ABSTRACT

Title of dissertation: THEORETICAL AND COMPUTATIONAL STUDIES OF HUMAN INTERPHASE CHROMOSOMES

Guang Shi
Doctor of Philosophy, 2019

Dissertation directed by: Professor Devarajan Thirumalai
Biophysics Program,
Institute for Physical Science and Technology

In this thesis, various aspects of dynamical and structural properties of human interphase chromosomes are studied using both theoretical and computational tools. In addition, the cooperative transport by the multi-motor system was investigated using a stochastic kinetic model.

First, I create the Chromosome Copolymer Model (CCM) by representing chromosomes as a copolymer. I first showed that the model is consistent with current experimental data. Using the CCM, I further investigated the dynamics of human interphase chromosomes. The model suggested that human interphase chromosome exhibit glassy-like dynamics characterized by sluggish movement, large loci-to-loci variations, and dynamical heterogeneity.

Furthermore, I predicted that human interphase chromosomes also display extensive structural heterogeneity. Using a theoretical framework I developed based on polymer physics, I am able to identify that the existence of subpopulations is the

reason for the Hi-C-FISH paradox. As an application of the theory, the information of subpopulations of cells can be readily extracted from experimental FISH data. The results suggest that heterogeneity is pervasive in genome organization at all length scales, reflecting large cell-to-cell variations.

Then I proceed to develop a method to reconstruct the three-dimensional genome structure directly from Hi-C data. By applying the theory combined with various manifold embedding methods to experimental Hi-C data, I am able to visualize the averaged global 3D organization of a single chromosome and also local structures such as Topological Associated Domains. The method provides a fast and simple way to help experimentalist visualize the genome organization from the measured Hi-C data.

Finally, I propose a kinetic model for the multi-motor system. I investigate the effect of mechanical coupling between multiple motors on their velocity and force-velocity behavior. Reduction of velocity is observed for coupled motor system especially when the coupling strength is strong. The model also shows that the multi-motors system is more efficient for transporting large cargo but is less efficient for transporting small cargo compared to a single motor.

THEORETICAL AND COMPUTATIONAL STUDIES OF HUMAN INTERPHASE CHROMOSOMES

by

Guang Shi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Devarajan Thirumalai, Chair/Advisor
Professor John Weeks, Co-chair
Professor Christopher Jarzynski
Professor Pratyush Tiwary
Professor Silvina Matysiak
Professor Jeffery Klauda, Dean's Representative

Dedication

To my wife,Jiameng Zheng, and my mother, Wen-ying Xie.

Acknowledgments

It has been a rocky and incredible journey for me since day one I arrived in the United States. For the past almost seven years, there are ups and downs. I appreciate all the things that happened to me as gifts and lessons. From all the wonderful people I have met, I would like to acknowledge people who have helped me to complete the doctoral degree here.

First, I want to express my gratitude to my advisor, Dr. Devarajan Thirumalali, for his support to me for the past five years. It is a privilege for me to work with him and be advised by him. I can't describe how much I have benefited through many insightful and intelligent discussions with him. His high standard for conducting good research has taught me how to do research and more importantly how to not to do research. Certainly, there were struggles due to his tough questions. Looking back, those experience have helped me become a better scientist and will continue to benefit me through my future career. Asides from research, I'm also deeply grateful for his support when I went through difficult times in my life.

I would also like to thank my colleagues in the group. I had many valuable discussions with Dr. Xin Li on both life and sciences. Through many insightful discussions with Dr. Mauro Mugnai, I also learned a lot in statistical mechanics, polymer physics, motors, etc. Dr. Marina Katava is a wonderful person to work with and I have been enjoying the collaboration with her. There are many other people I want to thank, Dr. Abdul Naseer, Dr. Yonathan Cwik, Sumit Sinha, and all the other current and former group members for giving me suggestions on my

research projects.

This journey would not have been possible without the support of my family. Thank to my cousins, aunts, uncles, and grandparents for their warm welcome whenever I visited China. To my father and mother, thank you for your unconditional love and encouragement in all of my pursuits. Especially, I want to thank my mother, Wen-ying Xie, for guiding me as a person, for teaching me to be loving and generous, and for supporting my passions in science.

At last, my deepest gratitude goes to my wife, Jiameng Zheng, who is always there for me. One thing I can be sure of is that without her relentless dedication, I would not finish my Ph.D.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3C	Chromosome Conformation Capture
CT	Chromosome Territory
FISH	Fluorescent In Situ Hybridization
TAD	Topological Associating Domain
SMC	Structural Maintenance Complex
CCM	Chromosome Copolymer Model

Chapter 1: Introduction

1.1 Genome organization in a nutshell

The history of the study of chromosomes begun with its discovery in the mid-to-late 19th century by numerous scientists [1]. Under the microscope, rod-shaped structures were identified during cell division. The observed structures, given the name "chromosomes", are correctly recognized as essential components of heredity, long before the structure of the basic unit, DNA, was discovered. Nowadays, it is well known that chromosomes are complex molecules formed by DNA and proteins, which adopt a variety of structures during different stages of the cell cycle. Although our understandings of chromosomes (sometimes called chromatin if it is referred as the chromosomes during the interphase stage) has advanced considerably since the 19th century [2, 3]. However, much remains unknown regarding their structures, dynamics, and biological functions.

We now know that the DNA in a mammalian cell is wrapped around the 10-nm sized nucleosomes and packaged in the micron-sized cell nucleus. At the length scale of about 10 nm, whether the chromatin fiber forms an ordered or disordered structure has been in debate for several decades. The main feature of chromatin on this length scale, which has been extensively studied using the bead-on-string model,

views it as nucleosomes that are regularly spaced along the DNA connected by linker DNA of 50 bps length. In the 1970s, the first transmission electron microscopy image of chromatin fiber showed that, *in vitro* under certain salt concentration, it adopts disordered 10-nm fiber conformation [4]. Later on, the 10-nm fiber was found to have the capability to fold into a higher-ordered structure with a diameter of roughly 30 nm in the presence of linker histone H1 or Mg^{2+} ions [4–6]. Several models have been proposed to explain the observation of 30-nm fiber [7,8]. However, it has been consistently debated (see [9] for detailed reviews) whether chromatin fiber adopts the 30-nm ordered structure *in vivo*. Cryo-EM and its variations were used to visualize the interphase chromosomes and no ordered structure was observed [10]. In a recent experiment [11], using a novel fluorescent dye, the authors overcome the difficulty of efficiently marking the interphase chromatin and found no evidence of ordered package of nucleosomes of any kind. Instead, it seems that the nucleosomes are dispersed randomly in the cell nucleus with large density fluctuations [11]. It was suggested that 30-nm chromatin fiber may be an artifact that might exist only under certain *in vitro* salt condition [9] and is absent *in vivo*.

On the other hand, the folding of the chromatin fiber on the length scales between several kilobase pairs (kbps) and hundreds of millions of base pairs has drawn increasing attention in the last decade, owing to the advances in the experimental techniques. These new instrumentations, such as the new imaging techniques with high throughput [12–14] and spatiotemporal resolution [15–17] and Chromosome Conformation Capture (3C) - based techniques [18–21], have provided many insights into our understanding of chromosome organization on both small and large

length scales (see [22], [23], and [24] for thorough reviews). Through the remarkable Hi-C experiments [21, 22, 25–28], glimpses of how the genome is organized in a number of species start to emerge. The power of Hi-C lies in its ability to detect the pairwise contacts between loci throughout the whole genome at a resolution as high as kbps. In spite of the loss of information by projecting 3D structure into a two-dimensional representation, Hi-C experiments, for the first time, provide a bird's eye view of the genome architecture, which lead to several significant findings such as CTCF loops, Topologically Associating Domains (TADs) and compartments (discussed below).

A complementary and more direct way to determine genome organization is to assess the spatial coordinates of the chromosome loci. The Fluorescence *In Situ* Hybridization (FISH) technique, although suffers from many limitations, can be used to visualize individual loci by labeling specific DNA sequences. By painting the whole chromosomes using multiple probes along the DNA chains, the global view of the distribution of chromosomes inside the cell nucleus is achieved [2]. More recently, a combination of super-resolution imaging with multiplex FISH allowed direct visualization of targeted chromatin segments [14, 29–31]. However, FISH-based methods cannot be used to probe dynamic information since it is performed on fixed cells. Instead, single-molecule tracking [17, 32] is used to assay the dynamic behavior of loci in real time, which revealed that the dynamics of chromatin loci can be largely described as subdiffusive [17, 33–37] with large heterogeneity [17, 34, 35, 38, 39], and can also be protein dependent [40]. In addition, the results from more recent experiments [41–44], in which the bulk dynamics across entire nucleus were

probed, suggested that the chromatin loci move coordinately over the length scale of hundreds of nanometers and the time scale of seconds. However, no experimental method is currently available for monitoring the dynamics of a large number of loci simultaneously with their genomic identification.

Physicists and chemists have a rich history of using ideas rooted in physics to understand various biological systems, and the chromosomes are no exception. It is well known that the properties of a polymer on a long length scale, measured in terms of the basic building block, do not depend on the chemical details [45]. Thus, it is natural to tackle the problems related to the structure and dynamics of chromosomes using polymer physics. The early work of this kind [46–51] dates back to early-1990 when the random walk model was used to explain the experimentally measured $R(s)$, which is the averaged spatial distance between two loci separated by a genomic distance s [46]. In the last decade, more focus is placed on explaining the experimental observation of Hi-C contact maps, and due to the complexity of the problem, coarse-grained computational polymer models are often used [52–70] when analytical solutions are not feasible. In spite of simplifying the problem as a polymer, which is necessary to make the problem tractable, computational models are great tools to provide valuable insights, especially considering the inherent limitations associated with the experimental studies. *De novo* polymer models have been developed to explain the compartments, TADs and CTCF loops observed in Hi-C data [56, 60–62, 66, 67]. The *de novo* models can provide direct biophysical insights to the problem. In addition to the *de novo* approach, numerous algorithms are also proposed to reconstruct 3D chromosome structure from both ensemble Hi-C [71–78]

and single-cell Hi-C contact map [79]. These methods are useful and convenient in practice to help visualizing 3D chromosomes using existing experiment data. For this particular topic, see [80] for an extensive review.

1.2 From imaging to Hi-C

Using fluorescence *in situ* hybridization (FISH) techniques, specific DNA sequences can be visualized in fixed cells. Chromosome Painting, which is developed based on FISH to detect the individual whole chromosomes, has shown unambiguously that the individual chromosomes occupy distinct territories instead of mixing with other chromosomes in single cells [2]. These distinct territories, termed as Chromosome Territories (CTs), are found to be distributed in the cell nucleus in a non-random fashion. Larger chromosomes are more likely located in the periphery of the nucleus, whereas the smaller chromosomes preferentially localized in the interior [2, 79]. At the same time, gene-poor chromosomes are more often found in the periphery and gene-rich chromosomes with similar sizes situate toward the center of the nucleus [79]. The CTs generally are not round-shaped domains but have irregular shapes depending on its gene richness. Gene-poor chromosomes are more compact and round shaped and gene-rich chromosomes adopt more expanded shape with protrusion [81]. Such gene content dependence is closely related to the epigenetic profile rather than the DNA sequence. The most prominent examples are the active and inactive X chromosome pairs (Xa and Xi). The inactive X chromosome is transcriptionally inactive with a round compact structure, and the active X

chromosome occupies larger volume with irregularly shaped protrusions [82] in spite of the similarity in their sequences.

Using the conventional microscope, chromosome staining reveals two types of form of chromatin. These two forms of chromatin are referred to as heterochromatin and euchromatin [2], as in heterochromatin referring to dark stained, densely packed chromatin and euchromatin referring to light stained, loosely packed chromatin. It is found that heterochromatin is distributed near the periphery of the cell nucleus and around the nucleolus and the interior of the nucleus is filled with euchromatin. These two terms were originally coined according to their physical forms and later were found to differ in their biological functions as well. Heterochromatin is largely composed of inactive and repressive loci whereas euchromatin comprises active genes and participate in the transcription activity. The physical separation between the two forms of chromatin are also manifested in the Hi-C data where the contact map can be decomposed into A/B compartments [21]. Even though the separation between these two forms of chromatin can be clearly observed under just conventional microscope, the finer details of the packaging of the chromatin fiber in heterochromatin and euchromatin still remain ambiguous. The 30-nm fiber model has long been suggested as the folding principle of the chromosome, especially in heterochromatin. In this model, the nucleosomes and linker DNAs package into an ordered structure with a diameter of about 30 nm and the resulting fiber further folds into heterochromatin or euchromatin. This scenario was supported in numerous *in vitro* experiments. However, it is becoming clear that such ordered 30-nm fiber is absent *in vivo*. A recent Cryo-EM experimental study using a novel high-efficiency

dye showed no evidence of 30-nm fiber but instead supports the picture of irregular packing of nucleosomes inside the cell nucleus [11]. The volume fraction (a similar quantity as mass density) is estimated to be around 40% to 50% for heterochromatin region and 10% to 20% for euchromatin region. However, the boundaries between the high- and low-density chromatin are not as clear as expected, but there are intermediate regions with a volume fraction covering the range from 20% to 40%. In the same experiment, mitotic chromosomes are found to be much more homogeneous and without any ordered structural units beyond a single nucleosome. The volume fraction of chromatin in mitotic chromosomes is found to be similar to that of heterochromatin, suggesting that chromatin folding may be similar between heterochromatin and mitotic chromosome.

The merit of imaging technique lies in its ability to directly measure the three-dimensional coordinates of the genomic loci. However, the power of imaging methods currently is largely limited due to low-throughput. How to determine the spatial and genomic coordinates of a large number of loci at the same time with genomic identification is an ongoing research area. About two decades ago, a very powerful non-imaging technique was proposed, called Chromosome Conformation Capture (3C) [18], which can detect the contact frequency between chromatin loci. The basic principle of the 3C technique is that when two chromatin loci are in physical proximity, their contact can be fixed using cross linking agents. Subsequently, the cross-linked chromatin loci pair is identified through sequencing. This procedure allows one, in principle, to measure the contact probability between any two loci. But it was not until a decade ago, Hi-C, which combines the high-throughput se-

quencing and 3C technique, was invented to map the contact map of the entire genome [21]. The fundamental logic of Hi-C experiment is that the organization of the genome can be inferred from the pattern of contacts (map). Such an idea is very important in protein folding where the native contact map is a direct measurement of protein structure. The Hi-C technique has been used extensively in the last decade to provide a glimpse of genome organization [21, 25–28] (also see [83] for a comprehensive review). Although the contact map is not a direct measurement of three-dimensional genome organization but rather a two-dimensional projection, important and previously unknown structural features are unveiled from Hi-C data. Three major findings from Hi-C data are 1) Chromosome Loops, 2) Topologically Associating Domains and 3) Compartments, each revealing a distinct organization principle for chromosome.

1.2.1 Loops and Topologically Associating Domains

Loops, as the term suggests, are looped structure between two genome loci. Such a structure, if prominent in cells, can be observed as peaks in 3C/4C/5C contact profile or as an interaction hotspot in the Hi-C contact map. Obviously, only the specific looping structure can be detected since any non-specific looping interactions will be smeared out in the ensemble averaged measurements. It is tempting to view these loop interaction as analogues of the native contacts in the protein structure. However, the chromosome loops are much more dynamic and heterogeneous compared to the native contact in protein folding. Thus, they cannot

be viewed as stable structures. It is a traditional view that in order for a gene to be expressed in Eukaryotic species, enhancers need to come into physical proximity with the promoter of their targeted gene [84, 85]. Such looping interactions are indeed observed in the 3C/4C/5C/Hi-C data [85, 86], supporting the contact model of transcription.

Contact profile for a specific loci can be measured using 3C. However, it is not until recently that the more complicated and higher order interaction patterns were revealed by Hi-C experiment. One important feature discovered are the Topologically Associating Domains (TADs) [25]. The TADs are the square patterns along the diagonal of the contact maps in which the probability of two loci being in contact is more probable than between two loci belonging to distinct TADs.

In their seminal work, Rao et. al. [28] showed that there is an underlying connection between TADs and loops - more specifically, CTCF loops. The CTCF loops, formed between pairs of CTCF motifs (which are genome segment of specific DNA sequences), previously unknown, were found to be much more prominent compared to the looping of the promoter-enhancer pair. They showed that there are thousands of CTCF loops distributed along each chromosome and that all the CTCF loops form between the boundary loci of the majority of TADs. The coexistence of the CTCF loops and TADs raises questions such as whether CTCF loops create TADs or the other way around. Chip-seq data shows that the boundary of TADs or loop anchors are enriched with cohesins, a ring-shaped protein from the Structural Maintenance Complex (SMC) family. Recently, the loop extrusion model was proposed to account for the formation of both TADs and CTCF loops [62, 87, 88]. According to

the loop extrusion model, one or multiple cohesins can encircle and move along two distant chromatin segments, thus enlarge the loops as they translocate along the DNA. When cohesins collide or encounter the roadblocks, such as CTCF motifs, they stop and act as boundaries of the TADs. In principle, cohesins can attach and detach from the chromatin stochastically, making the TADs intrinsically fluctuating objects rather than stable structures. Recent Hi-C experiments [89, 90] show that cohesins are indeed essential for the formation of TADs. The acute depletion of cohesins leads to the almost complete disruption of the TADs.

Whether such an extrusion process is driven by ATP-dependent motor activity or thermal fluctuation is still under debate [91]. There are numerous studies showing that condensin, which is also in the SMC family, has the capability to slide [92], compact DNA [93], and extrude loops along the double helix DNA strand in an ATP-dependent manner [94]. It has been shown that cohesin does need ATP to load on DNA [95,96]. However, single molecule experiments showed no evidence of cohesin's motor activity but that it slides diffusively along DNA [97, 98]. Currently, it still remains unclear what is the physical mechanism of the formations of CTCF looping interaction due to the difficulty of directly visualizing the dynamics of chromatin segment between two CTCF motifs.

The biological significance of TAD lies in its connections to gene regulation. It is speculated that the TADs can either up-regulate or down-regulate the transcription of genes, by enhancing the contact probability between the promoter-enhancer pair or by insulating the formation of loops between the promoter and enhancer outside the TADs [99–102]. More interestingly, the translocation of RNA polymerase

along the DNA might be partially responsible for the formation of TADs by pushing the cohesins, and thus extruding the loops [103], indicating that the transcription can also affect the formation of TADs. However, Rao et al. [89] found contrary results, which suggest that disruption of TADs leads to only moderate changes in gene expressions, meaning that the connection between the TADs and gene regulation can be minimal. It is important to note that the subtle change in gene expression sometimes can results in substantial change in the phenotype.

1.2.2 Compartments

I have discussed two important organization principles - loops and TADs - both of which occur on the length scale smaller than megabase pairs (Mbps). Hi-C Contact maps also revealed that chromosomes are organized into compartments on the genomic length scales exceeding Mbps [21, 28]. The partitioning of the structure into compartments are highly correlated with the histone markers in the chromatin loci [28, 89], implying that contacts are enriched within the same compartment and depleted between different compartments. The loci associated with active histone markers and those associated with repressive histone markers are localized in different compartments. The compartment formation, observed in the Hi-C contact maps, is likely a manifestation of spatial separation between heterochromatin and euchromatin observed using the microscopy. The physical mechanism of such compartment formation is speculated to be due to microphase separation [56, 63, 66, 68, 104] - chromatin loci with similar histone markers preferentially interact with each other,

possibly through direct nucleosome-nucleosome interactions [105–107] or through histone binding proteins [108]. Interestingly, two recent studies show that HP1a, which binds to H3K9me markers - a typical heterochromatin marker, form liquid droplets both *in vivo* and *in vitro* [109, 110]. Similarly, Polycomb Repressive Complex 1 (PRC1), which is found to modify histone markers, also forms droplet *in vitro* under the physiological conditions [111]. These studies provide evidence for the model that distant heterochromatic loci form a cluster with heterochromatin binding proteins acting as bridges between them. As for the euchromatin, although direct evidence of phase separation type mechanism is lacking, it is possible that distant euchromatic loci are brought in physical proximity by the transcription hubs formed from RNA polymerases and coactivators [112–114].

The interplay between the different layers of genome organization is an ongoing topic. The bulk Hi-C experiments suggest that CTCF loop formation counter plays the compartmentalization [68, 89]. This can be understood by noting that many loop pairs are actually located within different compartments. Thus, two loci with different histone markers, which prefer to be in physical separation, are constrained in proximity by their direct looping interaction. Such looping interactions increase the mixing between different compartments. As a result, the disruption of TADs leads to the enhancement of compartmentalization. However, for *Drosophila*, the TADs and compartments are possibly two sides of the same coin, both reflecting the underlying epigenetic states [115]. It is, hence, intriguing to understand how does the interplay between TADs and compartments matter? Why are loops needed in the mammalian world and why do TADs counteract but not assist the formation of

compartments? These questions clearly request further investigation.

1.2.3 Single-cell genome organization

It is important to understand the distinction between bulk and single-cell experiments. The contact profiles in the bulk Hi-C experiment [21, 28] are measured from millions of cells. Thus, the results are ensemble/population averaged. In contrast, the single-cell Hi-C [79, 116, 117] or FISH experiments [14, 29–31, 118] take measurements from individual cells. Thus, they can generate snapshots of genome organization at the single-cell level. Furthermore, the single molecular tracking of genome loci in principle allows one to obtain both the spatial and temporal information of the genomes, in spite of the current limitation of relative low throughput of the technique. The importance of single-cell experiments is its ability to quantitatively measure the extent of cell-to-cell variations. Such variations seem to be a prevalent feature observed in essentially all dimensions of biological systems [119], from the molecular level such as the gene expression profiles [120], to community level such as tumour heterogeneity [121]. In this thesis, I will discuss that the structural and dynamical heterogeneity dominates genome organization. The merit of cellular heterogeneity may be that it provides a mechanism for cells to be more adaptive to changing environments by allowing a large range of responses, a mechanism analogous to how the diverse species are beneficial to an ecosystem. This, of course, is speculative and requires quantification in the future.

Recent single-cell Hi-C experiments show that the genome organization indeed

exhibits large variations between cells of different types [116, 117] and between cells in different stages in the cell cycle [122], indicating that chromosomes are dynamical objects that undergo substantial conformational changes through the cell cycle. More interestingly, genome organization also displays extensive cell-to-cell heterogeneity even for the same type at the same stage of the cell cycle as demonstrated by single Hi-C experiment [79] as well as imaging experiments [14, 24, 30, 118]. Both Stevens et al. [79] and Bintu et al. [30] showed that the TADs are not conserved structural units but rather adopt different structures from cell to cell. Bintu [30] further show that even with the depletion of cohesins, which are essential in preserving the TADs at the ensemble level, the TADs-like structures can still be observed at the single cell, although the preferential location of the boundaries of the TADs observed in the ensemble Hi-C contact maps are obsolete. Wang et al. [14] showed that the compartments are preserved structures even in an individual cell and confirmed that the A/B compartments defined from Hi-C contact map are indeed physically separated, and arranged in a polarized fashion. However, a detailed analysis of their data [66] reveals that, although the physical separation between the two compartments is observed in every cell, the exact conformations of the chromosome exhibit a widespread continuous distribution without falling into a small number of subpopulations. Finn et al. [118], in a high throughput experiment, obtained a large dataset of measurements of distances for about 200 pairs of loci in human fibroblast cells. They found that the distributions of distances between any pair of loci are widespread, indicating an extensive heterogeneity in genome organization on both small (0.1 Mbps) and large length scales (100 Mbps). In this thesis, I will provide

a theory which can be used to analyze such data and extract information about the distribution of subpopulations of cells.

The complexity of genome organization on the single-cell level can be revealed by observing snapshots of many cells or by monitoring a small set of cells over a long time. In the language of statistical mechanics, they correspond to ensemble average and time average, respectively, and become identical only if the system is ergodic. However, it is still unclear if the biophysical properties of chromosomes are ergodic. To resolve this question, the dynamical information is required, a topic I will explore in the following sections. This question is also addressed in Chapter 3, in which I provide evidence of non-ergodicity in genome dynamics by means of coarse-grained simulations.

It is also worth noting that the thermal fluctuations may play a non-negligible role in the observed variations in genome architecture. It is well known that any polymer has its intrinsic continuous distribution of conformations, which can be widespread (e.g. an ideal chain or a self-avoiding chain) simply due to the thermal fluctuations. Hence, it is important for one to be able to discern the variations caused by fluctuations and those of other origins. Such a question, at the present, remains largely a mystery and unexplored. In this thesis, I developed a theoretical framework using polymer physics and hope to provide some insights to this question.

1.3 Chromosome Dynamics

It is not hard to realize that the dynamics is an essential part of all biological functions. A genome where everything is fixed in space cannot perform gene expressions, DNA repair, and many other functions. Thus it is of paramount to understand how does the chromosome loci move in the confined space of cell nucleus. FISH technique and its derivative (e.g. multiplex FISH) as well as Hi-C techinque can not probe the dynamics of genomes since they are performed on fixed cells. Live-cell imaging techniques [17, 32], mostly relying on labeling chromatin loci with fluorescence tags, are required to investigate the dynamical aspects of genomes.

1.3.1 Some theoretical concepts in polymer dynamics

Before delving into an overview of key experimentally-observed results regarding the genome dynamics, I will briefly discuss a few key theoretical concepts in polymer dynamics. The simplest polymer model is an ideal chain or sometimes also referred to as Gaussian chain. It is a polymer in which all interactions but chain connectivity are neglected. Rouse model [123] shows that the mean square displacement (MSD) of a monomer in a single ideal chain in a thermal bath behaves as MSD $\sim Dt^{1/2}$ for small t and exhibit normal diffusion (MSD $\sim t^1$) at long time. In general, MSD for a monomer can be written as MSD $\sim Dt^{\alpha}$ where α is the diffusion exponent and D is the diffusion coefficient. If $\alpha < 1$, the diffusion process is termed as sub-diffusive and is referred to super-diffusive if $\alpha > 1$. Both cases sometimes are referred as anomalous diffusion.

1.3.1.1 Connection between structure and dynamics

The dynamics of a polymer is intrinsically connected to its structure [45, 123]. This can be illustrated using the following scaling argument. The characteristic time scale of relaxation of a polymer segment, τ_r, is given by $\tau_r = R^2/D_r$ where R is the characteristic length scale of the segment and D_r is the diffusion coefficient of the center of mass (COM) of the segment. The dynamic of monomers is sub-diffusive for $t \ll \tau_r$ and exhibits normal diffusion for $t \gg \tau_r$. For an ideal chain, all monomers contribute to the diffusion of the COM. Thus, D_r must scale as N^{-1} where N is the number of monomers. Instead, let's consider a compact globular structure formed by a single polymer and assume its internal motion is sluggish such that the chain moves like a rigid body on the time scale of τ_r. Under these condition, $D_r \sim N^{-2/3}$ because only the surface monomers contributes to the diffusion of the whole chain. In general, we have $D_r \sim N^{-\theta}$ where θ is the characteristic exponent quantifying how many monomers contribute to the global motion of the COM. Geometrically, θ can be viewed as a measurement of the surface roughness of the segment. Combined with the relation $R \sim N^{2\nu}$ where ν is the Flory exponent, we have $\tau_r \sim N^{2\nu+\theta}$. The length scale of a monomer's diffusion at time τ_r must coincide with the length scale of the chain itself, leading to $\tau_r^{\alpha} \sim R^2 \sim N^{2\nu}$. Hence, the diffusion exponent of a monomer for time $t < \tau_r$ is given be $\alpha = 2\nu/(2\nu + \theta)$. For an ideal chain, using $\nu = 1/2$ and $\theta = 1$, we have $\alpha = 1/2$ which recovers the predictions from Rouse model. For the crumpled globule, proposed by Grosberg [48], it can be shown that $\alpha = 0.4$ (with $\nu = 1/3$ and $\theta = 1$), which is close to the value measured for human

interphase chromosome [17, 33–37].

1.3.1.2 Dynamics of contact formation

The distribution of distance between two arbitrary monomers along a chain has a general form $P(r) \sim r^{2+g}\exp(-Br^{\delta})$ where B is some constant depending on the specific polymer model [124, 125], and g is the "correlation hole" exponent, and δ is related to Flory exponent ν by $\delta = 1/(1-\nu)$. The exponents g and δ govern the small and long length scales of polymer conformation, respectively. In the previous section, I showed that the dynamical property - the diffusion exponent of monomers on the small time scale - is directly related to its structural property - the Flory exponent ν. More interestingly, the dynamic of contact formation between two monomers is connected to both g and ν values. Based on the arguments presented by Toan and colleagues [126], combined with the scaling argument $\alpha = 2\nu/(2\nu+\theta)$, it can be shown that the compactness of the exploration of the space between two monomers before they come into contact can be assessed by the parameter $\gamma = (3+g)\nu/(2\nu+\theta)$ where 3 is the dimensions of space. When $\gamma > 1$, the exploration is non-compact and the mean first passage time of contact formation, τ_c, is $\sim N^{\nu(3+g)}$. For $\gamma < 1$, the compact exploration of the conformations between two monomers leads to $\tau_c \sim N^{2\nu+\theta}$. The dynamics of contact formation is particularly interesting because it is the first step in the gene regulations in which the enhancer and promoter form contacts [99–102]. Hence, the compactness of the searching process has direct biological consequences.

1.3.2 Single loci movements

Early experimental attempts for investigating interphase chromosome dynamics showed that individual loci undergo constrained sub-diffusive motion in living cells on the time scale as long as a hundred seconds [127, 128]. Such constrained diffusion is likely a direct result of chromosome territories (CTs). Individual CT is relatively immobile at its fixed position in the cell nucleus and consequently, chromatin loci can only dangle within the territory, move as far as the size of territory which is sub-micron. The sub-diffusivity can be largely explained by the polymeric nature of chromosome. More recent experiments show the average diffusion exponent, α, of chromatin loci lies within $(0.4, 0.5)$ [17, 33–37]. However, the difference between 0.4 and 0.5 is subtle and hence it is hard to conclude a confident value of α. It is likely that the genome organization does not follow any generic polymer model such as fractal globule or ideal chain, but rather adopts complex structures with different folding principles at different length scales. Nevertheless, the fact that the value of α lies around 0.5 indicates that the chromosome dynamics, to a large extent, originates from generic polymer effect.

A more biologically significant question arises in regard to the dynamics of promoter and enhancer and their communication in the context of gene transcription. In two recent experiments, combined live-cell imaging and CRISPR-based technique was used to address this very issue. Chen and colleagues [129] used multi-color labeling to directly monitor the enhancer, its promoter as well as the transcription activity simultaneously. It provides direct evidence that promoter-enhancer phys-

ical contact is directly coupled with the transcription of the targeted gene in real time. More importantly, they showed that the contact is not stable but rather transient, meaning that the contact breaks and forms dynamically. Gu and colleagues showed [130] that enhancer and promoters are sub-diffusive, even in the presence of transcription, and the transcription activity increases the apparent diffusion coefficient of enhancer and promoters but hardly affects the diffusion exponent which was found to be about 0.5. In addition, by directly inhibiting the RNA polymerase II initiation or elongation, the mobility of enhancer/promoter is decreased. Furthermore, a large variation is observed in the single loci trajectories, which seems not to be explained by statistical fluctuations. Such dynamical heterogeneity will be further discussed in the following section.

1.3.3 Global motions of chromosomes

Most of the live-cell imaging experiments monitor a handful of chromatin loci by tagging them with fluorescence probes. They provide great insights into understanding how individual loci move but suffer from the limitation of accessing the global dynamic view of the chromosomes territories and the cell nucleus. A recently developed experimental technique based on the correlation spectroscopy of time-resolved imaging using particle imaging velocimetry (PIV) has been used to tackle this very question [41]. In their work, the global motions of chromosomes within an entire cell nucleus can be measured from the displacement vector field which is inferred from a series of fluorescence images. An important finding in their work is

that chromatin movement was found to be coherent on the length scale of microns. The chromatin loci develop dynamical correlation - meaning that the loci close in space also move along each other in similar directions and with similar magnitudes. Such coherent motions build up to its maximum around time scale of several seconds and vanished at longer times. Later on, Nozaki and colleagues [43], using super-resolution live-cell imaging, identified dynamically coherent chromosome domains with an average size of 160nm, which is comparable to the length scale of an average CTCF loops [28]. In another study, the coherent motions were found to be transcription-dependent [44], by showing that the long-range dynamical correlation between chromatin loci is diminished in the absence of RNA elongation. I have shown, using a coarse-grained computational polymer model [66], that the observed long-range dynamical correlation is a consequence of glass-like dynamics of the chromosomes. On the contrary, an activity-based mechanism was proposed [131] according to which the coherent motions are driven by the active force exerted on the loci, potentially by RNA polymerase II, helicase, and topoisomerase.

1.3.4 Dynamical heterogeneity

With the growing experimental evidence of structural heterogeneity of genomes [14, 24, 30, 79, 118], it is natural to wonder if the chromosome also show dynamical heterogeneity. In the light of recent experimental data, the answer to this question becomes increasingly clear that the chromosome loci indeed exhibit extensive variations in their dynamics. The large variance in the single loci trajectories has been

found in both *E.coli* and human cells [17, 34, 35, 38, 39], reflected by the widespread distribution of apparent diffusion coefficients and exponents of individual loci [40]. The mobility of individual loci can vary from each other by several orders of magnitude, suggesting a coexistence of both fast and slow loci. The heterogeneities in the loci mobility can be very well captured by the Van Hove function, which is the distribution of displacements of individual loci at a certain time lag. For an ideal chain or self-avoiding chain, the Van Hove function is a Gaussian. However, if there are large variations among loci's mobilities, the Van Hove function has a fatter tail. Lampo and colleagues [132] showed that the Van Hove function of the mRNA molecules in the cytoplasm of both *E.coli* and *Saccharomyces cerevisiae* cells is best fit by a Laplace distribution rather than a Gaussian. Interestingly, in another work, nucleoid-structuring (H-NS) proteins' movements are monitored as a proxy of chromatin loci dynamics due to its ability of binding to DNA [133] and the authors found that the distribution of displacements of H-NS is of Pearson Type VII which has a power-law tail. Clearly, it is interesting to see future experiment works reporting similar measurements for chromosome loci.

The physical origin of the observed dynamical heterogeneity of chromatin loci is of great interest. It should be noted that the variation of chromatin loci dynamics can be a result of multiple effects rather than just one cause. In principle, the potential origins of dynamical heterogeneity can be classified into two mechanisms. One is that the differences between individual loci's dynamics simply is a reflection of differences in their intrinsic properties. It is intuitive that a euchromatin loci probably should behave differently from a heterochromatin loci due to the dif-

ference in their chemical compositions and the macro-environment. This intuition is indeed supported in experiments that euchromatin loci diffuse faster than the heterochromatin [42].

The other possibility is that dynamical heterogeneity can emerge near the glass transition of both passive and active material (see [134–136] for detailed reviews of glass transition), which is accompanied by slowing down of particles. This can be understood as a direct result of crowding - due to confinement of cell nucleus - and/or effective attraction between chromatin loci [105–107, 109, 110]. The distinction between these two types of dynamical heterogeneity is that in a glassy material, the fast and slow particles exchange their diffusivities over the time scale longer than the relaxation time scale of the system. On the contrary, if the differences between the dynamics of two loci are caused by their underlying intrinsic property, the fast particle will always be fast and vice versa. Thus, in principle, one can potentially differentiate between these two mechanisms by monitoring individual loci over a long time.

1.3.5 The role of active forces

Unlike the protein folding which is usually governed by equilibrium thermodynamics, the non-equilibrium effect may play an important role in both genome organization and dynamics. The active force has been argued to be the cause of the observed super-diffusive motion in *E.coli* [38]. However, it is hard to approach this question in an experimental set up since the elimination of ATP-activity can lead

to disruption of various biological functions. Thus, the direct causation between the active force and genome dynamics and organization is difficult to be established. Theoretical and computational models in this context is thus of great use and can provide valuable insights. A recent computational study [137] showed that a single polymer undergoes a coil-to-globule transition when an active force parallel to the backbone of the chain is injected into the system. This is a surprising result since the active forces are usually considered to increase the effective temperature of the system, which should cause the chain to expand not the other way around. In a more chromosome-specific context, Weber and colleagues [138] argued that random motions of chromatin loci, to a large extent, are driven by the ATP-dependent activity, which leads to an Arrhenius dependence of diffusion coefficient on temperature. I and others showed that by including isotropic active noise on euchromatin chromatin loci, the phase separation between heterochromatin and euchromatin can be further enhanced [67], which is in coordinance with the activity-induced phase separation observed in various computational studies [139–141]. In addition, David et al. [131] showed that anisotropic active force along the chromatin fiber can give rise to the observed coherent motions of chromatin loci on micron length scale.

1.4 Theoretical and computational models for chromosomes

1.4.1 Early theoretical models

With its polymeric nature being a potential determinant of the biophysical property of chromosomes, many theoretical and computational models for chromo-

somes have been developed using the concepts in polymer physics [46–70]. The first attempt in the field is to model the interphase chromosomes using a simple Gaussian chain [46], in which the volume exclusions are neglected. The study found that the distributions of distances between probed loci pairs, within reasonable accuracy, can be fit by the theoretical formula for a Gaussian chain. Presented in this pioneering work, an important quantity used to describe the chromosome organization is the mean spatial distance as a function of the genomic distance, $R(s)$, where s is the genomic distance. The authors found that $R(s)$ scales as $s^{1/2}$ for small $s < 2$Mbps and reaches a plateau for larger s. Later experimental studies with larger data set showed that $R(s)$ has a finer structure with different scaling regimes at different length scales [14], and is epigenetic state dependent [29]. To explain the plateau behavior of $R(s)$, the random-walk/giant-loop model [49] was developed in which the chromosome segment form fixed loops with an average length 3Mbps. A similar early study [51] models the interphase chromosome as a micelle-like structure whose core is consisted of high GC content and the surface is formed by low GC content chromosome segments. In retrospect, this study provides two valuable insights of genome organizations. One is the compactness of chromosomes reflected by micelle-like structure. The other is the copolymeric nature of chromosomes by treating the low CG and high CG content regions as two different types of chromosome segments.

From a different perspective, Grosberg and colleagues [48] first discussed the effect of entanglements on genome organization. They correctly recognized crowding and confinements as two important features of chromosomes. The authors argued that the interphase chromosomes must be unknotted, at least to some extent, to per-

form its biological functions properly. One of the scenarios is that the searching of the targeted promoter by the enhancer will not be hindered if the local organization is disentangled. Based on this assumption, a theoretical argument of how the chromosomes package is provided and the concept of the "crumpled/fractal" globule - a non-equilibrium structure - was born. The crumpled globule is a compact structure which is also self-similar on all length scales. From classical Flory theory, we know that polymer segments inside a globule tend to adopt ideal chain conformation. Hence the crumpled globule can only be realized by specific attractive interactions between chromosome loci, by being in a long-lived non-equilibrium state, or by a combination of both. One of the distinct features of the "crumpled/fractal" globule is that $R(s)$ scales as $s^{1/3}$ and the contact probability $P(s)$ scales as s^{-1}. Here, $P(s)$ is the contact probability of two loci separated by the genomic distance s. The scalings are drastically different from those of ideal chain ($R(s) \sim s^{1/2}$ and $P(s) \sim s^{-3/2}$) or self-avoiding chain in good solvent ($R(s) \sim s^{0.6}$ and $P(s) \sim s^{-2.17}$). Thanks to the Hi-C technique, $P(s)$ can be obtained from the Hi-C contact maps and it is found indeed behave differently from a simple Gaussian chain [21] and was found approximately scales as s^{-1} [21, 22]. Based on this finding it is argued that $P(s)$ can be explained by the "crumpled/fractal" globule model [142]. However, the high resolution Hi-C experiment [28] further showed that the $P(s)$ exhibits different scaling regimes with $s^{-0.75}$ for $s < 0.5$Mbps and $s^{-1.25}$ for $s > 0.5$Mbps. Multiphasic behavior of $P(s)$ indicates that the organization on different length scales may be governed by different folding principles. This is in line with the growing experimental evidence that the TADs (small scale organization) and compartment

(large scale organization) observed in the contact maps have different origins and can exist independent of each other [89, 90].

1.4.2 Homopolymer model

Due to the complexity of genome organization, the theoretical models for chromosomes have their limitations. That being said, in the past decade, many computational models for chromosomes have been developed [52–70], partially owing to the increasing amount of data acquired from Hi-C and imaging exeperiments. Rosa and colleagues [52], using computer simulations of coarse-grained polymer, showed that many properties of interphase chromosomes originate from its generic polymer property. In particular, they argued that due to the confinement of cell nucleus and the topological constraints - polymer chain cannot cross each other - the long chromosomes are unlikely able to reach to equilibrium within the time scale of one cell cycle. The story is different for Yeast, whose chromosomes are much shorter and can be well modeled as polymer chain in equilibrium for both its structures [143,144] and dynamics [37]. In their first Hi-C experiment work, Lieberman et al. [21] also demonstrated that the apparent scaling $P(s) \sim s^{-1}$ can be obtained in a single polymer chain by quenching the system into a spherical confinement. Their simulations, for the first time, showed that the "crumpled/fractal" globule can be realized through a non-equilibrium process. The early coarse-grained models showed that some of the features in the contact maps, such as the contact probability $P(s)$ or the mean spatial distance $R(s)$ measured in FISH experiment, can be reproduced

using a homopolymer model [21, 52] without accounting for the epigenetic states, whereas fine structures such as TADs and compartments require more complicated models [55, 56, 60–70].

1.4.3 Copolymer/Heteropolymer-based model

In their pioneering work, Jost and colleagues [56] used a heteropolymer model with four different types of monomers representing active, Polycomb, HP1 and black chromatin to describe the formation of TADs in *Drosophila* genome. The underlying assumption of using heteropolymer to model chromosome is that the genome organization is largely driven and maintained by the interactions between the loci of similar epigenetic states. Based on this assumption, there are two kinds of approaches to tackle the modeling of chromosomes. One is the bottom-up approach, where the existing data on epigenetic states are used as input to determine the Hamiltonian of the system [56, 60, 63, 66]. The other one is a reverse-engineering approach, where Hi-C contact maps are instead used as inputs to determine the Hamiltonian of the system [24, 61, 104]. As I discussed in the previous section, the interactions between chromatin loci can come from two origins: the direct nucleosome-nucleosome interactions [105–107] and the effective interaction due to the various bridging proteins [109, 110]. The strings and binders switch model [55, 60] takes these DNA binding proteins directly into consideration and models the chromosome as a system with both polymer - representing the chromosome itself - and the free particles - representing the binding proteins. The polymer is set to have different types of

binding sites which can be bound by their protein counterparts. The difference between the binders model and the copolymer/heteropolymer model is subtle since the first can be viewed as an effective heteropolymer with renormalized loci-loci interactions. The differences between these types of models have received little attention and need further investigation.

1.4.4 Loop extrusion model

The copolymer/heteropolymer models, with the epigenetic states of chromatin loci being correctly represented, can faithfully reproduce the compartments observed in Hi-C contact maps [56, 61, 66]. The accuracy of such models is particular good for human interphase chromosome organization on the length scale beyond 10Mbps. They are also good enough to model the *Drosophila* where the fine TADs structures seem to be mostly driven by the underlying epigenetic states [145]. However, the finer structure such as TADs and sub-TADs in mammalian cells cannot be fully reproduced by just epigenetic markers. The fact that the TADs structure in mammalian cells exhibit complicated pattern such as nested loops/domains suggests that a parallel mechanism is responsible for the chromosome organization on the length scale below Mbps. The loop extrusion model (introduced above) [62, 87, 88, 146] has been used to simulate the human TADs structure and is able to reproduce the experimental measured TADs pattern to great accuracy. The extrusion is assumed to be an active process in the original model [87], where the cohesin translocates along the chromatin by consuming ATPs. Other types of loop extrusion models

have been proposed, including the transcription-driven model in which the cohesin is pushed by the supercoiling as a result of RNA Polymerase II translocation [103] and ATP-independent model in which the extrusion process is driven by osmotic pressure [147, 148]. In addition, to explain the formation of TADs in mammalian cells, loop extrusion model was also proposed to be the main mechanism of compaction of chromosomes before cells enter the mitosis [149], the segregation between chromosomes during the mitosis [150], and the disentanglement of genome organization [151]. Currently, the molecular mechanism of extrusion process remains largely unknown. Marko and colleagues [152] propose a rachet-like kinetic model for coehsin (and other SMC family proteins). According to the model, the cohesin's motor activity is coupled with the random DNA looping driven by thermal fluctuation and thus its kinetics depends strongly on the looping probability of DNA. As a result, the extruding velocity strongly depends on the tension of the DNA which is supported by experiments [94]. A different mechanism, called tether-inchworm model [153], was also proposed. In this model, the cohesin is thought to perform inchworm-like motion by opening and closing its ring. It is important to bear in mind that an appropriate model for cohesin/condensin should account for its abilities both to translocate along a tethered DNA [92], to drive DNA compaction [93], and to extrude the loops from a flexible DNA [94]. It is difficult to explain these observations in one model using the conventional picture of other molecular motors such as Kinesins, Myosins or Dynes.

1.5 Outline of Thesis

In Chapters 2 and 3, I present the chromosome copolymer model (CCM) by representing chromosomes as a copolymer with two epigenetic loci types corresponding to euchromatin and heterochromatin. Using novel clustering techniques, I establish quantitatively that the simulated contact maps and topologically associating domains (TADs) for chromosomes 5 and 10 and those inferred from Hi-C experiments are in very good agreement. Chromatin exhibits glassy dynamics with coherent motion on micron scale. The broad distribution of the diffusion exponents of the individual loci, which quantitatively agrees with experiments, is suggestive of highly heterogeneous dynamics. This is reflected in the cell-to-cell variations in the contact maps. Chromosome organization is hierarchical, involving the formation of chromosome droplets (CDs) on genomic scale, coinciding with the TAD size, followed by coalescence of the CDs, reminiscent of Ostwald ripening.

In Chapter 4, I propose a theoretical framework based on Generalized Rouse Model to solve the FISH-Hi-C paradox and provide the insights of understanding the heterogeneity of genome organization. Hi-C experiments are used to infer the contact probabilities between loci separated by varying genome lengths. Contact probability should decrease as the spatial distance between two loci increases. However, studies comparing Hi-C and FISH data show that in some cases the distance between one pair of loci, with larger Hi-C readout, is paradoxically larger compared to another pair with a smaller value of the contact probability. The FISH-Hi-C paradox arises because the cell population is highly heterogeneous, which means that

a given contact is present in only a fraction of cells. Insights from the GRMC is used to construct a theory, without any adjustable parameters, to extract the distribution of subpopulations from the FISH data, which quantitatively reproduces the Hi-C data. Heterogeneity is pervasive in genome organization at all length scales, reflecting large cell-to-cell variations.

In Chapter 5, based on the theory proposed in Chapter 4, I prove that there exist a theoretical lower bound to connect both quantities by a simple power law relation. Hence the inverse engineering problem - inferring spatial organization from Hi-C map - can be solved approximately in spite of the presence of heterogeneity. Using simulations, I show that the overall organization can be captured by constructing distance map from contact map justifying the use of the lower bound. Finally, by applying our method combined with various manifold embedding methods to experimental Hi-C data, I am able to visualize the averaged global 3D organization of single chromosome, and also local structures such as Topological Associated Domains (TADs). In the end, discussion on the limitation of Hi-C map as an ensemble average measurement is provided.

In Chapter 6, on a side project, I present a kinetic model for coupled motor system. The simplicity of the model allows me to investigate the effect of mechanical coupling between multiple motors on their velocity and force-velocity behavior. Reduction of velocity are observed for coupled motor system especially when the coupling strength is strong. I found that the velocity in the absence of load only weakly depends on the number of motors n and reaches to limiting value when $n \to \infty$. Stall force is shown to be given by $F_s = nF_s^0$ but velocity vanishes at a

smaller apparent stall force F_a. In addition, I found that multi-motors system is more efficient for transporting large cargo but less efficient for transporting small cargo compared to the single motor. The model presented in this study is general and could be extended to study cooperation and tug-of-war between motors.

Chapter 7 provides the conclusion and future aspects of the studies presented in the thesis.

Chapter 2: Chromosome Copolymer Model

2.1 Introduction

The work presented in Chapter 2 and Chapter 3 was published [66] and the copyright was obtained to reuse the content in [66] in this thesis.

Contact maps [21, 28] of interphase chromosomes show that they are partitioned into genome-wide compartments, displaying plaid (checkerboard) patterns. If two loci belong to the same compartment they have a higher probability to be in contact than if they are in different compartments. Although finer classifications are possible, compartments [21] can be categorized broadly into two (open (A) and closed (B)) classes associated with distinct histone markers. The open compartment is enriched with transcriptional activity-related histone markers, such as H3K36me3, whereas the closed compartment is enriched with repressive histone markers, such as H3K9me3. Chromatin segments with repressive histone markers have effective attractive interactions, which models HP1 protein-regulated interactions between heterochromatin regions [109, 110, 154, 155]. In CCM, I assume that chromatin fiber, with active histone markers, also has such a similar attraction. From these considerations, it follows that the minimal model for human chromosome should be a copolymer where the two types of monomers represent active and repressive chro-

matin states. To account for the two states, I introduce the Chromosome Copolymer Model (CCM) as a self-avoiding polymer with two kinds of beads. A similar genre of models have been proposed in several recent studies [56,60,61,63] to successfully decipher the organization of genomes.

In this chapter, I will describe the details of the construction of the Chromosome Copolymer Model (CCM). The simulation results of the CCM will be presented and discussed in Chapter 3.

2.2 The construction of the model

2.2.1 The Hamiltonian of the model

For reasons explained in both Chapter 1 and the introduction section in this chapter, the interphase chromosome is modeled as a self-avoiding AB-copolymer with A (B) type beads representing the active (repressive) chromatin (Fig. 2.1). Note that in many of the polymer models developed to reproduce Hi-C contact maps, self-avoidance is not strictly enforced, which is partially justified because Topoisomerase facilitates chain crossing. I do not find it necessary to impose this restriction. The chromosome copolymer model (CCM) potential energy is,

$$U(\mathbf{r}_1, ..., \mathbf{r}_N) = \sum_{i=1}^{N-1} U_i^{\mathrm{S}} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} U_{i,j}^{\mathrm{P}} + \sum_{\{p,q\}} U_{\{p,q\}}^{\mathrm{L}} \tag{2.1}$$

For the bond stretch potential, U_i^s, I use the FENE (Finite Extensible Non-

linear Elastic) potential given by,

$$U_i^{\mathrm{S}} = -\frac{1}{2} K_{\mathrm{S}} R_0^2 \ln\Big[1 - \Big(\frac{|\mathbf{r}_{i+1} - \mathbf{r}_i|}{R_0}\Big)^2\Big] \tag{2.2}$$

where R_0 is the equilibrium bond length, and K_{S} is the spring constant.

The interaction between beads accounting for steric repulsion and attraction is given by the Lennard-Jones potential with different parameters for the distinct bead types. The potential between the active locus and repressive locus is,

$$U_{i,j}^{\mathrm{P}} \equiv U_{\alpha\beta}(r = |\mathbf{r}_i - \mathbf{r}_j|) = 4\epsilon_{\alpha\beta}\Big[\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big] \tag{2.3}$$

where α and β can be either A (active/euchromatin) or B (repressive/heterochromatin). For simplicity, I assume that the σ value for the active state (A) and the repressive state (B) is identical.

The interaction between the loop anchors is modeled using a harmonic potential,

$$U_{\{p,q\}}^{\mathrm{L}} = K_{\mathrm{L}}(|\mathbf{r}_p - \mathbf{r}_q| - a)^2 \tag{2.4}$$

where $\{p, q\}$ is the index of the loop, and a is the equilibrium bond length between the loop pairs. The indices of loop anchors, modeling the role of CTCF motifs, taken from the Hi-C data [28], are listed in Table 2.1. The values of all the parameters in the CCM model of the chromosome are given in Table 2.2.

2.2.2 Setting the length scale

In CCM, each monomer represents 1,200 base pairs (bps), with six nucleosomes connected by six linker DNA segments. The size of each monomer, σ, is estimated by considering two limiting cases. If it is assumed that nucleosomes are compact then the value of σ may be obtained by equating the volume of the monomer to $6v$ where v is the volume of a single nucleosome. This leads to $\sigma \approx 6^{1/3} R_{\mathrm{N}} \approx 20$ nm where $R_{\mathrm{N}} \approx 10$ nm is the size of each nucleosome [156]. Another limiting case may be considered by treating the six nucleosome array as a worm-like chain. The persistence length of the chromatin fiber is estimated to be $\sim 1{,}000$ bps [88], which is about the size of one monomer. The mean end-to-end distance of a worm like chain whose persistence length is comparable to the contour length L is $R \approx L\sqrt{2/e}$. The value of L for a six nucleosome array is $6(16.5 + R_{\mathrm{N}})$nm where the length of a single linker DNA is 16.5 nm. This gives us the upper bound of σ to be 130 nm. Thus, the two limiting values of σ are 20 nm and 130 nm. I assume that the value of σ is an approximate mean, yielding $\sigma = 70$ nm.

2.2.3 Identification of the monomer type and loop anchors from experimental data

The epigenetic state of each bead is determined using the Broad ChromHMM track [157–159]. There are a total of 15 chromatin states in the track. For simplicity, I assign states 1-11 to be in the active state (A) and states 12-15 to be

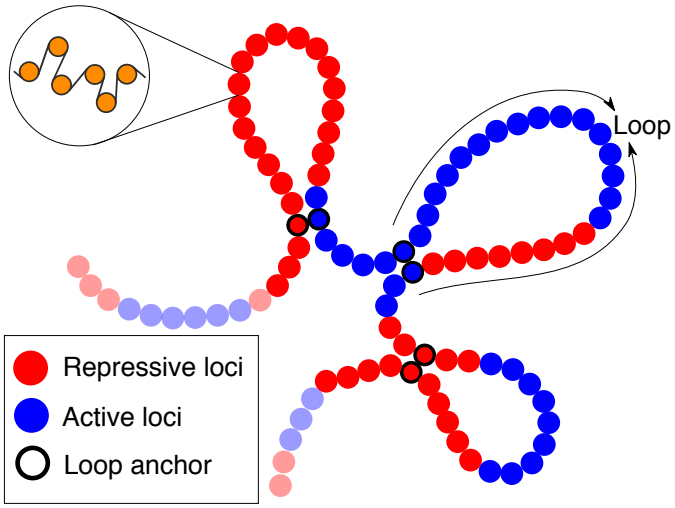


Figure 2.1: The sketch of the Chromosome Copolymer Model (CCM). Each bead represents $1,200$ basepairs (representing roughly six nucleosomes (orange circles) connected by linker DNAs). Red (Blue) corresponds to repressive (active) chromatin. The three pairs of loop anchors (in this cartoon) are marked by beads with black boundaries. A crucial aspect of the model, based on the experimental observation [28] is that the loops are consecutive and do not overlap with each other. The CCM accounts for two epigenetic states and the locations of the loop anchors. These two criteria are sufficient to reproduce all the subtle structural features noted in the Hi-C and super-resolution experiment.

