fluorescent labeling of nascent rna reveals transcription

- by rna polymerase ii in domains scattered throughout the nucleus. *J Cell Biol*, 122(2):283–93, 1993.
- [69] 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 ( Pt 6):1427–36, 1996.
- [70] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet*, 36(10):1065–1071, 2004.
- [71] E. V. Volpi, E. Chevret, T. Jones, R. Vatcheva, J. Williamson, S. Beck, et al. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J Cell Sci*, 113 ( Pt 9):1565–76, 2000.
- [72] S. Chambeyron and W. A. Bickmore. Chromatin decondensation and nuclear reorganization of the hoxb locus upon induction of transcription. *Genes Dev*, 18(10):1119–30, 2004.
- [73] A. E. Wiblin, W. Cui, A. J. Clark, and W. A. Bickmore. Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *J Cell Sci*, 118(Pt 17):3861–8, 2005.
- [74] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, 42(1):53–61, 2010.
- [75] M. Xu and P. R. Cook. Similar active genes cluster in specialized transcription factories. *J Cell Biol*, 181(4):615–23, 2008.
- [76] J. Bartlett, J. Blagojevic, D. Carter, C. Eskiw, M. Fromaget, C. Job, et al. Specialized transcription factories. *Biochem Soc Symp*, (73):67–75, 2006.
- [77] P. J. Verschure, I. van Der Kraan, E. M. Manders, and R. van Driel. Spatial relationship between transcription sites and chromosome territories. *J Cell Biol*, 147(1):13–24, 1999.

- [78] L. Parada and T. Misteli. Chromosome positioning in the interphase nucleus. *Trends Cell Biol*, 12(9):425–32, 2002.
- [79] H. A. Foster and J. M. Bridger. The genome and the nucleus: a marriage made by evolution. genome organisation and nuclear architecture. *Chromosoma*, 114(4):212–29, 2005.
- [80] D. C. Allison and A. L. Nestor. Evidence for a relatively random array of human chromosomes on the mitotic ring. *J Cell Biol*, 145(1):1–14, 1999.
- [81] S. A. Lesko, D. E. Callahan, M. E. LaVilla, Z. P. Wang, and P. O. Ts'o. The experimental homologous and heterologous separation distance histograms for the centromeres of chromosomes 7, 11, and 17 in interphase human t-lymphocytes. *Exp Cell Res*, 219(2):499–506, 1995.
- [82] R. Nagele, T. Freeman, L. McMorrow, and H. Y. Lee. Precise spatial positioning of chromosomes during prometaphase: evidence for chromosomal order. *Science*, 270(5243):1831–5, 1995.
- [83] R. G. Nagele, T. Freeman, J. Fazekas, K. M. Lee, Z. Thomson, and H. Y. Lee. Chromosome spatial order in human cells: evidence for early origin and faithful propagation. *Chromosoma*, 107(5):330–8, 1998.
- [84] R. G. Nagele, T. Freeman, L. McMorrow, Z. Thomson, K. Kitson-Wind, and H. Lee. Chromosomes exhibit preferential positioning in nuclei of quiescent human cells. *J Cell Sci*, 112 ( Pt 4):525–35, 1999.
- [85] H. B. Sun, J. Shen, and H. Yokota. Size-dependent positioning of human chromosomes in interphase nuclei. *Biophys J*, 79(1):184–90, 2000.
- [86] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol*, 145(6):1119–31, 1999.
- [87] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet*, 10(3):211–9, 2001.