43(B),396(B)	43(B),582(B)	143(B),396(B)	948(B),1110(B)
1407(B),1570(B)	1628(A),2120(B)	2355(B),2562(A)	2409(B),2562(A)
2622(A),2917(B)	3059(A),3106(B)	3307(A),3378(A)	3307(A),3630(B)
3307(A),3471(B)	4131(B),4175(A)	4131(B),4307(A)	4445(A),5012(B)
4445(A),4710(B)	5058(B),5548(B)	6318(B),6766(B)	6318(B),6408(B)
6318(B),6595(A)	6408(B),6595(A)	6647(B),6766(B)	7605(B),8907(A)
7917(B),8644(B)	7917(B),8743(A)	7917(B),8907(A)	8743(A),8907(A)
8921(B),9008(B)	9157(B),9396(A)	9481(A),9562(A)	9510(A),9562(A)

Table 2.1: Loop anchor indices for Chromosome 5 (Chr 5) derived from the experimental data [28] for use in the CCM. Each pair of numbers represents single loop corresponding to the locations of the loop anchors along the backbone of the copolymer. The letter A (B) after each number indicates the type of the loop anchor. The number of loops in the simulation using the CCM is thirty-two. Fifteen out of thirty-two pairs have loop anchors formed from loci of the same type.

in the repressive state (B). This is reasonable since all the states between 1 and 11 are related to gene transcription, and hence can be modeled as euchromatin. States 12 to 15 are polycomb repressed, heterochromatin or repetitive region, which I modeled as heterochromatin. ChromHMM track has a resolution of 200bps which is smaller than 1,200bps representing one monomer in the CCM. I first count the number of basepairs of state A and B within the 1,200 basepairs segment

$K_{\mathrm{S}}/k_{\mathrm{B}}T\sigma^{-2}$	R_0/σ	$K_{\mathrm{L}}/k_{\mathrm{B}}T\sigma^{-2}$	$\epsilon_{AA}/k_{\mathrm{B}}T$	$\epsilon_{BB}/k_{\mathrm{B}}T$	$\epsilon_{AB}/k_{\mathrm{B}}T$	a/σ
30	1.5	300	1.0	1.0	0.82	1.13
30	1.5	300	2.0	2.0	1.64	1.13
30	1.5	300	2.4	2.4	1.96	1.13
30	1.5	300	2.7	2.7	2.21	1.13

Table 2.2: Parameters values in the CCM for Chr5 and 10. Energy is in the unit of $k_{\mathrm{B}}T$ (k_{B} is the Boltzmann constant and T is the room temperature 300K), bead diameter σ is used as a measure of length. Without loss of generality, I choose $\epsilon_{AA} = \epsilon_{BB} = \epsilon$.

represented by each monomer. Then the type of each monomer is assigned as the state with a larger number of basepairs. Such a coarse-graining procedure may not be appropriate when the number of bps of the two types has a similar value in the 1,200bps segment. Although this is a possible outcome, I found that most of the 1,200bps long segments are overwhelmingly occupied by only one state, corresponding to either active or repressive state. For loop anchors, I directly used the Hi-C data [28]. The locations of loops are provided in the file *GSE63525_GM12878_primary+replicate_HiCCUPS_looplist_with_motifs.txt.gz* under the GEO accession number GSE63525. I only selected the loops with CTCF motifs "uniquely" called at both anchors (see Section VI.e.7 of the Extended Experimental Procedures of [28]). For each pair of CTCF loop anchors, I assign a harmonic constraint (Eq. 2.4) between the two corresponding loci.

2.2.4 Simulation details

I use both low friction Langevin Dynamics (LD) and Brownian dynamics (BD) to simulate the equilibrium and dynamical properties of the chromosome. The equation of motion for the i^{th} locus is given by,

$$m_i\ddot{\mathbf{r}}_i = \mathbf{F}_i - \xi\dot{\mathbf{r}}_i + \mathbf{R}_i(t), \tag{2.5}$$

where ξ is the friction coefficient, $\mathbf{F}_i$ is the systematic force $-\frac{\partial U}{\partial \mathbf{r}_i}$ experienced by each bead, and $\mathbf{R}_i(t)$ is the random force mimicking the thermal fluctuation of

the surrounding environment. In Eq. 2.6, $\mathbf{R}_i(t)$ is the Gaussian random force that satifies the fluctuation-dissipation theorem $\langle R_i \rangle = 0$ and $\langle \mathbf{R}_i(t) \cdot \mathbf{R}_j(t^{'}) \rangle = 6k_{\mathrm{B}}T\xi\delta(t-t^{'})\delta_{ij}$. The LD simulations are performed using the Molecular Dynamics software LAMMPS [160] in which the equation of motion are integrated using the velocity-Verlet algorithm. The sampling of the conformations are accelerated in LD, as was shown previously [161], and hence, I use LD simulations to generate well-equilibrated conformations.

The equation of motion for BD, derived by neglecting the inertial term in Eq. 2.6, is,

$$\dot{\mathbf{r}}_i = \frac{1}{\xi}\mathbf{F}_i + \frac{1}{\xi}\mathbf{R}_i(t). \tag{2.6}$$

I modified the LAMMPS software to perform the BD simulations, thus allowing us to obtain a realistic description of the dynamics. The use of BD also allows us to calculate the timescales for the chromosome dynamics, which can be directly compared to experiments. I employed the Euler algorithm to integrate the equation of motion in Eq. 2.6.

For BD, the relevant time scale is $\tau_{\mathrm{B}} = \sigma^2/D$ where $D = kT/\xi$, and $\xi = 6\pi\eta\sigma/2$. I choose our integration time step to be $\Delta t_{\mathrm{B}} = 0.0001\tau_{\mathrm{B}}$. With the choice of $\sigma = 70$nm, I obtain $D = \frac{kT}{6\pi\eta\sigma/2} \approx 7.0\mu m^2/s$ with $\eta = 0.89 \times 10^{-3} Pa \cdot s$. Thus, the value of $\tau_{\mathrm{B}} \approx 0.0007s$. For the LD simulations, I use the time step $\Delta t_{\mathrm{L}} = 0.01\tau_{\mathrm{L}}$ where $\tau_{\mathrm{L}} = \sqrt{m\sigma^2/kT}$.

2.2.5 Generation of the initial conformations and production runs

The copolymer is initially prepared as a rod. After determining the positions of the loop anchors, I performed simulations using LD with temperature $T = 1.0$ (measured in the unit of $k_{\mathrm{B}}T$) using the WCA potential with the same parameter values regardless of the bead type. I used the WCA potential,

$$U_{\alpha\beta}(r = |\mathbf{r}_i - \mathbf{r}_j|) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \epsilon, & \text{if } 0 < r < 2^{1/6}\sigma. \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

with $\epsilon = 1.0k_BT$ and $\sigma = 1$. Since all the loop anchor pairs initially are spatially well-separated, I first performed simulations using a small time step ($\Delta t_L = 10^{-6}\tau_{\mathrm{L}}$) to avoid numerical instabilities. After a certain number of time steps, all the loop pair beads are in proximity fluctuating around their equilibrium bond distance. At this stage, I increased the value of the time step to $\Delta t_{\mathrm{L}} = 0.01\tau_{\mathrm{L}}$, and turned on the attractive pairwise interaction (Eq. 2.3), and continued the simulations for an additional $10^8\Delta t_{\mathrm{L}}$. I monitored the radius of gyration, R_g, to ensure that R_g fluctuates around a mean value as one indication of thermalization (Fig. 2.2a). In addition, the potential energy (Fig. 2.2b) has reached plateau values, which is a necessary condition indicating that the copolymer has adequately sampled a large number of distinct conformations. I also computed the evolution of $P(s)$ during the pre-production run (Fig. 2.3). The negligible change in $P(s)$ also suggest convergence

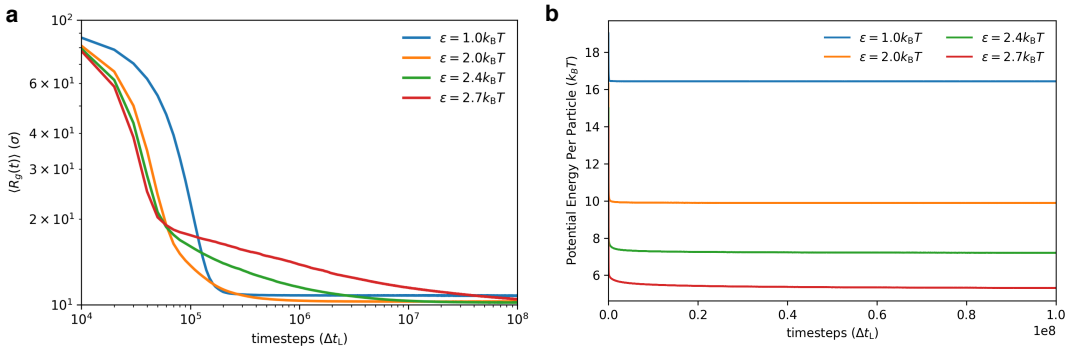


Figure 2.2: Preparation of the initial conformations. (**a**) The ensemble average radius of gyration $\langle R_{\rm g}(t)\rangle$ as function of time step t after the attractive interactions are turned on. $\langle R_{\rm g}(t)\rangle = (1/{\rm M})\sum_{i=1}^{\rm M} R_{\rm g}^{(i)}(t)$ where i is the $i^{\rm th}$ trajectory, and M is total number of independent trajectories. $R_{\rm g}^{(i)}(t)$ is the radius of gyration of trajectory i at time t. In our simulations, M = 90. (**b**) The average potential energy per bead as a function of the number of time steps t after the attractive interactions are turned on. The average is over the 90 independent trajectories. The plateau in (**a**) and (**b**) suggest that the polymer conformations are well sampled.

of our simulations from the perspective of structural measures. I then performed LD simulations for an additional $10^8\Delta t_{\rm L}$ to compute the static structural properties. The final chain conformations obtained at the end of production runs are used as initial conformations in the subsequent BD simulations.

2.3 Discussion

The virtue of the CCM is that it has essentially only one energy scale ϵ given that I have assumed that $\epsilon_{AA} = \epsilon_{BB}$. Explicit simulations show that this is sufficient to capture not only the compartments and TADs in the contact map reasonably well but also the chromosome dynamics (results presented in Chapter 3). The inclusion

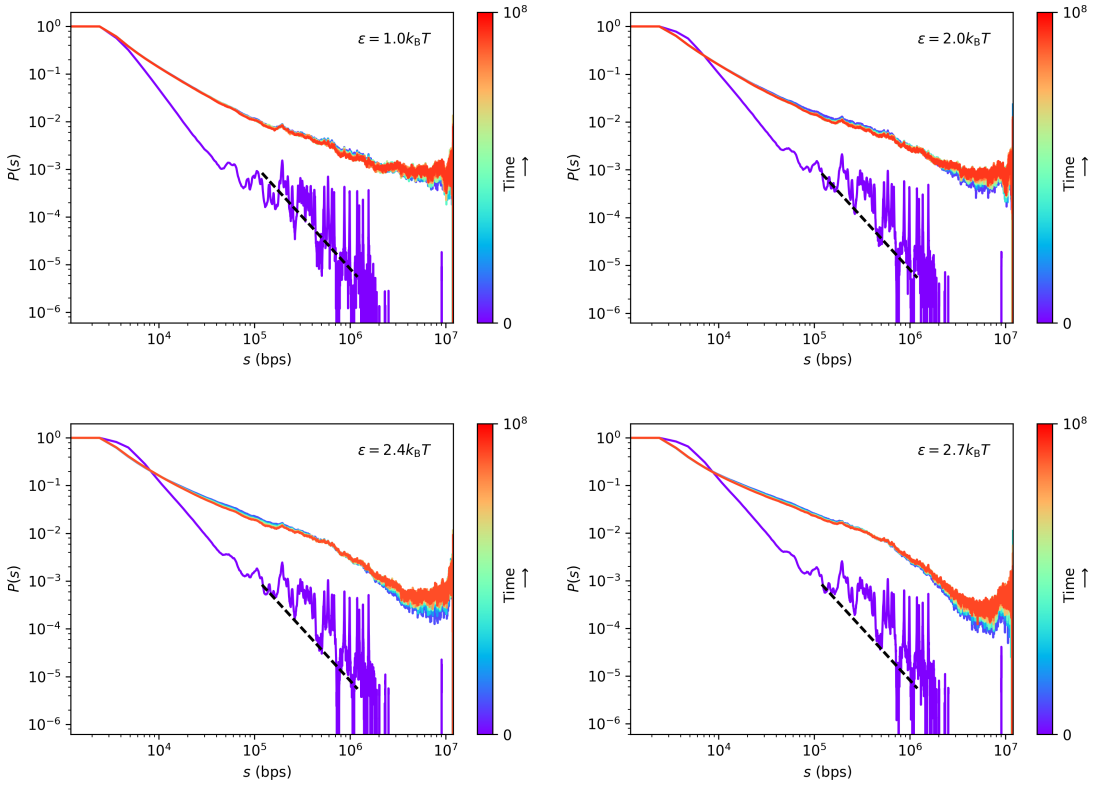


Figure 2.3: The top panel shows typical structures of the simulated folded Chr5 for $\epsilon = (1.0, 2.0, 2.4, 2.7)k_BT$ from left to right. The color indicates the index of the locus, from the $5'$ to $3'$-end. The spatial distance map and the corresponding Ward Linkage Matrices (WLMs) are shown in the middle and bottom panels, respectively.

of other epigenetic states identified in experiments may produce better agreement with the contact map inferred from Hi-C experiment but comes at the expense of introducing additional parameters. It is unlikely that such a model would alter the chromosome dynamics, which is the focus of the study here. I should also note that by independently tuning three parameters ϵ_{AA}, ϵ_{BB} and ϵ_{AB} separately, the model could be further optimized in the comparison with experiment Hi-C data. For the simplicity and because the errors in Hi-C data are hard to quantify, I make the assumption that $\epsilon_{\mathrm{AA}} = \epsilon_{\mathrm{BB}} \equiv \epsilon$ and fix the ratio $\epsilon_{\mathrm{AB}}/\epsilon$. In practice, I varied ϵ while keeping the ratio $\epsilon_{\mathrm{AB}}/\epsilon$ a constant. The results in Chapter 3 suggest that even with this simplification, CCM produces near quantitative agreement with Hi-C data.

With the assumption that $\epsilon_{AA} = \epsilon_{BB} = \epsilon$, the only free parameter in the CCM is ϵ_{AB}. The only physical requirement for choosing a specific value of ϵ_{AB} is that loci with distinct epigenetic state should segregate in order to capture the compartment feature that is prominent in the Hi-C maps. For the interaction parameter values listed in the third row of Table 2.2 which is most appropriate for interphase chromosomes 5 and 10, loci A and B do not mix. In other words they phase separate. This can be rationalized by adopting a Flory-Huggins type argument, which involves calculating the second virial coefficient, $B_{2,\alpha\beta} = 2\pi \int \mathrm{d}r r^2[1 - e^{-U_{\alpha\beta}/(k_{\mathrm{B}}T)}]$. I find that for the parameters in the third row of Table 2.2, $|B_{2,AB}| < |B_{2_AA}|$, which implies that A and B loci tend not to mix. Note that $B_{2_AA} = B_{2_BB}$ in the CCM. This argument shows that for any value of ϵ_{AB} for which the inequality $|B_{2,AB}| < |B_{2_AA}|$ is satisfied, the copolymer would exhibit micro-phase separation. However, the extent of segregation will depend on the precise numerical values. For our energy function,

the values listed in third row of Table 2.2 is optimal, because simulations using them provide the best agreement with the measured Hi-C contact maps.

I have made references to copolymer models [56, 60, 61, 63], which have been previously used to study chromatin organization. The one that is most similar to CCM is the block copolymer model used to describe the architecture of the roughly one Mbps *Drosophila* genome [56]. In their model micro-phase separation results by adjusting a non-specific energy scale between all loci pairs to induce global chain compaction and specific interaction (the analog of ϵ_{AA}, ϵ_{BB} and ϵ_{AB} in the CCM) between identical epigenetic states. A related minimal model, with three epigenetic states, was recently introduced in [63] that accounts for active, inactive, and unmarked states. These models, along with CCM, show that many aspects of chromosome organization can be captured using a minimum number of free parameters.

The folding of chromatin is simulated starting from extended conformations (see section 2.2.5). Due to the slow relaxation process, theoretically predicted in a previous study [58], and topological constraints [52], long polymers such as human interphase chromosomes are unlikely to come to equilibrium even on the time scale of a single cell cycle. Thus, the initial conformations could in principle affect the organization of genomes. Although the folding from an extended conformation is unlikely to occur for chromosome as a whole *in vivo*, I believe that the folding process investigated in this work provides insights into gene activation because it involves only local folding or unfolding [162–164].

2.4 Conclusions

Here, I present a copolymer model to describe both the structure and dynamics of human interphase chromosomes based on the assumption that the large-scale organization of human interphase chromosome is largely driven and maintained by the interactions between the loci of similar epigenetic states. Similar models, that differ greatly in details, have been developed to model the 3D structure of Drosophila chromosomes [56, 60]. Jost et al. [56] used a heteropolymer model with four different types of monomers representing active, Polycomb, HP-1 and black chromatin to describe the formation of TADs in *Drosophila* genome. Michieletto et al. [63] constructed a heteropolymer with three epigenetic states (acetylated, methylated, and unmarked) to probe how the epigenetic states are maintained. A very different reverse-engineering approach, with Hi-C contact maps as inputs, was used to construct an energy function with twenty-seven parameters [61]. I take a "bottom-up" approach to incorporate the epigenetic states into the polymer model similar in spirit to the previous studies [56, 60].

I performed simulations using both Langevin Dynamics (low friction) and Brownian Dynamics (high friction) using a custom modified version of the molecular dynamics package LAMMPS. The use of Langevin Dynamics accelerates the sampling of the conformational space [161], needed for reliable computation of static properties. Realistic value of the friction coefficient is used in Brownian Dynamics simulations to investigate chromosome dynamics, thus allowing us to make direct comparisons with experiments.

The results using CCM described here are presented in Chapter 3.

Chapter 3: Structures and dynamics of human interphase chromosome: a study using Chromosome Copolymer Model

3.1 Overview

The work presented in Chapter 2 and Chapter 3 was published [66] and the copyright was obtained to reuse the content in [66] in this thesis.

In this chapter, I present the results from Chromosome Copolymer Model (CCM) described in Chapter 2. I show that in order to capture the structural features faithfully, at least two types of beads, representing active and repressive loci are needed. Simulations of the resulting Chromosome Copolymer Model (CCM) for human interphase chromosomes 5 and 10 show that the structural characteristics, such as the scaling of $P(s)$ as a function of s, compartments, and TADs indicated in the Hi-C contact maps are faithfully reproduced. I use sophisticated clustering algorithms to quantitatively compare the simulated contact maps and those inferred from Hi-C experiments. The compartment feature noted in the Hi-C contact map is due to micro-phase separation between chromosome loci associated with different epigenetic states, implying that a copolymer model is needed for characterizing large-scale genome organization. The TADs emerge by incorporating experimentally

inferred positions of the loop anchors, whose formation is facilitated by CTCF motifs. The only free parameter in the CCM, the optimal loci-loci interaction strength between loci belonging to the same epigenetic states, is adjusted to give a good description of the Hi-C contact map. Using simulations based on the resulting CCM I show that chromosome dynamics is highly heterogeneous and exhibits many of the characteristics of out of equilibrium glassy dynamics, with coherent motion on μm scale, including stretched exponential decay of the scattering function ($F_{\rm s}(k,t)$), a non-monotonicity behavior in the time dependence of the fourth order susceptivity associated with fluctuations in $F_{\rm s}(k,t)$. Of particular note is the remarkable cell-to-cell and loci-to-loci variation in the time (t) dependence of the mean square displacement, $\Delta_i(t)$, of the individual loci. The distribution $P(\alpha)$ of the exponent associated with the increase in $\Delta_i(t) \sim t^{\alpha}$ is broad. The simulated and experimentally measured $P(\alpha)$s are in excellent agreement. Our work shows that chromosomes structures are highly dynamic exhibiting large cell-to-cell variations in the contact maps and dynamics. The rugged chromosome energy landscape, with multiple minima separated by large barriers, is perhaps needed to achieve a balance between genomic conformational stability and dynamics for the execution of a variety of biological functions.

3.2 Results

3.2.1 Choosing the energy scale in the Chromosome Copolymer Model

I fixed N, the size of the copolymer to $N = 10,000$, modeling a 12 Mbps (megabases) chromatin fiber, corresponding to a selected region of the Human Cell line GM12878 Chromosome 5 (Chr 5) from 145.87 Mbps to 157.87 Mbps. In the CCM (Fig. 2.1), the only unknown parameter is ϵ, characterizing the strength of the interaction between the loci (Table 2.1). I chose a ϵ value that reproduces the contact maps that is near quantitative agreement with the Hi-C data. As ϵ increases the structures of the chromosome are arranged in such a way that segments with small genomic distance s are more likely to be in spatial proximity (see the section **Chromosome Structures in terms of Ward Linkage Matrix (WLM)** below). This is also illustrated in Fig. 3.1, which shows that higher values of ϵ lead to clearer segregation between the loci with different colors. The colors encode the genomic locations. The snapshots of the organized chromosome, the good agreement between the calculated and Hi-C contact maps (see Fig. 3.2d and section 3.2.2), and the accurate description of the spatial organization as assessed by the Ward Linkage Matrix (WLM) (see section 3.2.5 and Appendix A.3) confirm that $\epsilon = 2.4k_{\mathrm{B}}T$ produces the closest agreement with experiments. Increasing ϵ beyond $2.4k_{\mathrm{B}}T$ leads to a worse description of segregation between loci with distinct epigenetic states.

Furthermore, $P(s)$ as a function of s obtained in simulations with $\epsilon = 2.4k_{\mathrm{B}}T$ is also consistent with experiments (see below). The s-dependent contact probability,

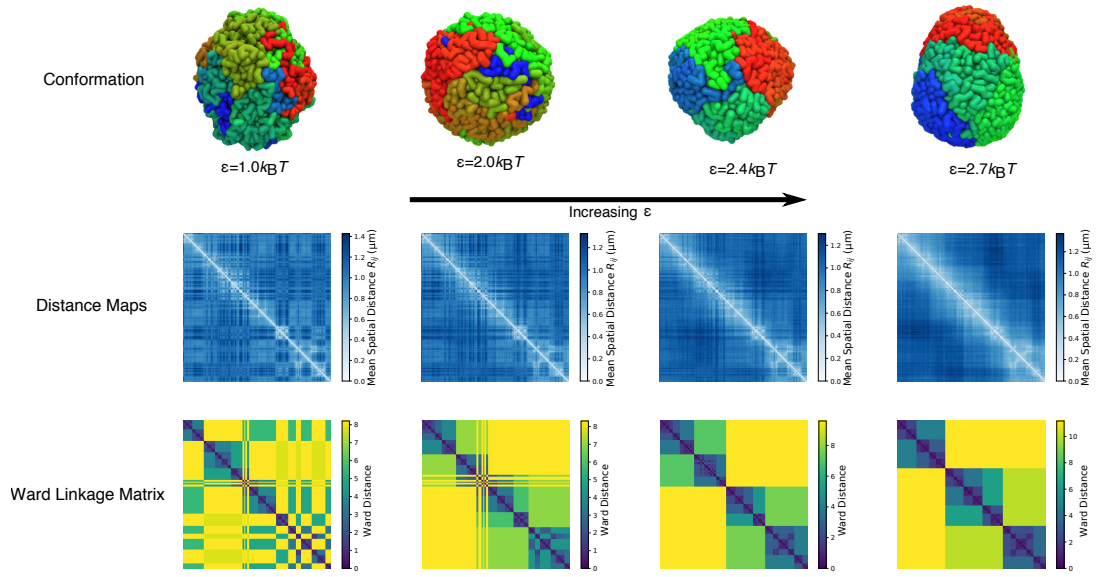


Figure 3.1: The top panel shows typical structures of the simulated folded Chr5 for $\epsilon = (1.0, 2.0, 2.4, 2.7)k_{\mathrm{B}}T$ from left to right. The color indicates the index of the locus, from the $5'$ to $3'$-end. The spatial distance map and the corresponding Ward Linkage Matrices (WLMs) are shown in the middle and bottom panels, respectively.

$P(s)$ in Fig. 3.2b, shows that there are two scaling regimes. As ϵ increases, the probability of short-range (small s) increases by several folds, while $P(s)$ for large s decreases by approximately an order of magnitude. In particular, for $\epsilon = 1.0k_{\mathrm{B}}T$, $P(s)$, decreases much faster compared to experiments at small s. In contrast, I find that at $\epsilon = 2.4k_{\mathrm{B}}T$, $P(s) \sim s^{-0.75}$ for $s < 0.5$ Mbps and when s exceeds ~ 0.5 Mbps, $P(s) \sim s^{-1.25}$ (red curve in Fig. 3.2b). Such a behavior, with $P(s)$ exhibiting two distinct scaling regimes, agrees with experiments (black line in Fig. 3.2b). It is worth pointing out that the two-scaling regimes in $P(s)$ is a robust feature of all 23 Human interphase chromosomes (Fig. 3.2c). It is clear the two scaling regimes in $P(s)$ with a crossover from one to another at $s \approx 3 \cdot 10^5$bps $\sim 6 \cdot 10^5$bps is universally found in all the chromosomes. Interestingly, our simulation suggests that the crossover scale in $P(s)$ coincides with the size of the chromosome droplets (see discussion).

3.2.2 Active and repressive loci micro-phase segregate

Comparison of the contact maps between simulations and experiments illustrates that compartment formation appearing as plaid or checkerboard patterns in Fig. 3.2d, shows good agreement with Hi-C data [21, 28]. The dashed rectangles mark the border of one such compartment enriched predominantly with interactions between loci of the same type, suggesting that compartments are formed through the clustering of the chromatin segments with the same epigenetic states. A previous experimental study ~~also~~ suggests that the chromatin structuring in Topologically

Associated Domains (TADs) is also driven by the epigenome feature [165]. In order to make the comparison precise, I treated the contact maps as probabilistic matrices and used a variety of mathematical methods to quantitatively compare large matrices. First, the checkerboard pattern in the contact map is more prominent when illustrated using the Spearman correlation map (see Appendix A.1 and Figs. A.1 and A.2). Second, to quantitatively compare the simulated results with experiments, I use the spectral co-clustering algorithm [166] to bi-cluster the computed Spearman correlation map (see Appendix A.1 and Appendix A.2). Other methods, such as PCA [21] and k-means [28], have been used to extract the compartment features. Finally, the similarity between the simulated and experimental data is assessed using the Adjusted Mutual Information Score (AMI) (Appendix A.2). The CCM model, based only on epigenetic information and the locations of the loop anchors, yields an AMI score that results in correctly reproducing $\approx 81\%$ of the compartments obtained from the experimental data. In contrast, a pseudo homopolymer model with $\epsilon_{\mathrm{AA}} = \epsilon_{\mathrm{BB}} = \epsilon_{\mathrm{AB}} = \epsilon$, which has the same loop anchors as the CCM, has an absolute AMI score that is 200 times smaller (Fig. A.3), and does not lead to the formation of compartments (correctly reproducing only $\approx 51\%$ of the compartments, no better than random assignments). Thus, the CCM is the minimal model needed to reproduce the essential features found in the contact map.

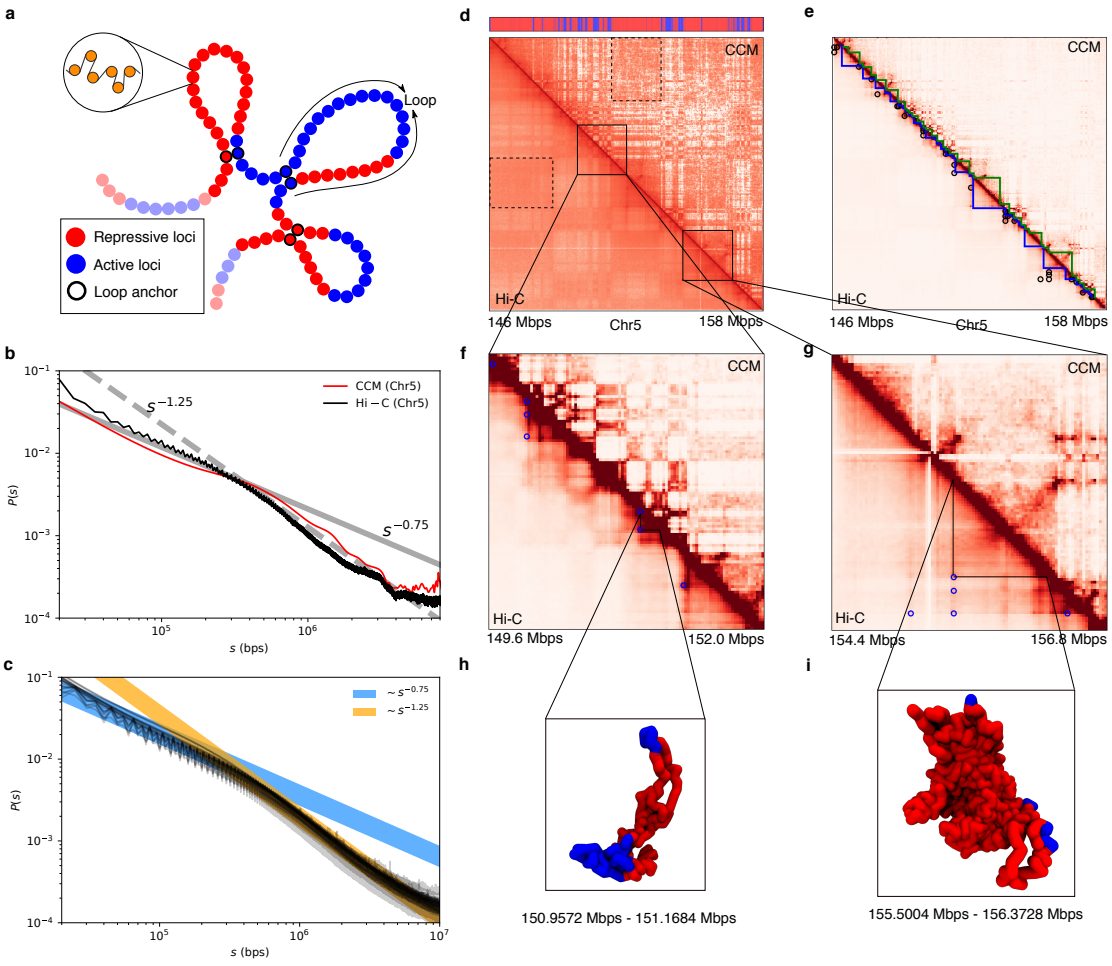

a
Loop
Repressive loci
Active loci
Loop anchor
b
CCM (Chr5)
Hi−C (Chr5)
$s^{-1.25}$
$s^{-0.75}$
P(s)
s (bps)
c
$\sim s^{-0.75}$
$\sim s^{-1.25}$
P(s)
s (bps)
d
CCM
Hi-C
146 Mbps
Chr5
158 Mbps
e
CCM
Hi-C
146 Mbps
Chr5
158 Mbps
f
CCM
Hi-C
149.6 Mbps
152.0 Mbps
g
CCM
Hi-C
154.4 Mbps
156.8 Mbps
h
150.9572 Mbps - 151.1684 Mbps
i
155.5004 Mbps - 156.3728 Mbps

Figure 3.2: Comparison between the simulated contact map and the Hi-C contact map. **(a)** A sketch of the Chromosome Copolymer Model (CCM). Each bead represents 1,200 basepairs (representing roughly six nucleosomes connected by linker DNAs). Blue (Red) corresponds to active (repressive) loci. The examples of three pairs of loop anchors (in this cartoon) are marked by beads with black boundaries. **(b)** Comparison between experimental data [28] (black) and simulated $P(s)$. Dashed and solid lines are plots of $s^{-1.25}$ and $s^{-0.75}$, respectively. The crossover point between the two scaling regimes at $s^* \sim 3 \cdot 10^5$ bps is noticeable in both the experimental and simulated results. **(c)** Experimental contact probability $P(s)$ for the 23 human interphae chromosomes calculated from the Hi-C data in [28]. Each black curve, all of which almost superimpose on each other, corresponds to one chromosome. Blue and orange lines are guides to the eye showing two scaling regimes. **(d)** Comparison of the contact maps inferred from Hi-C experiment [28] (lower triangle) and obtained from simulations (upper triangle) results. For easier visualization, the values of the contact probability are converted to a $\log_2$ scale. The bar above the map marks the epigenetic states with blue (red) representing active (repressive) loci. The dashed black box is an example of a compartment. Such compartment-like structures emerge due to contacts between loci separated by large genomic distances, which gives rise to spatial order in the organized chromosome. **(e)** Illustration of Topologically Associated Domains (TADs). The blue and green triangles are from experiments and simulations, respectively. The black circles mark the positions of loops detected from experiment data, which are formed by two CTCF motifs. **(f)** The zoom in of the diagonal region for the chromosome segment between 149.6 Mbps to 152.0 Mbps. The blue circle marks the positions of CTCF loops found in the experiment [28]. **(g)** Same as **(f)** except for 154.4 Mbps to 156.8 Mbps. **(h)** and **(i)**. Snapshots of two TADs, marked by the blue triangles in **(f)** and **(g)**, respectively.

The inset in Fig. 3.3a, displaying a typical snapshot of the condensed chromosome, reveals that active (A, blue) and repressive (B, red) loci are clustered together, undergoing micro-phase separation (see Methods for definition of active and repressive loci). The tendency to segregate is vividly illustrated in the radial distribution functions $g_{\rm AA}(r)$, $g_{\rm BB}(r)$ and $g_{AB}(r)$, which shows (Fig. 3.3a) that $g_{\rm AA}(r)$ and $g_{\rm BB}(r)$ have much higher values than $g_{\rm AB}(r)$ implying that active and repressive loci form the clusters of their own, and do not mix. Such a micro-phase separation between the A-rich and B-rich regions directly gives rise to compartments in the

contact map. Interestingly, the normalized radial density (Fig. 3.3b) shows that active chromatin exhibits a peak at large radial distance, r implying that the active loci localize on the periphery of the condensed chromosome whereas repressive chromatin is more homogeneously distributed. Visual inspection of the simulation trajectories also suggests that active and repressive chromatins are often separated in a polarized fashion, in accord with a recent experimental study [14], which shows that the two compartments are indeed similarly spatially arranged.

3.2.3 Spatial organization of the compact chromosome

In order to illustrate the spatial organization of the chromosome, I introduce the distance function,

$$R(s) = \left\langle \sum_{i<j}^{N} \frac{(\mathbf{r}_i - \mathbf{r}_j)^2 \delta(s - |i-j|)}{N - s} \right\rangle^{1/2} \tag{3.1}$$

where $\langle \cdot \rangle$ denotes both an ensemble and time average. I calculated $R(s)$, the mean end-to-end distance between the loci, by constraining the genomic distance $|i-j|$ to s. If the structured chromosome is maximally compact on all length scales, I expect $R(s) \sim s^{1/3}$ for all s. However, the plot of $R(s)$ on a log-log scale shows that in the range $10^5 \lesssim s \lesssim 10^6$ bps, $R(s) \sim s^{0.2}$. The plateau at large s arises due to s reaching the boundary of the compact structure. The inset in Fig. 3.4a, comparing the simulation result and experimental data [14], both show the same scaling for $R(s)$ as a function of s. Note that in [14] spatial distances are measured between centroids of TADs domains rather than individual loci.

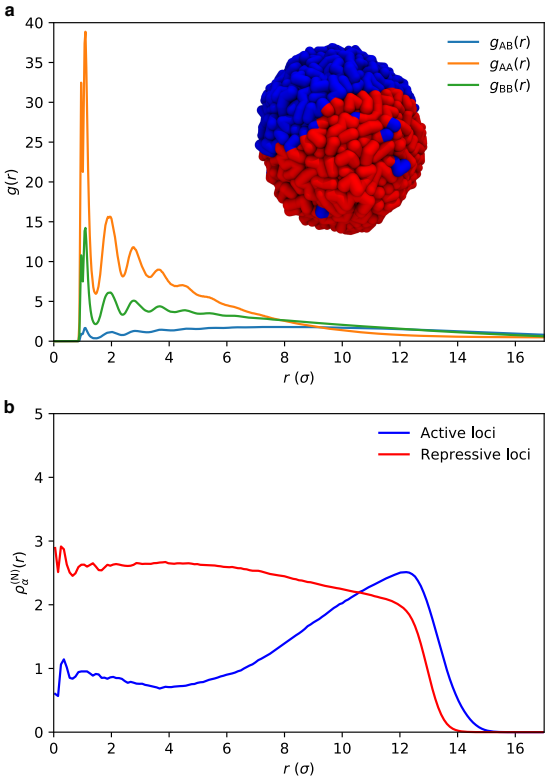


Figure 3.3: Micro-phase separation between active and repressive loci. **(a)** Radial distribution functions, $g(r)$, as a function of r (in the unit of σ) between active-active loci ($g_{\mathrm{AA}}(r)$), repressive-repressive loci ($g_{\mathrm{BB}}(r)$) and active-repressive loci ($g_{\mathrm{AB}}(r)$). The inset shows the typical conformation of the compact chromosome. Blue and red segments correspond to active and repressive loci, respectively. The structure vividly reveals micro-phase separation between active and repressive loci. **(b)** The normalized radial density, $\rho_\alpha^{(\mathrm{N})}(r) = \langle N_\alpha(r)\rangle/(4\pi r^2 \Delta r N_\alpha)$, where $N_\alpha(r)$ is the number of loci of given type α found in the spherical shell between r and $r + \Delta r$, N_α is the total number of loci of that type. The bracket $\langle\cdot\rangle$ is the ensemble average, V is the volume of the globule, given by $(4/3)\pi r_{\mathrm{max}}^3$ where $r_{\mathrm{max}} = 17\sigma$; $\rho_\alpha^{(\mathrm{N})}(r)$ shows that the active loci are predominantly localized on the periphery of the condensed chromosome. The repressive loci are more uniformly distributed.

By a systematic analysis of the FISH data, Wang et al [14] established that the probability of contact formation between loci i and j, P_{ij}, is inversely proportional to a power of their mean spatial distance $R_{ij} = \langle|\mathbf{r}_i - \mathbf{r}_j|\rangle$, with the latter providing a direct picture of the spatial organization. Similarly, in this work, I explored the relation between C_{ij} and R_{ij} where $C_{ij}(= P_{ij}\sum_{i<j} C_{ij} \propto P_{ij})$ is the number of contacts between loci i and j that are recorded in the simulations. The heat map of $(1/C_{ij}, R_{ij})$ in Fig. 3.4b shows that the two matrices are proportional to each other. In accord with the FISH data [14], I find that $1/C_{ij} \propto R_{ij}^{\lambda}$ where $\lambda \approx 4$, suggesting that larger mean spatial distance between loci i and j implies smaller contact probability, which is the usual assumption when experimental Hi-C data is used to infer three-dimensional chromosome organization. The decrease of C_{ij} with increasing R_{ij} with a large value of λ, is unexpected but is an important finding needed to link contact maps and spatial structures.

The slope of the dashed line in Fig. 3.4b obtained using the data in [14], is 4.1, which coincides with our simulation results. Mean field arguments [167] suggest that $P(s) \sim R(s)^{-3}$, which follows from the observation that the end of the chain is uniformly distributed over a volume $R^3(s)$. This is neither consistent with our simulations nor with experiments, implying that the distribution of the chain ends is greatly skewed. Although both the simulated and experimental results establish a strong correlation between $R(s)$ and $P(s)$, such a correlation is only valid in an ensemble sense (see Chapter 4 and 5 as well as [168]).

3.2.4 Topologically Associated Domains and their shapes

Our model reproduces Topologically Associated Domains (TADs), depicted as triangles along the diagonal in Fig. 3.2e, of an average length of 200 kbps along the diagonal of the contact map in which the interactions between the loci are greatly enhanced. It has been noted [28] that in a majority of cases, boundaries of the TADs are marked by a pair of CTCF motifs with a high probability of interaction between them. They are visualized as peaks in the Hi-C map (Fig. 3.2e). To quantitatively detect the boundaries of the TADs, I adopt the procedure described in [25] to identify the position of each TAD. The boundaries of the TADs, shown in blue (Hi-C data) and green (simulations) are reproduced by the CCM (Fig. 3.2e).

To investigate the sizes and shapes of each individual TADs (defined as CTCF loops in the simulations), I calculated the radii of gyration, $R_{\rm g}$, the relative shape anisotropies κ^2, as well as the shape parameters, S, for 32 TADs (see Appendix A.4 for details). These TADs are typical representations of all TADs. The genomic size of the 32 TADs is similar to the genome-wide distribution. The results are shown in Fig. A.4. The mean $R_{\rm g}$ for each individual TADs scales as their genomic length with exponent 0.27, which is an indicator of the compact structures for the TADs. However, unlike compact globular objects, their shapes are far from being globular and are much more irregular with smaller TADs adopting more irregular shapes compared to the larger TADs (see $\langle\kappa^2\rangle$ and $\langle S\rangle$ as a function of TAD size in Fig. A.4). Such compact but irregularly shaped nature of TADs are vividly illustrated by typical snapshots for the two TADs (Figs. 3.2h, i). How can I understand this

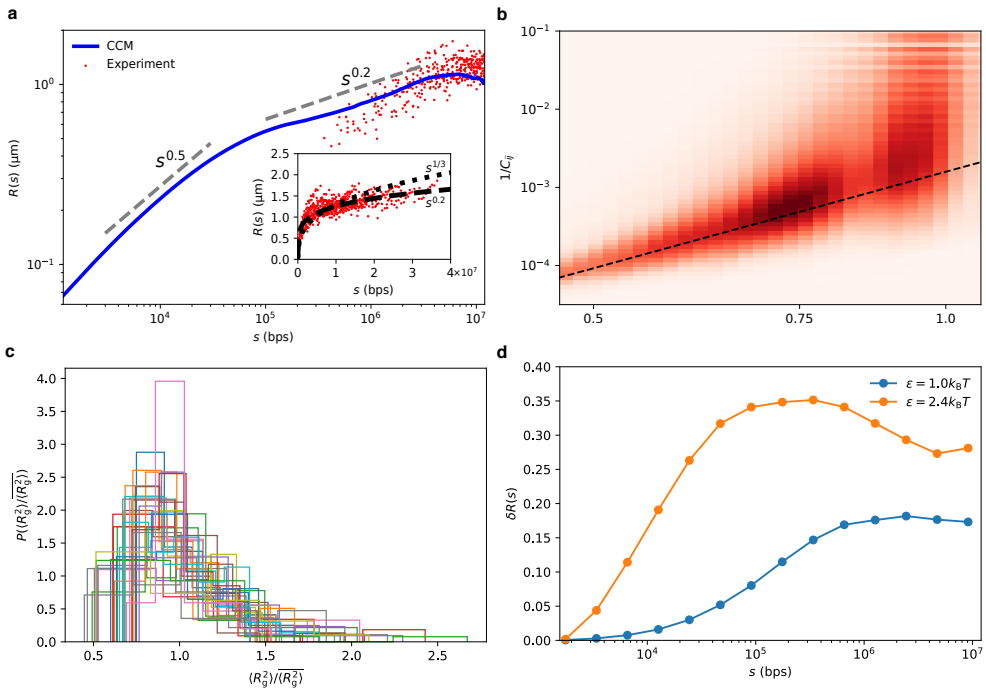


Figure 3.4: Organization and fluctuations of the chromosome structures. **(a)** The dependence of the spatial distance $R(s)$ (Eq.1) on the genomic distance, s. Grey dashed lines, indicating the slopes, are guides to the eye. The red dots are experimental data taken from [14] for $s < 1.2\times10^7$bps. The inset shows the complete set of experimental data. Short dashed and long dashed lines are $s^{1/3}$ and $s^{0.2}$, respectively. At small s ($s < 10^5$bps), $R(s) \sim s^{0.5}$ implying that chromatin behaves as almost an ideal chain. **(b)** The heatmap of the 2D histogram of $(R_{ij}, 1/C_{ij})$. The dashed black line is the curve with scaling exponent 4.1, which coincides with the value obtained by fitting the experimental data [14]. **(c)** Distribution $P(\langle R_g^2\rangle/\overline{\langle R_g^2\rangle})$, where $\langle R_g^2\rangle$ is the time average value of the squared radius of gyration of a single trajectory and $\overline{\langle R_g^2\rangle}$ is the mean value averaged over all independent trajectories. Different colors represent $P(\langle R_g^2\rangle/\overline{\langle R_g^2\rangle})$ for the thirty-two individual TADs. The distribution is surprisingly wide which suggests that TAD structures vary from cell-to-cell. **(d)** Coefficient of variation $\delta R(s) = (\langle R^2(s)\rangle - \langle R(s)\rangle^2)^{1/2}/\langle R(s)\rangle$, computed from simulations, shows a non-monotonic dependence on s for $\epsilon = 2.4k_BT$, increasing till $s \sim 10^5$bps and decreases at larger s.

non-trivial highly aspherical shapes of the TADs when the chromosome is spherical on long length scales (several Mbps)? Since TADs are constrained by the CTCF loops, they may be viewed locally as ring polymers. Ring polymers in a melt are compact [169] objects but adopt irregular shapes, consistent with our prediction for TADs.

I then wondered if TADs in each individual cells have similar sizes and shapes. I computed the dispersion in R_g, κ and S (Fig. 3.4c and Figs. A.4 and A.5) among different trajectories. Fig. 3.4c shows the $P(\langle R_g^2\rangle/\overline{\langle R_g^2\rangle})$, of the mean square radius of gyration $\langle R_g^2\rangle$ for the thirty-two Chr 5 TADs in each trajectory normalized by the average $\overline{\langle R_g^2\rangle}$ of each individual TAD. The bracket (bar) is the time (ensemble) average. The large dispersion in $P(\langle R_g^2\rangle/\overline{\langle R_g^2\rangle})$ (Fig. 3.4c) as well as $P(\langle\kappa\rangle/\overline{\langle\kappa\rangle})$ and $P(\langle S\rangle/\overline{\langle S\rangle})$ (Fig. A.5) suggest that TADs are fluctuating objects, which exhibit substantial cell-to-cell variations. Our result supports the recent FISH [170] and single-cell Hi-C experimental findings [79, 171] and imaging experiments [30], showing that individual TAD compaction varies widely from highly extended to compact states among different cells. To decipher how the variation of the structure of the chromosome changes as a function of s, I calculated the coefficient of variation, $\delta R(s) = (\langle R_s^2\rangle - \langle R_s\rangle^2)^{1/2}/\langle R(s)\rangle$. Interestingly, $\delta R(s)$ first increases with s up to $s \approx 10^5 \sim 10^6$ bps and then decreases as s further increases (Fig. 3.4d). Higher resolution experiments are needed to resolve the variance for $s < 10^5$ bps. The predicted non-monotonic dependence of $\delta R(s)$ on s is amenable to experimental test.

3.2.5 Chromosome Structures in terms of the Ward Linkage Matrix

To quantitatively analyze the spatial organization of the compact chromosome, I use the unsupervised agglomerative clustering algorithm to reveal the hierarchy organization on the different length scales. A different method, which is also based on clustering techniques, has recently been applied to Hi-C contact map [172]. I use the Ward Linkage Matrix (WLM) (see Appendix A.3 for details), which is directly applicable to the spatial distance matrix, $\mathbf{R}$ in which the element, $R_{ij} = \langle|\mathbf{r}_i - \mathbf{r}_j|\rangle$, is the mean spatial distance between the loci i and j. I also constructed the experimental WLM by converting the Hi-C contact map to a distance map by exploiting the approximate relationship between R_{ij} and P_{ij} ($\propto R_{ij}^{-4.1}$) discussed previously (also see Fig. 3.4b). The advantages of using distance matrices instead of contact maps are two folds. First, matrix $\mathbf{R}$ is a direct depiction of the three-dimensional organization of the chromosome. The WLM, constructed from $\mathbf{R}$ is a cophenetic matrix, which can be used to reveal the hierarchical nature of the chromosome organization. Second, the contact map matrix elements do not obey triangle inequality. Therefore, it is not a good indicator of the actual 3D spatial arrangement of the loci. I show the WLM for the two ϵ values (upper panel in Fig. 3.5) and the comparison between WLM computed based on experimental data and WLM for $\epsilon = 2.4k_{\mathrm{B}}T$ (lower panel in Fig. 3.5). Visual inspection of the WLMs for $\epsilon = 2.4k_{\mathrm{B}}T$ shows distinct segregation in the spatial arrangement of the loci. It is clear from Fig. 3.5 that the experimentally inferred WLM, constructed from Hi-C data, and simulations result with $\epsilon = 2.4k_{\mathrm{B}}T$ are almost identical. From the WLMs

for both $\epsilon = 1.0k_BT$ and $\epsilon = 2.0k_BT$ (Fig. 3.1), I surmise that loci with large genomic separation s are in spatial proximity, which is inconsistent with the experimental WLM. The Pearson correlation coefficient between experimental result and CCM using $\epsilon = 2.4k_BT$ is 0.96 (0.53 for $\epsilon = 1.0k_BT$, 0.84 for $\epsilon = 2.0k_BT$ and 0.75 for $\epsilon = 2.7k_BT$). Thus, the poorer agreement between the simulated WLM (Fig. 3.1) as well as Spearman correlation matrix (Fig. A.1) using $\epsilon = (1.0, 2.0, 2.7)k_BT$ and experiments, compared to $\epsilon = 2.4k_BT$, further justifies the latter as the optimum value in the CCM. I find it remarkable that the CCM, with only one adjusted energy scale (ϵ) is sufficient to produce such a robust agreement with experiments.

3.2.6 Cell-to-cell variations in the WLM

To assess the large structural variations between cells, I calculated the WLM for individual cells. I obtain the single cell WLM using time averaged distance map of individual trajectories. Fig. 3.6 shows that there are dramatic differences between the WLM for individual cells, with the ensemble average deviating greatly from the patterns in individual cells. Thus, the chromosome structure is highly heterogeneous. These findings are reflected in the small mean value of Pearson correlation coefficients ρ between all pairs of cells (Fig. 3.6b). The distribution $P(\rho)$ has mean $\bar{\rho} = 0.2$ with a narrow shape, implying little overlap in the WLMs between any two cells.

In order to make quantitative comparisons to experimental data, with the goal of elucidating large-scale variations in the spatial organizations of human interphase

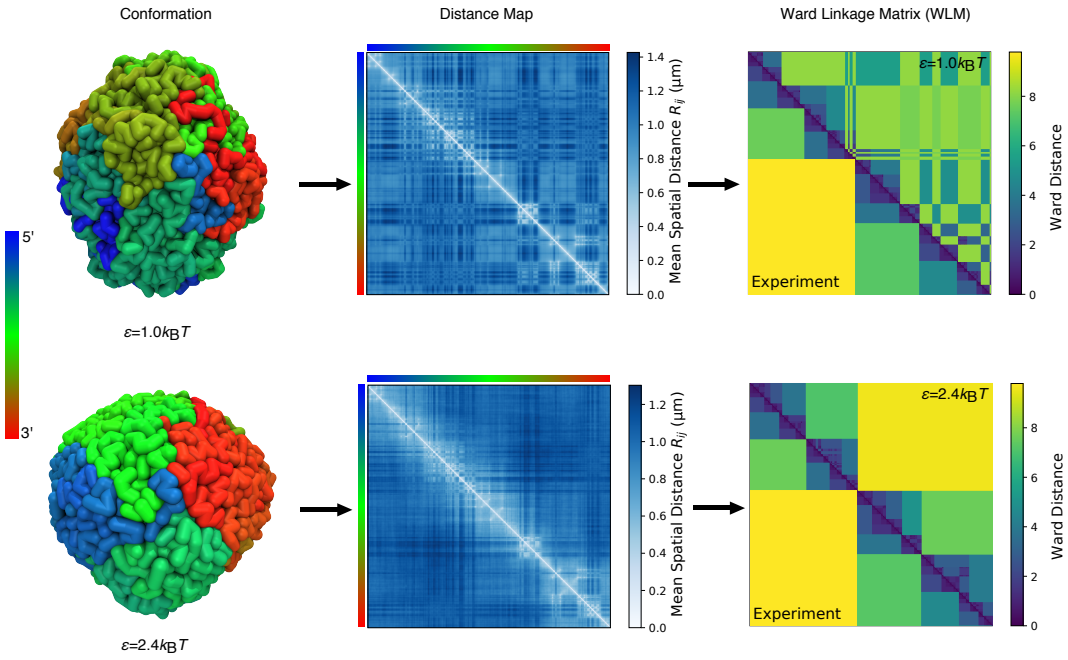


Figure 3.5: Chromosome structure in terms of Ward Linkage Matrix (WLM). **(Left)** Typical conformations of the organized chromosome for $\epsilon = 1.0k_{\mathrm{B}}T$ (upper) and $2.4k_{\mathrm{B}}T$ (bottom). The coloring corresponds to genomic distance from one endpoint, ranging from red to green to blue. **(Middle)** The ensemble averaged distance maps. **(Right)** Comparison between the simulated Ward Linkage Matrices (WLMs) (upper triangle) and the experiment WLM (lower triangle) inferred from Hi-C contact map. Ward distance is defined in the Appendix A.3

chromosomes, I constructed single cell WLMs for Chr 21 using the spatial distance data provided in [14] and computed the corresponding $P(\rho)$ (Fig. 3.6b). The results show that the experimentally organization of Chr 21 *in vivo* also exhibits large variations manifested by the distribution $P(\rho)$ covering a narrow range of low values of ρ with a small mean $\bar{\rho} = 0.25$. Comparison to simulated result suggest that Chr 21 shows a slightly lower degree of structural heterogeneity (with a modestly larger mean $\bar{\rho} = 0.25$) compared to Chr 5 investigated using CCM. Nevertheless, both the simulated and experimental results indicate that human interphase chromosomes do not have any well-defined "native structure". To investigate whether Chr 5 has a small number of spatially distinct structures, I show two-dimensional t-SNE (t-distributed stochastic neighboring embedding) representation of 90 individual WLMs of the metric $\sqrt{1-\rho}$ (Fig. 3.6c). It is clear that there is no dominant cluster, indicating that each Chr 5 in single cells is organized differently rather than belonging to a small subset of conformational states. Such large cell-to-cell variations in the structures, without a small number of well defined states, is another hallmark of glasses, which are also revealed in recent experiments [79,118]. The presence of multiple organized structures has profound consequences on the chromosome dynamics (see below).

3.2.7 Chromosome dynamics is glassy:

I probe the dynamics of the organized chromosome with $\epsilon = 2.4 k_{\mathrm{B}} T$, a value that yields the best agreement with the experimental Hi-C contact map. I first

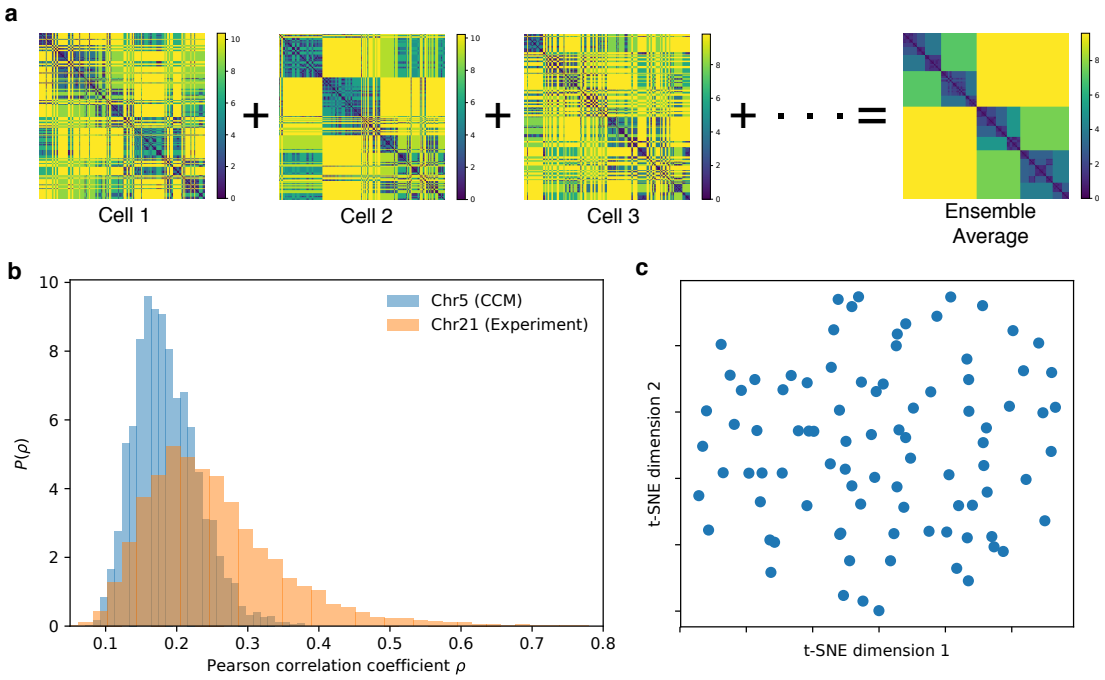


Figure 3.6: Structural heterogeneity in the chromosome. **(a)** Ward Linkage Matrices of different individual cells. The single cell WLM is the time average result over a single trajectory. The ensemble average WLM (rightmost) and the experimental WLM are in clear quantitative agreement (Fig. 3.5). However, the spatial organization show large variations from cell to cell. Each cell has very different WLM, implying their structures are distinct. (**b**) The distribution of ρ, $P(\rho)$, with a mean $\bar{\rho} = 0.2$ (blue curve), where ρ is the pearson correlation coefficient between WLMs of any two cells. The $P(\rho)$ distribution, spanning the low range of ρ values, is a further demonstration of structural heterogeneity in individual cells. In yellow I plot $P(\rho)$ with $\bar{\rho} = 0.25$ for 120 individual human interphase Chr 21, computed using the single cell WLMs constructed from experimental measured spatial distance data provided in [14]. **(c)** Two-dimensional t-SNE (t-distributed stochastic neighboring embedding) visualizations of WLM of simulated individual Chr 5 using the distance metric $\sqrt{1-\rho}$.

calculated the incoherent scattering function, $F_{\rm s}(k,t) = (1/N)\left\langle \sum_{j=1}^{N} \mathrm{e}^{\mathrm{i}\mathbf{k}(\mathbf{r}_j(t)-\mathbf{r}_j(0))} \right\rangle$ where $\mathbf{r}_j(t)$ is the position of j^{th} loci at time t. The decay of $F_{\rm s}(k,t)$ (orange line in Fig. 3.7a) for $k \sim 1/r_{\rm s}$ ($r_{\rm s}$ is the position of the first peak in the radial distribution function ($g_{\rm AA}(r)$ and $g_{\rm BB}(r)$) (Fig. 3.3a)) is best fit using the stretched exponential function, $F_{\rm s}(k,t) \sim e^{-(t/\tau_\alpha)^\beta}$ with a small stretching coefficient, $\beta \approx 0.27$. The stretched exponential decay with small β is another hallmark of glassy dynamics. For comparison, $F_{\rm s}(k,t)$ decays exponentially for $\epsilon = 1.0 k_{\rm B}T$, implying liquid-like dynamics (blue line in Fig. 3.7a).