- [88] A. C. Chandley, R. M. Speed, and A. R. Leitch. Different distributions of homologous chromosomes in adult human sertoli cells and in lymphocytes signify nuclear differentiation. *J Cell Sci*, 109 ( Pt 4):773–6, 1996.
- [89] A. R. Leitch, J. K. Brown, W. Mosgoller, T. Schwarzacher, and J. S. Heslop-Harrison. The spatial localization of homologous chromosomes in human fibroblasts at mitosis. *Hum Genet*, 93(3):275–80, 1994.
- [90] I. Alcobia, R. Dilao, and L. Parreira. Spatial associations of centromeres in the nuclei of hematopoietic cells: evidence for cell-type-specific organizational patterns. *Blood*, 95(5):1608–15, 2000.
- [91] M. Ferguson and D. C. Ward. Cell cycle dependent chromosomal movement in pre-mitotic human t-lymphocyte nuclei. *Chromosoma*, 101(9):557–65, 1992.
- [92] C. Vourc’h, D. Taruscio, A. L. Boyle, and D. C. Ward. Cell cycle-dependent distribution of telomeres, centromeres, and chromosome-specific subsatellite domains in the interphase nucleus of mouse lymphocytes. *Exp Cell Res*, 205(1):142–51, 1993.
- [93] S. Goetze, J. Mateos-Langerak, H. J. Gierman, W. de Leeuw, O. Giromus, M. H. Indemans, et al. The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Mol Cell Biol*, 27(12):4475–87, 2007.
- [94] J. R. Chubb, S. Boyle, P. Perry, and W. A. Bickmore. Chromatin motion is constrained by association with nuclear compartments in human cells. *Curr Biol*, 12(6):439–45, 2002.
- [95] P. Heun, T. Laroche, K. Shimada, P. Furrer, and S. M. Gasser. Chromosome dynamics in the yeast interphase nucleus. *Science*, 294(5549):2181–6, 2001.
- [96] A. Taddei, C. Maison, D. Roche, and G. Almouzni. Reversible disruption of pericentric heterochromatin and centromere function by inhibiting deacetylases. *Nat Cell Biol*, 3(2):114–20, 2001.

- [97] C. S. Osborne, L. Chakalova, J. A. Mitchell, A. Horton, A. L. Wood, D. J. Bolland, et al. Myc dynamically and preferentially relocates to a transcription factory occupied by igh. *PLoS Biol*, 5(8):e192, 2007.
- [98] D. A. Kleinjan and V. van Heyningen. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet*, 76(1):8–32, 2005.
- [99] A. G. West and P. Fraser. Remote control of gene transcription. *Hum Mol Genet*, 14 Spec No 1:R101–11, 2005.
- [100] M. Bulger and M. Groudine. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev*, 13(19):2465–77, 1999.
- [101] J. D. Engel and K. Tanimoto. Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. *Cell*, 100(5):499–502, 2000.
- [102] G. Stamatoyannopoulos. Control of globin gene expression during development and erythroid differentiation. *Exp Hematol*, 33(3):259–71, 2005.
- [103] D. R. Higgs, D. Garrick, E. Anguita, M. De Gobbi, J. Hughes, M. Muers, et al. Understanding alpha-globin gene regulation: Aiming to improve the management of thalassemia. *Ann N Y Acad Sci*, 1054:92–102, 2005.
- [104] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6):1453–65, 2002.
- [105] D. Vernimmen, M. De Gobbi, J. A. Sloane-Stanley, W. G. Wood, and D. R. Higgs. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J*, 26(8):2041–51, 2007.
- [106] E. Anguita, J. Hughes, C. Heyworth, G. A. Blobel, W. G. Wood, and D. R. Higgs. Globin gene activation during haemopoiesis is driven by protein complexes nucleated by gata-1 and gata-2. *EMBO J*, 23(14):2841–52, 2004.

- [107] G. L. Zhou, L. Xin, W. Song, L. J. Di, G. Liu, X. S. Wu, et al. Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes. *Mol Cell Biol*, 26(13):5096–105, 2006.
- [108] I. W. Duncan. Transvection effects in drosophila. *Annu Rev Genet*, 36:521–56, 2002.
- [109] N. Xu, C. L. Tsai, and J. T. Lee. Transient homologous chromosome pairing marks the onset of x inactivation. *Science*, 311(5764):1149–52, 2006.
- [110] C. P. Bacher, M. Guggiari, B. Brors, S. Augui, P. Clerc, P. Avner, et al. Transient colocalization of x-inactivation centres accompanies the initiation of x inactivation. *Nat Cell Biol*, 8(3):293–9, 2006.
- [111] C. G. Spilianakis and R. A. Flavell. Long-range intrachromosomal interactions in the t helper type 2 cytokine locus. *Nat Immunol*, 5(10):1017–27, 2004.
- [112] S. Lomvardas, G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland, and R. Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–13, 2006.
- [113] K. Bystricky, T. Laroche, G. van Houwe, M. Blaszczyk, and S. M. Gasser. Chromosome looping in yeast: telomere pairing and coordinated movement reflect anchoring efficiency and territorial organization. *J Cell Biol*, 168(3):375–87, 2005.
- [114] C. Lanzuolo, V. Roure, J. Dekker, F. Bantignies, and V. Orlando. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat Cell Biol*, 9(10):1167–74, 2007.
- [115] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–11, 2002.

- [116] C. D. 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):879–86, 2009.
- [117] S. Horike, S. Cai, M. Miyano, J. F. Cheng, and T. Kohwi-Shigematsu. Loss of silent-chromatin looping and impaired imprinting of *dlx5* in rett syndrome. *Nat Genet*, 37(1):31–40, 2005.
- [118] S. Cai, C. C. Lee, and T. Kohwi-Shigematsu. *Satb1* packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet*, 38(11):1278–88, 2006.
- [119] P. P. Kumar, O. Bischof, P. K. Purbey, D. Notani, H. Urlaub, A. Dejean, and S. Galande. Functional interaction between *pml* and *satb1* regulates chromatin-loop architecture and transcription of the *mhc class i* locus. *Nat Cell Biol*, 9(1):45–56, 2007.
- [120] Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, et al. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341–7, 2006.
- [121] H. Wurtele and P. Chartrand. Genome-wide scanning of *hoxb1*-associated loci in mouse es cells using an open-ended chromosome conformation capture methodology. *Chromosome Res*, 14(5):477–95, 2006.
- [122] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299–309, 2006.
- [123] V. K. Tiwari, L. Cope, K. M. McGarvey, J. E. Ohm, and S. B. Baylin. A novel 6c assay uncovers polycomb-mediated higher order chromatin conformations. *Genome Res*, 18(7):1171–9, 2008.
- [124] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, et al. An oestrogen-receptor- $\alpha$ -bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

- [125] Wikipedia, the free encyclopedia. chromosome conformation capture technology.jpg. [http://upload.wikimedia.org/wikipedia/commons/d/d2/Chromosome\\_Conformation\\_Capture\\_Technology.jpg](http://upload.wikimedia.org/wikipedia/commons/d/d2/Chromosome_Conformation_Capture_Technology.jpg). [Online; accessed January-2012].
- [126] J. D. Rowley. Chromosome translocations: dangerous liaisons revisited. *Nat Rev Cancer*, 1(3):245–50, 2001.
- [127] T. H. Rabbitts. Chromosomal translocations in human cancer. *Nature*, 372(6502):143–9, 1994.
- [128] F. Mitelman, B. Johansson, and F. Mertens. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet*, 36(4):331–4, 2004.
- [129] W. A. Bickmore and P. Teague. Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosome Res*, 10(8):707–15, 2002.
- [130] A. E. Murmann, J. Gao, M. Encinosa, M. Gautier, M. E. Peter, R. Eils, et al. Local gene density predicts the spatial position of genetic loci in the interphase nucleus. *Exp Cell Res*, 311(1):14–26, 2005.
- [131] M. Kuroda, H. Tanabe, K. Yoshida, K. Oikawa, A. Saito, T. Kiyuna, et al. Alteration of chromosome positioning during adipocyte differentiation. *J Cell Sci*, 117(Pt 24):5897–903, 2004.
- [132] J. J. Roix, P. G. McQueen, P. J. Munson, L. A. Parada, and T. Misteli. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet*, 34(3):287–91, 2003.
- [133] M. R. Branco and A. Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol*, 4(5):e138, 2006.
- [134] L. A. Parada, P. G. McQueen, P. J. Munson, and T. Misteli. Conservation of relative chromosome positioning in normal and cancer cells. *Curr Biol*, 12(19):1692–7, 2002.