In the context of relaxation in supercooled liquids, it has been shown that the fourth order susceptibility [173], $\chi_4(k,t) = N[\langle F_{\rm s}(k,t)^2\rangle - \langle F_{\rm s}(k,t)\rangle^2]$ provides a unique way of distinguishing between fluctuations in the liquid and frozen states. As in structural glasses, the value of $\chi_4(k,t)$ increases with t reaching a peak at $t = t_{\rm M}$ and decays at longer times. The peak in the $\chi_4(k,t)$ is an indication of dynamic heterogeneity, which in the chromosome is manifested as dramatic variations in the loci dynamics (see below). For $\epsilon = 2.4 k_{\rm B}T$, $\chi_4(k,t)$ reaches a maximum at $t_M \approx 1s$ (Fig. 3.7b), which surprisingly, is the same order of magnitude ($\sim 5s$) in which chromatin movement was found to be coherent on a length scale of $\approx 1\mu$m [41]. The dynamics in $F_{\rm s}(k,t)$ and $\chi_4(k,t)$ together show that human interphase chromosome dynamics is glassy [58], and highly heterogeneous.

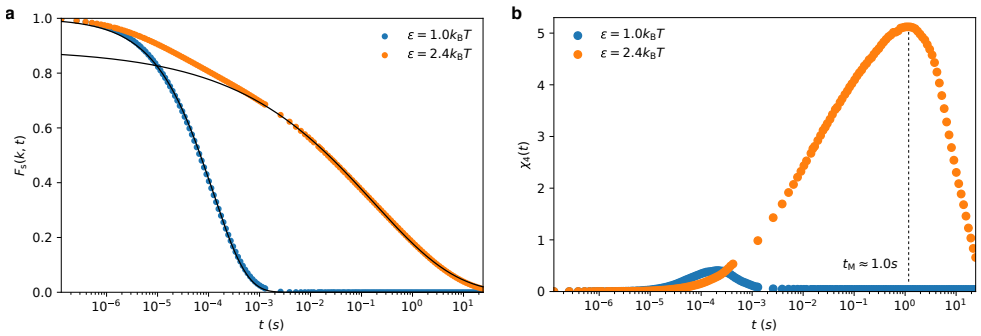


Figure 3.7: Chromosomes exhibit glassy dynamics. **(a)** Intermediate scattering function obtained for $\epsilon = 1.0k_BT$ (blue) and $\epsilon = 2.4k_BT$ (orange). The line shows an exponential function fit, $F_s(k,t)$, for $\epsilon = 1.0k_BT$. For $\epsilon = 2.4k_BT$, $F_s(k,t) \sim e^{-(t/t_\alpha)^\beta}$ with $\beta = 0.27$, for t exceeding a few milliseconds (black curve). **(b)** The fourth order susceptibility, $\chi_4(t)$, used as a function to demonstrate dynamic heterogeneity. The peak in $\chi_4(t)$ for $\epsilon = 2.4k_BT$ around $t_M \approx 1$s is a signature of heterogeneity.

3.2.8 Single loci Mean Square Displacements are heterogeneous:

In order to ascertain the consequences of glassy dynamics at the microscopic level, I plot the MSD, $\Delta(t) = \frac{1}{N}\langle\sum_{j=1}^{N}((\mathbf{r}_j(t) - \mathbf{r}_{com}(t)) - (\mathbf{r}_j(0) - \mathbf{r}_{com}(0))^2\rangle$ in Fig. 3.8 where $\mathbf{r}_{com}$ is the position of center of mass of the whole chromosome, from which a few conclusions can be drawn.

1. Because of the polymeric nature of the chromosome, the maximum excursion in $\Delta(t \to \infty) = 2R_g^2$, where $R_g \approx 0.7$ μm is the radius of gyration of Chr 5. Consequently, for both $\epsilon = 1.0k_BT$ (red) and $\epsilon = 2.4k_BT$, $\Delta(t)$ in the long time limit is smaller than $2R_g^2$. For $\epsilon = 2.4k_BT$ (green), $\Delta(t)$ shows a crossover at $t \approx 10^{-2}$s from slow to a faster diffusion, another indicator of glassy dynamics [174]. The slow diffusion is due to caging by neighboring loci, which is similar

to what is typically observed in glasses. The plateau in $\Delta(t)$ (Fig. 3.8a) is not pronounced, suggesting that the compact chromosome is likely on the edge of glassiness. The crossover is more prominent in the time-dependence of the mean squared displacement of single loci (see below). The slow diffusion predicted from the CCM is in accord with a number of experiments (Fig. 3.9). In contrast, diffusion coefficients measured in experiments are one or two orders of magnitude smaller than the system exhibiting liquid-like behavior, which further supports the glassy dynamics for mammalian chromosomes predicted here.

2. The two dashed lines in Fig. 3.8a show $\Delta(t) \sim t^{\alpha}$ with $\alpha = 0.45$. The value of s is close to 0.5 for the condensed polymer, which can be understood using the following arguments. The total friction coefficient experienced by the whole chain is the sum of contributions from each of the N monomers, $\xi_T = N\xi$. The time for the chain to move a distance $\approx R_{\rm g}$ is $\tau_R = R_{\rm g}^2/D_R \sim N^{2\nu+1}$. Let us assume that the diffusion of each monomer scales as Dt^{α}. If each monomer moves a distance on the order of $R_{\rm g}$ then the chain as a whole will diffuse by $R_{\rm g}$. Thus, by equating $D\tau_R^{\alpha} \sim R_{\rm g}^2$, one get $\alpha = 2\nu/(2\nu+1)$. For an ideal chain $\nu = 0.5$, which recovers the prediction by Rouse model, $\alpha = 0.5$. For a self-avoiding chain, $\nu \approx 0.6$, one get $\alpha \approx 0.54$. For a condensed chain, $\nu = 1/3$, one get $\alpha = 0.4$, thus rationalizing the findings in the simulations. Similar arguments have been reported recently for dynamics associated with fractal globule [59] and for the $\beta-$polymer model [175]. Surprisingly, $\alpha = 0.45$ found in simulations

is in good agreement with recent experimental findings [42]. I also obtained a similar result using a different chromosome model [67], when the dynamics were examined on a longer length scale. The finding that there is no clear Rouse regime ($\alpha = 0.5$) is also consistent with several other experimental results (Fig. 3.9). I should note that distinguishing between the difference, 0.4 and 0.5, in the diffusion exponent is subtle. Additional experiments are needed to determine the accurate values of the diffusion exponents of Human interphase chromatin loci in different time regimes.

3. I also calculated the diffusion of a single locus (sMSD) defined as $\Delta_i(t) = \langle(\mathbf{r}_i(t_0 + t) - \mathbf{r}_i(t_0))^2\rangle_{t_0}$, where $\langle\cdot\rangle_{t_0}$ is the average over the initial time t_0. Distinct differences are found between the polymer exhibiting liquid-like and glassy-like dynamics. The variance in single loci MSD is large for $\epsilon = 2.4k_{\mathrm{B}}T$, illustrated in Fig. 3.8b, which shows 10 typical trajectories for $\epsilon = 1.0k_{\mathrm{B}}T$ and $\epsilon = 2.4k_{\mathrm{B}}T$ each. For glassy dynamics, I found that the loci exhibiting high and low mobilities coexist in the chromosome, with orders of magnitude difference in the values of the effective diffusion coefficients, obtained by fitting $\Delta_i(t) = D_\alpha t^{\alpha_i}$. Caging effects are also evident on the timescale as long as seconds. Some loci are found to exhibit caging-hopping diffusion, which is a hallmark in glass-forming systems [176, 177]. Interestingly, such caging-hopping process has been observed in Human cell some time ago [33].

4. The large variance in sMSD has been found in the motion of chromatin loci in both *E.coli* and Human cells [17, 34, 35, 38, 39]. To further quantify het-

erogeneities in the loci mobilities, I calculated the Van Hove function $P(\Delta x)$, $P(\Delta x|\Delta t) = \langle (1/N) \sum_{i=1}^{N} \delta(\Delta x - [x_i(\Delta t) - x_i(0)]) \rangle$. Figs. 3.8c, d show the $P(\Delta x|\Delta t)$ and normalized $P(\Delta x/\sigma|\Delta t)$ for $\epsilon = 2.4 k_{\mathrm{B}}T$ at different lag times Δt. For $\epsilon = 1.0 k_{\mathrm{B}}T$, Van Hove function is well fit by a Gaussian at different lag times Δt. In contrast, for chromosome with glassy dynamics, all the $P(\Delta x|\Delta t)$ exhibit fat tail, which is well fit by an exponential function at large values of Δx (Figs. 3.8c, d) at all δt values, suggestive of the existence of fast and slow loci [177].

5. The results in Fig. 3.8 allow us to make direct comparisons with experimental data to establish signatures of dynamic heterogeneity. I calculated the distribution of effective diffusion exponent α_i, $P(\alpha)$, where α_i is obtained by fitting the sMSD to $\sim t^{\alpha_i}$ within some lag time (Δt) range. Fig. 3.8e shows that $P(\alpha)$ calculated from simulations is in good agreement with experiments [40] in the same lag time range (0.42 s $< \Delta t <$ 10 s). The $P(\alpha)$ distribution in the range 10^{-6} s $< \Delta t <$ 0.42 s shows two prominent peaks, further validating the picture of coexisting fast and slow moving loci. The good agreement between the predictions of the CCM simulations with data, showing large variations of mobilities among individual loci *in vivo*, further supports our conclusion that organized chromosome dynamics is glassy. Interestingly, a recent computational study in which Human interphase chromosomes are modeled as a generalized Rouse chain suggests that the heterogeneity of the loci dynamics measured in live cell imaging is due to the large variation of cross-linking

sites from cell to cell [64]. Our model implies a different mechanism that the heterogeneity observed is a manifesto of the intrinsic glassy dynamics of chromosomes.

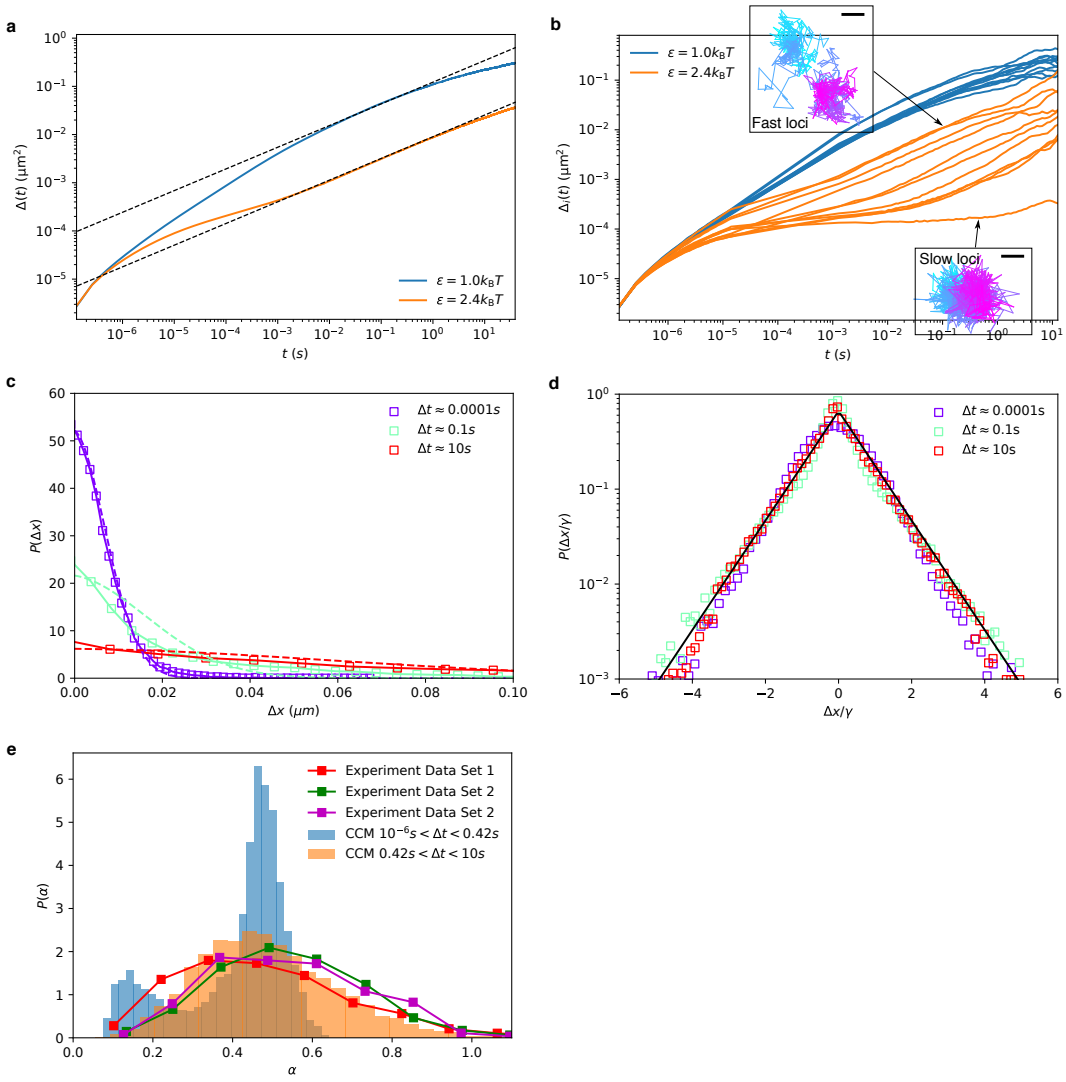


Figure 3.8: Dynamic heterogeneity of individual loci. (top) **(a)** Mean Square Displacement, $\Delta(t)$, as a function of time, t. The effective diffusion coefficients, D, computed from the fitted dashed lines are $0.122\mu\mathrm{m}^2/\mathrm{t}^{0.45}$and $0.009\mu\mathrm{m}^2/\mathrm{t}^{0.46}$ for $\epsilon = 1.0k_{\mathrm{B}}T$ and $\epsilon = 2.4k_{\mathrm{B}}T$, respectively. **(b)** Time dependence of 10 single loci MSD (sMSD, $\Delta_i(t)$) corresponding to $1^{st}, 1000^{th}, ..., 10,000^{th}$ loci for $\epsilon = 1.0k_{\mathrm{B}}T$ and $\epsilon = 2.4k_{\mathrm{B}}T$. The insets show $\Delta_i(t)$ for two trajectories for fast (top) and slow (bottom) loci. Cyan (Magenta) indicates short (long) lag times. The scale bar is 35 nm(0.07 nm) for fast (slow) loci. Caging effect can be clearly observed as the plateau in $\Delta_i(t)$ for $\epsilon = 2.4k_{\mathrm{B}}T$. (**c**) The Van Hove function $P(\Delta x)$ for $\epsilon = 2.4k_{\mathrm{B}}T$ at lag times $\Delta t = (0.0001, 0.1, 10)$s. $P(\Delta x)$ has heavy tail at large Δx and cannot be fit by a Gaussian (color dashed lines) except for $\Delta t = 0.0001$s at small Δx. **(d)** Same as **(c)** except displacement Δx is normalized by its standard deviation γ. $P(\Delta x/\gamma)$ for different lag times collapse onto a master curve. The black line is an exponential fit, $\sim e^{-\eta(\Delta x/\gamma)}$ with $\eta \approx 1.3$. (**E**) Distribution, $P(\alpha)$, of the effective diffusion exponent α. Comparison to experimental data [40] are shown. The values of α are extracted from single loci trajectories by fitting sMSD, $\Delta_i(t) \sim t^{\alpha}$. The lag time range $0.42\mathrm{s} < \Delta t < 10\mathrm{s}$ is in the approximate same range probed in the experiment. Experimental data set 1, 2, 3 are from Fig.2b, 2c, and Fig.S5 of [40], respectively. The results from our simulation (orange) agree well with experimental data, shown as orange. The blue bar plot is $P(\alpha)$ for small lag times 10^{-6} s $< \Delta t < 0.42$ s. It shows two peaks, indicating the coexistence of two populations of loci with distinct mobilities.

3.2.9 Active loci has higher mobility:

Fig. 3.10a shows MSD for active and repressive loci. For $\epsilon = 1.0k_{\mathrm{B}}T$, there is no difference between active and repressive loci in their mobilities. However, in the glassy state active loci diffuses faster than the repressive loci. The ratio between the effective diffusion coefficients (the slope of the dashed line) of the active and repressive loci is $0.0116/0.008 \simeq 1.45$, in good agreement with experimental estimate $0.018/0.013 \simeq 1.38$ [42]. These variations are surprising since the parameters characterizing the A-A and B-B interactions are identical. To investigate the origin of the differences between the dynamics of A and B loci, I plot the displacement

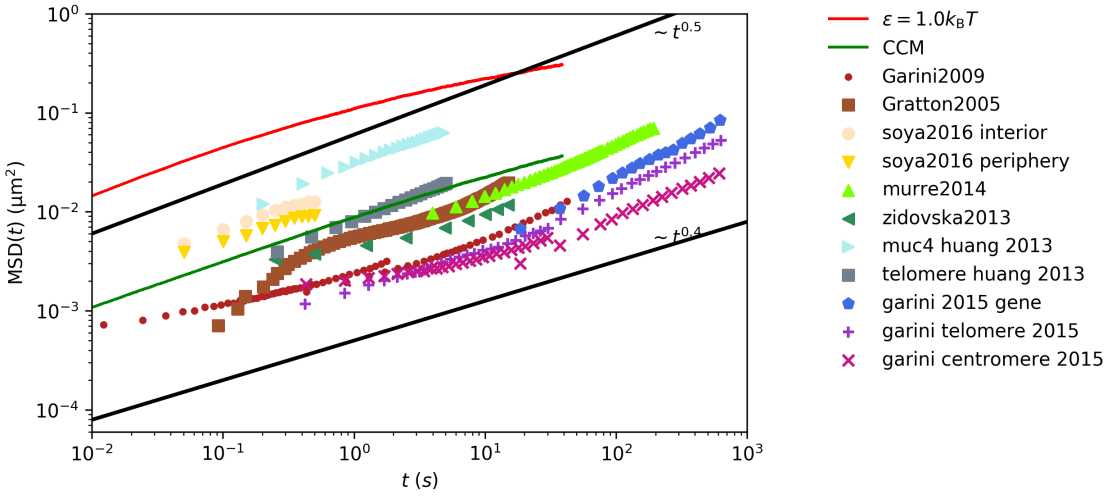


Figure 3.9: Chromosomes exhibit glassy dynamics.]MSD(t) experimental data collected from a number of works for human interphase cells. The simulation data for $\epsilon = 1.0k_BT$ and CCM is also plotted for comparison. The experimental data are taken from Bronstein et al., 2009 [34], Levi et al., 2005 [33], Shinkai et al., 2016 [42], Lucas et al., 2014 [39], Zidovska et al., 2013 [41], Chen et al., 2013 [17] and Bronshtein et al., 2015 [40]

vectors of the loci across the cross-section of the condensed chromosome (Fig. 3.10b) for a time window $\Delta t = 0.1$s. The loci on the periphery have much greater mobility compared to the ones in the interior. In sharp contrast, the fluid-like state exhibits no such difference in the mobilities of A and B (Fig. 3.10d). To quantify the dependence of the mobility on the radial position of the loci, I computed the amplitude of the displacement normalized by its mean, as a function of the radial position of the loci, r (Fig. 3.10c). For the chromosome exhibiting glass-like behavior, the mobility increases sharply around $r \approx 0.7$ μm whereas it hardly changes over the entire range of r in the fluid-like system. Because the active loci are mostly localized on the periphery and the repressive loci are in the interior (Fig. 3.3b), the results in Fig. 3.10 suggest that the differences in the mobilities of the loci with different epigenetic states are due to their preferred locations in the chromosome. It is intriguing that glassy behavior is accompanied by a position-dependent mobility, which can be understood by noting that the loci in the interior are more caged by the neighbors, thus restricting their movement. In a fluid-like system, the cages are so short-lived that the apparent differences in the environments the loci experience are averaged out on short timescales. Note that in the experimental result [42] comparison is made between the loci in the periphery and interior of the nucleus. It is well known that the nucleus periphery is enriched with heterochromatin (repressive loci) and the interior is enriched with euchromatin (active loci). However, for an individual chromosome, single cell Hi-C study [79] and other experimental studies [2, 81, 82, 178] suggest that the active loci are preferentially localized at the surface of the chromosome territory.

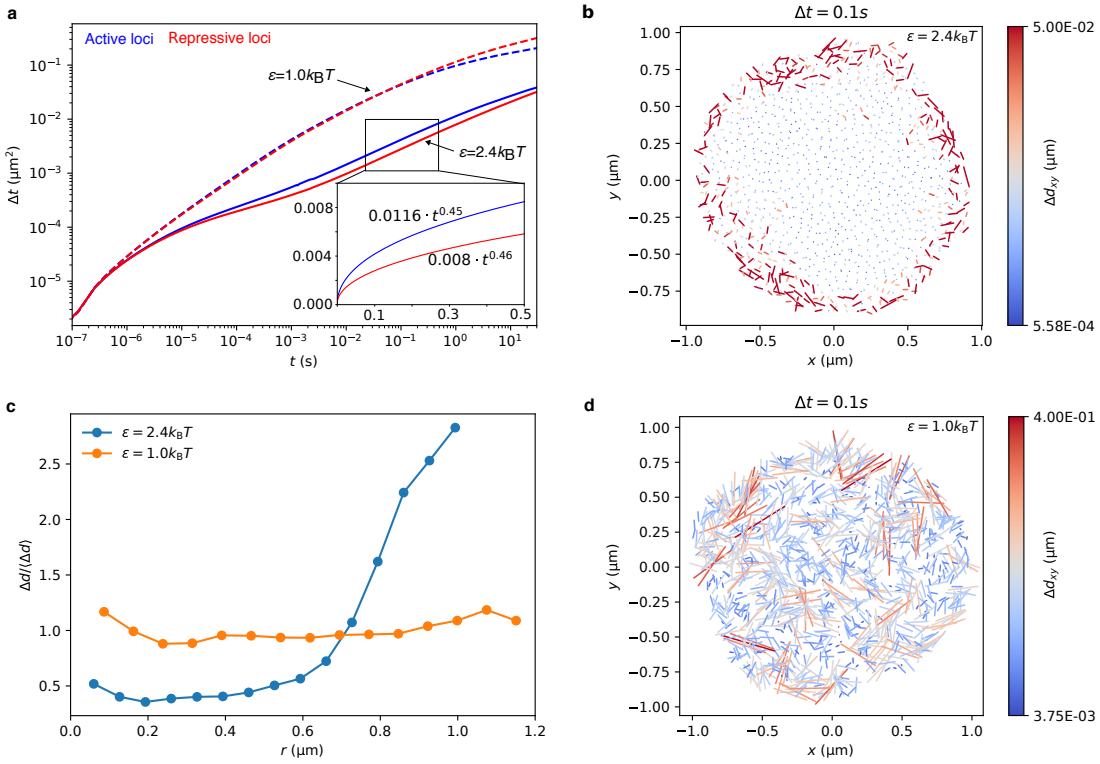


Figure 3.10: Mobility of active and repressive loci. **(a)** The Mean Square Displacement for active loci and repressive loci. The equation shown in the inset is the fit using Dt^{α}, where D is the diffusion coefficient and α is the diffusion exponent. **(b)** The displacement vectors of the loci within the equator cross-section of the structured chromosome for $\epsilon = 2.4k_{\mathrm{B}}T$. The displacements are computed for time window $\Delta t = 0.1$ s. The color bars on the right show the magnitudes of the displacements. **(c)** Displacement Δd normalized by its mean as a function of radial position, r, of the loci.**(d)** Same as (**b**) except the results are obtained using $\epsilon = 1.0k_{\mathrm{B}}T$.

3.3 Discussion

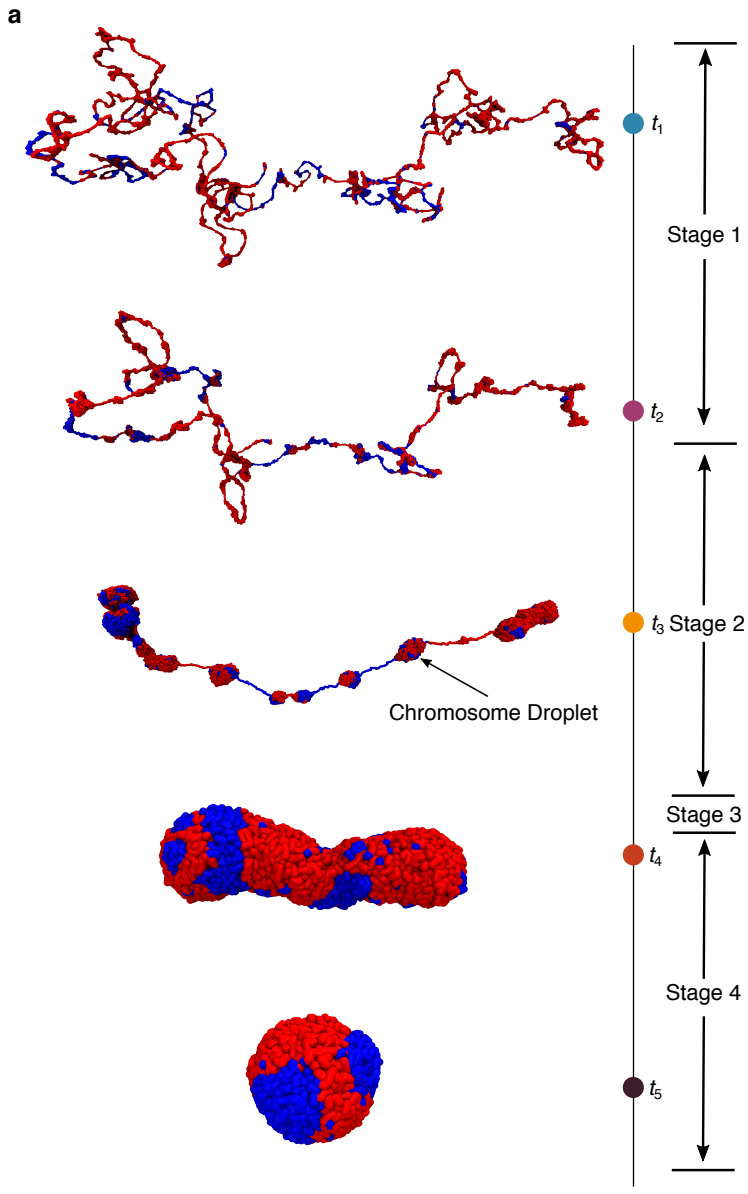
In order to demonstrate the transferability of the CCM, I simulated Chr 10 using exactly the same parameters as for Chr 5 (Appendix A.5). Fig. A.6 compares the WLM obtained from simulations for different ϵ values and the computed WLM using the Hi-C contact map. The contact map is translated to the distance R_{ij} by assuming that $P_{ij} \propto R_{ij}^{-4.1}$ holds for Chr 10 as well. It is evident that the CCM nearly quantitatively reproduces the spatial organization of Chr 10 (Fig. A.6). Thus, it appears that the CCM could be used for simulating the structures and dynamics of other chromosomes as well.

Two scaling regimes in $P(s)$ is suggestive of scale-dependent folding of genome. In order to reveal how chromosome organizes itself and to link these processes to the experimentally measurable $P(s)$, I calculated the time-dependent change in $P(s)$ as a function of t. At scales less (above) than $s^* \approx 5 \times 10^5$bps, $P(s)$ decreases (increases) as the chromosome becomes compact. The $P(s) \sim s^{-0.75}$ scaling for $s < s^*$ (see also Fig. 3.2b) is the result of organization on the small genomic scale during the early stage of chromosome condensation (Fig. 3.11a). In the initial stages compaction starts by forming $\approx s^*$ sized chromosome droplets (CDs) as illustrated in Fig. 3.11a. In the second scaling regime, $P(s) \sim s^{-1.25}$, global organization occurs by coalescence of the CDs (Fig. 3.11a). Thus, our CCM model, which suggests a hierarchical chromosome organization on two distinct scales, also explains the two scaling in $P(s)$.

The pictorial view of chromosome organization (Fig. 3.11a) shows that chro-

mosome structuring occurs hierarchically with the formation of CDs and subsequent growth of the large CDs at the expense of smaller ones. I quantitively monitored the growth of CDs during the condensation process and found that the size of CD grows linearly with time during the intermediate stage (Fig. 3.11b). Such a condensation process is reminiscent of the Lifshitz-Slazov mechanism [179] used to describe Ostwald ripening.

Our simulations show that the average TAD size and the crossover scale (s^*) the dependence of $P(s)$ on s coincide. In addition, the size of the CDs is also on the order of s^*, which is nearly the same for all the chromosomes (Fig. 3.2c). I believe that this is a major result. The coincidence of these scales suggests that both from the structural and dynamical perspective, chromosome organization takes place by formation of TADs, which subsequently arrange to form structures on larger length scales. Because gene regulation is likely controlled by the TADs, it makes sense that they are highly dynamic. I hasten to add that the casual connection between TAD size and s^* as well as the CDs size has to be studied further. If this picture is correct then chromosome organization, at length scales exceeding about 100 kbps, may be easy to describe.


a
t_1
Stage 1
t_2
t_3 Stage 2
Chromosome Droplet
Stage 3
t_4
Stage 4
t_5

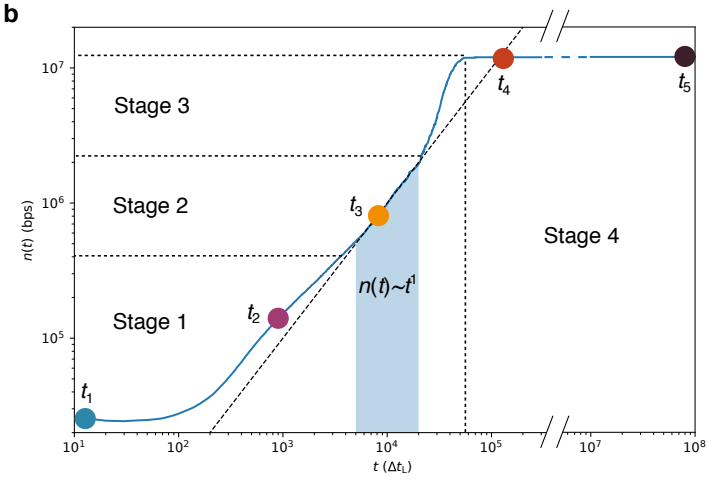

b
Stage 3
Stage 2
Stage 1
Stage 4
t_1
t_2
t_3
t_4
t_5
$n(t) \sim t^1$
$n(t)$ (bps)
t (Δt_L)

Figure 3.11: Dynamics of chromosome organization. **(a)** Typical conformations sampled during the chromosome organization process. After the short initial folding process (Stage 1, t_1 and t_2), the chromosome droplets (CDs) connected by "tension strings" begin to form (stage 2, t_3). The average size of CDs at the onset of CD formation is about $s \sim 4 \cdot 10^5$ bps, which coincides with approximate value of s^*, the typical size of TADs (Fig. 3.2b). At the later stage (stage 3, conformation not shown here), CDs merge to form larger cluster, eventuall form the final condensed structure (stage 4, t_4 and t_5). Red (Blue) represents represive (active) loci. **(b)** The time-dependent growth of CDs, $n(t)$, which is the average number of base pairs in a CD at t. The dashed line is a fit in the time window indicated by the shaded area, yielding $n(t) \sim t^1$. The roughly linear increase of $n(t)$, over a range of times, is consistent with the Lifshitz-Slazov growth mechanism [179].

In summary, I developed the Chromosome Copolymer Model (CCM), a self-avoiding polymer with two epigenetic states and with fixed loop anchors whose locations are obtained from experiment to describe chromosome dynamics. The use of rigorous clustering techniques allowed us to demonstrate that the CCM nearly quantitatively reproduces Hi-C contact maps, and the spatial organization gleaned from super-resolution imaging experiments. It should be borne in mind that contact maps are probabilistic matrices that are a low dimensional representation of the three-dimensional organization of genomes. Consequently, many distinct copolymer models are likely to reproduce the probability maps encoded in the Hi-C data. In other words, solving the inverse problem of going from contact maps to an energy function is not unique (see [180])

Chromosome dynamics is glassy, with correlated dynamics on scale $\approx 1\mu$m, implying that the free energy landscape has multiple equivalent minima. Consequently, it is likely that in genomes only the probability of realizing these minima is meaningful, which is the case in structural glasses. The presence of multiple minima

also leads to cell-to-cell heterogeneity with each cell exploring different local minimum in the free energy landscape. I speculate that the glass-like landscape might also be beneficial in chromosome functions because only a region on size $\sim s^*$ needs to be accessed to carry out a specific function, which minimizes large-scale structural fluctuations. In this sense, chromosome glassiness provides a balance between genomic conformational stability and mobility.

Chapter 4: Solution of the FISH-Hi-C paradox for Human Interphase Chromosomes

4.1 Introduction

Because chromosome lengths are extremely large, ranging from tens of million base pairs in yeast to billion base pairs in human cells, they have to fold into highly compact structures in order to be accommodated in the cell nucleus. This requires that loci that are well separated along the one-dimensional genome sequence be close in three-dimensional (3D) space, which is made possible by forming a large number of loops. The high throughput Hi-C technique and its variants are used to infer the probability of genome-wide contact formation between loci. In order to determine the contact probabilities between various loci in a genome, Hi-C experiments are performed in an ensemble of millions of cells. The readout of the Hi-C experiment are contact frequencies between a large number of loci from instantaneous snapshots of each cell, which are then used to construct the contact maps (Hi-C maps). The contact map is a matrix (2D representation) in which the elements represent the probability of contact between two loci that are separated by a specified genomic distance. A high contact count between two loci means that they interact with each

other more frequently compared to ones with low contact count.

A complementary and potentially a more direct way to determine genome organization is to measure spatial distances between loci using a low throughput Fluorescence *In Situ* Hybridization (FISH) technique [14, 30]. In addition to providing 3D distances in fixed cells, recently developed CRISPR–dCas9 FISH can be used to assay the dynamic behavior of loci in real time [17, 181, 182]. However, due to the limitation of number of distinct color probes, currently this method provides distance distribution information for only a small number of loci.

FISH and Hi-C, which are entirely different experimental techniques, provide data on different aspects of genome organization. As noted in recent reviews [183, 184], there are problems associated with each method. It is difficult to reconcile Hi-C and FISH data for the following reasons. In interpreting the Hi-C contact map, one makes the intuitive assumption that loci with high probability contact must also be spatially close. However, it has been demonstrated using Hi-C and FISH data on the same chromosome that high contact frequency does not always imply proximity in space [168, 183, 185, 186]. It should be noted that in most cases, the Hi-C and FISH measurements agree very well [14,24,30,118]. However from a purely theoretical perspective even a single contradiction is intriguing if the experimental errors can be ruled out. An outcome of our theory is that the discordance between FISH and Hi-C data arises because of extensive heterogeneity, which is embodied by the presence of a variety of conformations adopted by chromosomes in each cell. There are a variety of reasons, including differing fixation conditions and presence of two or more subpopulation of cells in which the chromosomes are present in

distinct conformations, which could give rise to the discordance between FISH and Hi-C data, as lucidly described recently [183, 184]. Contact between two loci could be a rare event, not present in all cells, which is captured in Hi-C experiment by performing an ensemble average. I show using a precisely solvable model that due to the absence of a contact between two specific loci in a number of cells, those with higher contact frequency could be spatially farther on an average than two others with lower contact frequency. In contrast, the probability of contact formation using the FISH method can only be obtained if the tail (small distance) of the distance distribution between locus i and j can be accurately measured. For a variety of reasons, including the size of the probe and the signal strength, this not altogether straightforward using FISH technique. Thus, in order to combine the data from the two powerful techniques, it is crucial to establish a theoretical basis with potential practical link, between the contact probability and average spatial distance.

Setting aside the conditions under which FISH and Hi-C are performed (see recommendations for comparing the results from the two techniques with minimum bias which are described elsewhere [183]) insights into the discordance between the two methods, when they occur, can be obtained using polymer physics concepts. Recently, Fudenberg and Imakaev [168] performed polymer simulations using a strong attractive energy between two labelled loci and a ten fold weaker interaction between two other loci that are separated by a similar genomic distance. In addition, they also reported simulations based on the loop extrusion model. Both these types of simulations showed there could be discordance between FISH and Hi-C, which I refer to as the FISH-Hi-C paradox. However, they did not provide any solution to

the paradox, which is the principle goal of this work.

Here, I develop a new and fully theoretical approach, which allows us to provide quantitative insights into the extent of heterogeneity in chromosome organization. From our theory, it follows that the resolution of the FISH-Hi-C paradox requires invoking the notion of heterogeneity, which implies multiple populations of chromosomes coexist. By using the concepts that emerge from the study of the Generalized Rouse Chromosome Model (GRMC), I demonstrate that the information of cell subpopulations can be extracted by fitting the experimental FISH data using our theory, thus allowing us to calculate the Hi-C contact probabilities from the theoretically calculated cumulative distribution function of spatial distance (CDF) - a quantity that can be measured using FISH and super resolution imaging methods. Our approach provides a theoretically based method to combine the available FISH and Hi-C data to produce a more refined characterization of the heterogeneous chromosome organization than is possible by using data from just one of the techniques. In other words, sparse data from both the experimental methods can be simultaneously harnessed to predict the 3D organization of chromosomes.

4.2 Methods

4.2.1 Generalized Rouse Model For Chromosoems (GRMC)

In order to derive an approximate relationship connecting contact probabilities between loci and the three dimensional distances, I use a variant of the random loop model [65, 180]. I first consider a minimal cross-linked phantom chain model,

which incorporates the presence of CTCF/cohein mediated loops [28]. The model, originally introduced for describing physical gels [180], and more recently used for chromosome dynamics in a number of insightful studies [64, 65], could be viewed as a Generalized Rouse Chromosome Model (GRMC) [187, 188]. The cross-links modeling the CTCF/cohein mediated loops here are not random. Their locations are predetermined by the Hi-C data [28].

The equations of motion for the GRMC is [123],

$$\xi \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{R} + \boldsymbol{F} \tag{4.1}$$

where ξ is the friction coefficient, $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N]^{\mathrm{T}}$ with $\boldsymbol{r}_i$ being the position of the i^{th} locus. The vector $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_N]^{\mathrm{T}}$ (T is the transpose), where $\boldsymbol{f}_i$ is the Gaussian random force acting on the i^{th} locus, characterized by $\langle f_n(t)\rangle$=0 and $\langle f_{n\alpha}(t) f_{m\beta}(t')\rangle = 2\xi k_{\mathrm{B}} T \delta_{nm}\delta_{\alpha\beta}\delta(t-t')$; $\boldsymbol{A}$ is the $N \times N$ connectivity matrix, embedding the information of chain connectivity and the location of the loops connecting two loci (Fig. 1(a));

$$A_{mn} = \begin{cases} -2\kappa - |\Sigma_m|\omega, & \text{if } m = n \neq 1 \text{ or } N \\ -\kappa - |\Sigma_m|\omega, & \text{if } m = n = 1 \text{ or } N \\ \kappa, & \text{if } |m-n| = 1 \\ \omega, & \text{if } |m-n| > 1, \text{and connected in } \Sigma \\ 0, & \text{if otherwise} \end{cases} \tag{4.2}$$

where Σ is the set of indices representing the loci pairs specifying the CTCF facil-

itated loop anchors, and $|\Sigma_m|$ is the number of loops connected to the m^{th} locus. The spring constant κ enforces chain connectivity, and ω is the associated spring constant for a CTCF pair. Note that the GRMC model does not account for excluded volume interactions, which in the modeling of chromatin is often justified by noting that topoisomerases enable chain crossing. Our purpose is to use GRMC to first illustrate concretely the challenges in going from the measured average contact map to spatial organization, precisely. More importantly, using the insights from the study of the GRMC, I solve the FISH-Hi-C paradox.

Since $\boldsymbol{A}$ in Eq. 4.2 is a real symmetric matrix, it can be diagonalized using the orthonormal matrix $\boldsymbol{V}$,

$$\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}^T = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_0, \lambda_1, ..., \lambda_{N-1}) \tag{4.3}$$

where $\lambda_0, \lambda_1, ..., \lambda_{N-1}$ are the eigenvalues of $\boldsymbol{A}$. By defining $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{R}$ and using $\boldsymbol{R} = \boldsymbol{V}^T\boldsymbol{X}$ and $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$, I obtain the equations of motion of the normal coordinates $\boldsymbol{X}$,

$$\xi\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t} = \boldsymbol{\Lambda}\boldsymbol{X} + \boldsymbol{f}. \tag{4.4}$$

Because $\boldsymbol{\Lambda}$ is a diagonal matrix, the normal coordinates of the GRMC $\boldsymbol{X}_p$ are decoupled. Using the normal modes, $\boldsymbol{X}$, the physical quantities associated with the polymer can be readily calculated. Therefore, for GRMC with a predetermined set of CTCF/cohein mediated loops, I can solve for the eigenvalues of the connectivity matrix $\boldsymbol{A}$, and the orthonormal matrix $\boldsymbol{V}$ numerically, and thus calculate the contact

probability and spatial distance precisely.

4.2.2 Relation between contact probability and mean spatial distance for GRMC

The vector between the positions of the m^{th} and the n^{th} loci may be written as,

$$\boldsymbol{R}_m - \boldsymbol{R}_n = \sum_{p=0}^{N-1} (V_{pm} - V_{pn}) \boldsymbol{X}_p \tag{4.5}$$

where V_{pm} and V_{pn} are the elements of orthonormal matrix $\boldsymbol{V}$. The equilibrium solution of Eq. 4.4 yields, $\lim_{t\to\infty} X_{p,\alpha}(t) \sim \mathcal{N}(0, -\frac{k_\mathrm{B}T}{\lambda_p})$, where $\alpha = x, y, z$, $\mathcal{N}$ is Gaussian distribution. Therefore,

$$\lim_{t\to\infty} R_{mn,\alpha}(t) \sim \mathcal{N}(0, -\sum_{p=0}^{N-1} (V_{pm} - V_{pn})^2 \frac{k_\mathrm{B}T}{\lambda_p}) \equiv \mathcal{N}(0, \sigma^2_{mn,\alpha}). \tag{4.6}$$

where $\sigma_{mn,\alpha} = -\sum_{p=0}^{N-1} (V_{pm} - V_{pn})^2 (k_\mathrm{B}T/\lambda_p)$. Since the model is isotropic, it follows that $\sigma^2_{mn,x} = \sigma^2_{mn,y} = \sigma^2_{mn,z} \equiv \sigma^2_{mn}$. The mean distance $\langle R_{mn} \rangle$ is related to σ_{mn} through $\langle R_{mn} \rangle = 2\sqrt{2/\pi}\sigma_{mn}$. The distribution of distance between the m^{th} and the n^{th} loci, $\lim_{t\to\infty} |\boldsymbol{R}_{mn}(t)| = \lim_{t\to\infty} \sqrt{\sum_\alpha R^2_{mn,\alpha}(t)}$ is a non-central chi distribution (I will neglect the notation $\lim_{t\to\infty}$ from now on),

$$P(R_{mn} = r) = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_{mn}} e^{-r^2/(2\sigma^2_{mn})} \frac{r^2}{\sigma^2_{mn}}. \tag{4.7}$$

The contact probability P_{mn}, for a given threshold r_c (contact exists if $r \leq r_c$),

computed using Eq. 4.7 yields,

$$\begin{aligned} P_{mn} &= \int_0^{r_c} \mathrm{d}r \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_{mn}} e^{-r^2/(2\sigma_{mn}^2)} \frac{r^2}{\sigma_{mn}^2} \\ &= \mathrm{Erf}\left(\frac{r_c}{\sqrt{2}\sigma_{mn}}\right) - \sqrt{\frac{2}{\pi}} e^{-\frac{r_c^2}{2\sigma_{mn}^2}} \frac{r_c}{\sigma_{mn}}. \end{aligned} \tag{4.8}$$

The mean spatial distance $\langle R_{mn} \rangle$ is given by,

$$\langle R_{mn} \rangle = \int_0^{\infty} \mathrm{d}r r \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_{mn}} e^{-r^2/(2\sigma_{mn}^2)} \frac{r^2}{\sigma_{mn}^2} = 2\sqrt{\frac{2}{\pi}} \sigma_{mn}. \tag{4.9}$$

Using Eqs.4.8 and 4.9, the desired relation between P_{mn} and $\langle R_{mn} \rangle$ becomes,

$$P_{mn} = \mathrm{erf}\left(\frac{2r_c}{\sqrt{\pi}\langle R_{mn} \rangle}\right) - \frac{4}{\pi} \frac{r_c}{\langle R_{mn} \rangle} e^{-\frac{4r_c^2}{\pi \langle R_{mn} \rangle^2}} \equiv R_0(\langle R_{mn} \rangle). \tag{4.10}$$

4.2.3 Generalized power law realtion between contact probability and mean spatial distance

A key goal in our theory is to theoretically establish a useful relationship between the contact probabilities and the mean spatial distances between the loci. I have shown in the section 4.2.2 that the contact probability is connected to mean spatial distance by a powerlaw for GRMC. In this section, I seek to generalize the power-law relation to real chromatin. Because long chromosomes are modeled as polymers, I look to rigorous results in polymer theory for the distance distribution function, $P(r|\langle R \rangle)$ between two loci separated by r with a mean distance $\langle R \rangle$.

Knowledge of $P(r|\langle R\rangle)$ is needed to construct the Cumulative Distribution Function, CDF$(R|\langle R\rangle)$. There are only few polymer models for which analytic results for $P(r)$ are known.

A particularly useful result for our purposes is $P(r|\langle R\rangle)$ for a self-avoiding homopolymer in a good solvent. In this case, the Redner- desCloizeaux [124, 125] distribution is given by,

$$P(r|\langle R\rangle) = A(r/\langle R\rangle)^{2+g}\exp(-B(r/\langle R\rangle)^{\delta}), \tag{4.11}$$

where $\langle R\rangle$ is the mean distance between two loci, and g is the "correlation hole" exponent, and δ is related to the Flory exponent ν by $\delta = 1/(1-\nu)$. In good solvents, $\nu \approx 0.588$. The constants A and B can be calculated using the normalization condition, $\int \mathrm{d}r P(r|\langle R\rangle) = 1$. Given the value of r_c, the threshold distance for contact formation, the contact probability P_c between the two loci is,

$$P_c = \int_0^{r_c} P(r|\langle R\rangle)\mathrm{d}r. \tag{4.12}$$

When the contact threshold is small compared to the size of the chain or the loop $r \ll \langle R\rangle$, the integral can be approximately evaluated using,

$$\begin{aligned} P_c &= \lim_{r_c\to 0}\int_0^{r_c} P(r)\mathrm{d}r \\ &= \lim_{r_c\to 0}\int_0^{r_c} A(r/\langle R\rangle)^{2+g}\exp(-B(r/\langle R\rangle)^{\delta})\mathrm{d}r, \\ &\sim \langle R\rangle^{-(3+g)}. \end{aligned} \tag{4.13}$$

Thus, the contact probability between two monomers, P, is related to the mean end-to-end distance $\langle R \rangle$ only through the scaling exponent $-(3+g)$. For ideal chain, $g = 0$, and thus I recover the asymptotically exact relation $P \sim \langle R \rangle^{-3}$. Note that $\langle R \rangle$ does depend on the genomic distance separating the two loci.

For a single polymer chain, there are three ways a contact between loci may be established [189]: i) the contact between two ends of the chain (Fig. 4.1a). ii) the contact between one end and a locus in the interior (Fig. 4.1b). iii) the contact between two loci in the interior of the chain (Fig. 4.1c). The correlation hole exponents corresponding to the three cases are $g_1 = 0.273$, $g_2 = 0.46$ and $g_3 = 0.71$ [189]. Thus, I have $P = \langle R \rangle^{-3.273}$, $P = \langle R \rangle^{-3.46}$ and $P = \langle R \rangle^{-3.71}$ for three cases. These rigorous values provide a bound for g, and should be viewed as a guide when considering the complicated case of chromosomes.

4.2.4 Simulations details

The energy function for the GRMC is,

$$U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) = \sum_{i=1}^{N-1} U_i^S + \sum_{\{p,q\}} U_{\{p,q\}}^L. \tag{4.14}$$

For the bonded stretch potential, U_i^S, I use,

$$U_i^S = \frac{\kappa}{2}(|\boldsymbol{r}_{i+1} - \boldsymbol{r}_i| - a)^2, \tag{4.15}$$

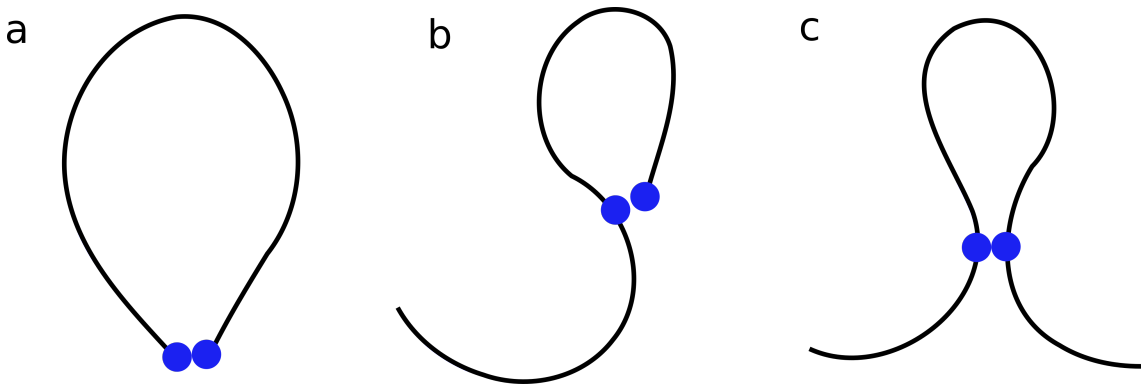


Figure 4.1: Three possibilities for contact formation between two loci in a polymer. **(a)** Contact formation between the two ends. **(b)** Contact formation between one end and a locus in the interior. **(c)** Contact formation between two loci located in the interior of a polymer. Although the relation between P and $\langle R \rangle$ decreases as power law the value of the exponent is different in the three scenarios (see the SI text for the precise values of a polymer in good solvent.

where a is the equilibrium bond length. The interaction between the loop anchors is also modeled using a harmonic potential,

$$U^L_{\{p,q\}} = \frac{\omega}{2}(|\boldsymbol{r}_p - \boldsymbol{r}_q| - a)^2 \tag{4.16}$$

where the spring constant is associated with the CTCF facilitated loops, and $\{p, q\}$ represent the indices of the loop anchors, which are taken from the Hi-C data [28] (see section 2.2.3). I simulate the chromosome segment from 146 Mbps to 158 Mbps of Chromosome 5. Each monomer represents 1200 bps, resulting total number of coarse-grained loci $N = 10,000$.

In order to accelerate conformational sampling, I performed Langevin Dynamics simulations at low friction [161]. The value of friction coefficient is 0.01 in LJ unit with mass of monomer set to be $m = 1$ and the equilibrium bond length $a = 1$.

I simulated each trajectory for 10^8 time steps, and saved the snapshots every $10,000$ time steps. I generated ten independent trajectories, which are sufficient to obtain reliable statistics.

4.3 Results

4.3.1 Relating contact probability to mean spatial distance for GRMC:

The exact relationship between P_{mn} (contact probability between m^{th} and n^{th} locus) and the corresponding mean spatial distance, $\langle R_{mn} \rangle$ for GRMC (see 4.2.1 for details of the derivation) is,

$$P_{mn} = \text{erf}\left(\frac{2r_c}{\sqrt{\pi}\langle R_{mn}\rangle}\right) - \frac{4}{\pi}\frac{r_c}{\langle R_{mn}\rangle}e^{-\frac{4r_c^2}{\pi\langle R_{mn}\rangle^2}} \equiv R_0(\langle R_{mn}\rangle). \tag{4.17}$$

The inverse of $R_0(\langle R_{mn}\rangle)$, the solution to Eq. 4.17, gives the mean spatial distance $\langle R_{mn}\rangle$ as a function of the contact probability P_{mn}. Note that m and n are arbitrary locations of any two loci, and thus Eq. 4.17 is general for any pair of loci.

A couple of conclusions, relevant to the application to the chromosomes, follow from Eq. 4.17. (i) Note that Eq. 4.17 is an exact one-to-one relation between the mean distance $\langle R_{mn}\rangle$ and the contact probability P_{mn} provided r_c is known, and if the contacts are present in all the cells, which is not the case in experiments. For small P_{mn}, it is easy to show from Eq. 4.17 that $\langle R_{mn}\rangle \approx r_c P_{mn}^{-1/3}$. For the ideal GRMC, this implies that for any m, n, k, l, if $P_{mn} < P_{kl}$ then $\langle R_{mn}\rangle > \langle R_{kl}\rangle$,

a consequence anticipated on intuitive grounds. (ii) If the value of the contact probability P and the threshold distance r_c are known precisely, then the distribution of the spatial distance can be readily computed by solving Eq. 4.17 numerically. In Fig. 1(b), I show the comparison between theory (Eq. 4.17) and simulations. The simulated curves are computed as follows: first collect $(P_{mn}, \langle R_{mn} \rangle)$ for every pair labeled (m, n) where P_{mn} and $\langle R_{mn} \rangle$ are computed. The total number of pairs is $N(N-1)/2$. I then binned the points over the values of P_{mn}. Finally, the mean value of $\langle R_{mn} \rangle$ for each bin, $\langle R \rangle = E[\langle R_{mn} \rangle]$, is computed where $\mathrm{E}[\cdots]$ is the binned average, which is computed using $(1/N_i) \sum_{j=1}^{N_i} \langle R_{mn} \rangle^j$ where N_i is the number of points in the i^{th} bin. The bin size, Δ, is centered at P_{mn}, spanning $P_{mn} - \Delta/2 \leq P_{mn} \leq P_{mn} + \Delta/2$. Using this procedure, I find (Fig. 1) that the theory and simulations are in perfect agreement, which validates the theoretical result.