- [135] M. Cremer, K. Kupper, B. Wagler, L. Wizelman, J. von Hase, Y. Weiland, et al. Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. *J Cell Biol*, 162(5):809–20, 2003.
- [136] Pierre Gilles de Gennes. *Scaling concepts in polymer physics*. Cornell University Press, Ithaca, N.Y., 1979.
- [137] Lothar Schäfer. *Excluded volume effects in polymer solutions, as explained by the renormalization group*. Springer, Berlin; New York, 1999.
- [138] Paul J. Flory. *Principles of polymer chemistry*. Cornell University Press, Ithaca,, 1953.
- [139] A. I. U. Grosberg and A. R. Khokhlov. *Statistical physics of macromolecules*. AIP Press, New York, 1994.
- [140] Michael Rubinstein and Ralph H. Colby. *Polymer physics*. Oxford University Press, Oxford ; New York, 2003.
- [141] Paul J. Flory. *Statistical mechanics of chain molecules*. Hanser Publishers ; Distributed in the U.S.A. by Oxford University Press, Munich ; New York New York, repr. edition, 1989.
- [142] Neal Noah Madras and Gordon Douglas Slade. *The self-avoiding walk*. Probability and its applications. Birkhäuser, Boston, 1993.
- [143] R. Lua, A. L. Borovinskiy, and A. Y. Grosberg. Fractal and statistical properties of large compact polymers: a computational study. *Polymer*, 45(2):717 – 731, 2004.
- [144] P. Virnau, Y. Kantor, and M. Kardar. Knots in globule and coil phases of a model polyethylene. *J Am Chem Soc*, 127(43):15102–6, 2005.
- [145] R. Metzler, A. Hanke, P. G. Dommersnes, Y. Kantor, and M. Kardar. Equilibrium shapes of flat knots. *Phys Rev Lett*, 88(18):188101, 2002.
- [146] A. Y. Grosberg. Critical exponents for random knots. *Phys Rev Lett*, 85(18):3858–61, 2000.

- [147] A. Y. Grosberg, S. K. Nechaev, and E. I. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal De Physique*, 49(12):2095–2100, 1988.
- [148] L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res*, 19(1):37–51, 2011.
- [149] I. Junier, O. Martin, and F. Kepes. Spatial and topological organization of dna chains induced by gene co-localization. *PLoS Comput Biol*, 6(2):e1000678, 2010.
- [150] J. Dorier and A. Stasiak. The role of transcription factories-mediated interchromosomal contacts in the organization of nuclear architecture. *Nucleic Acids Research*, 38(21):7410–21, 2010.
- [151] S. de Nooijer, J. Wellink, B. Mulder, and T. Bisseling. Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei. *Nucleic Acids Research*, 37(11):3558–68, 2009.
- [152] T. Cremer, C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders. Analysis of chromosome positions in the interphase nucleus of chinese hamster cells by laser-uv-microirradiation experiments. *Hum Genet*, 62(3):201–9, 1982.
- [153] K. E. Handwerger and J. G. Gall. Subnuclear organelles: new insights into form and function. *Trends Cell Biol*, 16(1):19–26, 2006.
- [154] M. G. Mayer and L. M. Floeter-Winter. Pre-mrna trans-splicing: from kinetoplastids to mammals, an easy language for life diversity. *Mem Inst Oswaldo Cruz*, 100(5):501–13, 2005.
- [155] R. Sorek and H. M. Safer. A novel algorithm for computational identification of contaminated est libraries. *Nucleic Acids Res*, 31(3):1067–74, 2003.
- [156] B. Lee and G. Shin. Cleanest: a database of cleansed est libraries. *Nucleic Acids Res*, 37(Database issue):D686–9, 2009.

- [157] R. E. Sutton and J. C. Boothroyd. Evidence for trans splicing in trypanosomes. *Cell*, 47(4):527–535, 1986.
- [158] K. E. Hastings. Sl trans-splicing: easy come or easy go? *Trends Genet*, 21(4):240–247, 2005.
- [159] T. Horiuchi and T. Aigaki. Alternative trans-splicing: a novel mode of pre-mrna processing. *Biol Cell*, 98(2):135–140, 2006.
- [160] M. Krause and D. Hirsh. A trans-spliced leader sequence on actin mrna in *c. elegans*. *Cell*, 49(6):753–761, 1987.
- [161] A. Rajkovic, R. E. Davis, J. N. Simonsen, and F. M. Rottman. A spliced leader is present on a subset of mrnas from the human parasite schistosoma mansoni. *Proc Natl Acad Sci U S A*, 87(22):8879–8883, 1990.
- [162] L. H. Tessier, M. Keller, R. L. Chan, R. Fournier, J. H. Weil, and P. Imbault. Short leader sequences may be transferred from small rnas to pre-mature mrnas by trans-splicing in euglena. *EMBO J*, 10(9):2621–2625, 1991.
- [163] R. E. Davis, H. Singh, C. Botka, C. Hardwick, M. Ashraf, el Meanawy, and J. Villanueva. Rna trans-splicing in fasciola hepatica. identification of a spliced leader (sl) rna and sl sequences on mrnas. *J Biol Chem*, 269(31):20026–20030, 1994.
- [164] Y. Chapdelaine and L. Bonen. The wheat mitochondrial gene for subunit i of the nadh dehydrogenase complex: a trans-splicing model for this gene-in-pieces. *Cell*, 65(3):465–472, 1991.
- [165] C. Caudevilla, D. Serra, A. Miliar, C. Codony, G. Asins, M. Bach, and F. G. Hegardt. Natural trans-splicing in carnitine octanoyltransferase pre-mrnas in rat liver. *Proc Natl Acad Sci U S A*, 95(21):12185–12190, 1998.
- [166] C. Finta and P. G. Zaphiropoulos. Intergenic mrna molecules resulting from trans-splicing. *J Biol Chem*, 277(8):5882–5890, 2002.

- [167] F. Mongelard, M. Labrador, E. M. Baxter, T. I. Gerasimova, and V. G. Corces. Trans-splicing as a novel mechanism to explain interallelic complementation in drosophila. *Genetics*, 160(4):1481–1487, 2002.
- [168] T. Horiuchi, E. Giniger, and T. Aigaki. Alternative trans-splicing of constant and variable exons of a drosophila axon guidance gene, *lola*. *Genes Dev*, 17(20):2496–2501, 2003.
- [169] C. Zhang, Y. Xie, J. A. Martignetti, T. T. Yeo, S. M. Massa, and F. M. Longo. A candidate chimeric mammalian mrna transcript is derived from distinct chromosomes and is associated with nonconsensus splice junction motifs. *DNA Cell Biol*, 22(5):303–315, 2003.
- [170] X. Li, L. Zhao, H. Jiang, and W. Wang. Short homologous sequences are strongly associated with the generation of chimeric rnas in eukaryotes. *J Mol Evol*, 68(1):56–65, 2009.
- [171] H. Li, J. Wang, G. Mor, and J. Sklar. A neoplastic gene fusion mimics trans-splicing of rnas in normal human cells. *Science*, 321(5894):1357–1361, 2008.
- [172] D. S. Rickman, D. Pflueger, B. Moss, V. E. VanDoren, C. X. Chen, A. de, la Taille, et al. Slc45a3-*elk4* is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res*, 69(7):2734–2738, 2009.
- [173] S. Janz, M. Potter, and C. S. Rabkin. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. *Genes Chromosomes Cancer*, 36(3):211–223, 2003.
- [174] M. Patel, J. M. Simon, M. D. Iglesia, S. B. Wu, A. W. McFadden, J. D. Lieb, et al. Tumor-specific retargeting of an oncogenic transcription factor chimera results in dysregulation of chromatin and transcription. *Genome Res*, Epub ahead of print, 2011.
- [175] M. N. Nikiforova, J. R. Stringer, R. Blough, M. Medvedovic, J. A. Fagin, and Y. E. Nikiforov. Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. *Science*, 290(5489):138–41, 2000.