4.3.2 Contact distance r_c affects the inferred value of the spatial distance:

However, in practice, the elements P_{mn} are measured with (unknown) statistical errors, and the value of the contact threshold r_c is only estimated. In the Hi-C experiments, contact probabilities and r_c by implication, are determined by a series of steps that start with cross-linking spatially adjacent loci using formaldehyde, chopping the chromatin into fragments using restriction enzymes, ligating the fragments with biotin, followed by sequence matching using deep sequencing methods

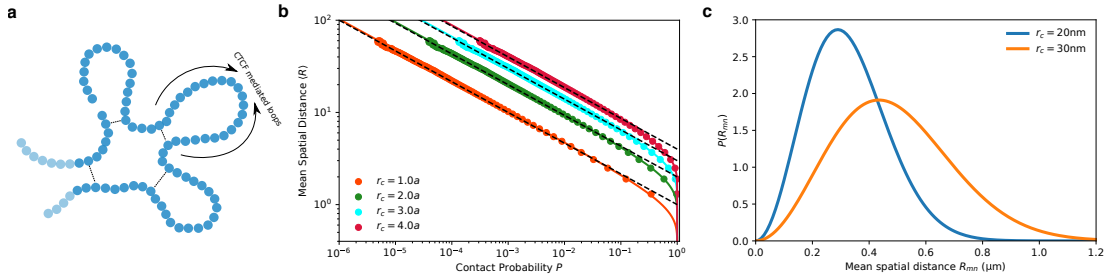


Figure 4.2: Simulations demonstrate the power law relation between contact probability and mean spatial distance and the effect of r_c on the inferred spatial distances. (**a**) A sketch of the Generalized Rouse Model for Chromosome (GRMC). Each bead represents a loci with a given resolution. Dashed lines represents harmonic bond between loop anchors. (**b**) Mean spatial distance $\langle R \rangle$ as a function of the contact probability P. The solid lines are obtained using Eq. 4.10 for different values of r_c (shown in the figure), the threshold distance for contact formation. The dots are simulation results. The agreement between simulations and theory is excellent. Asymptotically $\langle R \rangle$ approaches $r_c P^{-1/3}$ (dashed lines). The threshold for contact is expressed in terms of a which is the equilibrium bond length in Eq. 4.15. (**c**) Illustration of the sensitivity of r_c in determining the mean spatial distance $\langle R \rangle$. Blue and yellow curves are computed by solving $\langle R \rangle$ (Eq. 4.17) for a given contact probability $P_{mn} = 10^{-3}$, and r_c. The calculated $\langle R_{mn} \rangle$ is used in Eq. 4.7 to obtain the distribution of the spatial distance $P(R_{mn})$. Blue and yellow curves are for the same value of P but different r_c values.

(see [22] for a review). Because of the inherent stochasticity associated with the overall Hi-C scheme, as well as the unavoidable heterogeneity (only a fraction of cells has a specific contact and the contact could be dynamic) in the cell population the relationship P_{mn} and $\langle R_{mn} \rangle$ is not straightforward.

To illustrate how the uncertainty in r_c affects the determination of the spatial distance in GRMC even when population is homogeneous (all cells have a specific contact), I plot the distributions of distance for $r_c = 0.02, 0.03$ μm in Fig. 4.2c. A small change in r_c (from 0.02 μm to 0.03 μm) completely alters the distance distribution $P(R)$, and hence the mean spatial distance (from $\approx 0.2\mu m$ to $\approx 0.3\mu m$). For the exactly solvable GRMC, this can be explained by noting that $\langle R_{mn} \rangle \approx r_c P_{mn}^{-1/3}$ for small P_{mn}. Because P_{mn} appears in the denominator, any uncertainty in r_c is amplified by P_{mn}, especially when P_{mn} is small.

Heterogeneity causes Fish-Hi-C "Paradox": The expectation that the contact probability should decrease as the mean distance between the loci increases, which is the case in the exactly solvable ideal GRMC ($P_{mn} \approx r_c \langle R_{mn} \rangle^{-3}$), is sometimes violated when the experimental data [28] is analyzed [168, 183]. The paradox is a consequence of heterogeneity due to the existence of more than one population of cells, which implies that in some fraction of cells, contact between two loci exist while in others it is absent. Each distinct population has its own statistics. For instance, the probability distribution of spatial distance between the m^{th} and the n^{th} loci, $P_{i,mn}(r)$, for one population of cells could be different from another population of cells $P_{j,mn}(r)$ where i and j are the indices for the two different populations (Fig.

4.3(a)). The Hi-C experiments yield only an average value of the contact probability. Let us illustrate the consequence of the inevitable heterogeneous mixture of cell populations by considering the simplest case in which only two distinct populations, one with probability η and the other $1-\eta$, are present (a generalization is presented below). For instance, in one population of cells, there is a CTCF loop between m and n, and it is absent in the other population. The probability distribution of spatial distance between the m^{th} and the n^{th} loci is a superposition of distributions for each population. Using Eq. 4.17, the mixed distribution can be written as,

$$P(R_{mn}=r)=\sqrt{\frac{2}{\pi}}\left(\eta\frac{r^2}{\sigma_{1,mn}^3}e^{-\frac{r^2}{2\sigma_{1,mn}^2}}+(1-\eta)\frac{r^2}{\sigma_{2,mn}^3}e^{-\frac{r^2}{2\sigma_{2,mn}^2}}\right) \tag{4.18}$$

where $\sigma_{1,mn}$ and $\sigma_{2,mn}$ are the parameters with different values characterizing the two populations. In the GRMC, $\sigma_{1,mn}$ and $\sigma_{2,mn}$ are related to the mean spatial distances in the two populations by $\langle R_{1,mn}\rangle = 2\sqrt{2/\pi}\sigma_{1,mn}$ and $\langle R_{2,mn}\rangle = 2\sqrt{2/\pi}\sigma_{2,mn}$. The mean spatial distance is, $\langle R_{mn}\rangle = \eta\langle R_{1,mn}\rangle + (1-\eta)\langle R_{2,mn}\rangle$, and the contact probability is $P_{mn} = \eta P_{1,mn} + (1-\eta)P_{2,mn}$ where $P_{1,mn}$ and $P_{2,mn}$ are the contact probabilities for each population, given by Eq. 4.17, which depends on the values of $\langle R_{1,mn}\rangle$ and $\langle R_{2,mn}\rangle$ as well as r_c.

If the values of $\langle R_{1,mn}\rangle$ and $\langle R_{2,mn}\rangle$ are unkown (as is the case in Hi-C experiments), and only the value of the contact probability between the two loci is provided, one can not uniquely determine the values of the mean spatial distances. This is the origin of the Hi-C and FISH data paradox. In Figs. 4.3b-e I show an

example of the paradox for a particular set of parameters $(\eta, \sigma_{1,mn}, \sigma_{2,mn})$. Pair #1 has a larger contact probability than pair #2, while also exhibiting a larger mean spatial distance. The GRMC explains in simple terms the origin of the paradox.

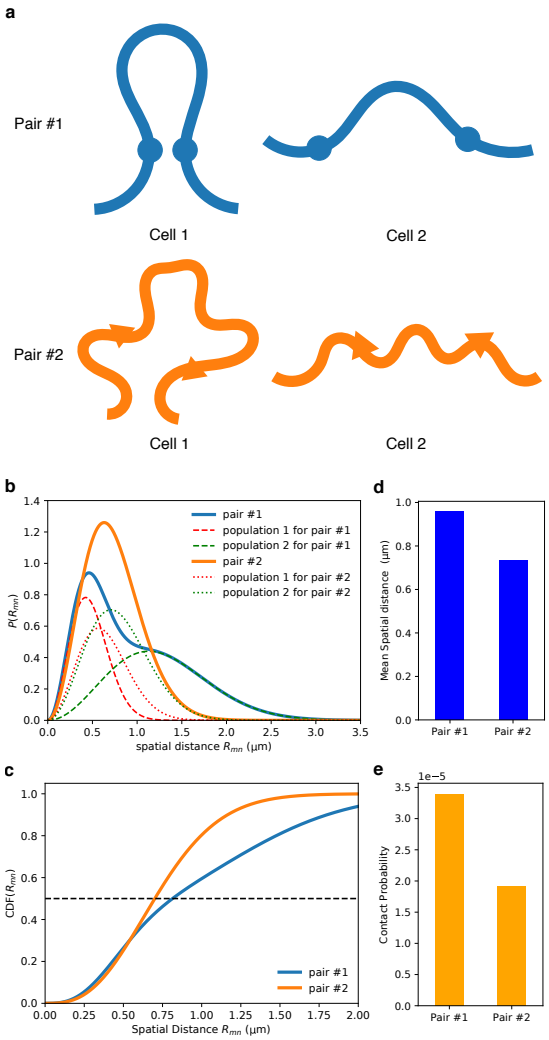


Figure 4.3: Illustrating the FISH-Hi-C ($[P_{mn}, \langle R_{mn}\rangle]$) paradox. **(a)** Schematic illustration of the populations of two cells. There are two pairs of loci, pair 1 and pair 2. Cells 1 and 2 belong to two distinct populations such that pair 1 and pair 2 have different distributions of distances in the two cells. Pair 1 is always in proximity (contact is formed) in cell 1, whereas it is spatially separated (mean distance $> r_c$) in cell 2. Pair 2 on the other hand has similar distributions of spatial distance in cells 1 and 2. Cell with two different populations gives rise to the paradoxical behavior, which is illustrated by choosing $\eta_1 = 0.4$ and $\eta_2 = 1 - \eta_1 = 0.6$. These are the probabilities for a cell belonging to population 1 and 2, respectively. The pair 1 has parameters $\sigma_1 = 0.3\mu m$ and $\sigma_2 = 0.8\mu m$. The pair 2 has parameters $\sigma_1 = 0.4\mu m$ and $\sigma_2 = 0.5\mu m$. See Eq. 4.18 for the definition of σ_1 and σ_2. **(b)** The distribution of distance for pair 1 (thick blue) and pair 2 (thick orange), respectively. The distributions for the two different populations are shown separately for pair 1 (dashed lines) and pair 2 (dotted lines). **(c)** Cumulative distribution of the spatial distance. The horizontal dashed line indicates the median distance. **(d)** Mean distances for pair 1 is larger than for pair 2. **(e)** Pair 1 has larger contact probability than 2, which is paradoxical since the distance between the loci in pair 1 is larger than in 2. The threshold for determining contact is $r_c = 20$ nm.

To systematically explore the parameter space, I display $\langle R_{mn}\rangle$ and P_{mn} as heat maps showing $\langle R\rangle_{1,mn}$ versus $\langle R\rangle_{2,mn}$ for different values of η (Fig. 4.4). When there is a single homogenous population ($\eta = 0.0$), the mean spatial distance $\langle R_{mn}\rangle$ and contact probability P_{mn} depend only on the value of $\langle R_{2,mn}\rangle$ (upper panel in Fig. 4.4). In this case, there is a precise one-to-one mapping between $\langle R_{mn}\rangle$ and P_{mn}. However, if $\eta \neq 0$ ($\eta = 0.3$, lower pannel in Fig. 4.4) then the relation between P_{mn} and $\langle R_{mn}\rangle$ is complicated. The contour lines for P_{mn} cross the contour lines of $\langle R_{mn}\rangle$, which implies that for a given value of P_{mn}, one cannot infer the value of $\langle R_{mn}\rangle$ without knowing the value of η, $\langle R_{1,mn}\rangle$ and $\langle R_{2,mn}\rangle$. For instance, the triangle and circle shown for $\eta = 0.3$ in Fig. 4.4 demonstrate an example of the paradox in which $\langle R(\blacktriangledown)\rangle (= 57a) > \langle R(\bullet)\rangle (= 40a)$ whereas $P(\blacktriangledown)(\approx 7.7 \times 10^{-4}) > P(\bullet)(\approx 3.9 \times 10^{-4})$.

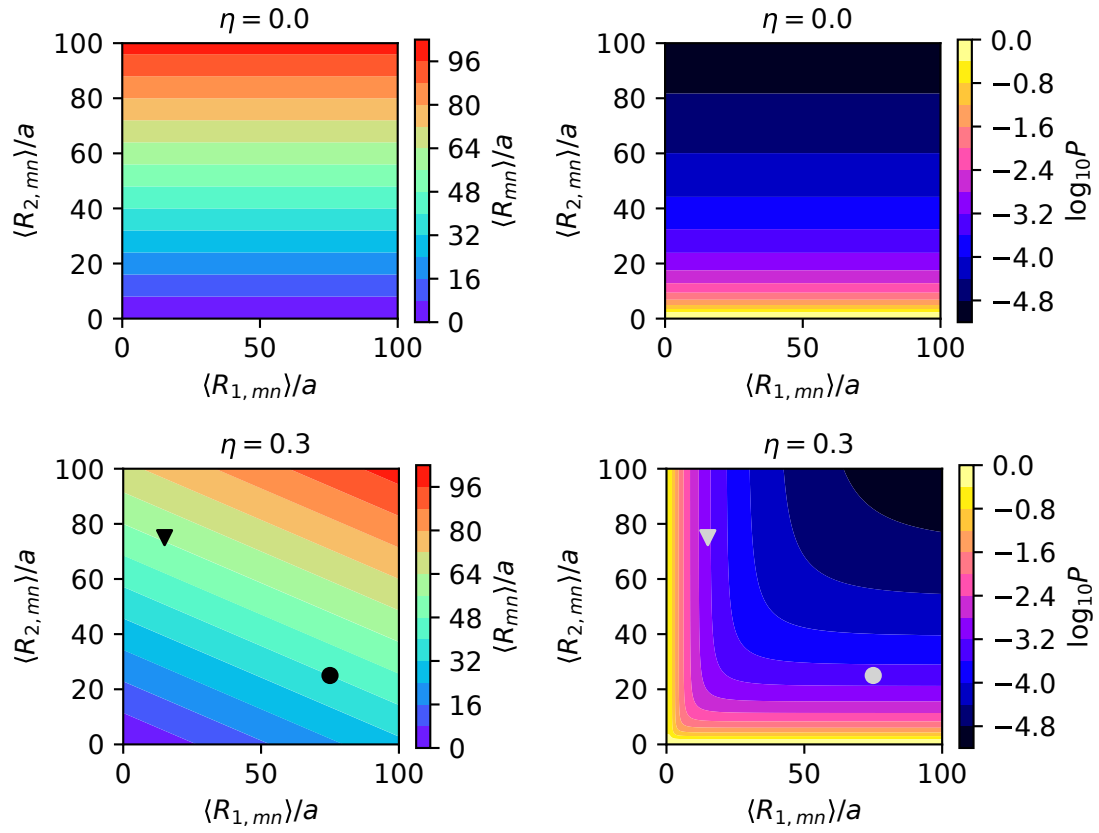


Figure 4.4: Plots of mean distance $\langle R_{mn}\rangle$ and the contact probability P_{mn} as heatmaps computed using $r_c = 2a$. The colorbars on the right show the values of $\langle R_{mn}\rangle$ and P_{mn}. The results for $\eta = 0 (\neq 0)$ is shown on top (bottom). Two specific pairs are marked as triangle and circle in the lower left panel These loci pairs illustrate the $[P_{mn}, \langle R_{mn}\rangle]$ paradox.

4.3.3 Extracting cell subpopulation information from FISH data:

Can one extract the information about subpopulations from experimental data so that the result from two vastly different techniques can be reconciled? To answer this question, I first generalize the theory derived from GRMC to real chromatins. The generalization of Eq. 4.18 is,

$$P(R_{mn} = r) = \eta P(r|\langle R_{1,mn}\rangle) + (1-\eta)P(r|\langle R_{2,mn}\rangle) \tag{4.19}$$

where $P(r|\langle R_{1,mn}\rangle)$ and $P(r|\langle R_{2,mn}\rangle)$ are the Redner-des Cloizeaux distribution of distances for polymers [124, 125] (section 4.2.3). The distribution $P(r|\langle R_{mn}\rangle)$ is rigorously known for self-avoiding homopolymer in good solvent, generalized Rouse model (Eq. 4.11 in section 4.2.3), and a semi-flexible polymer [190, 191]. However, a simple analytic expression for chromosomes is not known. By assuming that the Redner-des Cloizeaux form for $P(r|\langle R_{mn}\rangle)$ also holds for chromosomes (see Eq. 4.11 for details), I find that $g = 1$ and $\delta = 5/4$ in Eq. 4.11. These parameters were previously extracted using experimental data [14], and the Chromosome Copolymer Model (CCM) for chromosomes [66]. The value of g is inferred from the scaling relationship between mean spatial distance $\langle R\rangle$ and contact probability P, $P \sim \langle R\rangle^{3+g}$. The value of δ is computed as $\delta = 1/(1-\nu)$. ν is inferred from scaling $\langle R(s)\rangle \sim s^{\nu}$ where s is the genomic distance.

The integral of Eq. 4.19 up to R, which is the cumulative distribution function CDF(R), can be used to fit the FISH data. Thus, the probability of contact forma-

tion can be computed as, $\int_0^{r_c} P(r|\langle R\rangle)\mathrm{d}r$ where r_c is the contact threshold. Using the data in [28], the CDF(R) for two pairs of loci are shown in Fig. 4.6(a). By fitting the two experimentally measured curves to the theoretical prediction (see Appendix B.1), I obtain $\eta \approx 0.42$ for peak4-loop and $\eta \approx 0.97$ for peak3-control. The parameters obtained can then be used to compute the contact probability. Since the Hi-C experiments measure the number of contact events instead of contact probability and the value of r_c is unknown, I compared the *relative contact frequency*, which is computed as $P_i/\langle P\rangle$ where P_i is the contact probability computed using the model or the contact number measured in Hi-C for the i^{th} pair and $\langle P\rangle$ is the mean value for all the pairs considered. First, I fit all the eight CDF(R) curves in [28]. The excellent agreement between theory and experiments is vividly illustrated in the Fig. 4.5 and also manifested by the Kolmogorov-Smirnov statistics (Table B.1). Second, I calculated their corresponding *relative contact frequency* (Fig. 4.6(b)). Comparison of the theoretical calculations with Hi-C measurements shows excellent agreement (Fig. 4.6(b)) with the Pearson correlation coefficient being 0.87. The contact probability is computed using $r_c = 10$ nm. It is important to note that fitting the FISH data with the assumption that cell population is homogeneous leads to unphysical values of g and δ and the Kolmogorov-Smirnov statistics are inferior (see Appendix B.2 and Table B.3).

Interestingly, the values of $\langle R_1\rangle$ obtained from fitting the four CTCF/cohesin mediated loops (peak(1,2,3,4)-loop) are all about $0.25 - 0.35\mu$m ($R_{1,\mathrm{peak1-loop}} \approx 0.24\mu m$, $R_{1,\mathrm{peak2-loop}} \approx 0.33\mu m$, $R_{1,\mathrm{peak3-loop}} \approx 0.35\mu m$,$R_{1,\mathrm{peak4-loop}} \approx 0.30\mu m$) regardless of their genomic separation (see Table B.1), suggesting that the mechanism

of looping between CTCF motifs are similar with a mean spatial distance ≈ 0.3 μm. The physically reasonable value of $\langle R_{mn} \rangle \approx 0.3\nu$m for all peak-loop pairs shows that these CTCF-mediated contacts describe molecular interactions between loci that are separated by a few hundred kilo base pairs. It has been shown that these contacts, referred to as "peaks" [28] are significantly closer in space than others that are separated by similar genomic distance. The peak-loop contacts correspond to chromatin loops with the loci in the peaks being the anchor points between a specific loop. In sharp contrast, the distances between peak$i$-control (i goes from 1 to 4), which are greater than the distances between peak loci, vary ranging from ≈ 0.47 μm to ≈ 0.67 μm (see Table B.1). It is likely that these contacts are more dynamic because they are not be anchored by CTCF binding proteins.

4.3.4 Fitting FISH data when heterogeneity is extensive

In the results presented in Figs. 4.5 and 4.6, I assumed that chromosomes with or without CTCF loops may be categorized into two subpopulations, each with a characteristic mean distance. Here, I generalize the theory using a continuous distribution of subpopulations, which is required in light of recent study [118]. Let us denote $P(\langle R \rangle)$ as the probability distribution of mean distance $\langle R \rangle$ characterizing a subpopulation. In Eq. 4.19, $P(\langle R \rangle)$ is assumed to be a linear combination of two Delta functions. This assumption, which is reasonable in the context of CTCF loops, may not hold in cases where cooperative interactions between loops are prominent

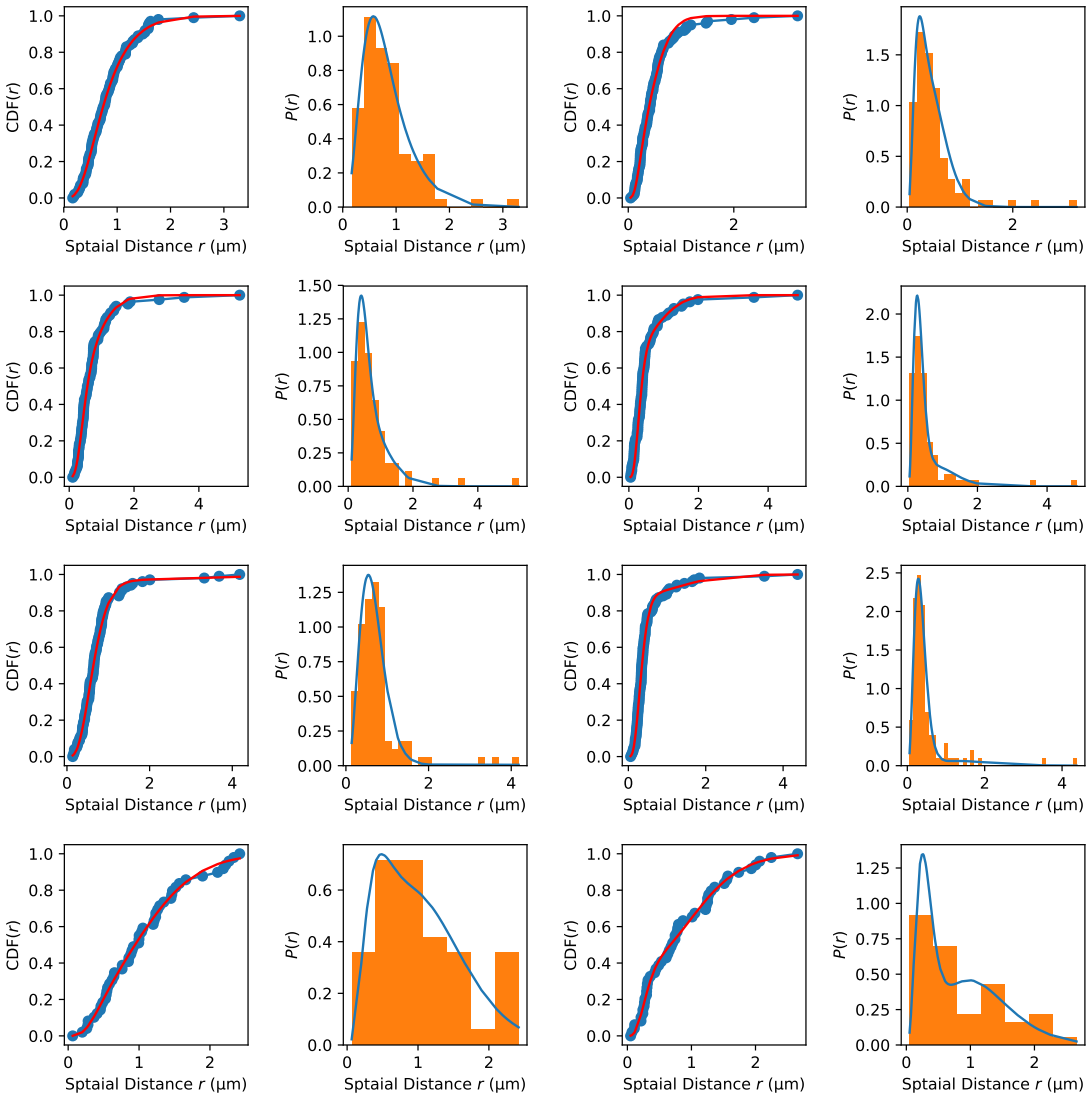


Figure 4.5: Fits of the CDF(R) (using Eq. 4.19 with values of $g = 1$ and $\delta = 5/4$) to the experimental data (blue dots) [28]. Orange lines are the fits. The parameters obtained from the fits for the eight loci pair are summarized in Table B.1. The probability density distribution (PDF) obtained using the fit parameters are also plotted along with experimental PDF. The excellent agreement between theory and experiments is self-evident.

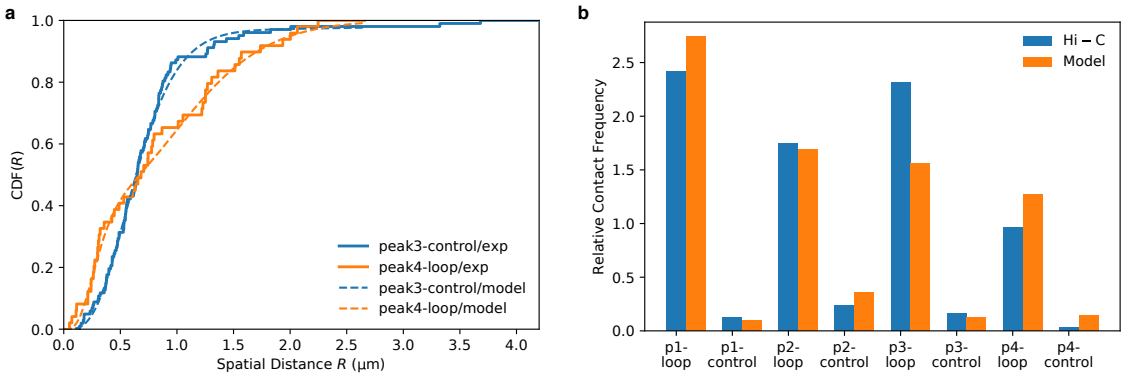


Figure 4.6: Extracting statistics of subpopulations from FISH data. **(a)** Cumulative distribution function of the spatial distance, CDF(R) for two pairs of loci, labeled peak3-control and peak4-loop in [28]. The excellent agreement between theory and experiments shows the usefulness of the relationship between P_{mn} and R_{mn} obtained using GRMC. The solid curves are the experiment data [28]. The dashed lines are the fits to $\int_0^R P(r)\mathrm{d}r$ (the needed expressions are in Eq. 4.19). The best fit parameters are: $\eta_{\mathrm{peak3-control}} \approx 0.97$, $\langle R_{1,\mathrm{peak3-control}}\rangle \approx 0.67$ μm, $\langle R_{2,\mathrm{peak3-control}}\rangle \approx 4.08$ μm, $\eta_{\mathrm{peak4-loop}} \approx 0.42$, $\langle R_{1,\mathrm{peak4-loop}}\rangle \approx 0.30$ μm and $\langle R_{2,\mathrm{peak4-loop}}\rangle \approx 1.21$ μm. **(b)** Relative Contact Frequency computed from the fits of CDF(R) for eight pairs of loci investigate experimentally [28] (orange bars). For each pair of loci, the contact probability is calculated as $P_{mn} = \int_0^{r_c} P(r)\mathrm{d}r$ (Eq. 4.19) using the parameters obtained by fitting CDF(R) with $r_c = 20$ nm. Blue bars are computed using the contact number from Hi-C measurements in [28]. The relative contact frequency is calculated as $P_i/\langle P\rangle$ where P_i is the contact probability computed using the model or the contact number measured in Hi-C for i^{th} pair, and $\langle P\rangle$ is the mean value for all the pairs considered. "p1-loop/p1-control/..." are the ones referred to "peak1-loop/peak1-control/..." in [28].

or when there is extensive conformational heterogeneity. Here, I discuss how to analyze the FISH data without making any prior assumption about $P(\langle R\rangle)$. The generalization of Eq. B.2 is,

$$\mathrm{CDF}(R) = \int_0^\infty \mathrm{d}\langle R\rangle P(\langle R\rangle)\mathrm{CDF}(R|\langle R\rangle). \tag{4.20}$$

If $P(\langle R\rangle) = \eta\delta(\langle R\rangle - \langle R_1\rangle) + (1-\eta)\delta(\langle R\rangle - \langle R_2\rangle)$, then one obtain Eq. B.2. However, extensive heterogeneity in chromosome organization implies that $\langle R\rangle$ could take arbitrary values with a distribution, $P(\langle R\rangle)$. The left side of Eq. 4.20 is the experimentally measured cumulative distribution function and $\mathrm{CDF}(R|\langle R\rangle)$ on the right hand side is given by Eq. B.4. The goal is to solve for $P(\langle R\rangle)$ in Eq. 4.20, which is Fredholm integral equation of the first kind. In this work, I solve Eq. 4.20 using a discretization scheme on grid points $(R_j, \langle R\rangle_i)$. Eq. 4.20 is replaced by a summation approximately,

$$\mathrm{CDF}(R_j) = \sum_i \omega_i P(\langle R\rangle_i)\mathrm{CDF}(R_j|\langle R\rangle_i) \tag{4.21}$$

where ω_i are the weight coefficients for a quadrature formula. If one use small and equal grid size $\Delta_{\langle R\rangle} \to 0$, one can replace ω_i with $\Delta_{\langle R\rangle}$. Eq. 4.21 can be solved as a system of linear equations using non-negative Tikhonov regularization (see Appendix B.3).

4.3.5 Accounting for massive heterogeneity in chromosome organization:

In a recent study [118], which combined Hi-C and high-throughput optical imaging to map contacts within single chromosomes in human fibroblasts, revealed massive heterogeneity. Such extensive existence of a large number of conformations, leading to multiple or nearly continuous distribution of subpopulations, was much greater than previously anticipated. Although, the results in Figs. 4.5 and 4.6 quantitatively reveal heterogeneity associated with CTCF loops by considering only two dominant subpopulations, the most recent experiment requires a generalization of the theory. In principle, our theory also applies to interactions of any nature, not only the CTCF loops. In doing so, it may be more reasonable to assume a continuous distribution of subpopulations, $P(\langle R\rangle)$, (see section 4.3.4 for generalization) instead of two discrete subpopulations, $\langle R_1\rangle$ and $\langle R_2\rangle$, which of course is much simpler and may suffice in many cases as the results in Fig. 4.5 illustrate. To show that our theory has a broader range of applicability, I used the FISH data from the recent study [118], which reports spatial distance measurements for 212 pairs of loci. I obtained the raw data from the 4D Nucleome data repository [192]. Using non-negative Tikhonov regularization (Appendix B.3), $P(\langle R\rangle)$ is solved for each of a total of 212 pairs of loci. To illustrate our results, I compare in Fig. 4.7 the predicted CDF(r) and the experimentally measured CDF(r), as well as the $P(\langle R\rangle)$ obtained by fitting for six pairs of loci as examples in Fig. 4.7. The results show substantial variations in $\langle R\rangle$, manifested by the multiple peaks and wide spread

variations in $P(\langle R\rangle)$. Remarkably, the calculated CDF(r) (without any adjustable parameters) and the measured CDF(r) are in excellent agreement for the six loci pairs, which were arbitrarily chosen for illustration purposes. The residual errors between the two, shown as insets in Fig. 4.7, are extremely small.

In Fig. 4.8 I show the calculated the normalized distributions $P(\langle R\rangle/\mu(\langle R\rangle))$ for each of the 212 pairs of loci. I expect that $P(\langle R\rangle/\mu(\langle R\rangle))$ should be narrowly distributed around value 1 if there is only one population. However, many $P(\langle R\rangle/\mu(\langle R\rangle))$ show multiple peaks with large variations. To further quantify the extent of heterogeneity, I calculated the coefficient of variation, CV $= \sigma(\langle R\rangle)/\mu(\langle R\rangle)$ where $\sigma(\langle R\rangle)$ and $\mu(\langle R\rangle)$ are the standard deviation and the mean of $\langle R\rangle$, respectively. If there is only one population associated with $\langle R\rangle$, CV should have a value of around zero. Fig. 4.8b shows the histogram of CV for all 212 pairs of loci. The CV values are widely distributed, suggesting that 3D structural heterogeneity is common and is associated with many pairs of loci rather than a few. Thus, the analyses of experimental data is not possible without taking heterogeneity into account. The theory presented here is sufficiently general and simple that it can be used to calculate the measurable quantities readily.

4.3.6 Loop extrusion as a possible physical mechanism for chromosome heterogeneity:

What is the origin of heterogeneity in the individual cell populations? There are two possibilities. The first one is "static heterogeneity": each subpopulation

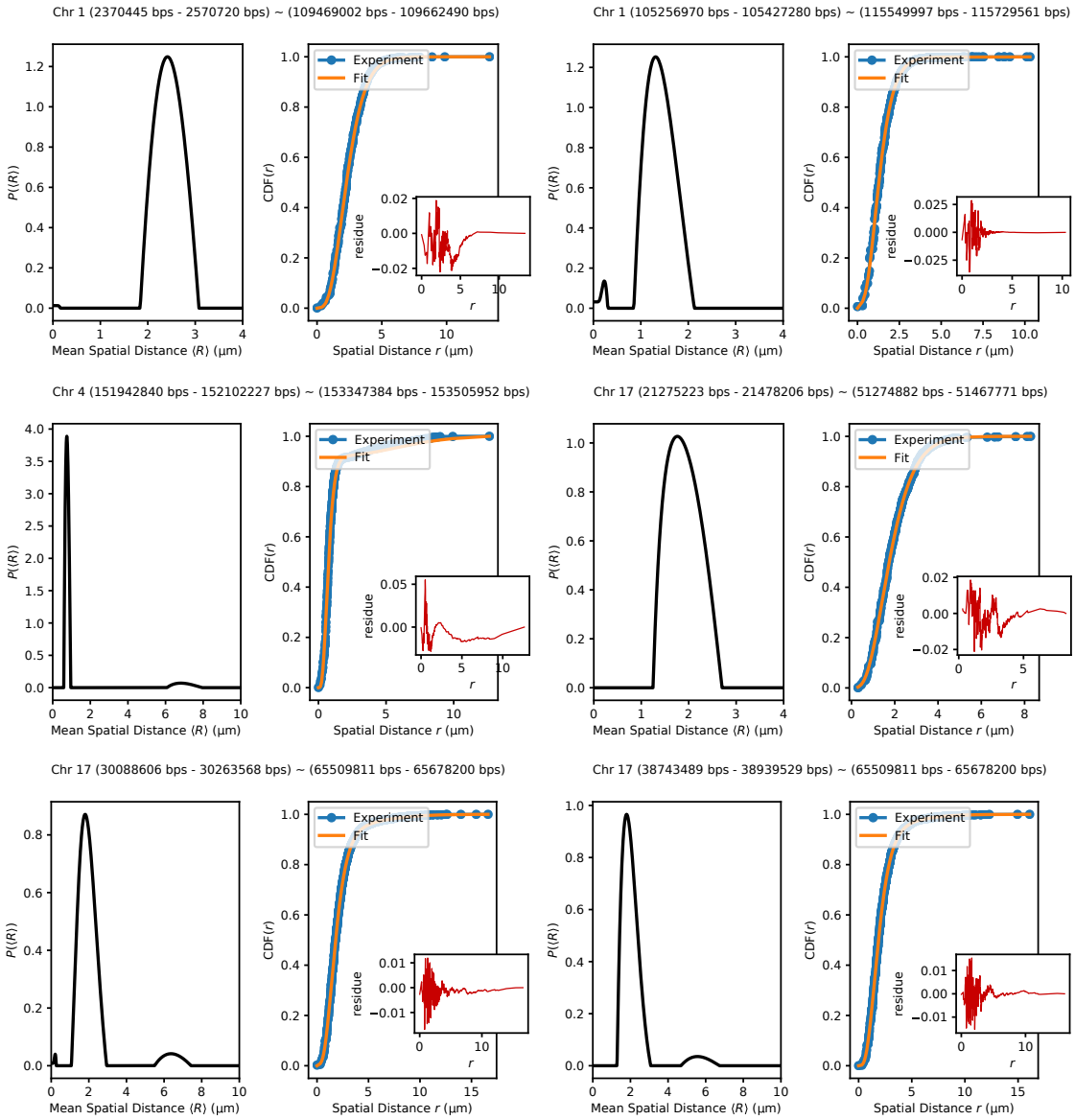


Figure 4.7: Exampled fits of CDF(r) using Eq. 4.20 to the experimental data [118]. The six exampled pairs of loci are indicated above each subfigure. Orange lines, showing the fits using our theory, is indistinguishable from experiment (the differences between fitted and experimental curve are shown in the insets). The distribution $P(\langle R\rangle)$ given in the integral equation (Eq. 4.20) is solved using non-negative Tikhonov Regularization (Appendix B.3). As shown here, $P(\langle R\rangle)$ have multi-peaks and are widespread, which is a manifestation of heterogeneity. I set $g = 1$ and $\delta = 5/4$.

explores a distinct region of the genomic folding landscape (GFL) (Fig. 4.9a). The second is "dynamic heterogeneity". Each cell explores a local minimum of the GFL before transiting to another local minimum (Fig. 4.9b). The only assumption in the application of our theory to genome organization is that there must be more than one population of cells, which does not violate the observation that the Hi-C experiment report only the average contact probability over millions of cells. Dynamic looping would be an example of the dynamic heterogeneity where the CTCF/cohein mediated loops are formed and broken dynamically on a fast time scale compared to the life time of a cell. Such a picture is supported by recent single-cell molecule experiment [193, 194]. The average residence time of CTCF/cohesin complex is shown to be in the range of a few to tens of minutes, which is much smaller compared to the time scale of cell cycle (15-30 hours). Loop extrusion model [62, 87, 88] is another possible origin of dynamic heterogeneity. In the loop extrusion model, it is thought that cohesins extrude loops along the chromsome fiber, which could detach stochastically. At any given time, there would be many subpopulations, each characterized by a distinct set of loops in the chromosome. Indeed, our analyses of the most recent high throughput optical imaging data lends credence to the notion that mutiple subpopulations in chromosomes arise because of massive dynamic heterogeneity. Our theory also gives an indirect theoretical justification for the work in [168] in which the authors found the loop extrusion model could lead to the $[P_{mn}, \langle R_{mn} \rangle]$ paradox.

Single-cell temporal information is necessary to determine whether the loops are static or dynamic or a combination of the two (Fig. 4.9c). Hence, the combi-

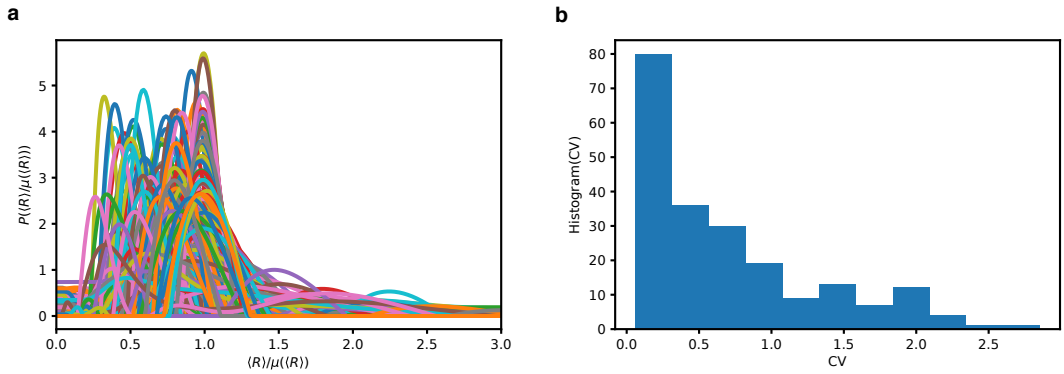


Figure 4.8: **(a)** Normalized distribution $P(\langle R\rangle/\mu(\langle R\rangle))$ ($\mu(\langle R\rangle)$ is the mean of $\langle R\rangle$) for all the 212 pairs of loci reported in [118]. For almost every pair of loci, the associated $P(\langle R\rangle/\mu(\langle R\rangle))$ has multiple peaks and is widespread. **(b)** Histogram of the coefficient of variations CV for all 212 pairs of loci probed in [118]. The CV values are calculated for each pair of loci, using CV $= \sigma(\langle R\rangle)/\mu(\langle R\rangle)$ where $\sigma(\langle R\rangle)$ is the standard deviation of $\langle R\rangle$. For a large number of loci pairs, CV exceeds 0.5, which is a quantitative measure of the extensive heterogeneity noted in experiment [118]

nation of the dynamic FISH technique such as CRISPR-dCas9 FISH and single-cell Hi-C would be crucial for us to fully understand the organization of genomes. Our theory provides a theoretically rigorous method based on polymer physics to connect the results from measurements using the two vastly different techniques.

4.4 Discussion

From polymer physics for single chains it follows that in a homogeneous system, the contact probability and mean 3D distances are linked, resulting in a power law relation connecting the two quantities that can be measured using Hi-C and FISH techniques. However, the one-to-one mapping does not hold in Hi-C experiments because of the presence of a mixture of distinct cell subpopulations each

characterized by its own statistics leads to heterogeneity, which in turn gives rise to the $[P_{mn},\langle R_{mn}\rangle]$ parodox. I have shown that the theory based on precisely solvable GRMC could be used to solve the paradox in practice. The theory can be readily used to analyze data from experiments, provided the FISH and Hi-C experiments are done under similar conditions [28]. The central result of the theory in Eq. 4.19 can be used to analyze the available sparse FISH data. I showed that the fraction of cell subpopulations (η in Eq. 4.19) and the generalization derived in section 4.3.4 can be extracted by fitting the FISH data using our theory. From Eq. 4.19 I calculated the Hi-C contact probabilities, thus establishing that the theory resolves the $[P_{mn},\langle R_{mn}\rangle]$ paradox.

In this work, I confined ourselves to two-point interactions, which allows us to consider one pair of loci at a time. However, recent experiments probing multi-point interactions have suggested that formations of loops are likely to be cooperative [30, 195], such that the formation of one loop could facilitate the formation of a nearby loop. Such cooperative loop formation was previously shown in an entirely different context involving folding of proteins directed by disulfide bond formation [196]. It can be shown within our framework that the formation of one loop can certainly increase the probability of formation of another loop.

The reconciliation of the FISH and Hi-C data using polymer physics concepts is the first key step in integrating the data from these experimental techniques to construct the 3D structures of chromosomes. The work described here provides a theoretical basis for accomplishing this important task. Finally, our results suggest that heterogeneity in contact formation is an intrinsic property of genome organiza-

tion, and hence acquisition of single-cell experimental data is crucial to further our understanding of both the dynamics and the heterogeneous structural organization of chromosomes.

4.5 Summary

Here, I first establish a relationship between the contact probability and the mean spatial distance using an analytically solvable Generalized Rouse Chromosome Model (GRMC), which incorporates the presence of CTCF/cohein mediated loops. The GRMC may be thought of as an ideal chromosome model, very much in the spirit of the Rouse model for polymers, in which conceptual issues such as the origin of the FISH-Hi-C paradox can be rigorously established. I first consider the solvable homogeneous limit in which contacts are present in all the cells. In this case, precise numerical and analytical results show that there is a simple relation between the contact probability, P, and the ensemble mean 3D distance $\langle R \rangle$. However, the unavoidable heterogeneity in the cell populations in Hi-C experiments, results in contacts between loci only in a fraction of cells. I first show that a direct consequence of the heterogeneity in both GRMC and chromosomes is that two loci (m and n) that have higher probability (P_{mn}) of being in contact relative to another two loci (k and l) does not imply a direct spatial correlation, a finding that has already been qualitatively established in previous studies [168, 183]. In other words, the average spatial distance between m and n ($\langle R_{mn} \rangle$) could be larger than $\langle R_{kl} \rangle$, the distance between loci k and l, even if $P_{mn} > P_{kl}$. These results provide a basis for

understanding the origin of FISH-Hi-C paradox.

By building on the GRMC results, I show that heterogeneity is the dominant feature of chromosome organization. Indeed, recent single-cell Hi-C [79, 116, 117] and imaging experiments [14, 24, 30, 118] have revealed that there are substantial cell-to-cell variations on genome organization. However, how to utilize the data reported in these experiments to enhance our understanding of 3D genome structural heterogeneity has not been unexplored. One approach is to create an appropriate polymer model based on Hi-C and imaging data, which would readily allow us to probe the structural variability using simulations [64,66,67,197]. Indeed, it has been shown, using Hi-C and FISH data as well simulations [197], that if the conformation of the chromatin fiber is taken to be homogeneous then trends observed in the FISH data could not be predicted. However, using simulations and including two levels chromatin organization (open and compact) qualitative trends observed in the FISH data could be recovered [197].

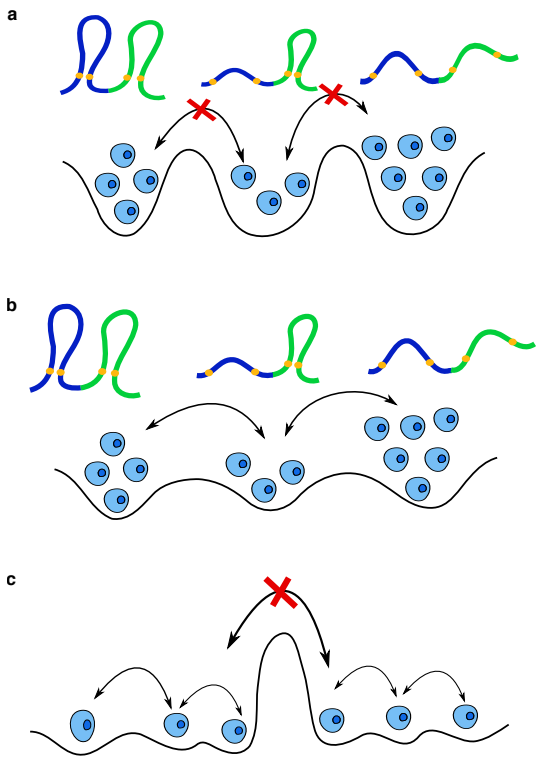


Figure 4.9: Schematic of the Genomic Folding Landscape (GFL). **(a)** Static heterogeneity: Cell subpopulation occupies distinct local minima in the GFL, with each minimum representing a stable organization. The energy barrier is too large for transition between different local minima on a biological time scale (one cell cycle). **(b)** Dynamical heterogenentiy: The energy barrier between local minima on the langscape is small enough which allows the dynamic transition between different subpopulations. **(c)** Combination of two different types of heterogeneity. In all three scenarios, the $[P_{mn}, \langle R_{mn} \rangle]$ paradox arises. The loci contacts are in orange. The polymer conformation sketches are not shown in this scenarios due to insufficient space.

Chapter 5: Reconstruction of three-dimensional chromosomes organization from Hi-C contact map

5.1 Introduction

Hi-C data describes the chromosome structures in statistical terms expressed approximately in terms of a matrix the element of which indicates the probability that two loci separated by a specific genomic distance are in contact. How can I go beyond the genomic contact information to 3D distances between loci, and eventually the spatial location of each locus is an important problem that has to be solved in order to exploit the available data quantitatively. Imaging techniques, such as Fluorescence *In Situ* Hybridization (FISH) and its variations, are the most direct way to measure the spatial distance and coordinates of genomic loci. But currently these techniques are limited in that they provide information on only a small number of loci in one experiment set up. Is it possible to harness the power of the two methods to construct, at least approximately, 3D structures of chromosomes? Here, I answer this question in the affirmative by building on the precise results for Generalized Rouse Chromosome Model and by using certain universal principles of polymer.

A number of data-driven approaches have been developed in order to go from Hi-C to 3D structure of genomes [71–78] (see the summary in [80] for additional related studies). However, no attention is paid to the theoretical aspects relating contact probability and 3D spatial distances from a polymer physics perspective. As I have shown in Chapter 4, the apparent difficulties in reconciling Hi-C (contact probability) and FISH data (spatial distances) is caused by the fact that the cell population is heterogeneous even though they are synchronized in the Hi-C experiment. As a result, a given contact is not present with a fixed probability in all the cells.

The purposes of the chapter 5 are two folds. (1) I first establish that there is a lower theoretical bound connecting the contact probability and the 3D distance. I test this concept by using the GRMC polymer for which accurate simulations can be performed. (2) However, distances, R_{ij}s, between the loci do not give the needed coordinates of each locus. In order to solve this problem, I rely on the lessons from GRMC and polymer physics concept and used them to obtain the individual 3D coordinates of the loci. The method allows us to go from the Hi-C contact map to the three-dimensional coordinates, $\boldsymbol{R}_i$ ($i = 1, 2, 3, \cdots, N_c$), where N_c is the length of the chromosome) may be summarized as follows. First, I construct the average distances $\langle R_{mn} \rangle$ between all m and n using a power-law relation P_{mn}, the probability m and n are in contact measured in Hi-C experiments, and $\langle R_{mn} \rangle$. The justification for the power law is established using GRMC and polymer physics concepts. I obtain, $\boldsymbol{R}_i$, the 3D coordinates for all the loci from $\langle R_{mn} \rangle$ using Multidimensional Scaling. The application of our theory to decipher the 3D structure of chromosomes

from any species is limited only by experimental resolution of the Hi-C technique. Comparison with experimental data are made to validate our theory.

5.2 Results

5.2.1 Inferring distance map (DM) from contact map (CM) in a homogeneous system:

In GRMC, the relation between the contact probability and mean spatial distance is given by (see section 4.2.1 for derivation),

$$P_{mn} = \mathrm{erf}\left(\frac{2r_c}{\sqrt{\pi}\langle R_{mn}\rangle}\right) - \frac{4}{\pi}\frac{r_c}{\langle R_{mn}\rangle}e^{-\frac{4r_c^2}{\pi\langle R_{mn}\rangle^2}} \equiv R_0(\langle R_{mn}\rangle). \tag{5.1}$$

where $\mathrm{erf}(x)$ is the error function. The equation above provides a way to infer the distance map (DM) directly from the contact map (CM), which is a matrix whose elements, P_{mn}, specifies the contact probability between loci m and n. The CM can be inferred approximately using Hi-C experiments. However, there are uncertainties in determining both r_c due to systematic uncertainties and P_{mn} due to inadequate sampling, thus restricting the use of Eq.5.1 in practice. In light of this, I address the following questions, which I answer using a precisely solvable model. (a) How accurately can one solve the inverse problem of going from the contact map to the distance map? (b) Does the inferred distance map faithfully reproduce the topology of the spatial organization of a model for chromosomes?

To answer these two questions, I first constructed the distance map by solving

Eq.5.1 for $\langle R_{mn} \rangle$ for every pair with contact probability P_{mn}. The CM is determined using simulations of the GRMC, as described in the Methods. For such a large polymer, some contacts are almost never formed even in long simulations, resulting in $P_{mn} \approx 0$ for some loci. This would erroneously suggest that $\langle R \rangle \to \infty$, as a solution to Eq.5.1. Indeed, this situation arises often in the Hi-C experimental contact maps where $P_{mn} \approx 0$ for many m, n. To overcome this practical problem of dealing with $P_{mn} \approx 0$ for several pairs, I apply the block average (a coarse graining procedure) to the CM, which decreases the size of the CM. The procedure decreases the problem of having to deal with the vanishingly small values of P_{mn} while preserving the information needed to solve the inverse problem using Eq.5.1.

The simulated and constructed distance maps are shown in the lower and upper triangle, respectively, for the purpose of better visual comparison (Fig. 5.1a). I surmise from Fig. 5.1(a) that the constructed and simulated distance maps are in excellent agreement. There is a degree of uncertainty for the loci pairs with large mean spatial distance (elements far away from the diagonal in Fig. 5.1(a)) due to the unavoidable noise in the CM. To assess the quality of the constructed distance map, I found that the Spearman correlation coefficient between the simulated and theoretically constructed maps is 0.97. However, a single correlation coefficient is not sufficient to capture the topological structure embedded in the distance map. To assess the global similarity between the DMs from theory and simulations, I used the Ward Linkage Matrix (see Appendix A.3). Fig. 5.1b shows that the constructed DM indeed reproduces the hierarchical structural information correctly. The results in Fig. 5.1 show that the DM, in which the elements represent the mean distance

between the loci can be calculated accurately, as long as the CM is determined unambiguously.

5.2.2 A bound for the spatial distance inferred from contact probability:

The results in Fig. 5.1 show that for a homogeneous system (specific contacts are presented in all realization), the mean 3D distance map can be faithfully inferred/reconstructed solely from the contact map. However, the discrepancies between FISH and Hi-C data suggests that the cell population are hetergenous, which means that contact between m and n loci are present in only a fraction of cells. In this case, which one has to contend with in practice, the one-to-one mapping between contact probability and mean 3D distances (Eq.5.1) does not hold. This leads to the paradox described in Chapter 4, which means that higher contact probability does not imply closer distance. This implies that given the contact probability, one can no longer determine the mean 3D distance uniquely, which implies that for certain loci the results of Hi-C and FISH must be discordant. For a mixed population of cells, the contact probability P_{mn} and mean spatial distance $\langle R_{mn} \rangle$ between two loci m and n, are given by,

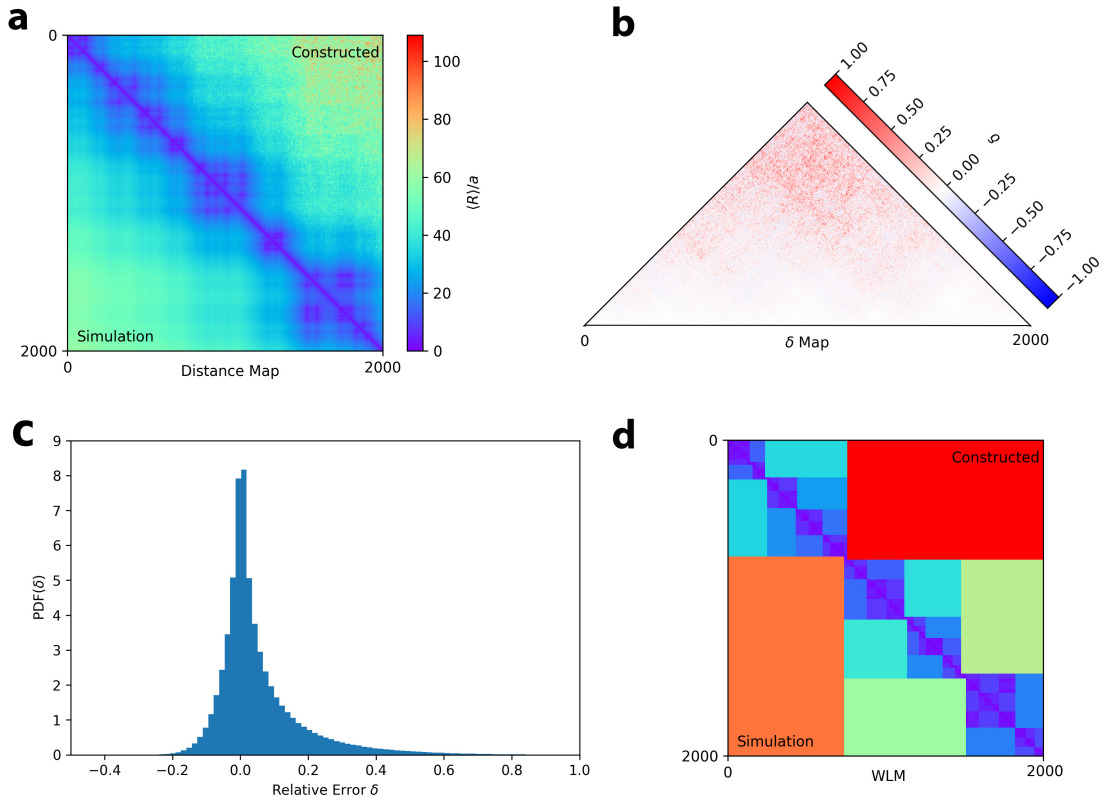


Figure 5.1: Comparison of the distance matrices (DMs) for GRMC. **(a)** The simulated DM (lower triangle) and constructed DM (upper triangle) are compared side by side. The colorbar indicates the value of the mean spatial distance, $\langle R_{mn} \rangle$. The constructed DM is obtained by solving Eq.5.1 using the CM. The value of $r_c = 2.0a$. The location of loop anchors are derived from experimental data [28] over the range from 146 Mbps to 158 Mbps for Chromosome 5 in the Human GM12878 cell. **(b)** Relative error δ as a map. The relative error is calculated as, $\delta = (d_{\text{inferred}} - d_{\text{sim}})/d_{\text{sim}}$ where d_{inferred} and d_{sim} are inferred and simulated distance, respectively; δ increases for loci with large genomic distance indicating tendency to overestimate the distances. **(c)** The distribution of the relative error, PDF(δ). The mean value of absolute relative error is 0.08 suggesting that on an average the inferred distance deviates from simulation by only 8% due to the statistical errors. **(c)** Ward Linkage Matrices (WLMs) from simulation and theoretical prediction, are shown in the lower and upper triangle, respectively, are in excellent agreement with each other.

$$\langle R_{mn} \rangle = \sum_{i}^{S} \eta_{i,mn} \langle R_{i,mn} \rangle \tag{5.2}$$

$$P_{mn} = \sum_{i}^{S} \eta_{i,mn} P_{i,mn} \tag{5.3}$$

where $\langle R_{i,mn} \rangle$ and $P_{i,mn}$ are the mean spatial distance and contact probability between m and n in i^th subpopulation, respectively. S is total number of distinct subpopulations and $\eta_{i,mn}$ is the fraction of the subpopulation i in the total population which satisfies the constraint $\sum_i^S \eta_{i,mn} = 1$. Although there exists a one-to-one relation between $P_{i,mn}$ and $\langle R_{i,mn} \rangle$ in each subpopulation i. It is no longer possible to determine P_{mn} solely from R_{mn} without knowing values of each $\eta_{i,mn}$ and *vice versa*.