- [176] L. Parada and T. Misteli. Chromosome positioning in the interphase nucleus. *Trends Cell Biol*, 12(9):425–32, 2002.
- [177] J. R. Savage. Cancer. proximity matters. *Science*, 290(5489):62–3, 2000.
- [178] J. R. Savage. Interchange and intra-nuclear architecture. *Environ Mol Mutagen*, 22(4):234–44, 1993.
- [179] J. E. Haber and W. Y. Leung. Lack of chromosome territoriality in yeast: promiscuous rejoining of broken chromosome ends. *Proc Natl Acad Sci U S A*, 93(24):13949–54, 1996.
- [180] D. R. Carter, C. Eskiw, and P. R. Cook. Transcription factories. *Biochem Soc Trans*, 36(Pt 4):585–589, 2008.
- [181] J. Q. Ling, T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, et al. Ctfc mediates interchromosomal colocalization between *igf2/h19* and *wsb1/nf1*. *Science*, 312(5771):269–272, 2006.
- [182] C. H. Chuang and A. S. Belmont. Close encounters between active genes in the nucleus. *Genome Biol*, 6(11):237, 2005.
- [183] P. Fraser. Transcriptional control thrown for a loop. *Curr Opin Genet Dev*, 16(5):490–495, 2006.
- [184] J. Dostie, Y. Zhan, and J. Dekker. Chromosome conformation capture carbon copy technology. *Curr Protoc Mol Biol*, Chapter 21:Unit 21 14, 2007.
- [185] P. Unneberg and J. M. Claverie. Tentative mapping of transcription-induced interchromosomal interaction using chimeric est and mrna data. *PLoS One*, 2(2):e254, 2007.
- [186] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–65, 2011.
- [187] The genbank database. <ftp://ftp.ncbi.nih.gov/genbank/>.
- [188] The est library report. <ftp://ftp.ncbi.nlm.nih.gov/repository/UniLib/library.report>.

- [189] The est library annotation. <ftp://ftp.ncbi.nih.gov/repository/UniGene/>.
- [190] The human genome reference. [ftp://ftp.ncbi.nih.gov/genomes/H\\_sapiens/Assembled\\_chromosomes/](ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes/).
- [191] The mouse genome reference. [ftp://ftp.ncbi.nih.gov/genomes/M\\_musculus/Assembled\\_chromosomes/](ftp://ftp.ncbi.nih.gov/genomes/M_musculus/Assembled_chromosomes/).
- [192] The ensembl human genome database. [ftp://ftp.ensembl.org/pub/release-64/mysql/homo\\_sapiens\\_core\\_64\\_37/](ftp://ftp.ensembl.org/pub/release-64/mysql/homo_sapiens_core_64_37/).
- [193] The ensembl mouse genome database. [ftp://ftp.ensembl.org/pub/release-64/mysql/mus\\_musculus\\_core\\_64\\_37/](ftp://ftp.ensembl.org/pub/release-64/mysql/mus_musculus_core_64_37/).
- [194] The ensembl ontology. [ftp://ftp.ensembl.org/pub/release-64/mysql/ensembl\\_ontology\\_64/](ftp://ftp.ensembl.org/pub/release-64/mysql/ensembl_ontology_64/).
- [195] The rebase database. <ftp://ftp.neb.com/pub/rebase/allenz.102>.
- [196] W. J. Kent. Blat—the blast-like alignment tool. *Genome Res*, 12(4):656–664, 2002.
- [197] E. Bon, S. Casaregola, G. Blandin, B. Llorente, C. Neuveglise, M. Munsterkotter, et al. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Research*, 31(4):1121–35, 2003.
- [198] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita. Coxpresdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*, 36(Database issue):D77–82, 2008.
- [199] The coxpresdb database. <http://coxpresdb.jp/download.shtml>.
- [200] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–81, 2007.

- [201] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, 2008.
- [202] Normalized hi-c contact maps from human lymphoblast. [http://compgenomics.weizmann.ac.il/tanay/?page\\_id=283](http://compgenomics.weizmann.ac.il/tanay/?page_id=283).
- [203] M. J. Zeitz, L. Mukherjee, S. Bhattacharya, J. Xu, and R. Berezney. A probabilistic model for the arrangement of a subset of human chromosome territories in wi38 human fibroblasts. *J Cell Physiol*, 221(1):120–129, 2009.
- [204] N. V. Marella, S. Bhattacharya, L. Mukherjee, J. Xu, and R. Berezney. Cell type specific chromosome territory organization in the interphase nucleus of normal and cancer cells. *J Cell Physiol*, 221(1):130–138, 2009.
- [205] F. R. Blattner, 3rd Plunkett, G., C. A. Bloch, N. T. Perna, V. Burland, M. Riley, et al. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–62, 1997.
- [206] V. F. Holmes and N. R. Cozzarelli. Closing the ring: links between smc proteins and chromosome partitioning, condensation, and supercoiling. *Proc Natl Acad Sci U S A*, 97(4):1322–4, 2000.
- [207] R. R. Sinden and D. E. Pettijohn. Chromosomes in living escherichia coli cells are segregated into domains of supercoiling. *Proc Natl Acad Sci U S A*, 78(1):224–8, 1981.
- [208] K. S. Jeong, J. Ahn, and A. B. Khodursky. Spatial patterns of transcriptional activity in the chromosome of escherichia coli. *Genome Biol*, 5(11):R86, 2004.
- [209] D. Marenduzzo, C. Micheletti, and P. R. Cook. Entropy-driven genome organization. *Biophys J*, 90(10):3712–21, 2006.
- [210] T. Vora, A. K. Hottes, and S. Tavazoie. Protein occupancy landscape of a bacterial genome. *Mol Cell*, 35(2):247–53, 2009.

- [211] G. R. Bowman, L. R. Comolli, J. Zhu, M. Eckart, M. Koenig, K. H. Downing, et al. A polymeric protein anchors the chromosomal origin/parb complex at a bacterial cell pole. *Cell*, 134(6):945–55, 2008.
- [212] G. Ebersbach, A. Briegel, G. J. Jensen, and C. Jacobs-Wagner. A self-associating protein critical for chromosome attachment, division, and polar organization in caulobacter. *Cell*, 134(6):956–68, 2008.
- [213] C. Sousa, V. de Lorenzo, and A. Cebolla. Modulation of gene expression through chromosomal positioning in escherichia coli. *Microbiology*, 143 ( Pt 6):2071–8, 1997.
- [214] B. J. Peter, J. Arsuaga, A. M. Breier, A. B. Khodursky, P. O. Brown, and N. R. Cozzarelli. Genomic transcriptional response to loss of chromosomal supercoiling in escherichia coli. *Genome Biol*, 5(11):R87, 2004.
- [215] D. Bates and N. Kleckner. Chromosome and replisome dynamics in e. coli: loss of sister cohesion triggers global chromosome movement and mediates chromosome segregation. *Cell*, 121(6):899–911, 2005.
- [216] H. Niki, Y. Yamaichi, and S. Hiraga. Dynamic organization of chromosomal dna in escherichia coli. *Genes Dev*, 14(2):212–23, 2000.
- [217] X. Wang, C. Possoz, and D. J. Sherratt. Dancing around the divi-some: asymmetric chromosome segregation in escherichia coli. *Genes Dev*, 19(19):2367–77, 2005.
- [218] X. Wang, X. Liu, C. Possoz, and D. J. Sherratt. The two escherichia coli chromosome arms locate to separate cell halves. *Genes Dev*, 20(13):1727–31, 2006.
- [219] K. E. Rudd. Ecogene: a genome sequence database for escherichia coli k-12. *Nucleic Acids Research*, 28(1):60–4, 2000.
- [220] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, et al. Regulondb version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (gensor units). *Nucleic Acids Research*, 39(Database issue):D98–105, 2011.