In Chapter 4, I show that paradox arises precisely because of the mixing of different subpopulations. The value $\eta_{i,mn}$ in principle can be extracted from distribution of $R_{i,mn}$ which can be measured using FISH technique. However this is usually unavailable which leads us to the question: despite of the lack of knowledge of the composition of cell populations, can I provide an approximate relation between P_{mn} and $\langle R_{mn} \rangle$? In other words, rather than answer question (a) precisely, as I did for the homogeneous GRMC, I are seeking an approximate solution. The GRMC calculations provide the needed insight to construct the approximate relation to calculate DM from CM. Here I demonstrate that there exists a theoretical lower bound of $\langle R_{mn} \rangle$ given the value of P_{mn} no matter what are the compositions

of the whole cell population. I demonstrate this by considering the case $S = 2$ in which $\langle R_{mn} \rangle = \eta \langle R_{mn} \rangle + (1 - \eta) \langle R_{2,mn} \rangle$ and $P_{mn} = \eta P_{1,mn} + (1 - \eta) P_{2,mn}$. When the value of the contact probability P_{mn} is known but the value of η is unknown, the possible values of $\langle R_{1,mn} \rangle$ and $\langle R_{2,mn} \rangle$ follows the contour lines (dashed line) in Fig. 5.2. There exists a contour line for $\langle R_{mn} \rangle$ which is tangent to the contour line for P_{mn} (green curve in Fig. 5.2) for all m and n. The tangent point (star in Fig. 5.2) corresponds to the minimum possible value for $\langle R_{mn} \rangle$. Thus, although one cannot precisely determine the mean spatial distance from the contact probability, the GRMC result suggests a precise lower bound to $\langle R_{mn} \rangle$, which can be calculated from P_{mn}. Detailed calculations (Appendix C.1) show that such tangent points are on the linear line $\langle R_{1,mn} \rangle = \langle R_{2,mn} \rangle$ (solid black line in Fig. 5.2). Remarkably, the lower bound is exactly the value of $\langle R_{mn} \rangle$ as if there is only a single homogenous population. It is important to emphasize that this lower bound holds generally for any number of subpopulations and any function form of R_0 as long as R_0 is a monotonic function of R_{mn} (Appendix C.1). This finding, which I proved here using precise numerical solution for the GRMC, is remarkably useful in predicting the approximate spatial organization of chromosomes from Hi-C contact map, as I demonstrated below. For the GRMC, I have $\langle R_{mn} \rangle \geq R_0(P_{mn})$. Thus, the precisely solvable model suggests that the approximate power law relating P_{mn} and R_{mn} could be used as a starting point in constructing spatial distance matrices using only the Hi-C contact map for chromosomes.

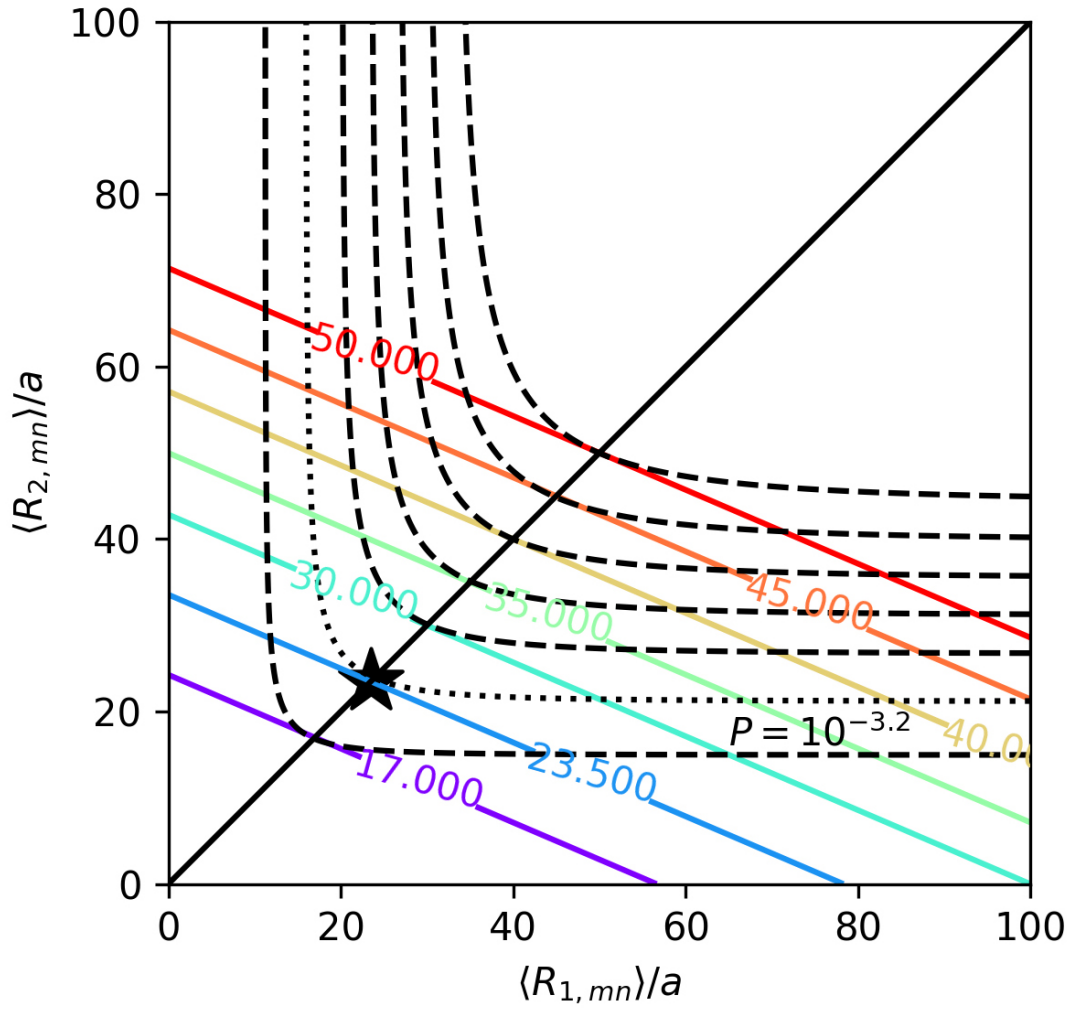


Figure 5.2: Lower Bound illustrated graphically. The colored lines are the contour lines of $\langle R_{mn}\rangle$ with their values marked. The dotted curve is the contour with constant $P_{mn} = 10^{-3.2}$. The line with $\langle R_{mn}\rangle = 23.5a$ is tangent to the constant P_{mn} are at the intersection marked by $(*)$, which gives the lower bound for $\langle R_{mn}\rangle$. The black line has slope 1. Several other contour lines are shown to illustreate that all the tangent points lie on this line (Appendix C.1).

5.2.3 Validating the lower bound between P_{mn} and R_{mn} when heterogeneity matters:

In order to investigate the effect of heterogeneity (contact between m and n for all (m, n) pairs does not exist in all the cells) on the quality of the constructed DM from CM, I simulated a model system where there are two distinct populations, one with all CTCF mediated loops present (fraction η), and the other being the polymer chain without any loop constraints (fraction $1-\eta$). I used the lower bound $R_0^{-1}(P_{mn})$ to infer $\langle R_{mn} \rangle$ from P_{mn}. The results, shown in Fig. 5.3(a), provide a numerical verification of the theoretical lower bound linking contact probability and mean spatial distance. Using the $R_0(P_{mn})$, the DMs shown in Fig. 5.3(b) are calculated from the the simulated CMs. Note that $\eta = 0.0$ and $\eta = 1.0$ correspond to the Rouse chain (no CTCF mediated loops) and the GRMC (all CTCF mediated loops are present), respectively. Interestingly, $\eta = 0.3$ results in variations in the simulated CM but has hardly any effect on the simulated DM (second column in Fig. 5.3(b)). The difference matrices between constructed and the simulated DMs are shown in the third column in Fig. 5.3(b). For $\eta = 0.3$, the difference between the constructed and simulated DMs is largest near the loops resulting in an underestimate of the spatial distances in the proximity of loops (Fig. 5.3(b)). This occurs because the constructed DM is computed from the simulated CM, which is sensitive to the heterogeneity of cell population. For the system with large $\eta = 0.7$, the constructed DM agrees better with the the simulated DM (see the difference matrix in the third column in Fig. 5.3(b)), suggesting that the accuracy of inferring the mean spatial

distance is less affected when the majority of the cells have CTCF mediated loops present during the measurement. The difference matrices for $\eta = 0.3$ and 0.7 show that although the constructed DMs underestimated the spatial distances around the loops most of pairwise distances are hardly affected, thus justifying the use of the lower bound as a practical way to construct DM.

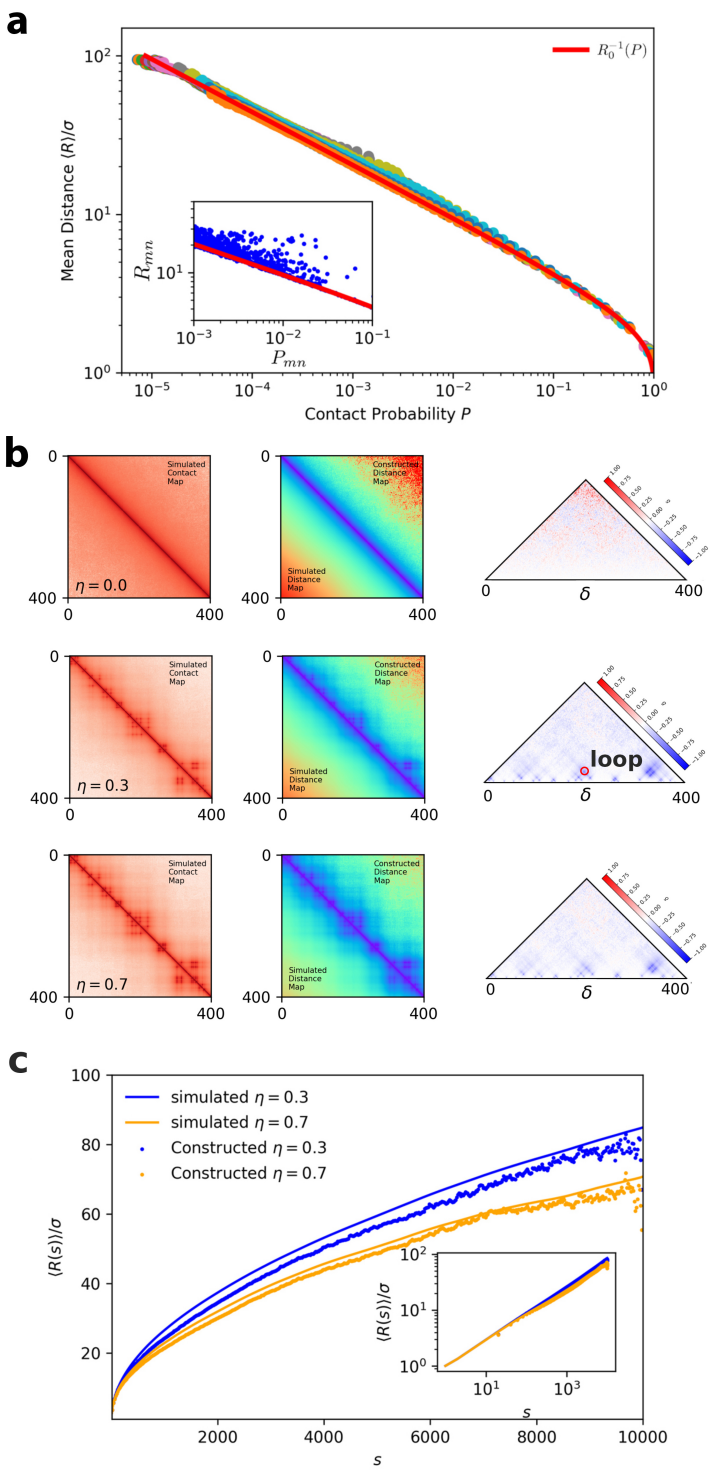

a
Mean Distance ⟨R⟩/σ
Contact Probability P
$R_0^{-1}(P)$
R_{mn}
P_{mn}
b
Simulated Contact Map
Constructed Distance Map
Simulated Distance Map
η = 0.0
η = 0.3
η = 0.7
loop
δ
c
simulated η = 0.3
simulated η = 0.7
Constructed η = 0.3
Constructed η = 0.7
⟨R(s)⟩/σ
s

Figure 5.3: **(a)** Mean spatial distance $\langle R \rangle$ as a function of the contact probability P. Red solid curve is given by $R_0^{-1}(P)$, which is the lower bound. Dots are computed from simulation using the binned average method described previously. Different colors represents different values of $\eta = (0.0, 0.005, 0.02, 0.04, 0.06, 0.08, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0)$. On an average, the simulated data is well described by the theoretical lower bound. The inset shows the mean spatial distance versus contact probability for each pairs at $\eta = 0.3$. Each blue dot represents one pair $(P_{mn}, \langle R_{mn} \rangle)$, and the red curve is $R_0(P)$. The simulated data is only slightly above the theoretical lower bound. **(b)** Simulated CM (left column), simulated DM and constructed DM side by side (middle column) and relative error δ (right column) for different values of $\eta = (0.0, 0.3, 0.7)$ for GRMC. The CMs are shown on $\log_{10}$ scale with darker red representing higher contact probability. In both the simulated and constructed DMs, darker red represents higher mean spatial distance. The constructed DM is obtained using $\langle R_{mn} \rangle = R_0^{-1}(P_{mn})$. Colorbars mark the value of δ in which the blue and red represents negative and positive values of δ. Negative values of δ indicates inferred spatial distance is smaller than the actual distance. The red circle in the δ matrix for $\eta = 0.3$ marks one loop. **(c)** Plots of $\langle R(s) \rangle$ as a function of the genomic distance, s, for $\eta = 0.3$ and 0.7. The inset shows the same data on a log-log scale; $\langle R(s) \rangle$ is calculated using $\langle R(s) \rangle = (1/TM) \sum_{a=1}^{M} \sum_{t=1}^{T} \left(|\boldsymbol{r}_i^{(a)}(t) - \boldsymbol{r}_j^{(a)}(t)| \delta(s - |i-j|)/(N-s) \right)$. The theoretical predictions are in agreement with simulations.

To show that the constructed DMs using the lower bound give good global description of the system, I also calculated the often-used quantity $\langle R(s) \rangle$, mean spatial distance as a function of the genomic distance s, as an indicator of average structure (Fig. 5.3(c)). The constructed $\langle R(s) \rangle$ differs only slightly from the simulation results. Notably the scaling of $\langle R(s) \rangle$ versus s is not significantly changed (inset in Fig. 5.3(c)), strongly suggesting that constructing the DMs using the lower bound gives a fairly good estimate of the average size of the chromosome segment.

5.2.4 Inferring 3D organization of interphase chromosomes from experimental Hi-C contact map:

To apply the insights from the study of GRMC to obtain 3D organization of chromosomes, I use the generalized power law relation relating P and $\langle R \rangle$ for chromatin. It is,

$$\langle R_{mn} \rangle = \Lambda P_{mn}^{-1/\alpha} \tag{5.4}$$

where α and Λ are unknown coefficients. For the GRMC, $\Lambda = r_c$ and $\alpha = 3.0$. For a self-avoiding polymer, $\alpha \approx 3.71$ for two interior loci that are in contact (see section 4.2.3). Based on experiments [14] and CCM results (see Fig. 3.4) a tentative suggestion could be made that $\alpha \approx 4.0$. I show below that the power law relation Eq.5.4 provides a way to infer the approximate 3D organization of chromosomes from experimental Hi-C contact map.

I first ask if the value of α could be determined from Hi-C contact map? To answer this question, I use the Multidimensional scaling (MDS) [198, 199], which is used to generate the coordinates of objects in such manner that the between-object distances ($\langle R_{mn} \rangle$ is our case) are preserved as precisely as possible. Recently, MDS has been specifically applied to reconstruct 3D chromosome structure [76, 77, 200]. When only the Euclidean distances among objects are known, it can be used to solve for the exact configurations from which these distances can be calculated using the methods described elsewhere [199]. Thus, it is reasonable to assume that the distance

map inferred using the correct α should give the most reasonable conformation, in the sense it would have the smallest Normalized Root Mean Square Error (nRMSE), $(\sum_{i<j}(R_{ij} - R_{ij}^{\mathrm{MDS}})^2 / \sum_{i<j} R_{ij}^2)^{1/2}$. Here, R_{ij} is the distance between the i^{th} and j^{th} loci in the inferred distance map calculated using Eq.5.4 and Hi-C contact map of 100 kbps resolution [28] and R_{ij}^{MDS} is the corresponding distance in the reconstructed 3D conformation using MDS.

Fig. 5.4(a) shows the nRMSE as a function of α for Human GM12878 Interphase Chromosome 1. The smallest nRMSE is obtained in the range $\alpha \approx 3.5 - 4.0$, with a minimum around $\alpha \approx 3.5 - 4.0$ for almost all the 23 chromosomes (Fig. C.1). Interestingly, the value of α inferred from MDS coincides with experimental data [14, 201] as well as the simulations based on the CCM (Fig. 3.4). Thus, I arrive at the important conclusion that the Human Interphase chromosome is best described by an exponent $\alpha \approx 3.5 - 4.0$. Without loss of generality, I use $\alpha = 4.0$ to reconstruct the 3D organization of all 23 human interphase chromosome. Fig. 5.4(b) shows the comparison between the inferred distance map and distance map from reconstructed configuration of Chromosome 1 (Chr1) using MDS. The Pearson correlation coefficient between R and R_{MDS} is 0.87 (Fig. 5.4(c)), a highly significant value.

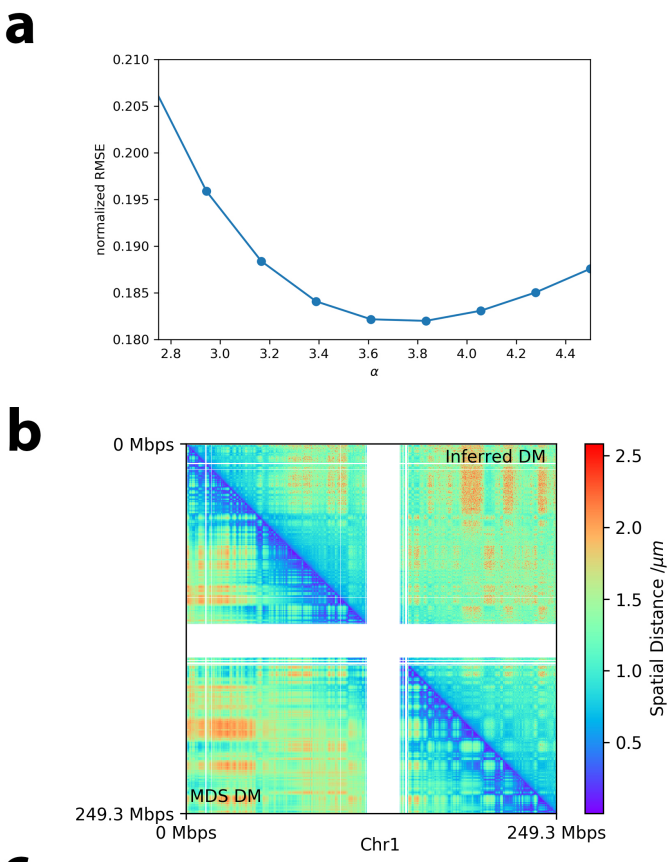

a
0.210
0.205
0.200
0.195
0.190
0.185
0.180
normalized RMSE
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
α
b
0 Mbps
Inferred DM
MDS DM
249.3 Mbps
0 Mbps
Chr1
249.3 Mbps
2.5
2.0
1.5
1.0
0.5
Spatial Distance /μm
c
2.5
2.0
1.5
1.0
0.5
0.0
MDS Distance R_MDS/μm
0.0
0.5
1.0
1.5
2.0
2.5
Inferred Distance R/μm

Figure 5.4: **(a)** Normalized Root Mean Squared Error (nRMSE) as a function α for Chromosome 1 (Chr1) in Human GM12878 cell; nRMSE is computed using, $(\sum_{i<j}(R_{ij} - R_{ij}^{\mathrm{MDS}})^2 / \sum_{i<j} R_{ij}^2)^{1/2}$, where $R_{ij} = \Lambda P_{ij}^{-1/4}$. I calculated R_{ij}^{MDS} using the coordinates $\boldsymbol{R}_i$ obtained from the distances R_{ij} using MDS. Note that nRMSE is a dimensionless quantity, thus the value of Λ has no effect on nRMSE. Here, I use $\Lambda = 1$. The Hi-C contact map of 100 kbps resolution is used [28]. **(b)** Side-by-side comparison between inferred distance matrix (upper triangle) and distance matrix of reconstructed structure using MDS (lower triangle). The white blank area corresponds to the missing data of the centromere in the Hi-C contact map [28]. Thus, the spatial distances for the centromere are not determined. **(c)** Scatter plot of $(R_{ij}, R_{ij}^{\mathrm{MDS}})$. The Pearson correlation coefficient between the two is 0.87. The dashed line is the linear fit with slope 1.08.

5.2.5 3D structure constructed using MDS:

The 3D configuration $\boldsymbol{R}_i$ ($i = 1, 2, 3, \cdots, N_c$ where N_c is the number of loci at a given resolution (the centromeres are discarded due to lack of information in Hi-C contact map). The values of N_c are given in Table.S1) of the 23 Human interphase chromosomes generated using MDS (with $\Lambda = 117$ nm; see below) are shown in Fig. 5.5. Color represents the genomic location of loci in which purple and red indicate the 5' and 3' ends, respectively. Fig. 5.6(a) and Fig. 5.5(b) shows the 3D and 2D MDS embedding reconstructed structrure of Chr1, respectively. These figures show that Chr1 folds hierarchically where the loci with small genomic distance (similar color) are also close in space. The long range intermingling between loci with large genomic distances (different color) is avoided. Such picture is remarkably consistent with the notion of crumpled globule [21,48], and also the recent single-cell Hi-C data [79]. Similar structural features are found for all 23 Human interphase chromosomes (Fig. 5.5). In addition, the reconstructed structure of Chr1 also shows clear

A/B compartments (Figs.5.6(c),(d)). Two compartments are spatially separated, suggesting the microphase separation between euchromatin and heterochromatin. Note that the arrangement of A/B compartments in a polarized manner is highly consistent with multiplexed FISH data [14] and single-cell Hi-C [79].

5.3 Experimental support

In order to further quantify the properties of the inferred 3D structure of chromosomes, I calculated the square of the radius of gyration of all 23 chromosomes using $R_g^2 = (1/2N_c^2)\sum_{i,j} R_{ij}^2$. The dashed line in Fig. 5.7(a) is a fit of R_g^2 as a function of chromosome size, which yields $R_g \sim N_c^{0.27}$ where N_c is the length of the chromosome. For a collapsed polymer, $R_g^2 \sim N_c^{2/3}$ and for an ideal polymer to be $R_g^2 \sim N_c$. To ascertain if the unusual value of 0.27 is reasonable, I computed the volume of each chromosome using $(4/3)\pi R_g^3$ and compared the results with experimental data [202]. The scaling of chromosome volumes versus N_c of inferred 3D chromosome structures are also in excellent agreement with the experimental data (Fig. 5.7(b)). The exponent of $0.27 \lesssim 1/3$ suggests the chromosomes overall adopt compact, space-filling structure, which is also vivdly illustrated in Fig. 5.5. Since the value of Λ (Eq.5.4) is unknown, I estimate it by minimizing the error between our inferred chromosome volumes and experimental measurements. I find that $\Lambda = 117$ nm, which gives an approximate size of locus of 100 kbps (the resolution of Hi-C map used in the analysis). It is noteworthy that genome density computed

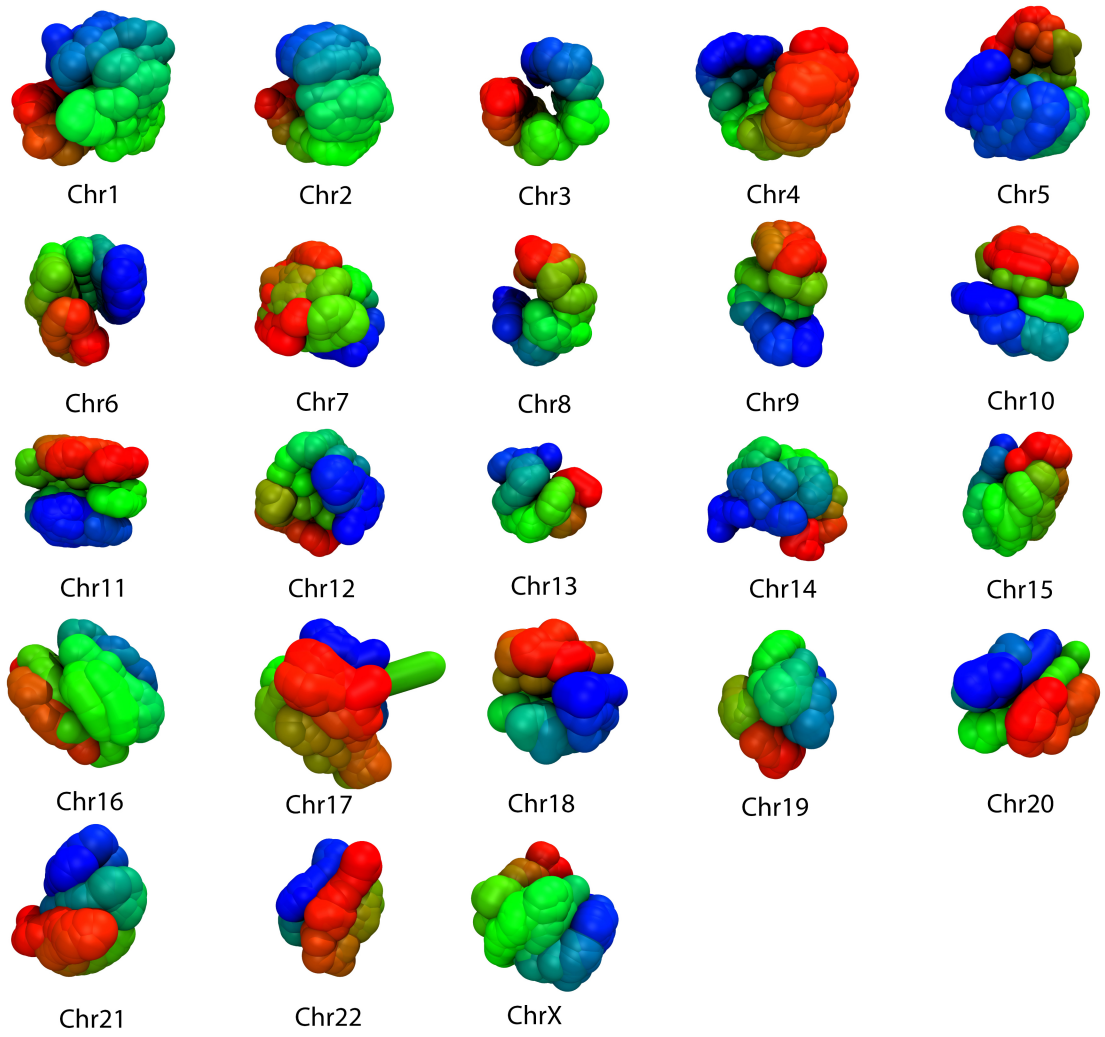


Figure 5.5: 3D reconstructed structure for all 23 Human interphase chromosomes using MDS with the inferred DM which is obtained using Eq.5.4 with with $\Lambda = 117$ nm and $\alpha = 4.0$. The colors encode the genomic position of the loci. The resolution of loci is 100 kbps. Red and purple represents 5' and 3' ends, respectively. The structures are rendered using VMD with bead radius of $\Lambda = 117$ nm.

using the value of Λ ($(100 * 10^3/(4/3)\pi\Lambda^3)\text{bps}\cdot\text{nm}^{-3} = 0.015\text{bps}\cdot\text{nm}^{-3}$) is consistent with the typical average genome density of Human cell nucleus $0.012\text{bps}\cdot\text{nm}^{-3}$ [52]. The value of Λ does not change the scaling but only the distances between the loci.

Topologically Associated Domains: It should be emphasized that the inferred distance is a metric but not Euclidean (see Appendix C.2 for details). One can only interpret the generated structures using MDS as average structures, which captures the global topology of chromosome organization. It is hard to infer the local structures like TADs in the MDS reconstructed structure. I find that the t-SNE embedding [203], which is known to preserve local structural variation [204] better than MDS, captures the polymer nature of the chromosome. The t-SNE embedding shows that the connected loci are constrained to form the backbone of the whole chromosome (Fig. 5.7(c)). In such a representation, TADs emerge as local structures represented by the small curls along the curve (Fig. 5.7(d)). To justify that these curls are indeed representations of TADs, which are maintained by CTCF/cohesin mediated loops, I apply two dimensional t-SNE embedding on GRMC simulations. With increasing η, the prominence of loops also increases. Fig. C.2 clearly shows that t-SNE embedding is able to capture the loops in the system. Thus, I conclude that the curls observed in Fig. 5.7(d) are actual representations of the TADs.

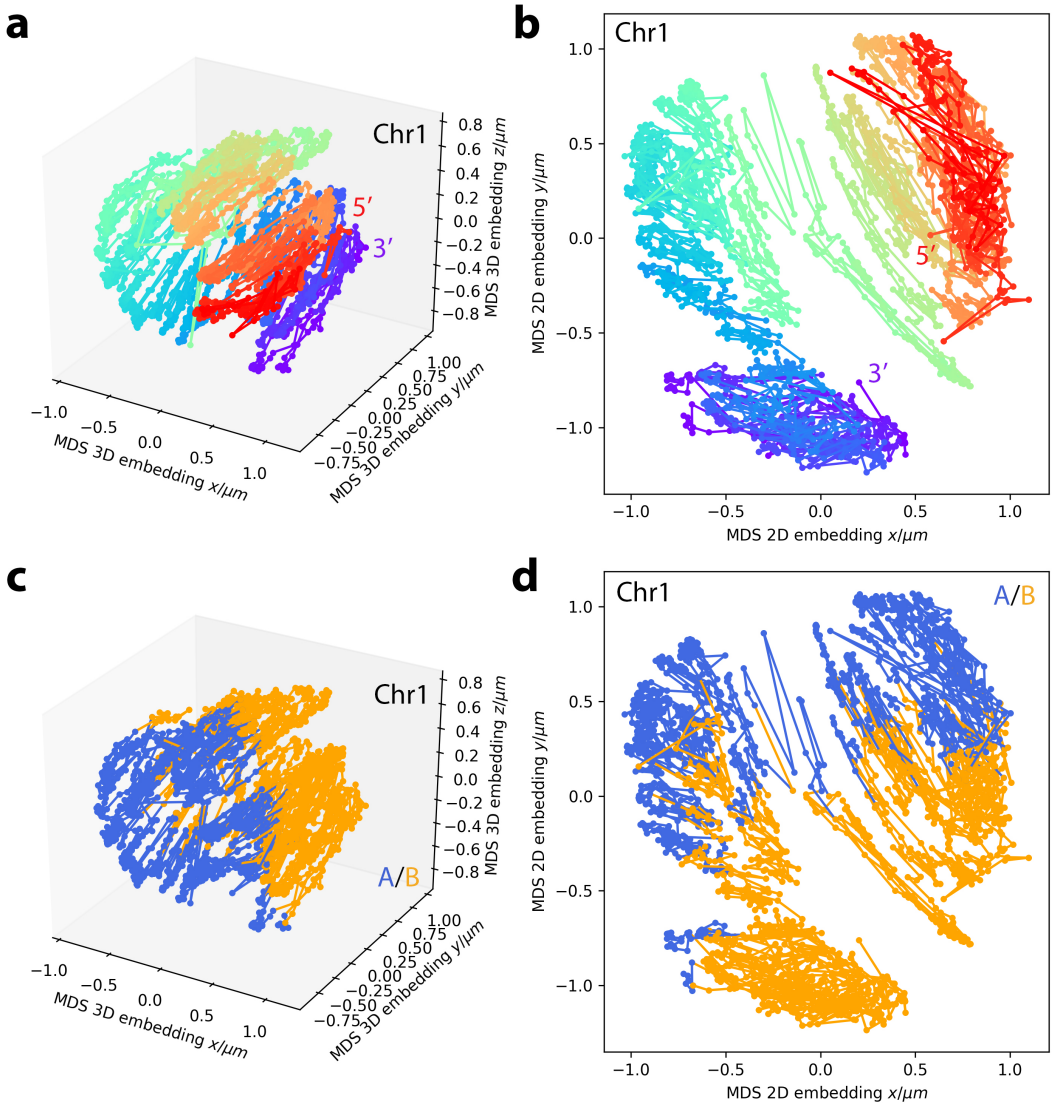


Figure 5.6: **(a)** 3D reconstructed structure for Chr1. Same as Fig. 5, but with point representation. The colors encode the genomic position of the loci. The resolution of loci is 100 kbps. Red and purple represents 5' and 3' ends, respectively. **(b)** Two-dimensional MDS embedded conformation. **(c)** A/B compartments of reconstructed Chr1 structure. Phase separation between two compartments are visually clear. A/B compartments are determined using spectral biclustering (Appendix A.2). **(d)** A/B compartments shown in two-dimensional MDS embedded structure.

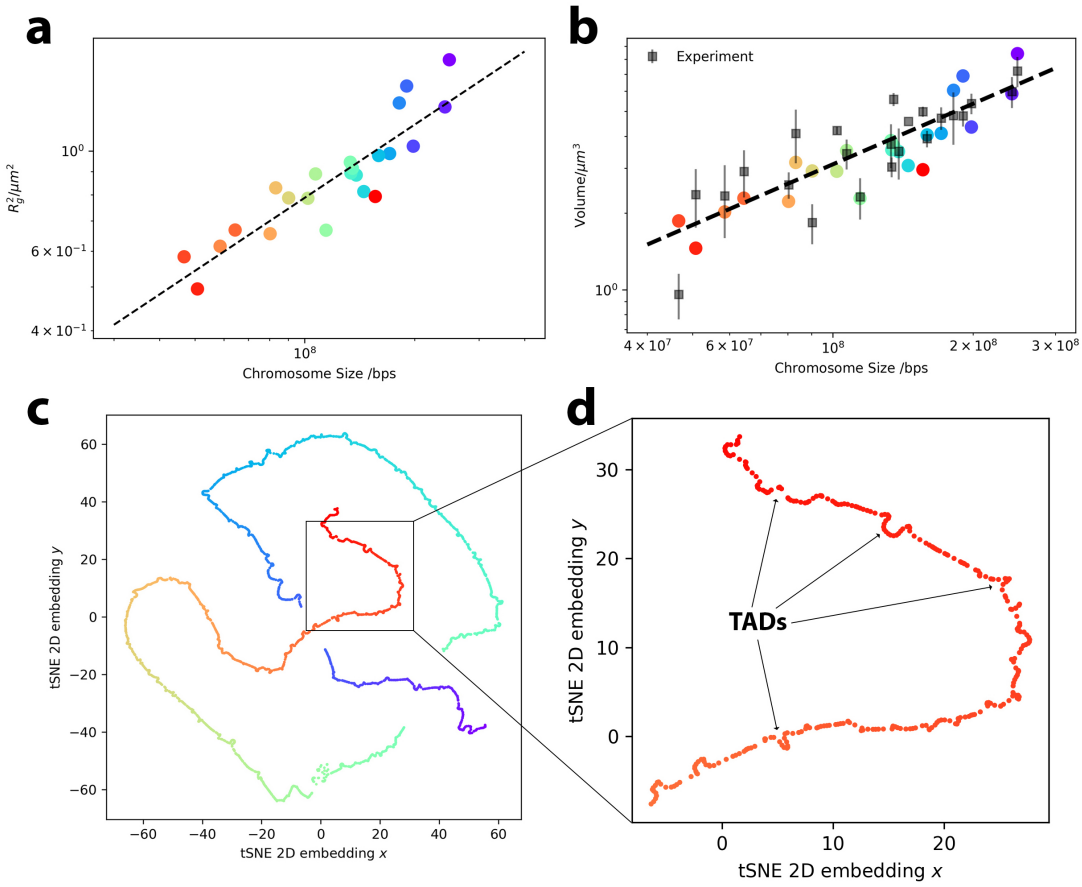


Figure 5.7: **(a)** Squared radius of gyration R_g^2 as a function of chromosome size. The dashed line is the fit to the data with the slope 0.54. **(b)** Volume of each chromosome versus the length in bps unit. Experimental values (black squares) are computed using the data in [202]. Dashed line is the fit to the experimental data with slope 0.8. Volume of each chromosome is calculated using λV_{nuc} where λ is the percentage of volume of nucleus volume V_{nuc}. The values of λ are provided in Fig. S5 in [202], and $V_{\text{nuc}} = (4/3)\pi r_{\text{nuc}}^3$ where $r_{\text{nuc}} = 3.5\mu m$ is the radius of Human lymphocyte cell nucleus [202]. Volumes of the reconstructed Chromosome using theory and computation are calculated using $(4/3)\pi R_g^3$ (color circles). The predicted and experimental values have a Pearson correlation coefficient 0.79. The excellent agreement validates the procedure to construct 3D organization. **(c)** tSNE 2D embedding from inferred DMs. **(d)** Local structures of TADs are preserved and illustrated as small curls along the backbone of the chromosome (shown in the figure).

5.4 Discussion and conclusion

Using the theoretical results and precise numerical simulations of a non-trivial model, I have provided an approximate solution to the problem of how to construct the three-dimensional coordinates of each locus from the measured probabilities (P_{mn}s) that two loci are in contact. The key finding that P_{mn} is related to $\langle R_{mn} \rangle$ through a power law, which is in accord with experiments as well as accurate models for interphase chromosomes. The distance measures are then used to obtain the coordinates of the loci using multidimensional scaling. This physically motivated procedure is self-consistently accurate for the precisely solvable GRMC, and was used to construct the 3D organization of the twenty three human chromosomes solely from Hi-C contact maps. I believe that our theory with sparse data from Hi-C and FISH experiment may be combined to produce 3D structure of chromosomes for any species.

The limitation of our current theoretical framework and many other ensemble-based approach is the inability to decipher the single-cell information. Due to the apparent heterogeneity present in the cell population (cite the other paper), Hi-C map as an ensemble average quantity has limited information regarding the organization of genomes, even though the experiments are remarkable. The Hi-C map and the derived DMs only characterize the average structure. In other words, there may not be a typical single cell genome that can be described by the Hi-C map and the DMs derived from it. Consider our simple mixture model system as an example. Each single trajectory can be described by either GRMC (CTCF mediated loops

present) or a chain devoid of loops. Therefore, averaging over an ensemble of cells may not be meaningful from an *in vivo* perspective. Nevertheless, the theoretical lower bound provides a way forward to obtain 3D organization from contact map alone, perhaps even from single cell data.

Chapter 6: Kinetic Model For Elastic Coupled Motors System

6.1 Introduction

Molecular motors are proteins which consume ATP to perform work in the cell. They play important roles in many biological processes, such as RNA polymerase in translocating along DNA to transcribe gene, Kinesins or dyneins in carrying a vesicle along the microtube, Myosins in generating a muscle contraction. Using single-molecule technique, how does a single motor, Kinesin in particular, move and transport cargo has been extensively studied [205–209] (also see [210] for a recent review). Meanwhile, due to the current limitation of the experiment technique, many details of mechanochemical cycles of motors are unclear. Theoretical models have proven very useful in our understanding of molecule motors (see [211] for an extensive review). Molecular motors are machines on the microscopic level, governed by the competition between thermal fluctuation and energy flux. The general theoretical aspects of such a system were reviewed in [212].

As much as what we know for a single motor, how do multiple motors work as a team is unclear. In fact, motors *in vivo* almost always work as teams, i,e a cargo is shared by multiple motors of the same kind or even different kind [213–215]. Such a system has been studied *in vitro* by attaching multiple motors to a soft

fluidlike vesicle [213] or to an elastic DNA origami [216]. It has been shown that the multi-motor system increases the run length significantly [216–220], which can be understood that it needs all motors to detach in order for the run stops [221]. Contradicting results regarding the velocity of the multi-motor system have been reported. It was reported that the velocity of the multi-motor system is similar to that of single motor [216, 218, 219]. Shubeita and colleagues [222] showed that an increased number of motors actually decrease the velocity *in vivo*. On the other hand, it has been reported that the velocity of cargo increases with the increase of the number of Myosin motor [214, 223, 224]. In addition, experimental studies have shown that multi-motors system exhibit fractional stepping [225] and coordinated stepping [226, 227] and coupling induced detachment [228, 229].

Numerous theoretical models have been proposed for the coupled motor system. Most of these models rely on the assumption of an equal share of load among motors [221, 230, 231] or are mean-field description rather than stochastic kinetic model [232–234].

Here I present a simple kinetic model for studying elastically coupled motor system. In this model, the chemical kinetic scheme for a single motor is a simple one-state model [235], allowing an analytical solution to the model. In addition, no assumption of an equal share of loading is made. This model is relevant to the chromosomes since it was suggested [92] (through private communications) that condensins also exhibit potential cooperative motor behavior. Thus the model presented here provide a basis for further investigation regarding condensin's loop extruding mechanism.

6.2 Model

6.2.1 Overview

In the Elastic Coupled Motor Model (ECMM), motors are mechanically coupled together. Mechanical coupling can be achieved either by sharing a cargo or attaching to a DNA origami scaffold (Fig. 6.1a). The latter has been used to study the multi-motors system *in vitro* [216, 227]. In the model, I assume the coupling between different motors is only mechanical, not affecting the chemical cycles of individual motors but only affecting the associated rates in appearance of external force. In principle, multiple motors can also be coupled by directly forming multimerization state. However, in such case, coupling is likely to be both mechanical and chemical. In order to provide insights to the problem in a way the system is analytical tractable and conceptually simple, I employ the one state model of single motor to study the system of n number of coupled motors. In a one state model, each individual motor, i, is characterized by its forward stepping rate k_i^+, backward stepping rate k_i^- and detachment rate γ_i. The mechanical coupling between motors results in the elastic tension energy E_i of the motor i and leads to external force exerted on each motor by other motors through the coupling. The elastic tension of the system is generated due to the deviation of the system from its relaxed state (Fig. 6.1b). In general, the elastic tension can be any function of the deviation. For simplicity, I assume the quadratic dependence of tension on the deviation Δx_i of the motor i, $E_i = (1/2)\kappa_i \Delta x_i^2$ with coupling strength κ_i. Another relevant choice

would be a Finite Extensible Nonlinear Elastic (FENE) potential which limits the maximum deviation.

It is clear that the deviation from each motor's relaxed position is due to the stepping of motors in the system. To illustrate this, I turn to the simplest case of two identical motors attached to a cargo separated a certain distance at their relaxed state (Fig. 6.1b). When the system is in a relaxed state, both motors experience zero force and the total elastic tension energy of the system is zero. At time t, the leading motor steps forward with the step size d. Due to the reposition of the cargo, it is straightforward to observe that its deviation from the relaxed position, Δx, equals $d/2$. Hence it experiences a resistant force. At the same time, the trailing motor also deviates from its relaxed position with $\Delta x = -d/2$. This leads to a assistant force exerted on the trailing motor. It should be noted that in the above picture, I assume that the cargo relaxes to the equilibrium position much faster than the stepping of the motors.

In addition to stepping, each motor can also detach from the track in a state dependent manner. For simplicity, I first consider the system with no detachment events. I reason that the system with detachment can be viewed as a system without detachment of a effective number of attached motors, $n_{\text{eff}} < n$. Within the framework of the model, I ask the following questions: how do the characteristics of the coupled motor system such as velocity, stall force depends on the characteristics of single motor and their coupling strength? I tackle this question using both analytical calculation and numerical simulations. The simulations are performed using Gillespie algorithm.

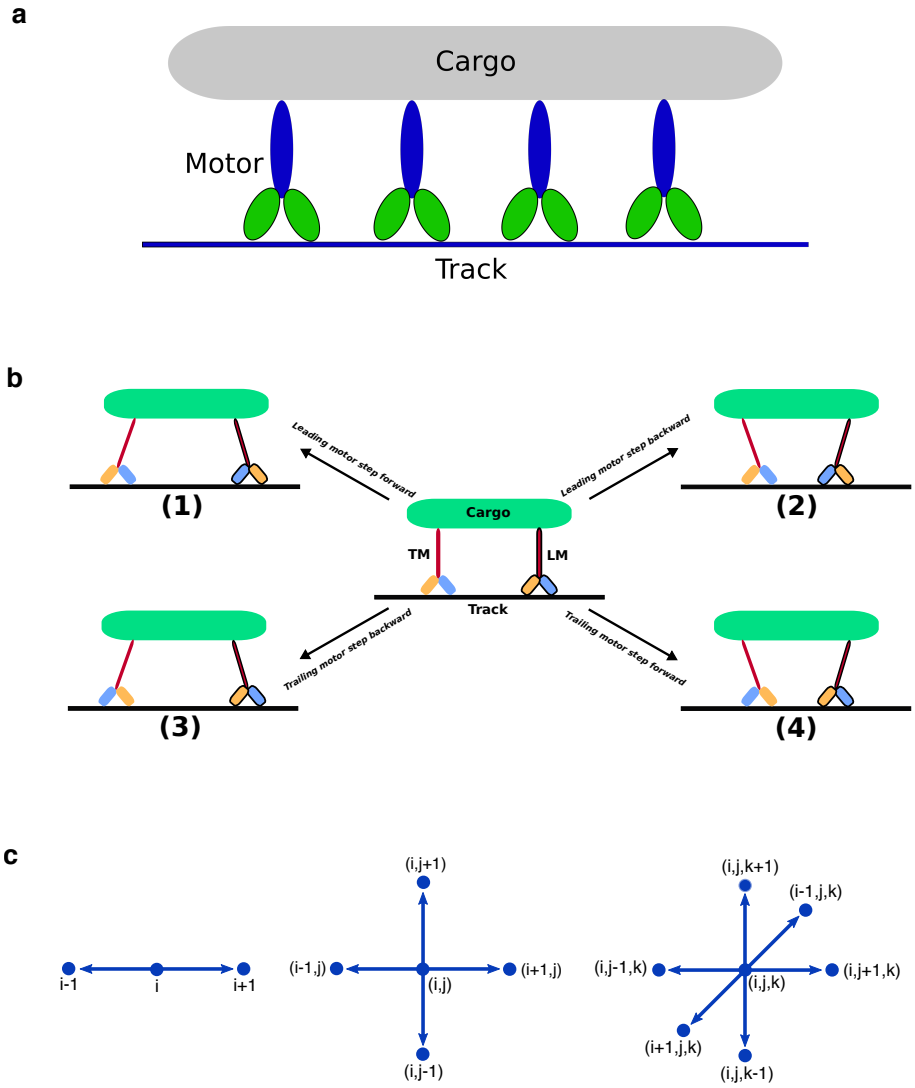


Figure 6.1: **(a)**. The sketch of the coupled motor system studied in this work. The system shown specially resemble the DNA origami experiments. **(b)**. The stepping of coupled motor system with two motors. The relaxed configuration is shown in the center. Neglecting the detachments, there are total number of four possible transitions from the relaxed state. The internal stress is created when the system deviates from its relaxed state. In the framework of the model, I assume internal stress depends on deformation linearly, satisfying Hooke's law. **(c)**. The n coupled motors system can be represented as a hopping process on a hypercubic lattice. From left to right, $n = 1$, $n = 2$ and $n = 3$ systems are shown. i, j, k quantifies the deviation of each motor from its initial relaxed position.

6.2.2 Derivation of mechanical coupling

In the ECMM, the mechanic coupling between motors is assumed to be elastic which originates from either the deformation of motors or it of cargo or a combination of two effects. The total effect is described by Hooke's law with coupling strength κ and deformation Δx. I now derive the equation to compute Δx. At time 0, let's assume that the coupled motor system is in relaxed state. Denote x_i^0 as the initial positions where i^{th} motor attaches to the track in its relaxed state. After some time t, the system is in a new state. Denote x_i as the position where i^{th} motor attaches to the track at time t. Denote x_i' as the relaxed position at time t. x_i' would be the position of i^{th} motor if it is released from the track under the condition that the position of the cargo remain fixed. Obviously, the relaxed position, x_i', are simply a translation transform from the initial relaxed position x_i^0. I write this as,

$$x_i' = x_i^0 + \Delta,\ \Delta \text{ is some constant for any } i. \tag{6.1}$$

It is important to point out that Δ is also the displacement of the mean position of the system.

In the model, I make the assumption that the cargo relaxes to mechanical equilibrium before the system takes the next step. It is equivalent to say that the force experienced by the cargo is always zero (force balanced condition). The force exerted by i^{th} motor on the cargo is $\kappa_i(x_i - x_i')$ where κ_i is the coupling strength for i^{th} motor. In the most general form, κ_i can be different for different motors.

The question now is that, given x_i and x_i^0, what is the solution of x_i' satisfying the force balanced condition $\sum_{i=1}^{n} \kappa_i(x_i - x_i') = 0$? Plug Eq. 6.1 into the force balanced condition equation, one have $\sum_{i=1}^{n} \kappa_i(x_i - (x_i^0 + \Delta)) = 0$. This yields $x_i' = x_i^0 + \Delta$ where $\Delta = \sum_i \kappa_i(x_i - x_i^0)/\sum_i \kappa_i$. The deformation of i^{th} motor, Δx_i is given by,

$$\Delta x_i = x_i - x_i' = x_i - x_i^0 - \Delta \tag{6.2}$$

Note that Eq. 6.2 indicates that once the current positions and relaxed positions of all the motors are known, the deformation of each motor can be computed and the rates associated with the stepping can be then evaluated. This gives the backbone of the simulation.

6.2.3 Coupled motor system can be represented as a hyper-cubic lattice random walk

It is convenient to define a new *normalized* position variable for motors, $\widetilde{x}_i = (x_i - x_i^0)/d_i$ where d_i is the step size of i^{th} motor. Since the motor only takes discrete steps, $\widetilde{x}_i$ can only take integer values. $\widetilde{x}_i$ quantifies how many steps the i^{th} motor is from its initial (equilibrium) position. The system can be fully described by discrete states $\widetilde{\boldsymbol{x}} \equiv (\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_n)$ where $\widetilde{x}_i = 0, \pm 1, \pm 2, \cdots, \pm\infty$ for $i \in (0, 1, \cdots, n)$ and the associated transition rates between states. Now the displacement of the mean position of the system, Δ, can be expressed in terms of the new variable,

$$\Delta = \sum_{i=1}^{n} \alpha_i \widetilde{x}_i \tag{6.3}$$

where $\alpha_i = \kappa_i d_i / \sum_i \kappa_i$.

For $n = 2$, one have $\widetilde{\boldsymbol{x}} \equiv (\widetilde{x}_L, \widetilde{x}_T)$ where subscript L and T represent leading motor and trailing motor, respectively. Suppose at some time t, the system is in state $\widetilde{\boldsymbol{x}} = (i, j)$. It then can takes one of four possible transitions i) leading motor steps forward (**L+**), $(i, j) \to (i+1, j)$; ii) trailing motor steps forward (**T+**), $(i, j) \to (i, j+1)$; iii) leading motor steps backward (**L-**), $(i, j) \to (i-1, j)$ and iv) trailing motor steps backward (**T-**), $(i, j) \to (i, j-1)$. Now I make another assumption regarding these transitions. I assume that the stepping of each motor is poisson point process and independent with each other. Thus the evolution of the system is Markovian that the transition to the next state only depends on the current state but not the path before. The rates characterizing these four transitions are k_{L}^{+}, k_{T}^{+}, k_{L}^{-} and k_{T}^{-}, respectively. For $n = 3$, there are six possible transitions $(i, j, k) \to (i+1, j, k)$, $(i, j, k) \to (i-1, j, k)$, $(i, j, k) \to (i, j+1, k)$, $(i, j, k) \to (i, j-1, k)$, $(i, j, k) \to (i, j, k+1)$, $(i, j, k) \to (i, j, k-1)$. It is useful to consider the evolution of coupled motor system as a general random walk on a lattice in which transition rates are site dependent. For $n = 2$, it is a random walk on a square lattice (Fig. 6.1c) and for $n = 3$, it is a random walk on a cubic lattice (Fig. 6.1c). For $n > 3$, it is a random walk on hyper-cubic lattice. Let's denote $\mathcal{P}(\widetilde{\boldsymbol{x}}, t|\widetilde{\boldsymbol{x}}_0, 0)$ as the probability that the system is at site $\widetilde{\boldsymbol{x}}$ at time t given that it starts from site $\widetilde{\boldsymbol{x}}_0$ at time 0. In general, the master equation for coupled motor system can be written

as,

$$\frac{\partial \mathcal{P}(\widetilde{\boldsymbol{x}}, t|\widetilde{\boldsymbol{x}}_0, 0)}{\partial t} = \sum_{\widetilde{\boldsymbol{y}} \in \Gamma(\widetilde{\boldsymbol{x}})} \Lambda_{\widetilde{\boldsymbol{y}}\widetilde{\boldsymbol{x}}} \mathcal{P}(\widetilde{\boldsymbol{y}}, t|\widetilde{\boldsymbol{x}}_0, 0) - \Lambda_{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{y}}} \mathcal{P}(\widetilde{\boldsymbol{x}}, t|\widetilde{\boldsymbol{x}}_0, 0) \tag{6.4}$$

where $\Gamma(\widetilde{\boldsymbol{x}})$ denotes the set of the nearest neighbors of site $\widetilde{\boldsymbol{x}}$ and $\Lambda_{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{y}}}$ is the transition rate from site $\widetilde{\boldsymbol{x}}$ to site $\widetilde{\boldsymbol{y}}$ which is site dependent in our case. Since each motor in the coupled motor system has two possible transitions (step forward and step backward). Then the number of nearest neighbors of any site is $|\Gamma(\widetilde{\boldsymbol{x}})| = 2n$ where n is the total number of coupled motors.