- [221] S. Jun and B. Mulder. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. *Proc Natl Acad Sci U S A*, 103(33):12388–93, 2006.
- [222] S. Jun and A. Wright. Entropy as the driver of chromosome segregation. *Nat Rev Microbiol*, 8(8):600–7, 2010.
- [223] H. J. Limbach, A. Arnold, B. A. Mann, and C. Holm. Espresso - an extensible simulation package for research on soft matter systems. *Computer Physics Communications*, 174(9):704–727, 2006.
- [224] F. Hommais, E. Krin, C. Laurent-Winter, O. Soutourina, A. Malpertuy, J. P. Le Caer, et al. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, h-ns. *Mol Microbiol*, 40(1):20–36, 2001.
- [225] E. P. Rocha. The organization of the bacterial genome. *Annu Rev Genet*, 42:211–33, 2008.
- [226] A. S. Carpentier, B. Torresani, A. Grossmann, and A. Henaut. Decoding the nucleoid organisation of bacillus subtilis and escherichia coli through gene expression data. *BMC Genomics*, 6:84, 2005.
- [227] I. A. Berlatzky, A. Rouvinski, and S. Ben-Yehuda. Spatial organization of a replicating bacterial chromosome. *Proc Natl Acad Sci U S A*, 105(37):14136–40, 2008.
- [228] X. Dong, C. Li, Y. Chen, G. Ding, and Y. Li. Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC Genomics*, 11:704, 2010.
- [229] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, 360(1):213–27, 2006.
- [230] S. de Nooijer, J. Wellink, B. Mulder, and T. Bisseling. Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei. *Nucleic Acids Research*, 37(11):3558–68, 2009.

- [231] M. Bohn and D. W. Heermann. Topological interactions between ring polymers: Implications for chromatin loops. *J Chem Phys*, 132(4):044904, 2010.
- [232] H. J. Gierman, M. H. Indemans, J. Koster, S. Goetze, J. Seppen, D. Geerts, et al. Domain-wide regulation of gene expression in the human genome. *Genome Res*, 17(9):1286–95, 2007.
- [233] J. Venema and D. Tollervey. Ribosome synthesis in *saccharomyces cerevisiae*. *Annu Rev Genet*, 33:261–311, 1999.
- [234] M. Fritsche, S. Li, D. W. Heermann, and P. A. Wiggins. A model for *escherichia coli* chromosome packaging supports transcription factor-induced dna domain formation. *Nucleic Acids Research*, 2011.
- [235] N. N. Batada, A. O. Urrutia, and L. D. Hurst. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet*, 23(10):480–4, 2007.
- [236] H. Schober, V. Kalck, M. A. Vega-Palas, G. Van Houwe, D. Sage, M. Unser, et al. Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. *Genome Res*, 18(2):261–71, 2008.
- [237] M. Thompson, R. A. Haeusler, P. D. Good, and D. R. Engelke. Nucleolar clustering of dispersed trna genes. *Science*, 302(5649):1399–401, 2003.
- [238] R. A. Haeusler, M. Pratt-Hyatt, P. D. Good, T. A. Gipson, and D. R. Engelke. Clustering of yeast trna genes is mediated by specific association of condensin with trna gene transcription complexes. *Genes Dev*, 22(16):2204–14, 2008.
- [239] S. Ide, T. Miyazaki, H. Maki, and T. Kobayashi. Abundance of ribosomal rna gene copies maintains genome integrity. *Science*, 327(5966):693–6, 2010.
- [240] The ensembl budding yeast genome database. [ftp://ftp.ensembl.org/pub/release-58/mysql/saccharomyces\\_cerevisiae\\_core\\_58\\_1j/](ftp://ftp.ensembl.org/pub/release-58/mysql/saccharomyces_cerevisiae_core_58_1j/).

- [241] A. B. Berger, G. G. Cabal, E. Fabre, T. Duong, H. Buc, U. Nehrbass, et al. High-resolution statistical mapping reveals gene territories in live yeast. *Nat Methods*, 5(12):1031–7, 2008.
- [242] G. Rustici, J. Mata, K. Kivinen, P. Lio, C. J. Penkett, G. Burns, et al. Periodic gene expression program of the fission yeast cell cycle. *Nat Genet*, 36(8):809–17, 2004.
- [243] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708, 2001.
- [244] J. A. Sawitzke and S. Austin. Suppression of chromosome segregation defects of escherichia coli muk mutants by mutations in topoisomerase i. *Proc Natl Acad Sci U S A*, 97(4):1671–6, 2000.
- [245] X. S. Xie, P. J. Choi, G. W. Li, N. K. Lee, and G. Lia. Single-molecule approach to molecular biology in living bacterial cells. *Annu Rev Biophys*, 37:417–44, 2008.
- [246] Z. Gitai. New fluorescence microscopy methods for microbiology: sharper, faster, and quantitative. *Curr Opin Microbiol*, 12(3):341–6, 2009.
- [247] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.
- [248] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.