What would be the site dependent rates associated with each transitions, $\Lambda_{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{y}}}$? The elastic energy of coupling associated with each state (site) is $E(\boldsymbol{\Delta x}) = (1/2)\sum_{i=1}^{n} \kappa_i \Delta x_i^2$ where $\Delta \boldsymbol{x}$ represents a specific configurations of the system $\boldsymbol{\Delta x} = (\Delta x_1, \Delta x_2, \cdots, \Delta x_n)$. Every time a motor steps, $E(\boldsymbol{\Delta x})$ changes correspondingly. Let's denote $\Delta E_i^+(\boldsymbol{\Delta x})$ as the change of E if i^{th} motor steps forward from configurations $\boldsymbol{\Delta x}$ and $\Delta E_i^-(\boldsymbol{\Delta x})$ as the change of E if i^{th} motor steps backward from configurations $\boldsymbol{\Delta x}$. If i^{th} motor steps forward, the position of i^{th} motor becomes $x_i^+ = x_i + d_i$ and the other motors' positions remain the same. Using Eq. 6.2, I obtain the new deformation $\Delta x_i^+ = \Delta x_i + (d_i - \alpha_i)$ where Δx_i is the deformation before the forward step. For other motors, $j \neq i$, one have $\Delta x_j^+ = \Delta x_j - \alpha_i$. Hence one have,

$$\begin{aligned}\Delta E_i^+(\boldsymbol{\Delta x}) &= \frac{1}{2}\sum_i \kappa_i \Delta x_i^{+^2} - \frac{1}{2}\sum_i \kappa_i \Delta x_i^2 \\ &= \kappa_i \Delta x_i d_i + \frac{1}{2}\kappa_i (d_i^2 - d_i \alpha_i) \\ &= \kappa_i d_i (d_i \widetilde{x}_i - \Delta) + \frac{1}{2}\kappa_i (d_i^2 - d_i \alpha_i)\end{aligned} \tag{6.5}$$

The last equality uses the relation $\Delta x_i = d_i \widetilde{x}_i - \Delta$. Similarly, I obtain,

$$\begin{aligned}\Delta E_i^-(\boldsymbol{\Delta x}) &= -\kappa_i \Delta x_i d_i + \frac{1}{2}\kappa_i (d_i^2 - d_i \alpha_i) \\ &= -\kappa_i d_i (d_i \widetilde{x}_i - \Delta) + \frac{1}{2}\kappa_i (d_i^2 - d_i \alpha_i)\end{aligned} \tag{6.6}$$

Given the changes of the coupling energy associated with the transitions, it is not unreasonable to set forward step and backward step rates of i^{th} motor to be ΔE dependent, given by,

$$k_i^+ = k_0^+ e^{-\beta \theta_i^+ \Delta E_i^+} \tag{6.7}$$

$$k_i^- = k_0^- e^{-\beta \theta_i^- \Delta E_i^-} \tag{6.8}$$

where θ_i^+, θ_i^- are the usual distribution factors of i^{th} motor and k_0^+ and k_0^- are the forward and backward rate under zero load. In this work, I employ the Local-Detailed Balance (LDB) principle [212] which constrains $\theta_i^+ + \theta_i^- = 1$. This constraint guarantee the thermodynamic consistency which I will discuss later. Such constraint

is also employed in [236] which the authors argued should hold in the vicinity of the equilibrium. In principle, the distribution factors can also be state dependent, $\theta_i^+ = \theta_i^+(\boldsymbol{\Delta x})$ and $\theta_i^- = \theta_i^-(\boldsymbol{\Delta x})$. For simplicity, I assume in this work θ_i^+ and θ_i^- are independent of the state $\boldsymbol{\Delta x}$.

6.2.4 Coupled motor system in the presence of external force

Now I will consider the system with external force F exerted on the cargo. I again make the assumption that the system is always in mechanical equilibrium. Denote δ as the change of cargo's position due to the presence of the external force F. The motor's deformation is $\Delta x_i - \delta$ where Δx_i is the deformation in the absence of the external force. The force balanced condition states,

$$F + \sum_{i=1}^{n} \kappa_i(\Delta x_i - \delta) = 0. \tag{6.9}$$

This yields $\delta = F/\sum_i \kappa_i$. Note that negative sign of F simply means it is a resisting force. The elastic energy in the presence of the external force $E(F)$ is obtained by,

$$\begin{aligned} E(F) &= \frac{1}{2}\sum_{i=1}^{n} \kappa_i(\Delta x_i - \delta)^2 \\ &= \frac{1}{2}\sum_{i=1}^{n} \kappa_i \Delta x_i^2 + \frac{F^2}{2\sum_i \kappa_i} = E(0) + \frac{F^2}{2\sum_i \kappa_i}. \end{aligned} \tag{6.10}$$

Hence the elastic energy in the presence of the external force is simply the energy in the absence of the external force plus a constant which only depends on F and

κ_i but not the micro-state of the system.

Denote the mechanical work done by the motors as W_{mech}. To obtain W_{mech}, one need to compute the displacement of the cargo due to a single step of a motor. From Eq. 6.3, one obtain the displacement due to forward and backward step of i^{th} motor are simply α_i and $-\alpha_i$, respectively. Therefore, $W^{+}_{\text{mech}} = -F\alpha_i$ if i^{th} motor steps forward and $W^{-}_{\text{mech}} = F\alpha_i$ if i^{th} motor steps backward. Finally, one obtain the rates in the presence of external force F,

$$k_i^{+} = k_0^{+} e^{-\beta\theta_i^{+}(\Delta E_i^{+} + W^{+}_{\text{mech}})} = k_0^{+} e^{-\beta\theta_i^{+}(\Delta E_i^{+} - F\alpha_i)} \tag{6.11}$$

$$k_i^{-} = k_0^{+} e^{-\beta\theta_i^{-}(\Delta E_i^{-} + W^{-}_{\text{mech}})} = k_0^{-} e^{-\beta\theta_i^{-}(\Delta E_i^{-} + F\alpha_i)} \tag{6.12}$$

6.3 Results

6.3.1 Two identical coupled motor system

Let's first focus on the case demonstrated in Fig. 6.1b in which only two *identical* motors are coupled together. For identical motors, I have only one set of parameters $k_0^{\pm}$, d, $\theta^{\pm}$. Under the condition the motors are identical, I have $\Delta = (d/n)\sum_i \widetilde{x}_i$. For $n = 2$, this leads to $\Delta = (d/2)(\widetilde{x}_{\text{L}} + \widetilde{x}_{\text{T}})$ where subscripts represent leading (L) and trailing (T) motor. Suppose at time $t = 0$, the system is in relaxed state. The evolution of the system can be represented by a random walk on square lattice (Fig. 6.1c). The rates associated with each transitions are given by $k^{+}_{\text{L}} = k_0^{+} e^{\beta\kappa\theta^{+}\Delta E^{+}_{\text{L}}}$, $k^{+}_{\text{T}} = k_0^{+} e^{-\beta\kappa\theta^{+}\Delta E^{+}_{\text{T}}}$, $k^{-}_{\text{L}} = k_0^{-} e^{\beta\kappa\theta^{-}\Delta E^{-}_{\text{L}}}$, and $k^{-}_{\text{T}} = k_0^{-} e^{\beta\kappa\theta^{-}\Delta E^{-}_{\text{T}}}$

where the subscripts and superscripts represent i) leading motor steps forward (**L+**) ii) trailing motor steps forward (**L+**) iii) leading motor steps backward (**L-**) iv) trailing motor steps backward (**T-**). $\Delta E_{\rm L}^{+}$, $\Delta E_{\rm T}^{+}$, $\Delta E_{\rm L}^{-}$ and $\Delta E_{\rm T}^{-}$ are the changes of the elastic energy of due to **L+**, **T+**, **L-** and **T-** respectively. Like discussed before, I employ the local-detailed balance principle [212] which constrains $\theta^{+} + \theta^{-} = 1$. Since the motors are identical, it is more convenient to describe the system using variable $i = \widetilde{x}_{\rm T} - \widetilde{x}_{\rm L}$ and $j = \widetilde{x}_{\rm T} + \widetilde{x}_{\rm L}$ where i represents the internal degree of freedom of the system (the separation between two motors) and j represents the displacement the system ($\Delta = (d/2)j$). The master equation of variables i and j is given by,

$$\begin{aligned}\frac{\partial P_t(i,j)}{\partial t} &= -P_t(i,j)\big[k_{\rm L}^{+}(i) + k_{\rm L}^{-}(i) + k_{\rm T}^{+}(i) + k_{\rm T}^{-}(i)\big] \\ &+ P_t(i-1,j+1)k_{\rm L}^{-}(i-1) + P_t(i+1,j-1)k_{\rm L}^{+}(i+1) \\ &+ P_t(i+1,j+1)k_{\rm T}^{-}(i+1) + P_t(i-1,j-1)k_{\rm T}^{+}(i-1)\end{aligned} \tag{6.13}$$

Observe that the transition rates in Eq. 6.13 only depends on i which relates the internal elastic tension of the system by $E = (1/4)\kappa d^2 i$. Sum over j on both sides of Eq. 6.13 leads to the master equation of variable i only.

$$\frac{\partial P_t(i)}{\partial t} = -P_t(i)(\omega_i^{+} + \omega_i^{-}) + P_t(i-1)\omega_{i-1}^{+} + P_t(i+1)\omega_{i+1}^{-} \tag{6.14}$$

where $\omega_i^{+} = k_{\rm L}^{+}(i) + k_{\rm T}^{-}(i)$ and $\omega_i^{-} = k_{\rm L}^{-}(i) + k_{\rm T}^{+}(i)$. The stationary distribution for

i (if exists) π_i^{s} is given by $\pi_i^{\mathrm{s}} = \pi_0^{\mathrm{s}} \frac{\prod_{j=0}^{i-1} \omega_j^+}{\prod_{j=1}^{i} \omega_j^-}$ for $i > 0$ and $\pi_i^{\mathrm{s}} = \pi_{-i}^{\mathrm{s}}$ for $i < 0$ and π_0^{s} is determined by normalization. Eq. 6.14 gives the master equation of internal degree of freedom i. The mean velocity of cargo v_2 (2 denotes I are considering two motors) is related to j by $v_2 = (1/2)d \lim_{t\to\infty} \langle j(t) \rangle / t$ where d is the step size. Sum over i on both sides of Eq. 6.13 yields,

$$\begin{aligned}
\frac{\partial P_t(j)}{\partial t} = &- \sum_i P_t(i,j) \left[k_{\mathrm{L}}^+(i) + k_{\mathrm{L}}^-(i) + k_{\mathrm{T}}^+(i) + k_{\mathrm{T}}^-(i) \right] \\
&+ \sum_i P_t(i-1, j+1) k_{\mathrm{L}}^-(i-1) \\
&+ \sum_i P_t(i+1, j-1) k_{\mathrm{L}}^+(i+1) \\
&+ \sum_i P_t(i+1, j+1) k_{\mathrm{T}}^-(i+1) \\
&+ \sum_i P_t(i-1, j-1) k_{\mathrm{T}}^+(i-1)
\end{aligned} \tag{6.15}$$

To solve Eq. 6.15, I now make the assumption that $P_t(i|j) = \pi_i^{\mathrm{s}} / \sum_{i \text{ is even}} \pi_i^{\mathrm{s}}$ for even values of j and $P_t(i|j) = \pi_i^{\mathrm{s}} / \sum_{i \text{ is odd}} \pi_i^{\mathrm{s}}$ for odd values of j. For a even (odd) j, i can only takes even (odd) values. I argue that this assumption holds true for large t since i relaxes to stationary distribution when $t \to \infty$. Plug $P_t(i,j) = P_t(i|j) P_t(j)$ into Eq. 6.15 and setting j to be even number, I obtain the master equation of variable j,

$$\frac{\partial P_t(j)}{\partial t} = -P_t(j)(\mu_2 + \lambda_1) + P_t(j-1)\lambda_2 + P_t(j+1)\mu_1 \tag{6.16}$$

where $\lambda_1 = (\sum_{i \text{ is even}} \pi_i^{\text{s}}\big(k_{\text{L}}^+(i) + k_{\text{T}}^+(i)\big))/(\sum_{i \text{ is even}} \pi_i^{\text{s}})$, $\lambda_2 = (\sum_{i \text{ is odd}} \pi_i^{\text{s}}\big(k_{\text{L}}^+(i) + k_{\text{T}}^+(i)\big))/(\sum_{i \text{ is odd}} \pi_i^{\text{s}})$, $\mu_1 = (\sum_{i \text{ is odd}} \pi_i^{\text{s}}\big(k_{\text{L}}^-(i) + k_{\text{T}}^-(i)\big))/(\sum_{i \text{ is odd}} \pi_i^{\text{s}})$, and $\mu_2 = (\sum_{i \text{ is even}} \pi_i^{\text{s}}\big(k_{\text{L}}^-(i) + k_{\text{T}}^-(i)\big))/(\sum_{i \text{ is even}} \pi_i^{\text{s}})$

Eq. 6.16 describes a one-dimensional periodic hopping process which was studied in [237]. The mean velocity and diffusion constant of such system is,

$$\begin{aligned} \widetilde{v} &= \lim_{t\to\infty} \frac{\langle j \rangle}{t} = \frac{2(\lambda_1\lambda_2 - \mu_1\mu_2)}{\lambda_1 + \lambda_2 + \mu_1 + \mu_2} \\ \widetilde{D} &= \lim_{t\to\infty} \frac{\langle j^2 \rangle - \langle j \rangle^2}{2t} \\ &= \frac{2(\lambda_1\lambda_1 + \mu_1\mu_2)}{\lambda_1 + \lambda_2 + \mu_1 + \mu_2} - \frac{4(\lambda_1\lambda_2 - \mu_1\mu_2)^2}{(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)^3} \end{aligned} \tag{6.17}$$

Hence the velocity and diffusion constant of the cargo are given by $v = \widetilde{v}(d/2)$ and $D = \widetilde{D}(d/2)^2$. The rates $\lambda_{1,2}$ and $\mu_{1,2}$ depends on coupling strength κ and single motor parameters in an complicated way. To the first order approximation, I only take the first term in the summation when compute $\lambda_{1,2}$ and $\mu_{1,2}$. I obtain the velocity of the cargo (see Appendix D.1 for details),

$$\hat{v}_2 = \frac{v_2}{v_1} = \frac{r+1}{re^{\epsilon\theta^+} + e^{\epsilon\theta^-}} \tag{6.18}$$

where $\epsilon = \beta\kappa d^2/4$ and $r = k_0^+/k_0^-$ and $\theta^- = 1 - \theta^+$. v_0 is the velocity of single motor under zero load $v_1 = (k_0^+ - k_0^-)d$ where d is the step size of a single motor. Eq. 6.18 shows that the normalized velocity $\hat{v}_2$ can be expressed in terms of three dimensionless quantities ϵ, r and θ^+. ϵ quantifies strength of the coupling of a single step. r is the ratio between forward and backward rates and directly related

to the energy released through ATP hydrolysis and θ^+ is the distribution factor characterizing the location of transition state.

Eq. 6.18 shows that velocity of two coupling motors under zero load is reduced compared to the single motor and monotonically decreases with increasing of coupling strength κ. At small κ, the motor hardly affect each other resulting the similar velocity of that of single motor. At large κ, the velocity vanishes. I reason that this is because that the internal tension is built when the system moves. With larger coupling strength κ, the energy associated with internal tension increases which makes the motor harder to step. At the limit of very large κ, although λ_2 increases λ_1 vanishes faster leading to vanishing of $\hat{v}_2$. It has been reported experimentally [223] that coupled motors system under zero load can exhibit larger velocity than the single motor under zero load. I surmise that this is possible when asymmetry behavior of the motor under assistant and resistant force is introduced to the model. In addition, it has been shown that the increase of velocity can also arises from attractive interaction between motors [234]. However such attractive interaction must involves the chemical binding between motors which is out of the scope of this work in which only mechanical coupling is considered. Under the framework of the model, additional interaction other than elastic coupling can be easily included.

According to Eq. 6.18, dimensionless quantities r and θ^+ also affect both the velocity and diffusion constant of the cargo. I compute exact $\hat{v}_2$ and $\hat{D}_2$ using Eq. 6.17 numerically to investigate the effect of these parameters. Fig. 6.2a shows the contourplot of $\hat{v}_2$ as a function of both ϵ and θ^+ for $r = 100$. Fig. 6.2a shows that there exists a θ^+ value at which the velocity of the cargo is maximized at a given

ϵ. This optimal θ^+ value also depends on r. The choice of θ^+ does not change the fact the velocity of two motors system monotonically decreases with increasing of coupling strength ϵ (κ). At large ϵ, the velocity vanishes for all values of r and θ^+.

However the diffusion constant can exhibit non-trivial behavior as a function of r, θ^+ and ϵ. Fig. 6.2b shows the $\hat{D}_2$ for $r = 100$ as a function of ϵ and θ^+. For weak coupling, it is expected the $\hat{D}_2 = 1/2$ since the variance is reduced exactly by half from adding a second motors. However Fig. 6.2b shows that there exists parameters space of (ϵ, θ^+) in which $\hat{D}_2 > 1/2$. The position of such parameter space does depends on value of r as well. Fig. 6.2b shows that the vanishing of the velocity is not because that the system becomes diffusive. The diffusion constant also vanishes at large κ (also refer to Eq. 6.17). This indicates that the system is "frozen" rather than becomes diffusive with large coupling strength.

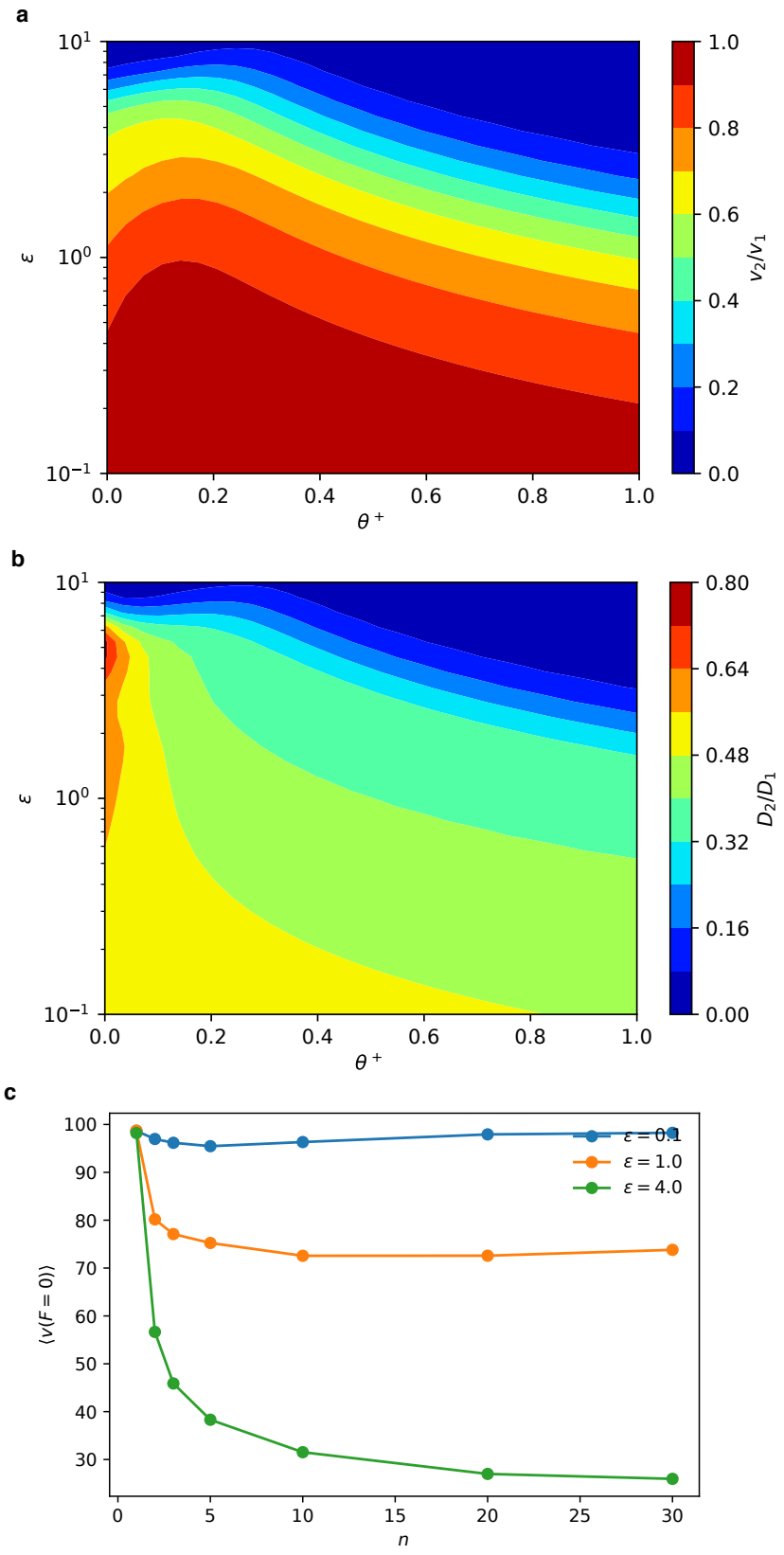

a
10^1
10^0
10^{-1}
ε
0.0
0.2
0.4
0.6
0.8
1.0
θ^+
1.0
0.8
0.6
0.4
0.2
0.0
v_2/v_1
b
10^1
10^0
10^{-1}
ε
0.0
0.2
0.4
0.6
0.8
1.0
θ^+
0.80
0.64
0.48
0.32
0.16
0.00
D_2/D_1
c
100
90
80
70
60
50
40
30
$\langle v(F=0)\rangle$
0
5
10
15
20
25
30
n
$\varepsilon = 0.1$
$\varepsilon = 1.0$
$\varepsilon = 4.0$

Figure 6.2: **(a)**. The normalized velocity of coupled motor system with two identical motors. The colorbar shows the value of v_2/v_1. The choice of parameters are $k_0^+ = 100.0$, $k_0^- = 1.0$. The dimensionless coupling strength ϵ is given by $\beta\kappa d^2/4$. Here I investigate the range $0.1 \leq \epsilon \leq 10$. This is a reasonable biological relevant range. For a realistic model, I have $d \approx 8 \sim 30$nm, $\kappa = 0.05 \sim 1$pN/nm. This leads to ϵ in the range between 0.2 and 50. **(b)**. The normalized diffusion coefficient for coupled motor system with two identical motors. The colorbar shows the value of D_2/D_1. If the motors operate independently, one would expect $D_2/D_1 = 1/2$.

6.3.2 Multi-motor system with $n > 2$

I show that for coupled motor system with two identical motors, the cargo undergoes a periodic one-dimensional hopping process of period two. It turns out that this can be generalized to the system with more than two identical motors ($n > 2$). For $n = 3$, see Appendix D.2 for the generalization. Generally speaking, the cargo of the coupled motor system with n number of identical motors undergoes a periodic one-dimensional hopping process of period of n. Let's denote the forward and backward rates of such a periodic random walk process are $\{\lambda_i\}$ and $\{\mu_i\}$ with $i \in (1, 2, \cdots, n)$, respectively. The solution for mean velocity is given by [237],

$$\widetilde{v}_n = \frac{n}{\sum_{i=1}^{n} r_i}\left[1 - \prod_{i=1}^{n}\left(\frac{\mu_i}{\lambda_i}\right)\right] \tag{6.19}$$

where r_i is given by $r_i = \frac{1}{\lambda_i}\left[1 + \sum_{j=1}^{n-1}\prod_{k=1}^{j}\left(\frac{\mu_{i+k}}{\lambda_{i+k}}\right)\right]$. The solution for diffusion constant $\widetilde{D}_n$ is not shown here, but is given in Derrida [237]. However the values of the rates $\{\lambda_i\}$ and $\{\mu_i\}$ depends on ϵ, r and θ^+ and are complicated to compute. Here I use Monte-Carlo kinetic simulation to study the system with $n > 2$.

The velocity of multi-motor system as a function number of motors n is shown in Fig. 6.2c. In general, the velocity v_n decreases with increasing of n and reaches to some values for $n \to \infty$. The saturating value of v_∞ depends on ϵ, r and θ^+. Fig. 6.2c also shows that with increasing of coupling strength ϵ, the velocity decreases which is consistent with case where $n = 2$. To understand the results in Fig. 6.2c, I turn to Eq. 6.18 which shows that the velocity of two motor system depends on three dimensionless parameters. I reason that r and θ^+ weakly depends on the number of motors n. However ϵ quantifies the increases of elastic tension associated with one step of a motor. For n identical motors, when one motor steps with a step size d. The cargo makes a displacement d/n. It can be obtained that the change of elastic tension energy E is $\epsilon g(n)$ where $g(n) = 2(n-1)/n$ and $\epsilon = (1/4)\beta\kappa d^2$. Then I can simply replace the ϵ in Eq. 6.18 by $\epsilon g(n)$ which yields,

$$\frac{v_n}{v_1} = \frac{r+1}{re^{g(n)\epsilon\theta^+} + e^{g(n)\epsilon\theta^-}} \tag{6.20}$$

Since $g(n)$ has limit 2, I obtain $\lim_{n\to\infty} v_n/v_1 = \frac{r+1}{re^{2\epsilon\theta^+}+e^{2\epsilon\theta^-}}$. Note Eq. 6.20 is just an approximation which only gives the correct qualitative but not quantitative behavior of v_n. Eq. 6.20 can also be used to fit the simulated results with r, ϵ and θ^+ as free parameters whose fitted values represent "effective" values for different values of n.

6.3.3 Stall force of the multi-motors system

The definition of the stall force, $F_s(n)$, here is the value of external force at which the velocity strictly equals zero. According to Eq. 6.19, the velocity becomes

zero only when $\prod_i \lambda_i(F_s) = \prod_i \mu_i(F_s)$. The question is that at what values of F_s, the equality holds. It is convenient to consider the process of translocation of the cargo on a circular track with a large number of sites M. Following the arguments presented in Qian [238] as well as Seifert [212], the mean entropy production rate for a circular Markovian chain with steady state can be written as,

$$\lim_{t\to\infty} \frac{\langle S(t)\rangle}{t} = \sigma = k_B T(j^+ - j^-) \ln\left(\frac{j^+}{j^-}\right) \tag{6.21}$$

where j^+ and j^- are the clockwise (forward translocation) and counter clockwise (backward translocation) probability fluxes in the steady state. $S(t)$ are the total mean entropy production up to time t and σ is the mean entropy production rate. At stall force F_s, the system's velocity vanishes, thus I have $\sigma = 0$. This becomes true only when $j^+/j^- = \prod_i \lambda_i / \prod_i \mu_i = 1$, which gives the same condition by observing Eq. 6.19. By thermodynamic consistency, one also have the equality [212],

$$\prod_i \frac{\lambda_i(F)}{\mu_i(F)} = e^{\Delta S} = e^{\beta(Fd - n\Delta\mu)}. \tag{6.22}$$

where $\Delta\mu$ is the chemical energy associated with one forward step of the motor. Since for n identical motors, the cargo's position changes by d/n when one motor steps. Fd is the work done by the motors along one cycle of n steps, and $n\Delta\mu$ is the energy consumed by motors during these n steps. Note $\Delta\mu$ should not be confused with the energy associated with ATP hydrolysis. $\Delta\mu$ here is an "effective" chemical energy which relates to stall force of a single motor by $F_s^0 = \Delta\mu/d = \ln r/(\beta d)$. From Eq. 6.22, it is easy to observe that $F_s(n) = nF_s^0$ which shows that the stall

force of n coupled motors is simply n times the stall force of a single motor (provided the motors are identical). Note that such equality can be reasoned when one assume the motor share the same load under external force. However such condition can be violated in the model which nevertheless gives the same result.

6.3.4 Force-velocity curve of multi-motors system

Using the framework given in the previous section, I am able to investigate the force-velocity curve of multi-motors sytem. First, I look at the case of two identical motors ($n = 2$). For this case, I can compute the force-velocity curves using Eq. 6.11 and Eq. 6.17. The general behavior of $v_2(F)$ is shown in Fig. 6.3a. $v_2(F)$ intersect with $v_1(F)$ at some critical force F_c (asterisk in Fig. 6.2a), below which $v_2(F) < v_1(F)$ and beyond which $v_2(F) > v_1(F)$. The value of F_c depends on all three dimensionless quantities r, θ^+ and ϵ. It is easy to see that $F_c \leq F_s^0$. Fig. 6.2b shows the dependence of F_c/F_s^0 on coupling strength ϵ and distribution factor θ^+. F_c increases with increasing of ϵ at any fixed θ^+ and approaches single motor's stall force F_s^0 at large ϵ. For a fixed ϵ, F_c can exhibits non-monotonic behavior as function θ^+ depending on the value of ϵ. Maximizing F_c for a fixed ϵ requires θ^+ takes value between 0 and 1. The same qualitative behavior is observed for different values of r. Fig. 6.3a shows that multi-motors system can move the cargo faster at large load compared to single motor but surprisingly move cargo at smaller velocity at small load. If I make the assumption that the friction the cargo experiences is given by Stokes-Einstein relation $F_{\text{friction}} = \gamma v$ where γ is the friction coefficient of

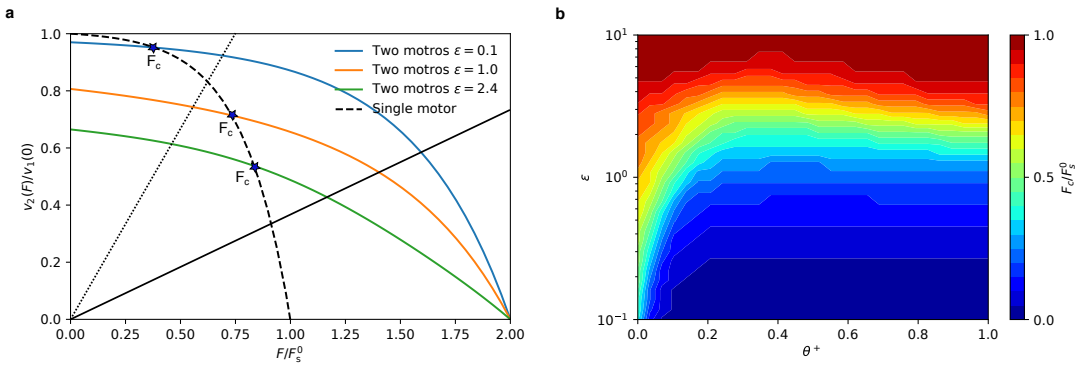


Figure 6.3: **(a)**.The normalized force-velocity curves for coupled motor system with $n = 2$. $v_1(0)$ is the velocity of the single motor in the absence of load. $F_{\rm s}^0$ is the stall force of the single motor. The choice of parameters are $k_0^+ = 100$, $k_0^- = 1$, $\theta^= = 0.0$. The dashed line is force-velocity curve for the single motor. Asterisk marks the interceptions between $v_2(F)$ and $v_1(F)$ which corresponds to the critical force $F_{\rm c}$. When $F > (<)F_{\rm c}$, $v_2(F) > (<)v_1(F)$. The dotted and solid lines are given by $F_{\rm friction} = \gamma v$, representing small and large cargo, respectively. The interception between the force-velocity curves gives the terminal velocity of the cargo transported by the coupled motor system. The velocity of cargo transported by two motors is faster than the velocity of the single motor for all three different values of ϵ. Whereas for small cargo (dotted line), two motor system transport the cargo slower than the single motor for $\epsilon = 1.0, 2.4$. **(b)**. The normalized critical force $F_{\rm c}/F_{\rm s}^0$ as a function of ϵ and θ^+. $F_{\rm c}$ increases with increasing of ϵ and exhibit non-monotonic dependence on θ^+.

the cargo. Assuming that γ is proportional to the size of the cargo, the results in Fig. 6.3a suggests that two-motor system translocate large cargo at higher velocity compared to single motor but at smaller velocity for small cargo compared to single motor.

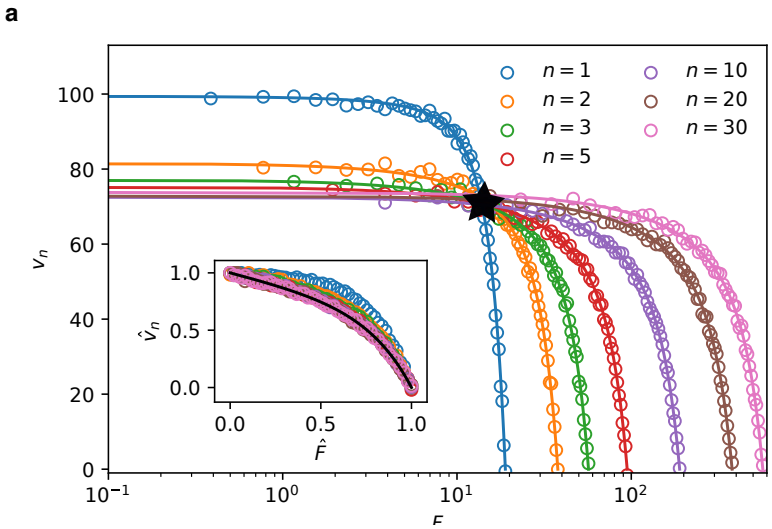
I then turn to the cases with $n > 2$. For a one state model for a single motor, the force-velocity curve can be expressed in dimensionless quantities,

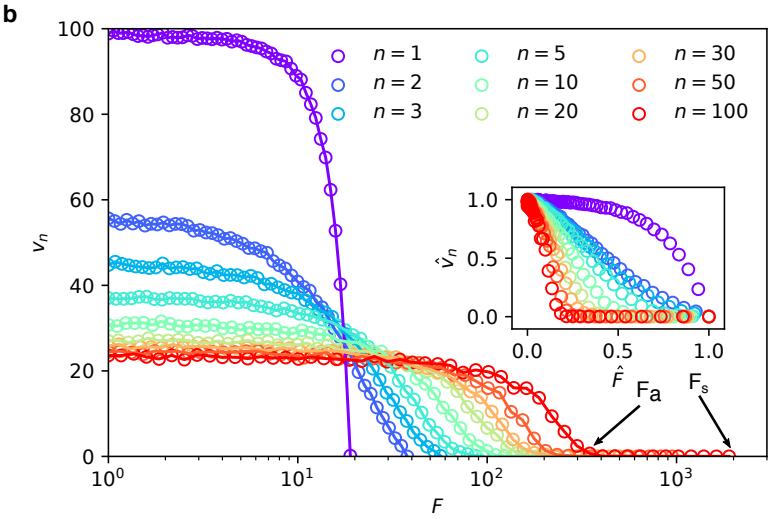
$$\hat{v} = \frac{r^{-\theta^+ \hat{F}}(r - r^{\hat{F}})}{r - 1} \tag{6.23}$$

where $\hat{v} = v(F)/v(0)$ and $\hat{F} = F/F_{\rm s}^0$.

For weak coupling strength, I find that this relation still holds for n coupled motor system by replacing $v(0)$ with $v_n(0,\epsilon)$ and rescaling $F_{\rm s}^0$ by $F_{\rm s} = nF_{\rm s}^0$. Fig. 6.4a show one example where the force-velocity curves for different values of n collapse on a master curve given by Eq. 6.23. However, for strong coupling, $v_n(F)$ do not collapse on a master curve. In fact, the shape of $v_n(F)$ changes for different values of n at large ϵ. Interestingly, I find that $v_n(F)$ with different values of n intercept approximately at the same point (Fig. 6.4a). This means that F_c only weakly depends on n.

In the previous section, I show that the stall force of n coupled motor system is simply $F_{\rm s} = nF_{\rm s}^0$. However, I find that multi-motors system with relatively strong coupling actually exhibits a much lower *apparent* stall force, $F_{\rm a} < F_{\rm s}$. When $F_{\rm a} < F < F_{\rm s}$, velocity vanishes, $v_n(F) \approx 0$. Fig. 6.4b shows the normalized force-velocity curves for different values of n at $\epsilon = 4.0$. It clearly shows that $\hat{v}$ decreases very close to zero at some value $\hat{F} < 1$ for large $n \leq 10$. However for weak coupling, Fig. 6.4a shows that the velocity vanishes exactly at stall force $F_{\rm s}$. I set $F_{\rm a}$ to be the force at which the velocity is less than 1 (1% of the single motor's velocity without load). Fig. 6.4c shows the comparison for $F_{\rm a}$ between strong and weak coupling. For weak coupling $\epsilon = 1.0$, $F_{\rm a} = F_{\rm s} = nF_{\rm s}^0$. Whereas $F_{\rm a}$ first coincidence with stall force for $n \leq 5$ and scales as $n^{0.55}$ for large n resulting a much lower *apparent* stall force.


a
$n=1$
$n=2$
$n=3$
$n=5$
$n=10$
$n=20$
$n=30$
v_n
$\hat{v}_n$
$\hat{F}$
F


b
$n=1$
$n=2$
$n=3$
$n=5$
$n=10$
$n=20$
$n=30$
$n=50$
$n=100$
v_n
$\hat{v}_n$
$\hat{F}$
Fa
F_s
F

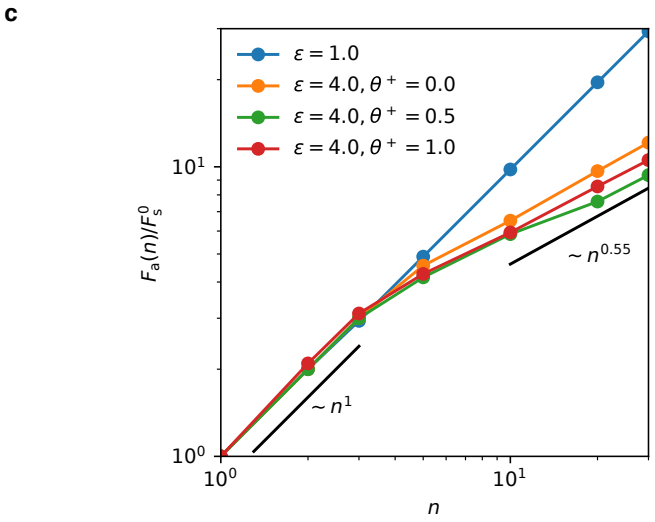

c
$\varepsilon=1.0$
$\varepsilon=4.0, \theta^+=0.0$
$\varepsilon=4.0, \theta^+=0.5$
$\varepsilon=4.0, \theta^+=1.0$
$F_a(n)/F_s^0$
$\sim n^1$
$\sim n^{0.55}$
n

Figure 6.4: **(a)**. Force-velocity curves for different values of n. Choice of parameters are $k_0^+ = 100$, $k_0^- = 1$, $\epsilon = 1.0$ and $\theta^= = 0.0$. The asterisk marks the interception between force-velocity curves. The inset shows that the normalized force-velocity curves collapse on a master curve. The solid line is the fit using the Eq. 6.23 with fitted parameters $r \approx 73$ and $\theta^+ \approx 0.085$. **(b)**. $v_n(F)$ for the coupled motor system with strong coupling, $\epsilon = 4.0$. The values of other parameters are the same as (**a**). The inset shows the normalized velocity $\hat{v}_n$ as the function of the normalized force $\hat{F}$. The curves for different values of n do not collapse. The *apparent* stall force F_a and stall force F_s are marked by the arrows. The velocity v_n vanished for $F_\mathrm{a} \leq F \leq F_\mathrm{s}$. **(c)**. The apparent stall force F_a as a function of number of motors n. For weak coupling, $F_\mathrm{a} = F_\mathrm{s} = nF_\mathrm{s}^0$ and $F_\mathrm{a} \ll F_\mathrm{s}$ for strong coupling.

6.3.5 Step Coordination of multi-motors system

It has been suggested experimentally that multiple motors working together coordinate their stepping in such a way that the stepping of multiple motors are synchronized [214, 227, 229]. This leads to experimental observation of step size of multiple motors to be the same as the step size of a single motor. From the framework of the model, the step size of the cargo transported by n identical motors is d/n where d is the step size of a single motor. This gives a fractional step size of multi-motor system which is also observed experimentally [225, 228]. I then ask the question that can I observe step coordination in the framework of the model? Fig. 6.5a shows three trajectories of $n = 30$ system for a set of parameter choice. The step size of a single motor is set to be 1. The apparent stepping of size 1 in Fig. 6.5a suggests that the motors indeed synchronize their stepping. Such coordination of stepping among motors is only observed for strong coupling strength but not weak coupling (Fig. 6.5b). As a consequence, the velocity of system is reduced by a large amount when the synchronization of stepping is achieved.

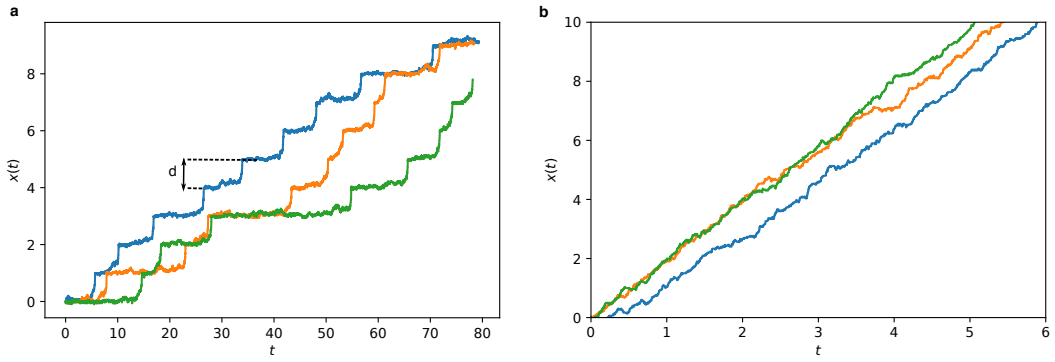


Figure 6.5: **(a)**. Three typical trajectories which show the coordination/synchronization of stepping of motors. The parameters are $n = 30$, $k_0^+ = 3.0$, $k_0^- = 1.0$, $d = 1.0$, $\theta^+ = 1.0$ and $\epsilon = 2.0$. **(b)**. Three typical trajectories which show no coordination/synchronization. The system is in the weak coupling regime with $\epsilon = 0.1$. Other paraemters have the same value as (**a**).

Consider the coupled motor system with $n = 2$ is at relaxed state at time 0. At some time t, either the leading or trailing motor steps forward leading to an elastic internal stress in the system. When such stress is strong enough, the rate of forward stepping of the second motor dominates all the other transitions ($\lambda_2 \gg \mu_1$). Hence whenever a motor steps forward, a second stepping of the other motor immediately follows giving rise to the synchronization. On the other hand, the strong coupling will reduce the rates of stepping from the relaxed state (λ_1 is reduced). This decrease of rates of stepping from the relaxed state counteract the synchronization, leading to an overall reduced velocity.

6.4 Discussion and Conclusion

I develop a kinetic model for elastic coupled multi-motor system to investigate its velocity as well as force-velocity relation. In this work, I neglect the detachment of motors from track. I reason that a coupled motor system with detachment can be effectively viewed as a system (no detachment) with an effective number of attached motor $n_{\text{eff}} < n$. Thus the velocity and velocity-force relations studied in the present study are still relevant for the system with detachment. Furthermore, the framework of the model is general and easy to extend to include the detachment which will be studied in the following works.

I show that the coupled motor system can be represented as a hopping process on a hypercubic lattice whose dimension is simply the number of coupled motors. In the model presented here, I assume the Markovian nature of such hopping process. Analytical solution of the case with $n = 2$ is provided. In principle the cases with $n > 2$ can also be analytically solved, but I do not find solutions of simple forms. Approximated solutions are given in this work to provide insights and physical arguments. For $n > 2$, I use Monte-Carlo simulation (Gillespie's algorithm) to simulate the trajectories of the system.

Furthermore, the results show that velocity of coupled motors largely depends on three dimensionless quantities, the ratio between forward and backward step rates of a single motor r, the distribution factor under force θ^+ and elastic coupling strength ϵ. I find that the velocity of coupled motors system is always smaller than the single motor velocity under zero load. The v_n decreases monotonically with

increasing of coupling strength ϵ and eventually the system becomes frozen at high ϵ. Such dependence of velocity on stiffness is reported experimentally for Myosins transporting actin filaments [226]. I do not find an increase of velocity compared to single motor velocity in the model. Future studies involving the force-dependent distribution factors and asymmetrical response of detachment under assistant and resistant force will be pursued. Consistent with experiment [226], I also find the velocity depends on n when n is small but only weakly depends on n when n is large, converging to a limit values v_∞ at $n \to \infty$ for a given set of r, θ^+ and ϵ.

I then show that the stall force of n coupled motor system is simply n times the stall force of a single motor, $F_\mathrm{s} = nF_\mathrm{s}^0$, regardless of choice of parameters. However behavior of the force-velocity curves highly depends on r, θ^+ and ϵ. Although the velocity of coupled motors system at zero load is reduced compared to single motor velocity, I find that the coupled motor system have higher velocity at high load. I show that there is a critical force F_c. When $F < F_\mathrm{c}$, $v_n(F) < v_1(F)$ and $v_n(F) > v_1(F)$ for $F > F_\mathrm{c}$. This result suggests that the multi-motor system is more efficient at transporting large cargo whereas single motor is more efficient at transporting smaller cargo. In addition, I find that in the strong coupling regime, the velocity of coupled motors system vanishes at force smaller than its stall force, leading to a smaller *apparent* stall force. $F_\mathrm{a} \approx F_\mathrm{s} = nF_\mathrm{s}^0$ for small values of ϵ. When ϵ is large, I find $F_\mathrm{a} \sim n^{0.55}$ leading to a much weaker dependence on n.

Since the model is based on stochastic stepping of motors. I am able to investigate the translocating of coupled motor system on the level of single steps. At strong coupling strength ϵ, the coordination/synchronization of steps among

motors are observed. However, such coordination does not lead to an enhancement of velocity. The decrease of rates of stepping from relaxed state due to elastic coupling counteract the coordination and results in a overall reduced velocity.

The stiffness of the coupled motor system depends on both the stiffness of motor and the stiffness of the cargo. the model set a single coupling strength ϵ. By first order approximation, the ϵ in the model corresponds to the stiffness of the softer one between motor and cargo. For very stiff cargo, the system is determined by the stiffness of the motor itself. On the other hand, if the stiffness of the motor is higher than the cargo, I predict that experimentally tuning the stiffness of the cargo will affect the behavior of the system. Interestingly, one study shows that the velocity of coupled motors is enhanced compared to unloaded single motor when the cargo is soft [223]. The model proposed here predicts that the velocity of multi-motors system can be higher than the velocity of a single motor transporting the identical cargo when the coupling is weak and cargo is large. However, the velocity is always lower than the unloaded single motor. An interesting model is suggested in [223] that the enhancement of velocity is due to the detachment of trailing motor and recentering of the cargo due to mechanical equilibrium. Such model can be easily included in the model by introducing asymmetric force dependent detachment rate of motors [239]. The results of such model will be studied in the following works.

Chapter 7: Conclusions and Future Perspectives

In this thesis, I first provided an introduction to several key aspects of genome organization discovered by both imaging and Hi-C experiments. The strengths and usages of different experimental techniques are discussed. The dynamics of the chromosomes are examined based on current literature as well as from theoretical arguments using polymer physics. In addition, an overview of computational studies on genome organization is presented.

In Chapter 2 and Chapter 3, I presented a copolymer model for human interphase chromosomes (CCM). I demonstrate that CCM is consistent with Hi-C data in terms of genome organization. Using the CCM, I am able to investigate various dynamical properties of human interphase chromosomes which are found to exhibit glassy-like dynamics.

The *minimal* model CCM allows further improvements and integration with additional components. No active force is considered in CCM. The inclusion of active force can be integrated with the CCM and it would be interesting to investigate how does the active force affect the structure and dynamics of the chromosomes. As I have introduced in Chapter 1, it has been suggested that the active force may be responsible for the observed super-diffusive chromatin loci [38] and the coherent

motions [131]. The force generated through the translocation of RNA polymerase may be modeled as forces parallel to the backbone of the chromosome. It would be worth investigating the difference between the passive CCM and active CCM by changing the directions and magnitudes of the active forces.

Since the CCM suggests that human interphase chromosomes have glassy-like dynamics, it is hence natural to ask that do chromosomes resemble soft glassy material? The response of chromosome segment [240] and nucleus [241] to mechanical stress has been studied by aspiration experiments. The results suggested that the nucleus has an elastic response at timescale below 10 seconds and behaves like a fluid at longer times. However, direct measurements of the mechanical response of individual chromosome have not yet been conducted experimentally. Thus, it would be interesting to investigate the mechanical properties of interphase chromosomes using CCM. The micro-dynamics - mean square displacements of chromatin loci can be related to macro-rheology - stress relaxation modulus through a Laplace transforms [242]. In addition, the force-extension curve can be computed using simulations in which an external force is applied to the chromosomes generated by CCM.

Although I want to keep CCM minimal by using only one controlled parameter ϵ, it is certainly important to investigate the dependence of the model on other effects. It has been suggested [58] confinement can induce glassy dynamics in the genome organization. In CCM, the glassy-like dynamics originate from the loci-loci interactions. Hence, it would be interesting to study the interplay between these two effects. Furthermore, in this thesis, a single chromosome is considered

in the simulation. Two or more chromosomes can be considered to study inter-chromosome interactions. In particular, it would be interesting to see whether two individual chromosomes segregate to occupy their own territories in the simulations using CCM.

In Chapter 4, I developed a theoretical framework to extract the distribution of subpopulations of cells using FISH data. As proof of concept, I demonstrate this method using two datasets. I demonstrate that heterogeneity of genome organization is extensive in human fibroblast cells. In the future, it is important to extend this method to different species other than human to investigate the extent of heterogeneity across the spectrum of species. In Chapter 5, a non-simulation method was proposed to reconstruct three-dimensional chromosome organization using Hi-C data. However, the reconstructed structure is an average one in which heterogeneity is neglected. It is possible to use single-cell Hi-C data as constraints to generate an ensemble of single-cell genome organization using Generalized Rouse model.

A more general question regarding the heterogeneity of genome organization is a biological one - does it matter? To be more specific, does heterogeneity play a direct role in gene regulation or it has little or none function significance? Of course, to answer this question requires further understanding of the cause of heterogeneity and data of correlations between structural measurements of chromosome and the gene expression profile on the level of single-cell.

In Chapter 6, a simple kinetic model is developed to study the force-velocity curve and stall force of coupled motors system. I found that the stall force of the multi-motor system is simply the summation of single motor stall forces. I also

found that two identical motors are more efficient at transporting large cargo but less efficient at transporting small cargo compared to a single motor. A crucial simplification - detachment of motors from the track is neglected - is made to make the model analytical tractable. I argued that the coupled motors system with detachment and reattachment can be viewed as a system with an effective number of always attaching motors. However, it would be important to directly incorporate the detachment and reattachment into the model to investigate what effect it has on the properties of the force-velocity curve and stall force. It has been suggested that the reattachment of motors play a dominant role in multi-motors cargo transport [243]. In particular, the asymmetrical dependence of detachment on the force exerted on the motor could be investigated, which has been observed in experiment [239].

Appendix A: Supplementary Information for Chapter 3

A.1 Spearman correlation map

In order to quantitatively assess the closeness of the simulated and experimental contact maps, we first calculated the Spearman correlation maps. When computing the Spearman correlation map, the contact map obtained from our simulations or the Hi-C data is first transformed to a log scale. For each entry, c_{ij}, in the transformed log scale contact map, we calculated the Z-Score using $z_{ij} = (c_{ij} - \langle c_s \rangle)/\sigma_s$ where $\langle c_s \rangle = (1/(N-s)) \sum_{i<j} \delta(s-(j-i)) c_{ij}$, and σ_s is the standard deviation of c_s. The Spearman correlation coefficient, ρ_{ij}, is calculated between the i^{th} row, $\mathbf{X}_i$, and the j^{th} column, $\mathbf{Y}_j$, of the matrix $\mathbf{Z}$ whose elements are z_{ij}. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. First, the raw vector $\mathbf{X}_i$ and $\mathbf{Y}_j$ are converted to rank variable $\mathbf{R}_{X_i}$ and $\mathbf{R}_{Y_j}$ by assigning a rank of 1 to the lowest value in the $\mathbf{R}_{X_i}$ and $\mathbf{R}_{Y_j}$ vectors, and 2 to the next lowest and so on. The Spearman correlation coefficient is the Pearson correlation coefficient between two rank variable vectors $\mathbf{R}_{X_i}$ and $\mathbf{R}_{Y_j}$, computed using $\rho_{ij} = \mathrm{cov}(\mathbf{R}_{X_i}, \mathbf{R}_{Y_j})/(\sigma_{\mathbf{R}_{X_i}} \sigma_{\mathbf{R}_{Y_j}})$ where $\mathrm{cov}(\mathbf{R}_{X_i}, \mathbf{R}_{Y_j})$ is the covariance between $\mathbf{R}_{X_i}$ and $\mathbf{R}_{Y_j}$ and $\sigma_{\mathbf{R}_{X_i}}$,$\sigma_{\mathbf{R}_{Y_j}}$ are the standard deviation of $\mathbf{R}_{X_i}$ and $\mathbf{R}_{Y_j}$. The elements in the Spearman correlation matrix, ρ_{ij}, are the Spearman

correlation coefficients between $\mathbf{X_i}$ and $\mathbf{Y_j}$.

A.2 Comparison of the Correlation Maps

We use quantitative measures to assess if the simulated and experimentally inferred contact and correlation maps are similar. In particular, we compare as precisely as possible, the compartment patterns suggested in the Hi-C map and the simulation results from the CCM. For a fixed genomic distance, the contact probability between two loci of the same type is greater than between two loci of different types. The task is to partition the loci based on the contact map such that each partition corresponds to one distinct loci type while being consistent with the observed checkerboard pattern. This relationship allows us to extract additional information about TAD organization than is possible from experiment alone [14]. It is carried out in two steps. (1) Since the contact probability is a function of both s and their epigenetic states, it is necessary to minimize the effect of the genomic separation in order to highlight the enrichment of contacts between loci of the same epigenetic state in the compartments. The correlation between the same type of loci is more transparent if the Spearman correlation matrix is used because it is based on the rank order (see above). (2) We treat the Spearman correlation matrix, $\mathbf{A}$, as an Adjacency matrix, where the vertices are the loci and the edge weight between the i^{th} and j^{th} loci is the Spearman correlation coefficient, ρ_{ij}.

With these two steps, our clustering problem can be solved by finding the

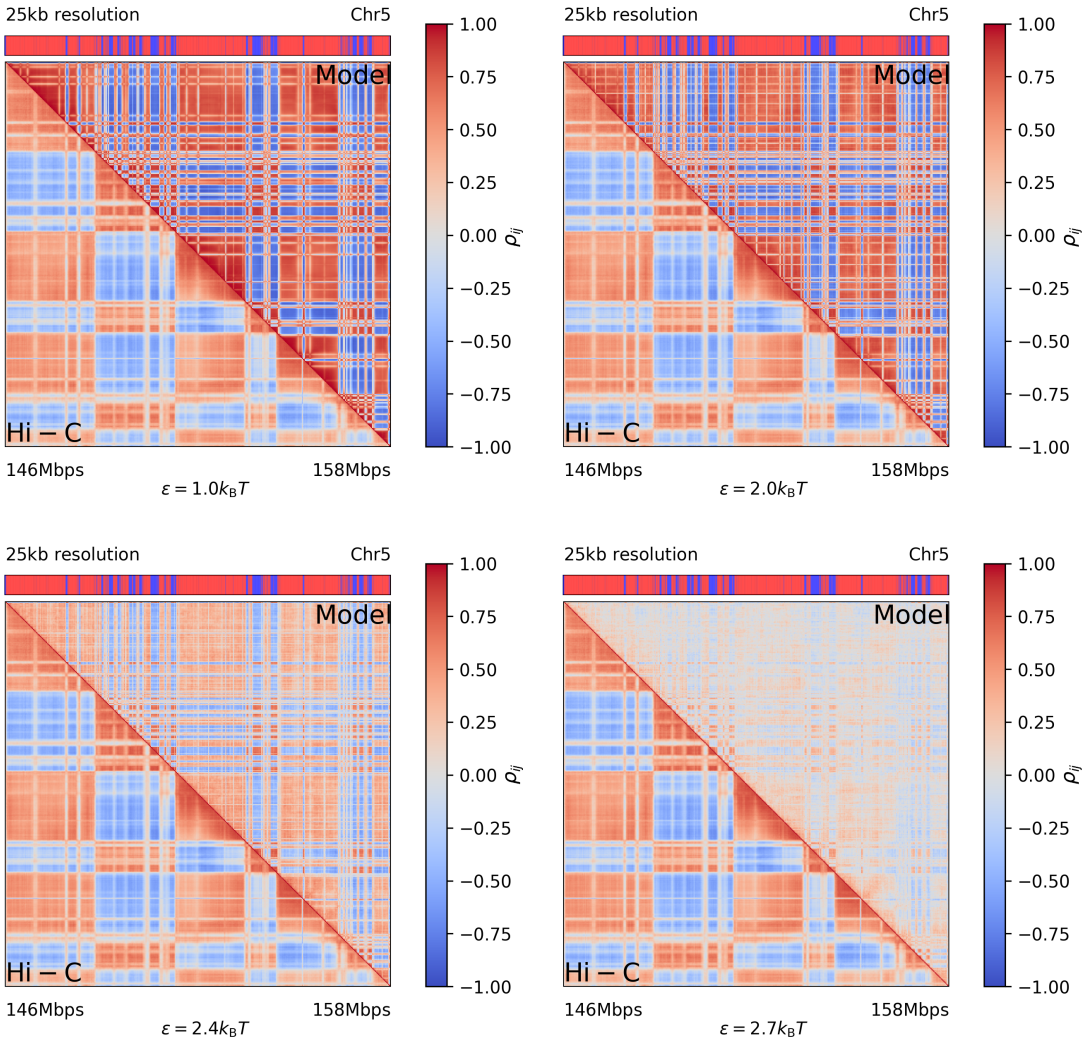


Figure A.1: Spearman correlation map computed for $\epsilon = (1.0, 2.0, 2.4, 2.7)\mathrm{k_B}T$. For each figure, left lower triangle is the Spearman correlation map computed from the Hi-C data, and the upper right triangle is the simulated map. The color bar shows the value of the Spearman correlation coefficient with the value of 1 (-1) indicating perfect correlation anti-correlation; 0 implies no correlation. When the copolymer goes from displaying liquid-like behavior ($\epsilon < 2.4\mathrm{k_B}T$) to exhibiting glassy dynamics ($\epsilon > 2.4\mathrm{k_B}T$), the distinction between anti-correlation (blue) and correlation (red) becomes less transparent. Note that the agreement between simulation and experiment is best for $\epsilon = 2.4\mathrm{k_B}T$.

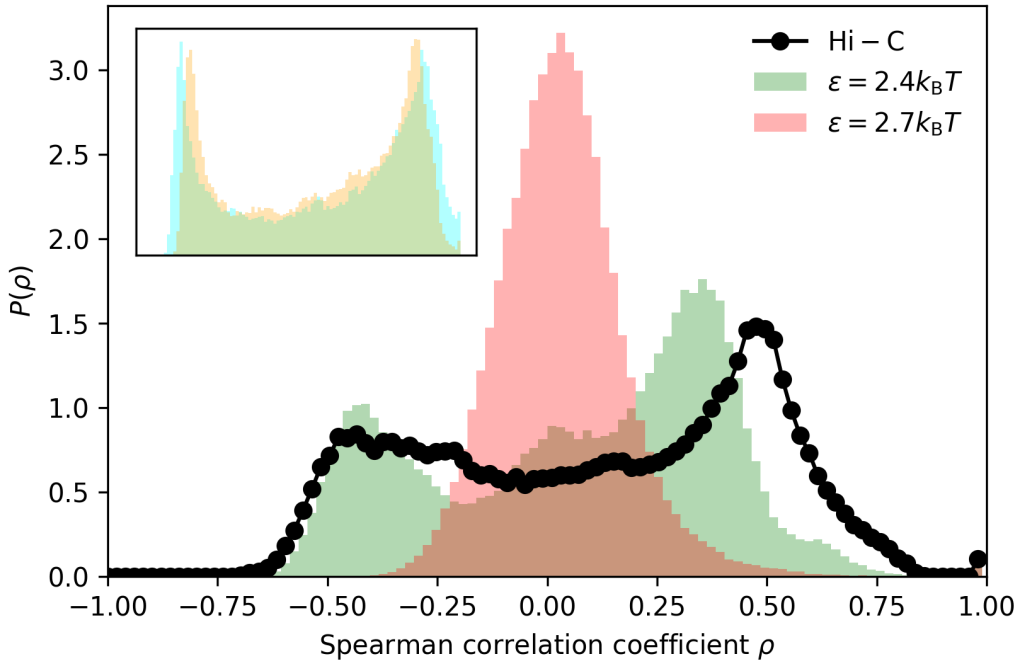


Figure A.2: Comparison of the histograms of the Spearman correlation coefficient, ρ, from simulations and experiment. We plot the distribution of ρ_{ij} for every pair of (i, j). The black line is from the Hi-C experiment. The bimodal shape of the distribution is a result of two different compartment patterns in the Hi-C map. The inset shows the distribution for $\epsilon = 1.0k_BT$ (cyan) and $\epsilon = 2.0k_BT$ (orange). As the dynamics becomes increasingly glassy, the extent of bimodality becomes weaker and exhibits only one peak for $\epsilon = 2.7k_BT$. The closest agreement between simulations and the experiment data occurs when $\epsilon = 2.4k_BT$, thus justifying this value in simulating Chr 5 and 10.

minimum cut vertex in a bipartite graph between the loci. This problem was solved by Dhillon using a spectral co-clustering algorithm [166] in a different context (clustering of documents and words). It is noteworthy that the underlying assumption of this method for our problem is that a pair of loci with positive Spearman correlation coefficient should be the same type. Similarly, a pair of loci with negative Spearman correlation coefficient should be distinct.

The Dhillon spectral biclustering algorithm is implemented as follows [166]:

1. Given the Spearman correlation map, $\mathbf{A}$, construct

$$\mathbf{A}_n = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{F1}$$

2. Compute the left and right second singular vectors of $\mathbf{A}_n$, $\mathbf{u}_2$ and $\mathbf{v}_2$ and form the vector z_2 using,

$$\mathbf{z}_2 = \begin{bmatrix} \mathbf{D}^{-1/2}\mathbf{u}_2 \\ \mathbf{D}^{-1/2}\mathbf{v}_2 \end{bmatrix}. \tag{F2}$$

3. Perform the k-means algorithm on the 1-dimensional data z_2 to obtain the needed clustering.

The matrix $\mathbf{D}$ where $D_{ii} = \sum_j A_{ij}$ and $D_{ij} = 0$ for $i \neq j$ is the degree matrix of the graph. Note that by definition A and A_n are symmetric matrices. The left and right second singular vectors $\mathbf{u}_2$ and $\mathbf{v}_2$ would be the same. Thus, the simpler algorithm is to construct $\mathbf{z}_2 = \mathbf{D}^{-1/2}\mathbf{u}_2$ in step 2, and run step 3. The reason for using k-means clustering is that the values in $\mathbf{u}_2$ and $\mathbf{v}_2$ should have a bi-modal distribution, which is an approximation of the optimal two-valued

partition vector [166]. The use of k-means algorithm allows us to find the two clusters corresponding to the bi-modal distribution.

Using the Dhillon's method, the Spearman correlation map $\mathbf{A}$ is bi-clustered into two clusters, with labeling vector $\mathbf{L}$, where $L_i = 1$ if the i^{th} loci belongs to one cluster and $L_i = 0$ if the i^{th} loci belongs to the other cluster. Note that swapping 0 and 1 in the labeling does not change the meaning.

The second step is to compare the cluster labeling between experiment and the prediction of the CCM. We denote the label assignment of the experimental data as $\mathbf{L}_{\text{exp}}$ and that extracted from the simulations as $\mathbf{L}_{\text{sim}}$. To measure the similarity between $\mathbf{L}_{\text{exp}}$ and $\mathbf{L}_{\text{sim}}$, we use the Adjusted Mutual Information score (AMI) measure. The Mutual Information score (MI) is,

$$\text{MI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}}) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(i,j)\log\Big(\frac{P(i,j)}{P(i)P'(j)}\Big). \tag{F3}$$

where $P(i) = |L^i_{\text{exp}}|/N$ is the probability that a loci (monomer) picked at random from $\mathbf{L}_{\text{exp}}$ falls into type i, L^i_{exp} is the set of loci (monomers) of type i, and N is the total number of loci (monomers). Similarly, $P'(j) = |L^j_{\text{sim}}|/N$. In the above equation, $P(i,j) = |L^i_{\text{exp}} \cap L^j_{\text{sim}}|/N$ is the probability that a locus picked at random belongs to both set L^i_{exp} and L^j_{sim}. Since the expected value of mutual information is non-zero, it is preferable to define the normalized AMI,

$$\text{AMI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}}) = \frac{\text{MI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}}) - E[\text{MI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}})]}{\max\{H(\mathbf{L}_{\text{exp}}), H(\mathbf{L}_{\text{sim}})\} - E[\text{MI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}})]} \tag{F4}$$

where $H(\mathbf{L}_{\text{exp}}) = -\sum_{i=1}^{2} P(i)\log\big(P(i)\big)$ and $H(\mathbf{L}_{\text{sim}}) = -\sum_{j=1}^{2} P'(j)\log\big(P'(j)\big)$. In the above equation, $E[\text{MI}(\mathbf{L}_{\text{exp}}, \mathbf{L}_{\text{sim}})]$ is the expected value of the mutual information, which can be calculated using the following equation [244],

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{Nn_{ij}}{a_i b_j}\right)\frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!} \tag{F5}$$

where $a_i = |L^i_{\text{exp}}|$ and $b_j = |L^j_{\text{sim}}|$, and $(a_i + b_j - N)^+$ denotes $\max(1, a_i + b_j - N)$.

Fig. A.3 compares the AMI between experimental data and results from simulations. The AMI scores for the CCM model are significantly higher than those for the homopolymer model. Thus the long-range compartment pattern can only be obtained using the minimal CCM or other copolymer model.

A.3 Ward Linkage Matrix

The method described in the previous section allows for a quantitative comparison of simulated and measured contact maps. However, it cannot be used as a measure to compare 3D structures (spatial patterns obtained in super-resolution experiments, for example) of chromosomes. In order to achieve this goal, we first relate the information contained in the contact maps to spatial distances. As shown in Fig. 3.4b, the contact probability is inversely proportional to a power of the spatial distance, $P(s) \propto R(s)^{-4.1}$ which provides a way to convert a Hi-C contact matrix to the spatial distance matrix, differing from the physical spatial distance matrix by

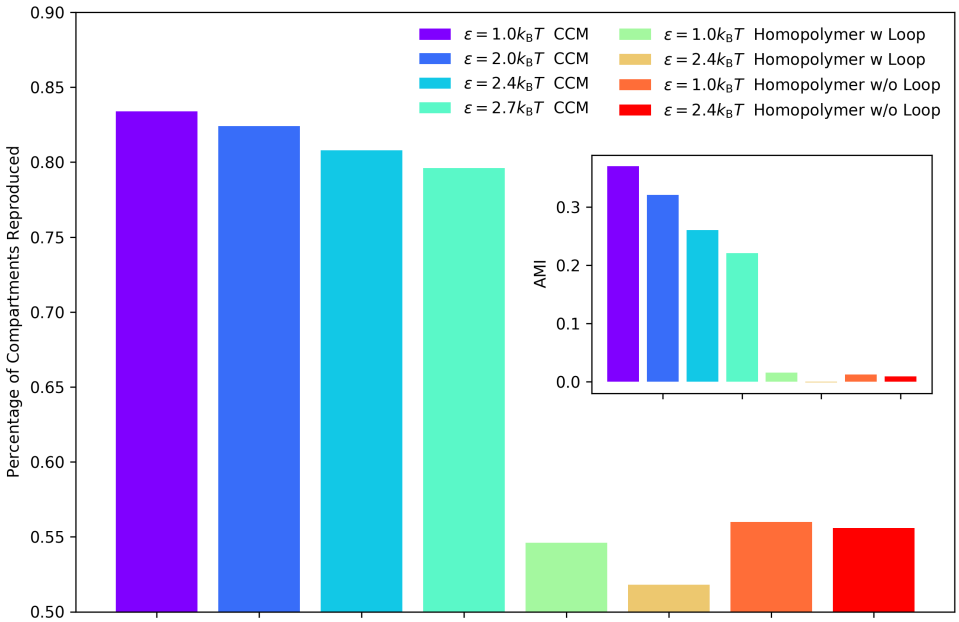


Figure A.3: Percentage of correctly predicted compartments based Adjusted Mutual Information score (AMI) for the CCM and homopolymer models. CCM correctly reproduces $\approx 83\%, 82\%, 81\%, 80\%$ of the compartments found in experiments for $\epsilon = (1.0, 2.0, 2.4, 2.7)k_BT$, respectively. The values for the homopolymer are very low, which implies that it cannot capture the structures of the chromosomes. The inset shows the AMI score for different cases. Note that AMI score is more sensitive compared to the percentage of compartments reproduced.

only a constant prefactor. We then compare the "pseudo" spatial distance matrix with our simulated spatial distance matrix. Needless to say that in simulations R_{ij} can be directly computed.

Matrix norm is often used to measure the distance between two matrices. However, it has severe drawbacks in the context of chromosome organization for two reasons. First, the element-wise differences cannot capture the similarities of higher order structure embedded in the matrix. Second, it suffers from "curse of dimensionality" [245], i.e. there is little difference in the distances between different pairs of matrices, which makes it impossible to differentiate between the experimentally inferred spatial distance matrix and the matrices obtained in the simulations with different parameters. To overcome these difficulties, we adopted the method described recently [246], which suggests treating the original matrix as a graph where the matrix element is a measure of the distance (which is naturally satisfied in our context), and transform it to a cophenetic matrix. In the process, the topological structure of the information embedded in the matrix is preserved. By adopting this method, we can compare the simulated structures of the folded chromosomes with that inferred from Hi-C data.

We converted the Hi-C contact matrix to a "pseudo" spatial distance matrix $\mathbf{R}_{\text{exp}}$, using the relation $R_{ij} = P_{ij}^{-1/4.1}$ ($|i-j| \propto s$). We constructed the Ward Linkage Matrix (WLM), $\mathbf{W}$, from $\mathbf{R}_{\text{exp}}$ and the simulated spatial distance matrices $\mathbf{R}_{\text{sim}}$. The algorithm to construct a WLM is the following. First, start with each locus in a cluster of its own. Second, find the pair of clusters with the smallest Ward distance (see below) and merge them. Third, repeat the second step until

there is only one cluster. Finally, the WLM is constructed as follows. Suppose i and j belong to two disjoint clusters S and T and are joined by a *direct* parent cluster. The entry of the WLM, w_{ij}, is the Ward distance between clusters S and T, given by,

$$d(S,T) = \left(\sum_{i \in S \cup T} ||\mathbf{x}_i - \mathbf{c}_{S\cup T}||^2 - \sum_{i \in S} ||\mathbf{x}_i - \mathbf{c}_S||^2 - \sum_{i \in T} ||\mathbf{x}_i - \mathbf{c}_T||^2 \right)^{1/2} = \left(\frac{n_S n_T}{n_S + n_T} ||\mathbf{c}_S - \mathbf{c}_T||^2 \right)^{1/2} \tag{G1}$$

where $\mathbf{c}_S$ and $\mathbf{c}_T$ are the centers of S and T, respectively; n_S (n_T) is the number of monomers in S (T), and $\mathbf{x}_i$ is the position of point i. The initial clustering occurs between singleton clusters (cluster on its own). The distance between two singleton clusters, i and j, is,

$$d(i,j) = ||\mathbf{x}_i - \mathbf{x}_j|| = (\mathbf{x}_i - \mathbf{x}_j)^{1/2} \tag{G2}$$

which in our case is simply R_{ij}, the spatial distance between the i^{th} and j^{th} loci.

In practice, we used the Lance-Williams recursive algorithms [247] to compute the Ward distance. Suppose we have three clusters C_i, C_j and C_k, and the Ward distances between them, $d(C_i, C_k)$, $d(C_j, C_k)$ and $d(C_i, C_j)$, are known. The Ward distance between the union of clusters i and j, $C_i \cup C_j$ and C_k, is obtained using the recursive equation,

$$d(C_i \cup C_j, C_k) = \left(\frac{n_i + n_k}{n_i + n_j + n_k} d^2(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d^2(C_j, C_k) - \frac{n_k + n_k}{n_i + n_j + n_k} d^2(C_i, C_j) \right)^{1/2} \tag{G3}$$

where C_i, C_j and C_k are disjoint clusters with sizes n_i, n_j and n_k.

A.4 Shape of TADs

We use shape parameters, to investigate the shape of the 32 TADs in Chr 5 (Table I) formed by loop anchors. We calculated three metrics to quantify the shapes, radius of gyration R_g, relative shape anisotropy, κ^2 and shape parameter, S. The value of R_g^2 is,

$$R_g^2 = \lambda_1 + \lambda_2 + \lambda_3, \tag{H1}$$

where λ_i are the eigenvalues of the gyration tensor; κ^2 is defined as,

$$\kappa^2 = \frac{3}{2}\frac{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}{(\lambda_1 + \lambda_2 + \lambda_3)^2} - \frac{1}{2}. \tag{H2}$$

The shape parameter, S, is,

$$S = 27 \prod_{i=1,2,3} (\lambda_i - \bar{\lambda})/\bar{\lambda} \tag{H3}$$

where $\bar{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$. The bounds for κ^2 is $0 \leq \kappa^2 \leq 1$, where 0 is for a highly symmetric conformation and 1 corresponds to a rod, S satisfies $-1/4 \leq S \leq 2$. If $-0.25 < S < 0$, then the shape is predominantly oblate and is prolate for $0 < S < 2$ [248,249]. The results in the Fig. A.4 (Left panel) show R_g^2, κ^2 and S measurements for the CCM for $\epsilon = 2.4 k_B T$. The top figure shows that R_g increases as the size of the TAD increases. Both the middle and bottom figures show that small TADs deviate from spherical shape (large value of κ^2 and S) but adopt a more spherical

shape as the size of TAD increases.

We also calculated the dispersion in the three measurements for each TAD among trajectories (right column in Fig. A.4). For instance, dispersion of the radius of gyration is defined as $\sigma_{R_g^2}/\mu_{R_g^2}$, where $\sigma_{R_g^2}$ and $\mu_{R_g^2}$ are the standard deviation and mean value of $\overline{R_g^2}$ over trajectories and the bar denotes the time average. The histograms of $\overline{R_g^2}$, $\overline{\kappa^2}$ and $\overline{S}$ normalized by their mean values are shown in Fig. A.5.

A.5 Chromsome 10

In order to check the transferability of the CCM, we obtained the contact maps of Chr 10 with the same set of parameters (Table 2.1) used for simulating Chr5. The locations of loop anchors used in Chr 10 simulation are summarized in Appendix Table A.1. The chromatin segments selected is from 70 to 82 Mbps. The types of monomers (chromatin loci) are determined using the Broad ChromHMM track, as described earlier. The numbers of active and repressive monomers are 3980 and 6020, respectively. About 80% of compartments inferred from experiment Hi-C contact map are correctly predicted by the CCM.

For precise comparison between the prediction of the CCM and experiment, we computed the Ward Linkage Matrices. Fig. A.6 shows the WLM inferred from experiment and computed directly from simulations. Just as for Chr 5, visual inspection suggests that $\epsilon = 2.4k_{\mathrm{B}}T$ and $\epsilon = 2.7k_{\mathrm{B}}T$ agrees best with experiments. To quantitatively compare the WLMs, we compute the Pearson correlation coef-

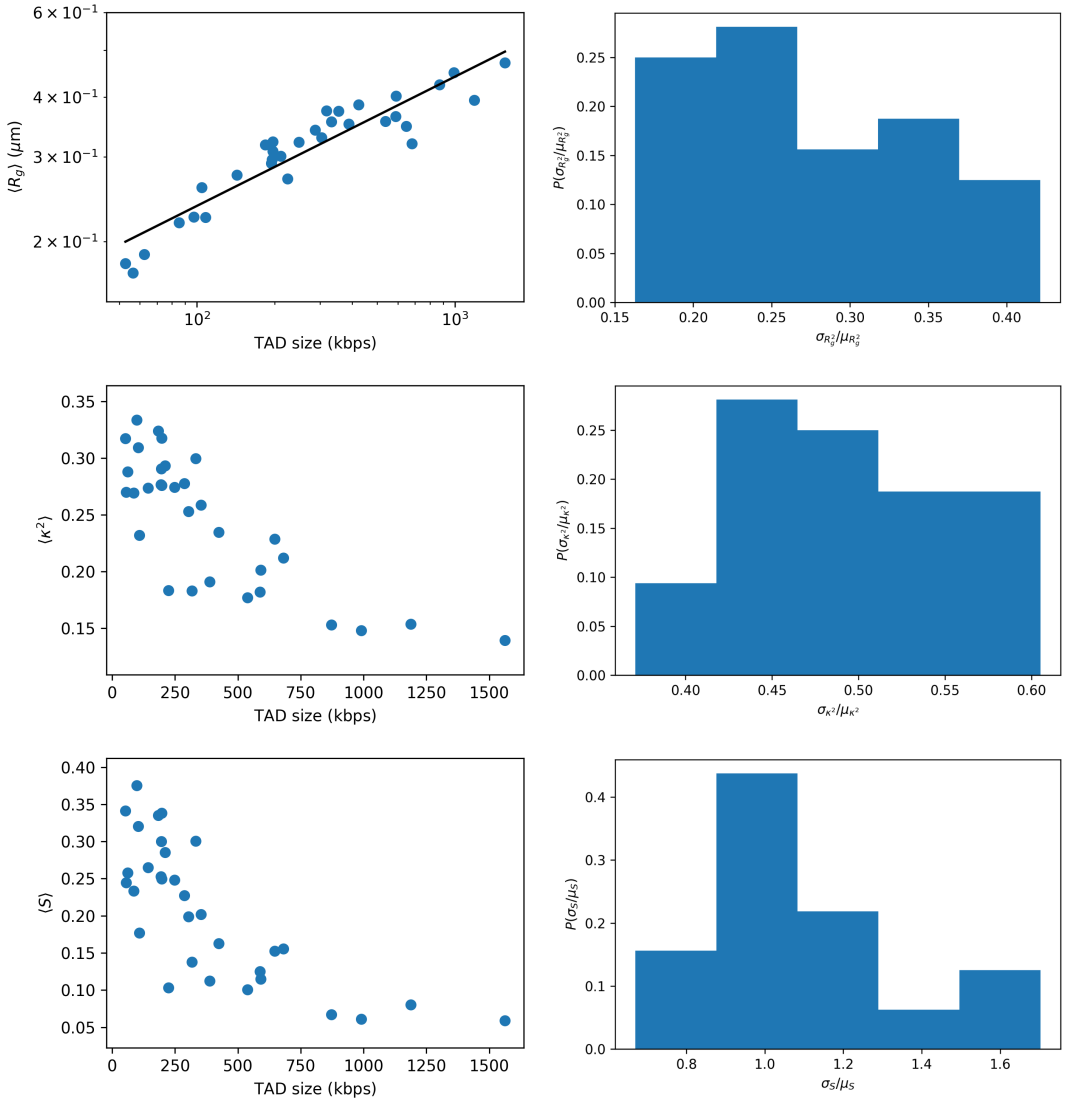


Figure A.4: **(Left panel)** $\langle R_g^2 \rangle$(top) (Eq.H1), $\langle \kappa^2 \rangle$(middle) (eq.H2) and $\langle S \rangle$(bottom) (eq.H3) for each TAD, where $\langle \cdot \rangle$ denotes both ensemble and time average. The black line in the top figure is the fit to the data, $\langle R(g) \rangle \sim l^{0.27}$, where l is the TAD size. **(Right panel)** Distribution $P(\sigma_{R_g^2}/\mu_{R_g^2})$(top), $P(\sigma_{\kappa^2}/\mu_{\kappa^2})$(middle) and $P(\sigma_S/\mu_S)$(bottom) over all TADs. $\sigma_{R_g^2} = \left[\left\langle \overline{R_g^2}^2 \right\rangle - \left\langle \overline{R_g^2} \right\rangle^2\right]^{1/2}$, $\sigma_{\kappa^2} = \left[\left\langle \overline{\kappa^2}^2 \right\rangle - \left\langle \overline{\kappa^2} \right\rangle^2\right]^{1/2}$ and $\sigma_{S^2} = \left[\left\langle \overline{S^2}^2 \right\rangle - \left\langle \overline{S^2} \right\rangle^2\right]^{1/2}$ where $\bar{\cdot}$ denotes time average over single trajectory and $\langle \cdot \rangle$ denotes ensemble average over different trajectories.

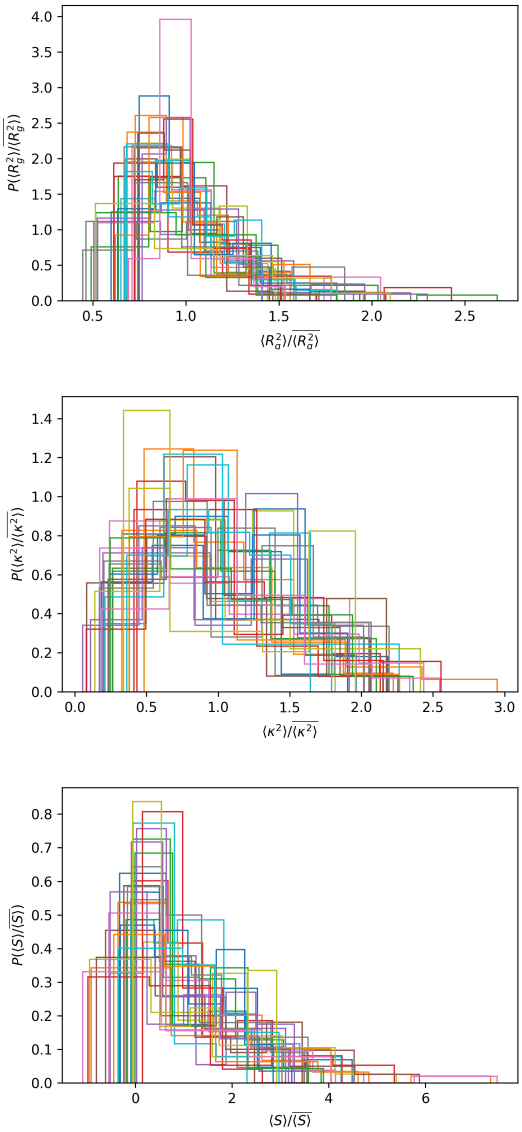


Figure A.5: The distribution of $\overline{R_g^2}/\langle\overline{R_g^2}\rangle$ for the thirty-two TADs in Chr5 where $\overline{R_g^2}$ is the time average value of the squared radius of gyration of single trajectory and $\langle\overline{R_g^2}\rangle$ is its mean value averaged over all independent trajectories. TADs are represented by different colors. Distribution of $\overline{\kappa^2}/\langle\overline{\kappa^2}\rangle$ and $\overline{S^2}/\langle\overline{S^2}\rangle$ are shown in the middle and bottom panels, respectively.

637(A),818(B)	637(A),960(A)	637(A),711(A)	831(B),960(A)
1924(A),2099(B)	1924(A),2199(A)	2146(B),2199(A)	2238(B),2474(B)
2620(B),2774(B)	2620(B),2890(B)	3096(B),3173(B)	3436(B),3828(A)
4186(B),4503(A)	4407(B),4503(A)	4674(A),4704(B)	4674(A),4750(B)
4704(B),4867(A)	4704(B),4750(B)	6922(B),7287(B)	9278(B),9356(B)
9435(A),9741(B)	9919(B),9985(A)	9940(A),9985(A)	

Table A.1: Loop anchor indices derived from the experimental data [28] for use in the CCM for Chr 10. Each pair of numbers represents single loop corresponding to the locations of the loop anchors along the backbone of the copolymer. The letter A (B) after each number indicates the type of the loop anchor.

ficient between experimental WLM and simulated WLMs. The values of Pearson correlation coefficients are 0.58, 0.75, 0.92 and 0.92 for $\epsilon = (1.0, 2.0, 2.4, 2.7)k_{\mathrm{B}}T$, respectively, suggesting that $\epsilon = 2.4k_{\mathrm{B}}T$ and $\epsilon = 2.7k_{\mathrm{B}}T$ give excellent agreement with experiments. This is consistent with our detailed study on Chr 5. These results show that the minimal CCM is sufficient to capture the features of the folded chromosomes. Applications of the CCM to other chromosomes are planned.

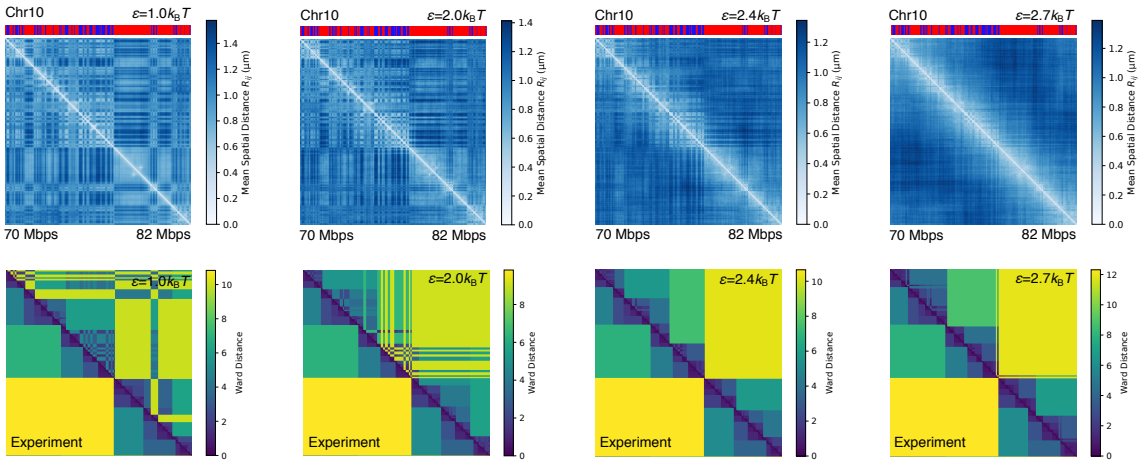


Figure A.6: Structural organization of Chromosome 10. **(Upper panel)** Distance maps of Chr10 for $\epsilon = (1.0, 2.0, 2.4, 2.7)k_{\mathrm{B}}T$. **(Lower panel)** Comparison between the experimental WLM (lower triangle) and the simulated Ward Linkage Matrix (WLMs) (upper triangle). Just as we found in Chr5, $\epsilon = 2.4k_{\mathrm{B}}T$ provides the best comparison with experiment, implying that the CCM is transferable.

Appendix B: Supplementary Information for Chapter 4

B.1 Procedure of fitting the FISH data

We use Eq. 4.11 to fit to the FISH data. The integration of Eq. 4.11 gives the CDF($R|\langle R\rangle$),

$$\begin{aligned}\mathrm{CDF}(R|\langle R\rangle) &= \int_0^R P(r|\langle R\rangle)dr \\ &= 1 - \frac{A\langle R\rangle}{\delta} B^{-\frac{3+g}{\delta}} \Gamma\left(\frac{3+g}{\delta}, B\left(\frac{R}{\langle R\rangle}\right)^{\delta}\right),\end{aligned} \tag{B.1}$$

where $\Gamma(s,x)$ is the lower incomplete gamma function. Thus, the integral of Eq. 4.19 with respect to r may be written as,

$$\mathrm{CDF}(R) = \eta \mathrm{CDF}(R|\langle R_{1,mn}\rangle) + (1-\eta)\mathrm{CDF}(R|\langle R_{2,mn}\rangle). \tag{B.2}$$

Given the values of g, δ and $\langle R\rangle$, the two constants A and B are determined using the conditions: (1) The distribution Eq.4.11 is normalized, $\int_0^\infty \mathrm{d}r P(r|\langle R\rangle) = 1$. (2) The calculated value of the second moment $\langle R\rangle^2$, $\int_0^\infty \mathrm{d}r r P(r|\langle R\rangle) = \langle R\rangle$ should equal the measured value. With these two constraints, we obtain

$$A = \frac{\delta}{\langle R \rangle} \frac{\Gamma^{3+g}((4+g)/\delta)}{\Gamma^{4+g}((3+g)/\delta)},$$
$$B = \frac{\Gamma^{\delta}((4+g)/\delta)}{\Gamma^{\delta}((3+g)/\delta)}, \tag{B.3}$$

where $\Gamma(z)$ is the gamma function.

Using Eq. B.3, Eq. B.1 can be further simplified as,

$$\mathrm{CDF}(R|\langle R \rangle) = 1 - \frac{\Gamma\big((3+g)/\delta, B(R/\langle R \rangle)^{\delta}\big)}{\Gamma((3+g)/\delta)}. \tag{B.4}$$

Using Eq. B.2 and B.4, we minimize the square of the difference between the calculated and the measured values of the CDFs. Using this procedure we calculated the values of η, $\langle R_{1,mn} \rangle$ and $\langle R_{2,mn} \rangle$ for all eight loci pairs for which FISH data were reported [28]. The best fit values of the three parameters are given in Table B.1. The goodness of the fits for different values of g and δ, corresponding to three different polymer models, are reported in Table B.2.

	η	$\langle R_{1,mn}\rangle(\mu m)$	$\langle R_{2,mn}\rangle(\mu m)$	Kolmogorov-Smirnov statistics	p-value
peak1-control	0.55	0.63	1.06	0.0350	0.999
peak1-loop	0.36	0.24	0.56	0.0619	0.834
peak2-control	0.63	0.47	1.04	0.0522	0.977
peak2-loop	0.75	0.33	1.07	0.110	0.248
peak3-control	0.97	0.67	4.08	0.0561	0.902
peak3-loop	0.91	0.35	1.64	0.0507	0.954
peak4-control	0.27	0.48	1.25	0.0633	0.988
peak4-loop	0.42	0.30	1.21	0.0657	0.982

Table B.1: Values of the optimal parameters obtained by fitting the FISH data using our theory, given by Eq.4.19. The parameters η, $\langle R_{1,mn}\rangle(\mu m)$, and $\langle R_{2,mn}\rangle(\mu m)$ are defined in the main text. It is interesting that the values of $\langle R_{1,mn}\rangle$ for the peak-loop positions are similar ($\approx 0.3\mu$m). Kolmogorov-Smirnov statistics and their p-values are also reported.

g	δ	RE
1	5/4	0.00392
0	2	0.00397
0.71	5/2	0.00762

Table B.2: Residual error (RE) for fits of theory to the FISH data using three different sets of g and δ (Eq. 4.11). We define RE as RE $= \sum_j (1/N_j) \sum_i^{N_j} (y_i - f(x_i))^2$ where y_i is the i^{th} value of the measured data, $f(x_i)$ is the fit value for y_i. The sum is over all the data points from all the eight curves marked by index j ($j = 1, 2, \cdots, 8$). N_j is the number of data points of j^{th} curve. The values of g and δ in the first row corresponds to chromosomes, while those in the second and third rows are for the Rouse model and a polymer in a good solvent, respectively. The smallest error is for the exponents describing the chromosome model. Surprisingly, the RE value for the unphysical Rouse model is also low.

B.2 Fitting FISH data by assuming homogenous cell population

To assess the quality of fits of the FISH data that assuming that the cell population is homogeneous, we use Eq.B.4 with g and δ as free parameters. Since for a homogenous population, the distribution of spatial distance can be normalized by dividing the spatial distance by the mean, which eliminates the parameter $\langle R \rangle$, leaving g and δ as the only free parameters. Table B.3 show the result of the fits and the values of g and δ. First, the Kolmogorov-Smirnov statistics are poorer than for for fits obtained using two subpopulations. Second, the values the extracted values of g and δ are unphysical. We, therefore, surmise that the FISH data cannot be reasonably fitted assuming the cell population is homogeneous.

	g	δ	Kolmogorov-Smirnov statistics	p-value
peak1-control	8.68	0.30	0.0346	0.999
peak1-loop	115.37	0.017	0.0716	0.678
peak2-control	111.38	0.017	0.0923	0.458
peak2-loop	108.70	0.012	0.118	0.180
peak3-control	204.82	0.015	0.113	0.132
peak3-loop	180.84	0.010	0.148	0.019
peak4-control	-0.99	1.42	0.0558	0.997
peak4-loop	-1.76	1.26	0.0863	0.850

Table B.3: Values of the optimal g and δ obtained by fitting the FISH data assuming that the cell population is homogenous (Eq.B.4). The best fit values of g and δ are unphysical. The results of the Kolmogorov-Smirnov test are are inferior to the results reported in Table B.1.

B.3 Non-negative Tikhonov regularization Method

In this section, we show how to solve $P(\langle R\rangle_i)$ in Eq.4.21 numerically. In order to simplify the notation, we denote $\mathrm{CDF}(R_j|\langle R\rangle_i) \equiv A$, $\mathrm{CDF}(R_j) \equiv b$ and $\Delta\langle R\rangle P(\langle R\rangle_i) \equiv x$. Then, Eq.4.21 is written as,

$$Ax - b = 0 \tag{B.5}$$

Since Eq.4.20 is an integral equation for which solutions may not be unique. For such an ill-posed problem, Eq.4.21 (or Eq.4.20) is usually solved using the Tikhonov regularization method, which is to solve,

$$\min(||Ax - b||_2^2 + \alpha^2||x||_2^2), \tag{B.6}$$

where α is a tuned parameter controlling the smoothness of solution x. For our problem, we also need an additional non-negative constraint on x since x is a probability density function. Thus, we want to solve Eq.B.6 subject to $x \geq 0$. Let us construct the matrix,

$$C = \begin{bmatrix} A \\ \alpha I \end{bmatrix} \tag{B.7}$$

and the vector,

$$d = \begin{bmatrix} b \\ \\ 0 \end{bmatrix} \tag{B.8}$$

where I is the identity matrix. Solving Eq.B.6 subject to $x \geq 0$ is equivalent to solving,

$$\min||Cx - d||_2^2, \text{ subject to } x \geq 0. \tag{B.9}$$

The above equation is a non-negative least square (NNLS) problem, which can be solved using an active set algorithm [250]. To solve the Eq.B.9, the value of α, which is a controlled parameter, needs to be provided. From a graphical perspective, α controls the smoothness of the solution, with x being smoother for larger α. The statistical significance of α lies in its ability to control the trade-off between the goodness of fit and the extent of over-fitting. To choose the value of α in a systematic way, we follow the procedure demonstrated in [251]. The goodness of fits is measured by the residue norm $||Ax - b||_2^2$ and the solution norm $||x||_2^2$ is used as a proxy to the extent of over-fitting. The L-curve is the function between $||Ax - b||_2^2$ and $||x||_2^2$ for different values of α. The optimal value of α is located where the L-curve has largest curvature. We solve Eq.B.9 using the optimal value of α. In practice, we use the function provided in PYTHON *scipy* package to solve Eq.B.9.

Appendix C: Supplementary Information for Chapter 5

C.1 Derivation of a lower bound of spatial distance

In this appendix, we prove the theoretical lower bound of $\langle R_{mn} \rangle$ given P_{mn}. We demonstrate this by considering the case there are two subpopulations, $S = 2$. In this case, we $\langle R_{mn} \rangle = \eta \langle R_{1,mn} \rangle + (1-\eta) \langle R_{2,mn} \rangle$ and $P_{mn} = \eta P_{1,mn} + (1-\eta) P_{2,mn}$. Note that $\langle R_{1,mn} \rangle = R_0(P_{1,mn})$ and $\langle R_{2,mn} \rangle = R_0(P_{2,mn})$. For simplicity, we denote $P_{1,mn} = x$ and $P_{2,mn} = y$. Given the value of the contact probability P_{mn}, we show that the lower bound for $\langle R_{mn} \rangle = R_0(P_{mn})$. This is equivalent to the optimization problem,

$$\begin{aligned} &\text{maximize } f(x,y) \\ &\text{subject to } g(x,y) = 0 \end{aligned} \tag{C.1}$$

where $f(x,y) = -\eta R_0(x) - (1-\eta) R_0(y)$ and $g(x,y) = \eta x + (1-\eta) y - P_{mn}$. The Lagrange multiplier is $L(x,y,\phi) = f(x,y) - \phi g(x,y)$. Using the condition that $\nabla_{x,y,\phi} \mathcal{L}(x,y,\phi) = 0$, it can be shown that $f(x,y)$ is maximized when $x = y$. Thus we proved that R_{mn} is minimized when $P_{1,mn} = P_{2,mn}$ and the its minimum is $R_0(P_{mn})$. It is important to point out that the proof shown here can be easily generalized for any form of function R_0 and any number of subpopulations S.

C.2 Mean spatial distances are metric but not Euclidean in 3D space

A metric must satifies the following condition: i) non-negativity ii) identiy of indiscernibles iii) symmetry iv) triangle inequality. It is clear that the mean spatial distance in a polymer systems satisfies the first three conditions. i) $\langle R_{mn}\rangle \geq 0$ for any m, n. ii) If $\langle R_{mn}\rangle = 0$, m and n loci has the same coordinates. And iii) $\langle R_{mn}\rangle = \langle R_{nm}\rangle$. Now I show that the triangle inequality is also satisfied, $\langle R_{mn}\rangle + \langle R_{nl}\rangle \geq \langle R_{ml}\rangle$. I can write $\langle R_{mn}\rangle$ as integral $\langle R_{mn}\rangle = \int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{mn}$ where $P(R_{mn}, R_{nl}, R_{ml})$ is the joint probability distribution for distance R_{mn}, R_{nl} and R_{ml}. Similarly I have $\langle R_{nl}\rangle = \int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{nl}$ and $\langle R_{ml}\rangle = \int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{ml}$. I have

$$\begin{aligned} &\int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})(R_{mn} + R_{nl}) \\ =&\int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{mn} + \\ &\int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{nl} = \langle R_{mn}\rangle + \langle R_{nl}\rangle \end{aligned} \tag{C.2}$$

Since R_{mn}, R_{nl} and R_{ml} are the spatial distances which satisfies the triangle inequality, $R_{mn}+R_{nl} \geq R_{ml}$. Thus I have $\langle R_{ml}\rangle = \int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})R_{ml} \leq \int \mathrm{d}R_{mn}\mathrm{d}R_{nl}\mathrm{d}R_{ml}P(R_{mn}, R_{nl}, R_{ml})(R_{mn}+R_{nl})$ which leads to the triangle inequality,

$$\langle R_{ml}\rangle \leq \langle R_{mn}\rangle + \langle R_{nl}\rangle \tag{C.3}$$

However mean spatial distances $\langle R \rangle$ are not Euclidean in 3D space. This can be illustrated by considering the ideal chain in which the mean spatial distance $\langle R_{mn} \rangle = |m-n|^{1/2}$ for any m, n. Hence the mean spatial distance matrix DM for a ideal chain is,

$$\begin{bmatrix} 0 & 1 & \sqrt{2} & \dots & \sqrt{N} \\ 1 & 0 & 1 & \dots & \sqrt{N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sqrt{N} & \sqrt{N-1} & \sqrt{N-2} & \dots & 0 \end{bmatrix} \tag{C.4}$$

where $N+1$ is the number of monomers. A distance matrix $\boldsymbol{D}$ with dimension $n \times n$ is Euclidean if there exists a configuration of set of n points in a Euclidean space of dimension p whose between-points distances exactly match $\boldsymbol{D}$. Schoenberg [252] showed that $\boldsymbol{D}$ is Euclidean iff

$$\boldsymbol{F} \equiv -(\boldsymbol{I} - \boldsymbol{e}\boldsymbol{s}^T)(\boldsymbol{D} \circ \boldsymbol{D})(\boldsymbol{I} - \boldsymbol{s}\boldsymbol{e}^T) \tag{C.5}$$

is positive semi-definite. $\circ$ is Hadamard product. $\boldsymbol{I}$ is the identity matrix. $\boldsymbol{e}$ is vector all of whose values are one and $\boldsymbol{s}$ is any vector such that $\boldsymbol{s}^T\boldsymbol{e} = 1$ and $(\boldsymbol{D} \circ \boldsymbol{D})\boldsymbol{s} \neq 0$. It can be shown that Eq.C.4 satisfies to condition (Eq.C.5) hence it is Euclidean in dimension p where p is unknown. It has been shown [253] that when the distance matrix $\boldsymbol{D}$ is Euclidean, its Euclidean dimension $p = \text{rank}(\boldsymbol{D}) - 1$. The rank of matrix in Eq.C.4 is $N+1$. Thus for $N \geq 4$, the mean spatial distance of a ideal chain is no longer Euclidean in three dimensional space. Thus there does not exist a

realization of conformation of monomers whose distances are the values in the mean spatial distance matrices.

It is noteworthy that from Eq.C.3 one can observe that the inverse of the contact probablity does not satisfy the triangle inequality and hence is not a metric. The inverse of contact probability is $1/P_{mn} \sim \langle R_{mn} \rangle^{\alpha}$ where $\alpha = 3$ for GRMC and ≈ 4.0 for Human Interphase Chromosome *in vivo*. It is easy to observe that the inequality $1/P_{ml} \leq 1/P_{mn} + 1/P_{nl}$ does not hold in general.

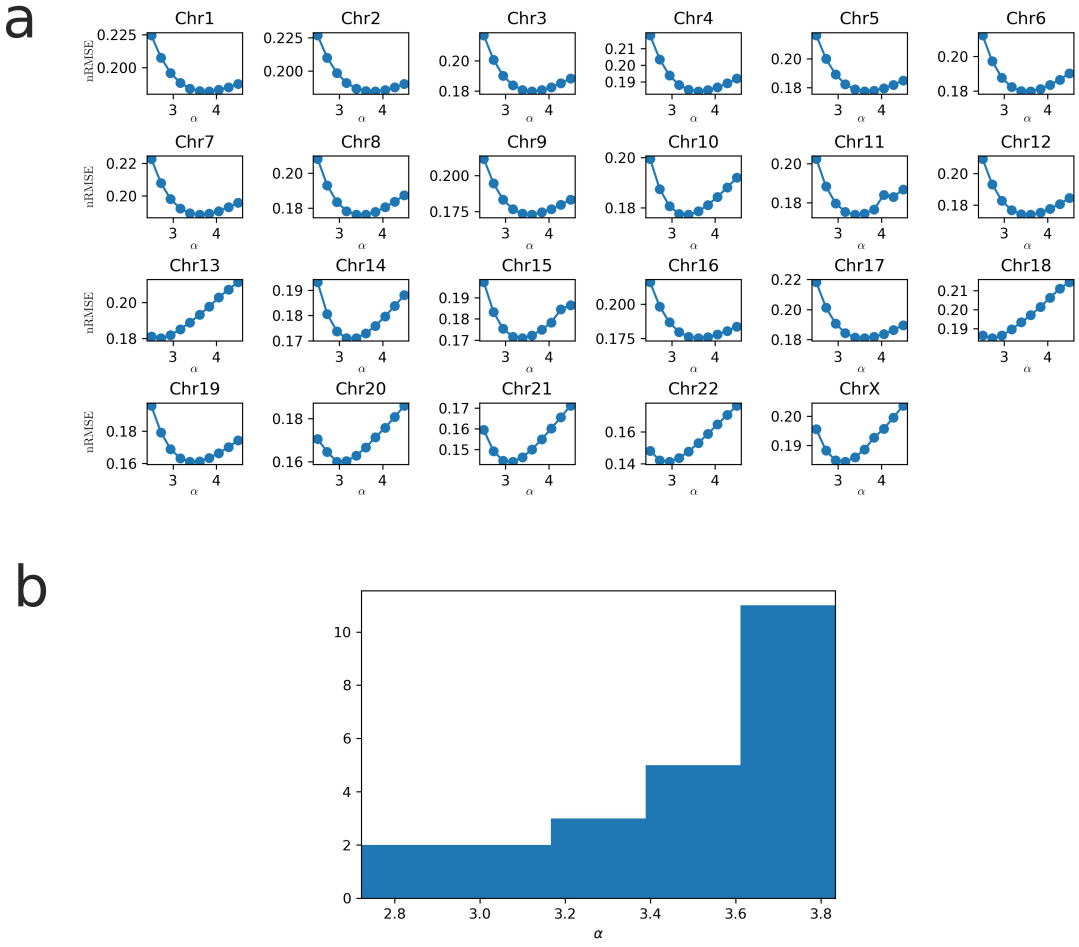
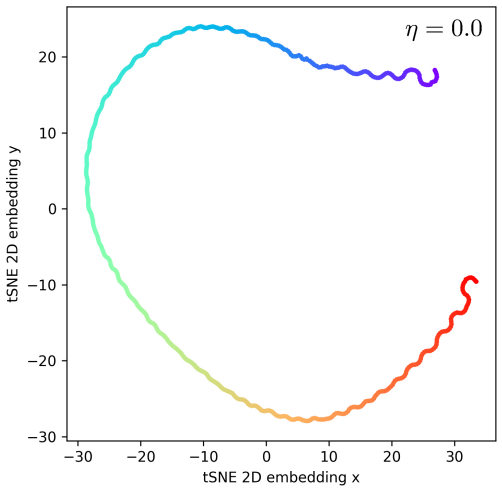
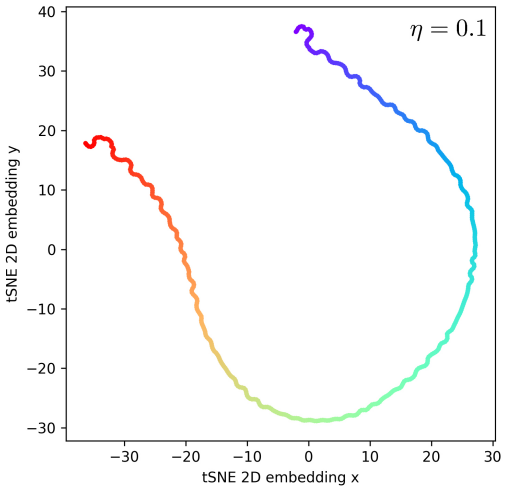
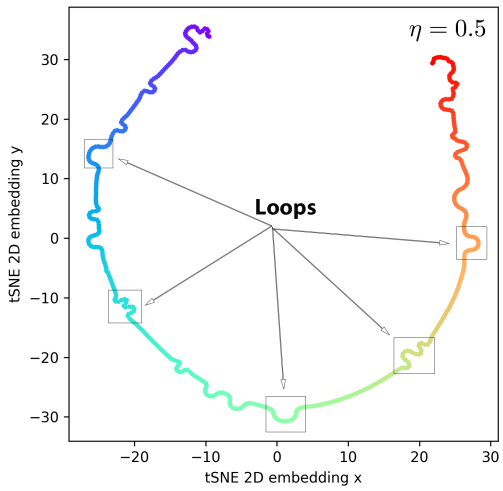
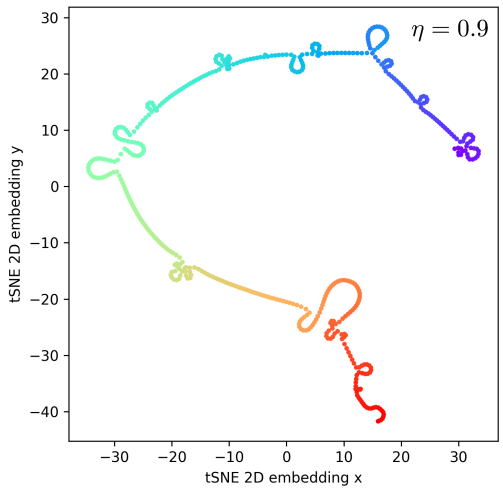


Figure C.1: **(a)** nRMSE as a function of α for all 23 Chromosomes of Human Interphase GM12878 Cell. **(b)** shows the histogram of values of α which minimizes the nRMSE for all 23 chromosomes.


$\eta = 0.0$
tSNE 2D embedding y
tSNE 2D embedding x


$\eta = 0.1$
tSNE 2D embedding y
tSNE 2D embedding x


$\eta = 0.5$
Loops
tSNE 2D embedding y
tSNE 2D embedding x


$\eta = 0.9$
tSNE 2D embedding y
tSNE 2D embedding x

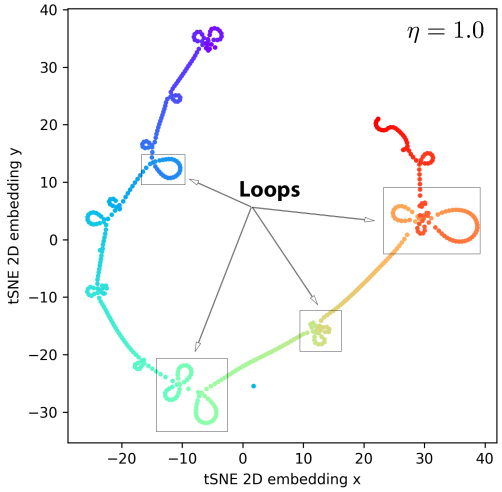

$\eta = 1.0$
Loops
tSNE 2D embedding y
tSNE 2D embedding x

Figure C.2: 2D t-SNE embedding for DMs of mixture system with $\eta = (0.0, 0.1, 0.5, 0.9, 1.0)$. Examples of loops are marked for $\eta = 0.5, 1.0$. It is clear that t-SNE 2D embedding is able to represent the loops present in the system and also is sensitive to the values of η. Note that the extent of loops for $\eta = 0.5$ is similar to what is observed for Human interphase chromosomes (Fig.5.7(f)). For $\eta = 1.0$, loops are present in all the cells, thus the anchors of loops are connected to each other in the tSNE representation. However in a mixture system, $\eta = 0.5$, loops are not present in all the cells, they are then represented by "open" curls/loops in the 2D t-SNE embedding.

Appendix D: Supplementary Information for Chapter 6

D.1 Limiting conditions with $\kappa \to 0$ and $\kappa \to \infty$

In this section, we return to the case $n = 2$ of identical motors. We have shown that the position of the system undergoes a one-dimensional periodic random walk with period of two. The forward and backward rates appear in Eq.6.16 as complicated summations. It is difficult to reduce the form in Eq.6.16 in simpler terms for arbitrary values of parameters. However the limiting cases of weak coupling ($\kappa \ll 1$) and strong coupling ($\kappa \gg 1$) can be investigated. First we look at what would be form of stationary distribution π_i^{s} at these two limiting cases. The stationary distribution π_i^{s} is given by,

$$\pi_i^{\mathrm{s}} = \pi_0^{\mathrm{s}} \frac{\prod_{j=0}^{i-1} \omega_j^+}{\prod_{j=1}^{i} \omega_j^-} \quad \text{for } i > 0 \tag{D.1}$$

$$\pi_i^{\mathrm{s}} = \pi_{-i}^{\mathrm{s}} \quad \text{for } i < 0 \tag{D.2}$$

and the ratio $\omega_i^+ / \omega_{i+1}^-$ is given by,

$$\frac{\omega_i^+}{\omega_{i+1}^-} = \frac{k_{\rm L}^+(i) + k_{\rm T}^-(i)}{k_{\rm L}^-(i+1) + k_{\rm T}^+(i+1)} \tag{D.3}$$

$$= \frac{e^{-\beta\theta^+\Delta E_i} + e^{-\beta\theta^-\Delta E_i}}{e^{\beta\theta^+\Delta E_i} + e^{\beta\theta^-\Delta E_i}} \tag{D.4}$$

1) Strong coupling, $\kappa \gg 1$. The cases $\theta^+ < \theta^-$, $\theta^+ = \theta^-$ and $\theta^+ > \theta^-$ needed to discussed separately.

i) $\theta^+ > \theta^-$. In this case, for $\kappa \gg 1$, we have $k_0^+ e^{-\beta\theta^+\Delta E_i} \gg k_0^- e^{-\beta\theta^-\Delta E_i}$ and $k_0^+ e^{\beta\theta^+\Delta E_i} \ll k_0^- e^{\beta\theta^-\Delta E_i}$. Eq.D.3 reduces to,

$$\frac{\omega_i^+}{\omega_{i+1}^-} \approx \frac{k_0^+ e^{-\beta\theta^+\Delta E_i}}{k_0^- e^{\beta\theta^-\Delta E_i}} = \frac{k_0^+}{k_0^-} e^{-\beta\Delta E_i} \tag{D.5}$$

The last equate uses the fact $\theta^+ + \theta^- = 1$. Then we obtain $\pi_i^{\rm s}$,

$$\pi_i^{\rm s} = \pi_0^{\rm s} \frac{\prod_{j=0}^{i-1} \omega_j^+}{\prod_{j=1}^{i} \omega_j^-} = \pi_0^{\rm s} \left(\frac{k_0^+}{k_0^-}\right)^i e^{-\frac{\beta\kappa d^2 i^2}{4}} \tag{D.6}$$

ii) $\theta^+ < \theta^-$. In this case, we have,

$$\pi_i^{\rm s} = \pi_0^{\rm s} \left(\frac{k_0^-}{k_0^+}\right)^i e^{-\frac{\beta\kappa d^2 i^2}{4}} \tag{D.7}$$

iii) $\theta^+ = \theta^- = 1/2$. In this case,

$$\pi_i^{\rm s} = \pi_0^{\rm s} e^{-\frac{\beta\kappa d^2 i^2}{4}} \tag{D.8}$$

2) Weak coupling, $\kappa \ll 1$. In this case, we can use taylor expansion on Eq.D.3

yielding,

$$\frac{\omega_i^+}{\omega_{i+1}^-} \approx \frac{k_0^+(1-\beta\theta^+\Delta E_i)+k_0^-(1-\beta\theta^-\Delta E_i)}{k_0^+(1+\beta\theta^+\Delta E_i)+k_0^-(1+\beta\theta^-\Delta E_i)} = e^{-2\eta\Delta E_i} \tag{D.9}$$

where $\eta = \frac{r\beta\theta^+ + \beta\theta^-}{r+1}$ and $r = k_0^+/k_0^-$. Finally, the stationary distribution π_i^{s} is obtained,

$$\pi_i^{\mathrm{s}} = \pi_0^{\mathrm{s}} e^{-\frac{2(r\theta^+ + \theta^-)}{r+1}\frac{\beta\kappa d^2 i^2}{4}} \tag{D.10}$$

Note that the factor $r\theta^+ + \theta^-$ can be smaller than zero which means π_i^{s} diverges with increasing of i. This indicates that there is no stationary distribution when $\theta^+ \leq -1/(r-1)$. For large r (Kinesin), this practically menas that $\theta^+ > 0$. For relative small r (Dynein), in principle θ^+ can take negative values.

For both cases, we show that the stationary distribution π_i^{s} decay exponentially fast with square of i. Then in computing rates λ_1, λ_2, μ_1 and μ_2, we can only consider first term, which leads to

$$\lambda_1 = 2k_0^+ e^{-\frac{\beta\theta^+\kappa d^2}{4}} \tag{D.11}$$

$$\lambda_2 = 2k_0^+ e^{-\frac{\beta\theta^+\kappa d^2}{4}} \cosh\left(\frac{\beta\theta^+\kappa d^2}{2}\right) \tag{D.12}$$

$$\mu_1 = 2k_0^- e^{-\frac{\beta\theta^-\kappa d^2}{4}} \cosh\left(\frac{\beta\theta^-\kappa d^2}{2}\right) \tag{D.13}$$

$$\mu_2 = 2k_0^- e^{-\frac{\beta\theta^-\kappa d^2}{4}} \tag{D.14}$$

Plugging Eq.D.11 into Eq.6.17 yields,

$$\frac{\langle v \rangle}{v_0} = \frac{r+1}{re^{\epsilon\theta^+} + e^{\epsilon\theta^-}} \tag{D.15}$$

where $\epsilon = \beta\kappa d^2/4$ and $r = k_0^+/k_0^-$ and $\theta^- = 1-\theta^+$. v_0 is the velocity of single motor under zero load $v_0 = (k_0^+ - k_0^-)d$ where d is the step size of a single motor. Eq.D.15 shows that the reduced velocity can be expressed in terms of three dimensionless quantities ϵ, r and θ^+. ϵ quantifies the energy associated with coupling. r is the ratio between forward and backward rates and directly related to the energy consumed through ATP hydrolysis and θ^+ is the distribution factor characterizing the location of transition state.

Eq.D.15 shows that velocity of two coupling motors is reduced compared to the single motor and monotonically decreases with increasing of coupling strength κ. At small κ, the motor hardly affect each other resulting the similar velocity of that of single motor. At large κ, the velocity vanishes. We reason that this is because that the internal tension is built when the system moves. With larger coupling strength κ, the energy associated with internal tension increases which makes the motor harder to step. At the limit of very large κ, it takes so much energy for either of the two motors to step that the velocity of the system vanishes.

The vanishing of the velocity is not because that the system becomes diffusive. The diffusion constant also vanishes at large κ (Eq.6.17). This indicates that the system is "frozen" rather than becomes diffusive with large coupling strength. At small κ, the diffusion constant of two coupled motors is simply half of that of a

single motor.

D.2 Coupled Motor System of identical motors with $n > 2$

In section 6.3.1, I showed that the coupled motor system of two identical motors can be mapped to a periodic one-dimensional random walk of period of two. The same argument can be extended to more number of motors $n > 2$ if they are identical.

For $n = 3$, using Eq.6.5, 6.6 and 6.7, we obtain that the rates,

$$k_1^+ = k_0^+ e^{-(1/3)\beta\theta^+\kappa d^2(2\widetilde{x}_1-\widetilde{x}_2-\widetilde{x}_3+1)} \tag{D.16}$$

$$k_1^- = k_0^- e^{-(1/3)\beta\theta^-\kappa d^2(-(2\widetilde{x}_1-\widetilde{x}_2-\widetilde{x}_3)+1)} \tag{D.17}$$

$$k_2^+ = k_0^+ e^{-(1/3)\beta\theta^+\kappa d^2(2\widetilde{x}_2-\widetilde{x}_1-\widetilde{x}_3+1)} \tag{D.18}$$

$$k_2^- = k_0^- e^{-(1/3)\beta\theta^-\kappa d^2(-(2\widetilde{x}_2-\widetilde{x}_1-\widetilde{x}_3)+1)} \tag{D.19}$$

$$k_3^+ = k_0^+ e^{-(1/3)\beta\theta^+\kappa d^2(2\widetilde{x}_3-\widetilde{x}_1-\widetilde{x}_2+1)} \tag{D.20}$$

$$k_3^- = k_0^- e^{-(1/3)\beta\theta^-\kappa d^2(-(2\widetilde{x}_3-\widetilde{x}_1-\widetilde{x}_2)+1)} \tag{D.21}$$

Like in the case with $n = 2$, it is useful to define new variables $\widetilde{s}_1 = 2\widetilde{x}_1 - \widetilde{x}_2 - \widetilde{x}_3$ and $\widetilde{s}_2 = 2\widetilde{x}_2 - \widetilde{x}_1 - \widetilde{x}_3$ and $\widetilde{s}_3 = \widetilde{x}_1 + \widetilde{x}_2 + \widetilde{x}_3$. The rates only depends on $\widetilde{s}_1$ and $\widetilde{s}_2$ explicitly but not $\widetilde{s}_3$. Similarly, summing the master equation over variables $\widetilde{s}_1$ and $\widetilde{s}_2$ leads to the master equation with only one variable $\widetilde{s}_3$ which measures the displacement of the system. The system can be mapped to an equivalent pe-

riodic random walk with period of 3 (Fig.D.1). The three layers are marked by Blue, Magenta and Red colors. Without loss of generality, set blue layer contains the initial relaxed state $(\widetilde{s}_1, \widetilde{s}_2, \widetilde{s}_3) = (0, 0, 0)$. $\widetilde{s}_3$ are orthogonal to other two variables $\widetilde{s}_1$ and $\widetilde{s}_2$. Ignoring the variable $\widetilde{s}_3$, the blue layer contains sites $(\widetilde{s}_1, \widetilde{s}_2) \in \Gamma_1 = \{(0,0), (0,3), (3,0), (0,-3), (-3,0), \cdots\}$ which are the sites of square lattice with lattice constant 3 with . Similarly, the magenta layer contains sites of $(\widetilde{s}_1, \widetilde{s}_2) \in \Gamma_2 = \{(-1,2), (-1,-1), (2,2), (2,-1), \cdots\}$ and red layer contains sites of $(\widetilde{s}_1, \widetilde{s}_2) \in \Gamma_3 = \{(-2,1), (1,1), (-2,-2), (1,-2), \cdots\}$. Provided that the stationary distribution $\pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)$ are known. The alternating rates associated with periodic one-dimensional random walk are given by,

$$\lambda_1 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_1} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^+(\widetilde{s}_1, \widetilde{s}_2) + k_2^+(\widetilde{s}_1, \widetilde{s}_2) + k_3^+(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_1} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.22}$$

$$\lambda_2 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_2} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^+(\widetilde{s}_1, \widetilde{s}_2) + k_2^+(\widetilde{s}_1, \widetilde{s}_2) + k_3^+(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_2} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.23}$$

$$\lambda_3 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_3} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^+(\widetilde{s}_1, \widetilde{s}_2) + k_2^+(\widetilde{s}_1, \widetilde{s}_2) + k_3^+(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_3} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.24}$$

$$\mu_1 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_1} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^-(\widetilde{s}_1, \widetilde{s}_2) + k_2^-(\widetilde{s}_1, \widetilde{s}_2) + k_3^-(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_1} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.25}$$

$$\mu_2 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_2} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^-(\widetilde{s}_1, \widetilde{s}_2) + k_2^-(\widetilde{s}_1, \widetilde{s}_2) + k_3^-(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_2} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.26}$$

$$\mu_3 = \frac{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_3} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)(k_1^-(\widetilde{s}_1, \widetilde{s}_2) + k_2^-(\widetilde{s}_1, \widetilde{s}_2) + k_3^-(\widetilde{s}_1, \widetilde{s}_2))}{\sum_{(\widetilde{s}_1,\widetilde{s}_2)\in\Gamma_3} \pi^{\mathrm{s}}(\widetilde{s}_1, \widetilde{s}_2)} \tag{D.27}$$

Generally speaking, coupled motor system with n number of identical motors can be mapped to a periodic random walk of period of n. Let's denote the forward

and backward rates of such a periodic random walk process are $\{\lambda_i\}$ and $\{\mu_i\}$ with $i \in (1, 2, \cdots, n)$, respectively. They are given by,

$$\lambda_i = \frac{\sum_{\widetilde{\boldsymbol{s}} \in \Gamma_i} \pi^{\mathrm{s}}(\widetilde{\boldsymbol{s}}) \sum_{j=1}^{n} k_j^{+}(\widetilde{\boldsymbol{s}})}{\sum_{\widetilde{\boldsymbol{s}} \in \Gamma_i} \pi^{\mathrm{s}}(\widetilde{\boldsymbol{s}})} \tag{D.28}$$

$$\mu_i = \frac{\sum_{\widetilde{\boldsymbol{s}} \in \Gamma_i} \pi^{\mathrm{s}}(\widetilde{\boldsymbol{s}}) \sum_{j=1}^{n} k_j^{-}(\widetilde{\boldsymbol{s}})}{\sum_{\widetilde{\boldsymbol{s}} \in \Gamma_i} \pi^{\mathrm{s}}(\widetilde{\boldsymbol{s}})} \tag{D.29}$$

$$\tag{D.30}$$

where Γ_i is the set of sites on the layer of indices i and $\widetilde{\boldsymbol{s}} = (\widetilde{s}_1, \widetilde{s}_2, \cdots, \widetilde{s}_n)$ and $\widetilde{s}_i = (n-1)\widetilde{x}_i - \sum_{j \neq i} \widetilde{x}_j$ for $i \neq n$ and $\widetilde{s}_n = \sum_i \widetilde{x}_i$. $\pi^{\mathrm{s}}(\widetilde{\boldsymbol{s}})$ is the stationary distribution for variables $(\widetilde{s}_1, \widetilde{s}_2, \cdots, \widetilde{s}_{n-1})$

The solution for mean velocity is given by [237],

$$\langle \widetilde{v}_n \rangle = \frac{n}{\sum_{i=1}^{n} r_i} \left[1 - \prod_{i=1}^{n} \left(\frac{\mu_i}{\lambda_i} \right) \right] \tag{D.31}$$

where r_i is given by,

$$r_i = \frac{1}{\lambda_i} \left[1 + \sum_{j=1}^{n-1} \prod_{k=1}^{j} \left(\frac{\mu_{i+k}}{\lambda_{i+k}} \right) \right] \tag{D.32}$$

The solution for diffusion constant $\widetilde{D}_n$ is not shown here, but is given in Derrida [237].

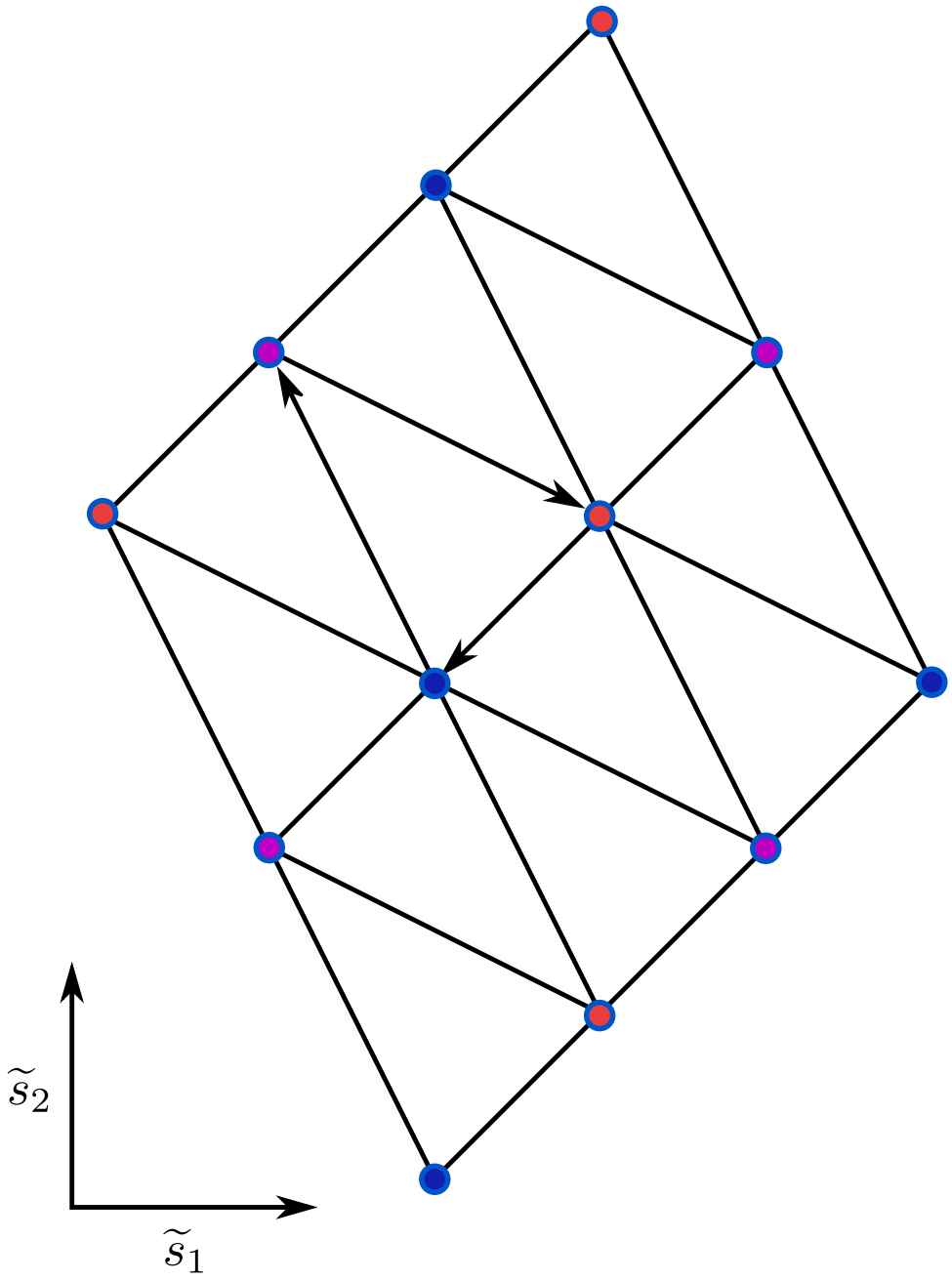


Figure D.1: The coupled motor system of 3 identical motors can be mapped to a one-dimensional random walk of period of 3. The variable $\widetilde{s}_3$ is orthogonal to $\widetilde{s}_1$ and $\widetilde{s}_2$. This figure shows the projection of $(\widetilde{s}_1, \widetilde{s}_2, \widetilde{s}_3)$ on to $(\widetilde{s}_1, \widetilde{s}_2)$ plane. The arrows marks transitions: $(\widetilde{s}_1, \widetilde{s}_2, \widetilde{s}_3) \to (\widetilde{s}_1 - 1, \widetilde{s}_2 + 2, \widetilde{s}_3 + 1) \to (\widetilde{s}_1 + 1, \widetilde{s}_2 + 1, \widetilde{s}_3 + 2) \to (\widetilde{s}_1, \widetilde{s}_2, \widetilde{s}_3 + 3)$. Note that $\widetilde{s}_1$ and $\widetilde{s}_2$ repeat themselves every three steps. The states with the same "phase" are marked with the same color.

Bibliography

[1] O'Connor. C. and Miko. I. Developing the chromosome theory. *Nature Education*, 1(1):44, 2008.

[2] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301, April 2001.

[3] T. Cremer and M. Cremer. Chromosome territories. *Cold Spring Harbor Perspectives in Biology*, 2(3):a003889–a003889, February 2010.

[4] Donald E. Olins and Ada L. Olins. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology*, 4(10):809–814, October 2003.

[5] J. T. Finch and A. Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6):1897–1901, June 1976.

[6] Christopher L Woodcock. A milestone in the odyssey of higher-order chromatin structure. *Nature Structural & Molecular Biology*, 12(8):639–640, August 2005.

[7] C. Woodcock. The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *The Journal of Cell Biology*, 99(1):42–52, July 1984.

[8] Ken van Holde and Jordanka Zlatanova. Chromatin fiber structure: Where is the problem now? *Seminars in Cell & Developmental Biology*, 18(5):651–658, October 2007.

[9] Kazuhiro Maeshima, Saera Hihara, and Mikhail Eltsov. Chromatin structure: does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, 22(3):291–297, June 2010.

[10] M. Eltsov, K. M. MacLellan, K. Maeshima, A. S. Frangakis, and J. Dubochet. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences*, 105(50):19732–19737, December 2008.

[11] Horng D. Ou, Sébastien Phan, Thomas J. Deerinck, Andrea Thor, Mark H. Ellisman, and Clodagh C. O'Shea. ChromEMT: Visualizing 3d chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349):eaag0025, July 2017.

[12] Sigal Shachar, Ty C. Voss, Gianluca Pegoraro, Nicholas Sciascia, and Tom Misteli. Identification of gene positioning factors using high-throughput imaging mapping. *Cell*, 162(4):911–923, August 2015.

[13] L. Kearney. Multiplex-FISH (m-FISH): technique, developments and applications. *Cytogenetic and Genome Research*, 114(3-4):189–198, 2006.

[14] Siyuan Wang, Jun-Han Su, Brian J. Beliveau, Bogdan Bintu, Jeffrey R. Moffitt, Chao ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, July 2016.

[15] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645, September 2006.

[16] Samuel T. Hess, Thanu P.K. Girirajan, and Michael D. Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272, December 2006.

[17] Baohui Chen, Luke A. Gilbert, Beth A. Cimini, Joerg Schnitzbauer, Wei Zhang, Gene-Wei Li, Jason Park, Elizabeth H. Blackburn, Jonathan S. Weissman, Lei S. Qi, and Bo Huang. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/cas system. *Cell*, 155(7):1479–1491, December 2013.

[18] J. Dekker. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.

[19] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature Genetics*, 38(11):1348–1354, October 2006.

[20] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, October 2006.

[21] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom,

B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[22] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403, May 2013.

[23] Jyotsana J. Parmar, Maxime Woringer, and Christophe Zimmer. How the genome folds: The biophysics of four-dimensional chromatin organization. *Annual Review of Biophysics*, 48(1), March 2019.

[24] Quentin Szabo, Daniel Jost, Jia-Ming Chang, Diego I. Cattoni, Giorgio L. Papadopoulos, Boyan Bonev, Tom Sexton, Julian Gurgo, Caroline Jacquier, Marcelo Nollmann, Frédéric Bantignies, and Giacomo Cavalli. TADs are 3d structural units of higher-order chromosome organization in drosophila. *Science Advances*, 4(2):eaar8082, February 2018.

[25] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, April 2012.

[26] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, February 2012.

[27] Fulai Jin, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, October 2013.

[28] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

[29] Alistair N. Boettiger, Bogdan Bintu, Jeffrey R. Moffitt, Siyuan Wang, Brian J. Beliveau, Geoffrey Fudenberg, Maxim Imakaev, Leonid A. Mirny, Chao ting Wu, and Xiaowei Zhuang. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586):418–422, January 2016.

[30] Bogdan Bintu, Leslie J. Mateo, Jun-Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N. Boettiger, and Xiaowei

Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, October 2018.

[31] Guy Nir, Irene Farabella, Cynthia Pérez Estrada, Carl G. Ebeling, Brian J. Beliveau, Hiroshi M. Sasaki, S. Dean Lee, Son C. Nguyen, Ruth B. McCole, Shyamtanu Chattoraj, Jelena Erceg, Jumana AlHaj Abed, Nuno M. C. Martins, Huy Q. Nguyen, Mohammed A. Hannan, Sheikh Russell, Neva C. Durand, Suhas S. P. Rao, Jocelyn Y. Kishi, Paula Soler-Vila, Michele Di Pierro, José N. Onuchic, Steven P. Callahan, John M. Schreiner, Jeff A. Stuckey, Peng Yin, Erez Lieberman Aiden, Marc A. Marti-Renom, and C. ting Wu. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLOS Genetics*, 14(12):e1007872, December 2018.

[32] C. C. Robinett. In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *The Journal of Cell Biology*, 135(6):1685–1700, December 1996.

[33] Valeria Levi, QiaoQiao Ruan, Matthew Plutz, Andrew S. Belmont, and Enrico Gratton. Chromatin dynamics in interphase cells revealed by tracking in a two-photon excitation microscope. *Biophysical Journal*, 89(6):4275–4285, December 2005.

[34] I. Bronstein, Y. Israel, E. Kepten, S. Mai, Y. Shav-Tal, E. Barkai, and Y. Garini. Transient anomalous diffusion of telomeres in the nucleus of mammalian cells. *Physical Review Letters*, 103(1), July 2009.

[35] Stephanie C. Weber, Andrew J. Spakowitz, and Julie A. Theriot. Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. *Physical Review Letters*, 104(23), June 2010.

[36] Saera Hihara, Chan-Gi Pack, Kazunari Kaizu, Tomomi Tani, Tomo Hanafusa, Tadasu Nozaki, Satoko Takemoto, Tomohiko Yoshimi, Hideo Yokota, Naoko Imamoto, Yasushi Sako, Masataka Kinjo, Koichi Takahashi, Takeharu Nagai, and Kazuhiro Maeshima. Local nucleosome dynamics facilitate chromatin accessibility in living mammalian cells. *Cell Reports*, 2(6):1645–1656, December 2012.

[37] Houssam Hajjoul, Julien Mathon, Hubert Ranchon, Isabelle Goiffon, Julien Mozziconacci, Benjamin Albert, Pascal Carrivain, Jean-Marc Victor, Olivier Gadal, Kerstin Bystricky, and Aurélien Bancaud. High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome. *Genome Research*, 23(11):1829–1838, September 2013.

[38] Avelino Javer, Zhicheng Long, Eileen Nugent, Marco Grisi, Kamin Siriwatwetchakul, Kevin D. Dorfman, Pietro Cicuta, and Marco Cosentino Lagomarsino. Short-time movement of e. coli chromosomal loci depends on coordinate and subcellular localization. *Nature Communications*, 4(1), June 2013.

[39] Joseph S. Lucas, Yaojun Zhang, Olga K. Dudko, and Cornelis Murre. 3d trajectories adopted by coding and regulatory DNA elements: First-passage times for genomic interactions. *Cell*, 158(2):339–352, July 2014.

[40] I. Bronshtein, E. Kepten, I. Kanter, S. Berezin, M. Lindner, Abena B. Redwood, S Mai, S. Gonzalo, R. Foisner, Y. Shav-Tal, and Y. Garini. Loss of lamin a function increases chromatin dynamics in the nuclear interior. *Nature Communications*, 6(1), August 2015.

[41] A. Zidovska, D. A. Weitz, and T. J. Mitchison. Micron-scale coherence in interphase chromatin dynamics. *Proceedings of the National Academy of Sciences*, 110(39):15555–15560, September 2013.

[42] Soya Shinkai, Tadasu Nozaki, Kazuhiro Maeshima, and Yuichi Togashi. Dynamic nucleosome movement provides structural information of topological chromatin domains in living human cells. *PLOS Computational Biology*, 12(10):e1005136, October 2016.

[43] Tadasu Nozaki, Ryosuke Imai, Mai Tanbo, Ryosuke Nagashima, Sachiko Tamura, Tomomi Tani, Yasumasa Joti, Masaru Tomita, Kayo Hibino, Masato T. Kanemaki, Kerstin S. Wendt, Yasushi Okada, Takeharu Nagai, and Kazuhiro Maeshima. Dynamic organization of chromatin domains revealed by super-resolution live-cell imaging. *Molecular Cell*, 67(2):282–293.e7, July 2017.

[44] Haitham A Shaban, Roman Barth, and Kerstin Bystricky. Formation of correlated chromatin domains at nanoscale dynamic resolution during transcription. *Nucleic Acids Research*, 46(13):e77–e77, April 2018.

[45] Pierre-Gilles De Gennes and Pierre-Gilles Gennes. *Scaling concepts in polymer physics*. Cornell university press, 1979.

[46] Ger Van den Engh, Rainer Sachs, and Barbara J Trask. Estimating genomic distance from dna sequence location in cell nuclei by a random walk model. *Science*, 257(5075):1410–1412, 1992.

[47] P. Hahnfeldt, J. E. Hearst, D. J. Brenner, R. K. Sachs, and L. R. Hlatky. Polymer models for interphase chromosomes. *Proceedings of the National Academy of Sciences*, 90(16):7854–7858, August 1993.

[48] A Grosberg, Y Rabin, S Havlin, and A Neer. Crumpled globule model of the three-dimensional structure of dna. *EPL (Europhysics Letters)*, 23(5):373, 1993.

[49] R. K. Sachs, G. van den Engh, B. Trask, H. Yokota, and J. E. Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences*, 92(7):2710–2714, March 1995.

[50] John F Marko and Eric D Siggia. Polymer models of meiotic and mitotic chromosomes. *Molecular biology of the cell*, 8(11):2217–2231, 1997.

[51] Joseph Ostashevsky. A polymer model for the structural organization of chromatin loops and minibands in interphase chromosomes. *Molecular Biology of the Cell*, 9(11):3031–3040, November 1998.

[52] Angelo Rosa and Ralf Everaers. Structure and dynamics of interphase chromosomes. *PLoS Computational Biology*, 4(8):e1000153, August 2008.

[53] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heermann, R. van Driel, and S. Goetze. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences*, 106(10):3812–3817, February 2009.

[54] Angelo Rosa, Nils B. Becker, and Ralf Everaers. Looping probabilities in model interphase chromosomes. *Biophysical Journal*, 98(11):2410–2419, June 2010.

[55] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, 109(40):16173–16178, September 2012.

[56] Daniel Jost, Pascal Carrivain, Giacomo Cavalli, and Cédric Vaillant. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Research*, 42(15):9553–9561, August 2014.

[57] Bin Zhang and Peter G. Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067, April 2015.

[58] Hongsuk Kang, Young-Gui Yoon, D. Thirumalai, and Changbong Hyeon. Confinement-induced glassy dynamics in a model for chromosome organization. *Physical Review Letters*, 115(19), November 2015.

[59] M. V. Tamm, L. I. Nazarov, A. A. Gavrilov, and A. V. Chertovich. Anomalous diffusion in fractal globules. *Physical Review Letters*, 114(17), April 2015.

[60] Chris A. Brackley, James Johnson, Steven Kelly, Peter R. Cook, and Davide Marenduzzo. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Research*, 44(8):3503–3512, April 2016.

[61] Michele Di Pierro, Bin Zhang, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113(43):12168–12173, September 2016.

[62] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. Formation of chromosomal domains by loop extrusion. *Cell Reports*, 15(9):2038–2049, May 2016.

[63] D. Michieletto, E. Orlandini, and D. Marenduzzo. Polymer model with epigenetic recoloring reveals a pathway for the de novo establishment and 3d organization of chromatin domains. *Physical Review X*, 6(4), December 2016.

[64] Ofir Shukron and David Holcman. Transient chromatin properties revealed by polymer models and stochastic simulations constructed from chromosomal capture data. *PLOS Computational Biology*, 13(4):e1005469, April 2017.

[65] O. Shukron and D. Holcman. Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data. *Physical Review E*, 96(1), July 2017.

[66] Guang Shi, Lei Liu, Changbong Hyeon, and D. Thirumalai. Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature Communications*, 9(1), August 2018.

[67] Lei Liu, Guang Shi, D. Thirumalai, and Changbong Hyeon. Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. *PLOS Computational Biology*, 14(12):e1006617, December 2018.

[68] Johannes Nuebler, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A. Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706, July 2018.

[69] Johan H. Gibcus, Kumiko Samejima, Anton Goloborodko, Itaru Samejima, Natalia Naumova, Johannes Nuebler, Masato T. Kanemaki, Linfeng Xie, James R. Paulson, William C. Earnshaw, Leonid A. Mirny, and Job Dekker. A pathway for mitotic chromosome formation. *Science*, 359(6376):eaao6135, January 2018.

[70] Lei Liu, Min Hyeok Kim, and Changbong Hyeon. Heterogeneous loop model to infer 3d chromosome structures from hi-c. March 2019.

[71] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C. Anthony Blau, and William S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, May 2010.

[72] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98, December 2011.

[73] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC Bioinformatics*, 12(1):414, 2011.

[74] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of Computational Biology*, 20(11):831–846, November 2013.

[75] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Computational Biology*, 9(1):e1002893, January 2013.

[76] N. Varoquaux, F. Ay, W. S. Noble, and J.-P. Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, June 2014.

[77] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature Methods*, 11(11):1141–1143, September 2014.

[78] Harianto Tjong, Wenyuan Li, Reza Kalhor, Chao Dai, Shengli Hao, Ke Gong, Yonggang Zhou, Haochen Li, Xianghong Jasmine Zhou, Mark A. Le Gros, Carolyn A. Larabell, Lin Chen, and Frank Alber. Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, 113(12):E1663–E1672, March 2016.

[79] Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G. S. Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D. Laue. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59–64, March 2017.

[80] Nan Hua, Harianto Tjong, Hanjun Shin, Ke Gong, Xianghong Jasmine Zhou, and Frank Alber. Producing genome structure populations with the dynamic and automated PGS software. *Nature Protocols*, 13(5):915–926, April 2018.

[81] A. Kurz. Active and inactive genes localize preferentially in the periphery of chromosome territories. *The Journal of Cell Biology*, 135(5):1195–1205, December 1996.

[82] C. M. Clemson, L. L. Hall, M. Byron, J. McNeil, and J. B. Lawrence. The x chromosome is organized into a gene-rich outer rim and an internal core containing silenced nongenic sequences. *Proceedings of the National Academy of Sciences*, 103(20):7688–7693, May 2006.

[83] M. Jordan Rowley and Victor G. Corces. Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, 19(12):789–800, October 2018.

[84] Adriana Miele and Job Dekker. Long-range chromosomal interactions and gene regulation. *Molecular BioSystems*, 4(11):1046, 2008.

[85] Ivan Krivega and Ann Dean. Enhancer and promoter interactions—long distance calls. *Current Opinion in Genetics & Development*, 22(2):79–85, April 2012.

[86] Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, September 2012.

[87] E. Alipour and J. F. Marko. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research*, 40(22):11202–11212, October 2012.

[88] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, October 2015.

[89] Suhas S.P. Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, Sarah E. Johnstone, Gavin D. Bascom, Ivan D. Bochkov, Xingfan Huang, Muhammad S. Shamim, Jaeweon Shin, Douglass Turner, Ziyi Ye, Arina D. Omer, James T. Robinson, Tamar Schlick, Bradley E. Bernstein, Rafael Casellas, Eric S. Lander, and Erez Lieberman Aiden. Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320.e24, October 2017.

[90] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H. Haering, Leonid Mirny, and Francois Spitz. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51–56, September 2017.

[91] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo. Extrusion without a motor: a new take on the loop extrusion model of genome organization. *Nucleus*, 9(1):95–103, January 2018.

[92] Tsuyoshi Terakawa, Shveta Bisht, Jorine M. Eeftens, Cees Dekker, Christian H. Haering, and Eric C. Greene. The condensin complex

is a mechanochemical motor that translocates along DNA. *Science*, 358(6363):672–676, September 2017.

[93] Jorine M Eeftens, Shveta Bisht, Jacob Kerssemakers, Marc Kschonsak, Christian H Haering, and Cees Dekker. Real-time detection of condensin-driven DNA compaction reveals a multistep binding mechanism. *The EMBO Journal*, 36(23):3448–3457, November 2017.

[94] Mahipal Ganji, Indra A. Shaltiel, Shveta Bisht, Eugene Kim, Ana Kalichava, Christian H. Haering, and Cees Dekker. Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384):102–105, February 2018.

[95] Prakash Arumugam, Stephan Gruber, Koichi Tanaka, Christian H. Haering, Karl Mechtler, and Kim Nasmyth. ATP hydrolysis is required for cohesins association with chromosomes. *Current Biology*, 13(22):1941–1953, November 2003.

[96] Stefan Weitzer, Chris Lehane, and Frank Uhlmann. A model for ATP hydrolysis-dependent binding of cohesin to DNA. *Current Biology*, 13(22):1930–1940, November 2003.

[97] Johannes Stigler, Gamze Ö. Çamdere, Douglas E. Koshland, and Eric C. Greene. Single-molecule imaging reveals a collapsed conformational state for DNA-bound cohesin. *Cell Reports*, 15(5):988–998, May 2016.

[98] Iain F Davidson, Daniela Goetz, Maciej P Zaczek, Maxim I Molodtsov, Pim J Huis in t Veld, Florian Weissmann, Gabriele Litos, David A Cisneros, Maria Ocampo-Hafalla, Rene Ladurner, Frank Uhlmann, Alipasha Vaziri, and Jan-Michael Peters. Rapid movement and transcriptional re-localization of human cohesin on DNA. *The EMBO Journal*, 35(24):2671–2685, October 2016.

[99] Michael H. Kagey, Jamie J. Newman, Steve Bilodeau, Ye Zhan, David A. Orlando, Nynke L. van Berkum, Christopher C. Ebmeier, Jesse Goossens, Peter B. Rahl, Stuart S. Levine, Dylan J. Taatjes, Job Dekker, and Richard A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, August 2010.

[100] Matthias Merkenschlager and Elphège P. Nora. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annual Review of Genomics and Human Genetics*, 17(1):17–43, August 2016.

[101] William A. Flavahan, Yotam Drier, Brian B. Liau, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvà, and Bradley E. Bernstein. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110–114, December 2015.

[102] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata

Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, May 2015.

[103] Dusan Racko, Fabrizio Benedetti, Julien Dorier, and Andrzej Stasiak. Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Research*, 46(4):1648–1660, November 2017.

[104] Michele Di Pierro, Ryan R. Cheng, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, 114(46):12126–12131, October 2017.

[105] Y. Cui and C. Bustamante. Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proceedings of the National Academy of Sciences*, 97(1):127–132, January 2000.

[106] Jonas J. Funke, Philip Ketterer, Corinna Lieleg, Sarah Schunter, Philipp Korber, and Hendrik Dietz. Uncovering the forces between nucleosomes using DNA origami. *Science Advances*, 2(11):e1600974, November 2016.

[107] Yuta Shimamoto, Sachiko Tamura, Hiroshi Masumoto, and Kazuhiro Maeshima. Nucleosome–nucleosome interactions via histone tails and linker DNA regulate nuclear rigidity. *Molecular Biology of the Cell*, 28(11):1580–1589, June 2017.

[108] Melanie A. Adams-Cioaba and Jinrong Min. Structure and function of histone methylation binding proteinsThis paper is one of a selection of papers published in this special issue, entitled CSBMCB's 51st annual meeting – epigenetics and chromatin dynamics, and has undergone the journal's usual peer review process. *Biochemistry and Cell Biology*, 87(1):93–105, February 2009.

[109] Adam G. Larson, Daniel Elnatan, Madeline M. Keenen, Michael J. Trnka, Jonathan B. Johnston, Alma L. Burlingame, David A. Agard, Sy Redding, and Geeta J. Narlikar. Liquid droplet formation by HP1 suggests a role for phase separation in heterochromatin. *Nature*, 547(7662):236–240, June 2017.

[110] Amy R. Strom, Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen. Phase separation drives heterochromatin domain formation. *Nature*, 547(7662):241–245, June 2017.

[111] Aaron J Plys, Christopher P Davis, Jongmin Kim, Gizem Rizki, Madeline M Keenen, Sharon K Marr, and Robert E Kingston. Phase separation and nucleosome compaction are governed by the same domain of polycomb repressive complex 1. November 2018.

[112] Won-Ki Cho, Jan-Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I. Cisse. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415, June 2018.

[113] Benjamin R. Sabari, Alessandra Dall'Agnese, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, Nancy M. Hannett, Alicia V. Zamudio, John C. Manteiga, Charles H. Li, Yang E. Guo, Daniel S. Day, Jurian Schuijers, Eliza Vasile, Sohail Malik, Denes Hnisz, Tong Ihn Lee, Ibrahim I. Cisse, Robert G. Roeder, Phillip A. Sharp, Arup K. Chakraborty, and Richard A. Young. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400):eaar3958, June 2018.

[114] Aaron J. Plys and Robert E. Kingston. Dynamic condensates activate transcription. *Science*, 361(6400):329–330, July 2018.

[115] Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay, and Suzana Hadjur. Comparative hi-c reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, 10(8):1297–1309, March 2015.

[116] Ilya M. Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, March 2017.

[117] Longzhi Tan, Dong Xing, Nicholas Daley, and X. Sunney Xie. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nature Structural & Molecular Biology*, 26(4):297–307, April 2019.

[118] Elizabeth H. Finn, Gianluca Pegoraro, Hugo B. Brandão, Anne-Laure Valton, Marlies E. Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515.e10, March 2019.

[119] Steven J. Altschuler and Lani F. Wu. Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559–563, May 2010.

[120] Jeffrey T. Leek and John D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[121] Corbin E. Meacham and Sean J. Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, September 2013.

[122] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos

Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, July 2017.

[123] Masao Doi and Samuel Frederick Edwards. *The theory of polymer dynamics*, volume 73. oxford university press, 1988.

[124] G Jannink and J des Cloizeaux. Polymers in solution. *Journal of Physics: Condensed Matter*, 2(1):1–24, January 1990.

[125] S Redner. Distribution functions in the interior of polymer chains. *Journal of Physics A: Mathematical and General*, 13(11):3525–3541, November 1980.

[126] Ngo Minh Toan, Greg Morrison, Changbong Hyeon, and D. Thirumalai. Kinetics of loop formation in polymer chains†. *The Journal of Physical Chemistry B*, 112(19):6094–6106, May 2008.

[127] W.F Marshall, A Straight, J.F Marko, J Swedlow, A Dernburg, A Belmont, A.W Murray, D.A Agard, and J.W Sedat. Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology*, 7(12):930–939, December 1997.

[128] Malte Wachsmuth, Waldemar Waldeck, and Jörg Langowski. Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy. *Journal of Molecular Biology*, 298(4):677–689, May 2000.

[129] Hongtao Chen, Michal Levo, Lev Barinov, Miki Fujioka, James B. Jaynes, and Thomas Gregor. Dynamic interplay between enhancer–promoter topology and gene activity. *Nature Genetics*, 50(9):1296–1303, July 2018.

[130] Bo Gu, Tomek Swigut, Andrew Spencley, Matthew R. Bauer, Mingyu Chung, Tobias Meyer, and Joanna Wysocka. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science*, 359(6379):1050–1055, January 2018.

[131] David Saintillan, Michael J. Shelley, and Alexandra Zidovska. Extensile motor activity drives coherent motions in a model of interphase chromatin. *Proceedings of the National Academy of Sciences*, 115(45):11442–11447, October 2018.

[132] Thomas J. Lampo, Stella Stylianidou, Mikael P. Backlund, Paul A. Wiggins, and Andrew J. Spakowitz. Cytoplasmic RNA-protein particles exhibit non-gaussian subdiffusive behavior. *Biophysical Journal*, 112(3):532–542, February 2017.

[133] Asmaa A. Sadoon and Yong Wang. Anomalous, non-gaussian, viscoelastic, and age-dependent dynamics of histonelike nucleoid-structuring proteins in live escherichia coli. *Physical Review E*, 98(4), October 2018.

[134] Ludovic Berthier and Giulio Biroli. Theoretical perspective on the glass transition and amorphous materials. *Reviews of Modern Physics*, 83(2):587–645, June 2011.

[135] T. R. Kirkpatrick and D. Thirumalai. Colloquium: Random first order transition theory concepts in biology and physics. *Reviews of Modern Physics*, 87(1):183–209, March 2015.

[136] Ludovic Berthier, Elijah Flenner, and Grzegorz Szamel. Perspective: Nonequilibrium glassy dynamics in dense systems of active particles. *arXiv preprint arXiv:1902.08580*, 2019.

[137] Valentino Bianco, Emanuele Locatelli, and Paolo Malgaretti. Globulelike conformation and enhanced diffusion of active polymers. *Physical Review Letters*, 121(21), November 2018.

[138] S. C. Weber, A. J. Spakowitz, and J. A. Theriot. Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. *Proceedings of the National Academy of Sciences*, 109(19):7338–7343, April 2012.

[139] Jan Smrek and Kurt Kremer. Small activity differences drive phase separation in active-passive polymer mixtures. *Physical Review Letters*, 118(9), March 2017.

[140] Michael E. Cates and Julien Tailleur. Motility-induced phase separation. *Annual Review of Condensed Matter Physics*, 6(1):219–244, March 2015.

[141] Joakim Stenhammar, Raphael Wittkowski, Davide Marenduzzo, and Michael E. Cates. Activity-induced phase separation and self-assembly in mixtures of active and passive particles. *Physical Review Letters*, 114(1), January 2015.

[142] Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, January 2011.

[143] Hua Wong, Hervé Marie-Nelly, Sébastien Herbert, Pascal Carrivain, Hervé Blanc, Romain Koszul, Emmanuelle Fabre, and Christophe Zimmer. A predictive computational model of the dynamic 3d interphase yeast nucleus. *Current Biology*, 22(20):1881–1890, October 2012.

[144] Jean-Michel Arbona, Sébastien Herbert, Emmanuelle Fabre, and Christophe Zimmer. Inferring the physical properties of yeast chromatin through bayesian analysis of whole nucleus simulations. *Genome Biology*, 18(1), May 2017.

[145] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4):eaaw1668, April 2019.

[146] Geoffrey Fudenberg, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A. Mirny. Emerging evidence of chromosome folding by loop extrusion. *Cold Spring Harbor Symposia on Quantitative Biology*, 82:45–55, 2017.

[147] Tetsuya Yamamoto and Helmut Schiessel. Osmotic mechanism of the loop extrusion process. *Physical Review E*, 96(3), September 2017.

[148] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo. Nonequilibrium chromosome looping via molecular slip links. *Physical Review Letters*, 119(13), September 2017.

[149] Anton Goloborodko, John F. Marko, and Leonid A. Mirny. Chromosome compaction by active loop extrusion. *Biophysical Journal*, 110(10):2162–2168, May 2016.

[150] Anton Goloborodko, Maxim V Imakaev, John F Marko, and Leonid Mirny. Compaction and segregation of sister chromatids via active loop extrusion. *eLife*, 5, May 2016.

[151] Sumitabha Brahmachari and John F. Marko. Chromosome disentanglement driven via optimal compaction of loop-extruded brush structures. April 2019.

[152] John F Marko, Paolo De Los Rios, Alessandro Barducci, and Stephan Gruber. DNA-segment-capture model for loop extrusion by structural maintenance of chromosome (SMC) protein complexes. May 2018.

[153] Michael H. Nichols and Victor G. Corces. A tethered-inchworm model of SMC DNA translocation. *Nature Structural & Molecular Biology*, 25(10):906–910, September 2018.

[154] Daniele Canzio, Maofu Liao, Nariman Naber, Edward Pate, Adam Larson, Shenping Wu, Diana B. Marina, Jennifer F. Garcia, Hiten D. Madhani, Roger Cooke, Peter Schuck, Yifan Cheng, and Geeta J. Narlikar. A conformational switch in HP1 releases auto-inhibition to drive heterochromatin assembly. *Nature*, 496(7445):377–381, March 2013.

[155] Peter J Mulligan, Elena F Koslover, and Andrew J Spakowitz. Thermodynamic model of heterochromatin formation through epigenetic regulation. *Journal of Physics: Condensed Matter*, 27(6):064109, January 2015.

[156] Kensal E. van Holde. *Chromatin*. Springer New York, 1989.

[157] Kate R. Rosenbloom, Cricket A. Sloan, Venkat S. Malladi, Timothy R. Dreszer, Katrina Learned, Vanessa M. Kirkup, Matthew C. Wong, Morgan Maddren, Ruihua Fang, Steven G. Heitner, Brian T. Lee, Galt P. Barber, Rachel A. Harte, Mark Diekhans, Jeffrey C. Long, Steven P. Wilder, Ann S. Zweig, Donna Karolchik, Robert M. Kuhn, David Haussler, and W. James

Kent. ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Research*, 41(D1):D56–D63, November 2012.

[158] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825, July 2010.

[159] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, March 2011.

[160] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, March 1995.

[161] J. D. Honeycutt and D. Thirumalai. The nature of folded states of globular proteins. *Biopolymers*, 32(6):695–709, June 1992.

[162] Waltraud G. Müller, Dawn Walker, Gordon L. Hager, and James G. McNally. Large-scale chromatin decondensation and recondensation regulated by transcription from a natural promoter. *The Journal of Cell Biology*, 154(1):33–48, July 2001.

[163] Noriko Sato, Masahito Nakayama, and Ken ichi Arai. Fluctuation of chromatin unfolding associated with variation in the level of gene expression. *Genes to Cells*, 9(7):619–630, July 2004.

[164] S. Chambeyron. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development*, 18(10):1119–1130, May 2004.

[165] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, April 2012.

[166] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 01*. ACM Press, 2001.

[167] Jonathan D Halverson, Jan Smrek, Kurt Kremer, and Alexander Y Grosberg. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Reports on Progress in Physics*, 77(2):022601, January 2014.

[168] Geoffrey Fudenberg and Maxim Imakaev. FISH-ing for captured contacts: towards reconciling FISH and 3c. *Nature Methods*, 14(7):673–678, June 2017.

[169] Jonathan D. Halverson, Won Bo Lee, Gary S. Grest, Alexander Y. Grosberg, and Kurt Kremer. Molecular dynamics simulation study of nonconcatenated ring polymers in a melt. i. statics. *The Journal of Chemical Physics*, 134(20):204904, May 2011.

[170] Luca Giorgetti, Rafael Galupa, Elphège P. Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, May 2014.

[171] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, September 2013.

[172] Noelle Haddad, Cédric Vaillant, and Daniel Jost. IC-finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Research*, page gkx036, January 2017.

[173] T. R. Kirkpatrick and D. Thirumalai. Comparison between dynamical theories and metastable states in regular and glassy mean-field spin models with underlying first-order-like phase transitions. *Physical Review A*, 37(11):4439–4448, June 1988.

[174] Wolfgang Götze. *Complex dynamics of glass-forming liquids: A mode-coupling theory*, volume 143. OUP Oxford, 2008.

[175] Assaf Amitai, Andrew Seeber, Susan M. Gasser, and David Holcman. Visualization of chromatin decompaction and break site extrusion as predicted by statistical polymer modeling of single-locus trajectories. *Cell Reports*, 18(5):1200–1214, January 2017.

[176] D. Thirumalai and Raymond D. Mountain. Activated dynamics, loss of ergodicity, and transport in supercooled liquids. *Physical Review E*, 47(1):479–489, January 1993.

[177] Pinaki Chaudhuri, Ludovic Berthier, and Walter Kob. Universal nature of particle displacements close to glass and jamming transitions. *Physical Review Letters*, 99(6), August 2007.

[178] S Dietzel. The 3d positioning of ANT2 and ANT3 genes within female x chromosome territories correlates with gene activity. *Experimental Cell Research*, 252(2):363–375, November 1999.

[179] I.M. Lifshitz and V.V. Slyozov. The kinetics of precipitation from supersaturated solid solutions. *Journal of Physics and Chemistry of Solids*, 19(1-2):35–50, April 1961.

[180] J. D. Bryngelson and D. Thirumalai. Internal constraints induce localization in an isolated polymer molecule. *Physical Review Letters*, 76(3):542–545, January 1996.

[181] Hanhui Ma, Ardalan Naseri, Pablo Reyes-Gutierrez, Scot A. Wolfe, Shaojie Zhang, and Thoru Pederson. Multicolor CRISPR labeling of chromosomal loci in human cells. *Proceedings of the National Academy of Sciences*, 112(10):3002–3007, February 2015.

[182] Hanhui Ma, Li-Chun Tu, Ardalan Naseri, Maximiliaan Huisman, Shaojie Zhang, David Grunwald, and Thoru Pederson. Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nature Biotechnology*, 34(5):528–530, April 2016.

[183] Luca Giorgetti and Edith Heard. Closing the loop: 3c versus DNA FISH. *Genome Biology*, 17(1), October 2016.

[184] James Fraser, Iain Williamson, Wendy A. Bickmore, and Josée Dostie. An overview of genome organization and how we got there: from FISH to hi-c. *Microbiology and Molecular Biology Reviews*, 79(3):347–372, July 2015.

[185] Wendy A. Bickmore and Bas van Steensel. Genome architecture: Domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284, March 2013.

[186] Iain Williamson, Soizik Berlivet, Ragnhild Eskeland, Shelagh Boyle, Robert S. Illingworth, Denis Paquette, Josée Dostie, and Wendy A. Bickmore. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & Development*, 28(24):2778–2791, December 2014.

[187] C. Hyeon, G. Morrison, and D. Thirumalai. Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proceedings of the National Academy of Sciences*, 105(28):9604–9609, July 2008.

[188] Edward P. O'Brien, Greg Morrison, Bernard R. Brooks, and D. Thirumalai. How accurate are polymer models in the analysis of förster resonance energy transfer experiments on proteins? *The Journal of Chemical Physics*, 130(12):124903, March 2009.

[189] J. des Cloizeaux. Short range correlation between elements of a long polymer in a good solvent. *Journal de Physique*, 41(3):223–238, 1980.

[190] Changbong Hyeon and D. Thirumalai. Kinetics of interior loop formation in semiflexible chains. *The Journal of Chemical Physics*, 124(10):104905, March 2006.

[191] Jan Wilhelm and Erwin Frey. Radial distribution function of semiflexible polymers. *Physical Review Letters*, 77(12):2581–2584, September 1996.

[192] `https://data.4dnucleome.org/publications/80007b23-7748-4492-9e49-c38400acbe60/`.

[193] Anders S Hansen, Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife*, 6, May 2017.

[194] Anders S. Hansen, Claudia Cattoglio, Xavier Darzacq, and Robert Tjian. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*, 9(1):20–32, December 2017.

[195] Sofia A. Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, Alexander A. Shishkin, Prashant Bhat, Yodai Takei, Vickie Trinh, Erik Aznauryan, Pamela Russell, Christine Cheng, Marko Jovanovic, Amy Chow, Long Cai, Patrick McDonel, Manuel Garber, and Mitchell Guttman. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. *Cell*, 174(3):744–757.e24, July 2018.

[196] C. J. Camacho and D. Thirumalai. Theoretical predictions of folding pathways by using the proximity rule, with applications to bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences*, 92(5):1277–1281, February 1995.

[197] Adam Buckle, Chris A. Brackley, Shelagh Boyle, Davide Marenduzzo, and Nick Gilbert. Polymer simulations of heteromorphic chromatin predict the 3d folding of complex genomic loci. *Molecular Cell*, 72(4):786–797.e11, November 2018.

[198] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, March 1964.

[199] *Modern Multidimensional Scaling*. Springer New York, 2005.

[200] Jonas Paulsen, Odin Gramstad, and Philippe Collas. Manifold based optimization for single-cell 3d genome reconstruction. *PLOS Computational Biology*, 11(8):e1004396, August 2015.

[201] Diego I. Cattoni, Andrés M. Cardozo Gizzi, Mariya Georgieva, Marco Di Stefano, Alessandro Valeri, Delphine Chamousset, Christophe Houbron, Stephanie Déjardin, Jean-Bernard Fiche, Inma González, Jia-Ming Chang, Thomas Sexton, Marc A. Marti-Renom, Frédéric Bantignies, Giacomo Cavalli, and Marcelo Nollmann. Single-cell absolute contact probability detection

reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nature Communications*, 8(1), November 2017.

[202] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, 4(5):e138, April 2006.

[203] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[204] Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391, 2009.

[205] Steven M. Block, Lawrence S. B. Goldstein, and Bruce J. Schnapp. Bead movement by single kinesin molecules studied with optical tweezers. *Nature*, 348(6299):348–352, November 1990.

[206] Koen Visscher, Mark J. Schnitzer, and Steven M. Block. Single kinesin molecules studied with a molecular force clamp. *Nature*, 400(6740):184–189, July 1999.

[207] S. M. Block, C. L. Asbury, J. W. Shaevitz, and M. J. Lang. Probing the kinesin reaction cycle with a 2d optical force clamp. *Proceedings of the National Academy of Sciences*, 100(5):2351–2356, February 2003.

[208] N. J. Carter and R. A. Cross. Mechanics of the kinesin step. *Nature*, 435(7040):308–312, May 2005.

[209] Roop Mallik, Brian C. Carter, Stephanie A. Lex, Stephen J. King, and Steven P. Gross. Cytoplasmic dynein functions as a gear in response to load. *Nature*, 427(6975):649–652, February 2004.

[210] Claudia Veigel and Christoph F. Schmidt. Moving into the cell: single-molecule studies of molecular motors in complex environments. *Nature Reviews Molecular Cell Biology*, 12(3):163–176, February 2011.

[211] Anatoly B. Kolomeisky and Michael E. Fisher. Molecular motors: A theorists perspective. *Annual Review of Physical Chemistry*, 58(1):675–695, May 2007.

[212] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, November 2012.

[213] C. Leduc, O. Campas, K. B. Zeldovich, A. Roux, P. Jolimaitre, L. Bourel-Bonnet, B. Goud, J.-F. Joanny, P. Bassereau, and J. Prost. Cooperative extraction of membrane nanotubes by molecular motors. *Proceedings of the National Academy of Sciences*, 101(49):17096–17101, November 2004.

[214] C. Kural. Kinesin and dynein move a peroxisome in vivo: A tug-of-war or coordinated movement? *Science*, 308(5727):1469–1472, June 2005.

[215] Steven P. Gross, Michael Vershinin, and George T. Shubeita. Cargo transport: Two motors are sometimes better than one. *Current Biology*, 17(12):R478–R486, June 2007.

[216] N. D. Derr, B. S. Goodman, R. Jungmann, A. E. Leschziner, W. M. Shih, and S. L. Reck-Peterson. Tug-of-war in motor protein ensembles revealed with a programmable DNA origami scaffold. *Science*, 338(6107):662–665, October 2012.

[217] M. Vershinin, B. C. Carter, D. S. Razafsky, S. J. King, and S. P. Gross. Multiple-motor based transport and its regulation by tau. *Proceedings of the National Academy of Sciences*, 104(1):87–92, December 2006.

[218] K. Furuta, A. Furuta, Y. Y. Toyoshima, M. Amino, K. Oiwa, and H. Kojima. Measuring collective transport by defined numbers of processive and nonprocessive kinesin motors. *Proceedings of the National Academy of Sciences*, 110(2):501–506, December 2012.

[219] L. Conway, D. Wood, E. Tuzel, and J. L. Ross. Motor transport of self-assembled cargos in crowded environments. *Proceedings of the National Academy of Sciences*, 109(51):20814–20819, December 2012.

[220] R. F. Hariadi, M. Cale, and S. Sivaramakrishnan. Myosin lever arm directs collective motion on cellular actin network. *Proceedings of the National Academy of Sciences*, 111(11):4091–4096, March 2014.

[221] S. Klumpp and R. Lipowsky. Cooperative cargo transport by several molecular motors. *Proceedings of the National Academy of Sciences*, 102(48):17284–17289, November 2005.

[222] George T. Shubeita, Susan L. Tran, Jing Xu, Michael Vershinin, Silvia Cermelli, Sean L. Cotton, Michael A. Welte, and Steven P. Gross. Consequences of motor copy number on the intracellular transport of kinesin-1-driven lipid droplets. *Cell*, 135(6):1098–1107, December 2008.

[223] S. R. Nelson, K. M. Trybus, and D. M. Warshaw. Motor coupling through lipid membranes enhances transport velocities for ensembles of myosin va. *Proceedings of the National Academy of Sciences*, 111(38):E3986–E3995, September 2014.

[224] Katerina Toropova, Miroslav Mladenov, and Anthony J Roberts. Intraflagellar transport dynein is autoinhibited by trapping of its mechanical and track-binding elements. *Nature Structural & Molecular Biology*, 24(5):461–468, April 2017.

[225] C. Leduc, F. Ruhnow, J. Howard, and S. Diez. Detection of fractional steps in cargo movement by the collective operation of kinesin-1 motors. *Proceedings of the National Academy of Sciences*, 104(26):10847–10852, June 2007.

[226] R. F. Hariadi, R. F. Sommese, A. S. Adhikari, R. E. Taylor, S. Sutton, J. A. Spudich, and S. Sivaramakrishnan. Mechanical coordination in motor ensembles revealed using engineered artificial myosin filaments. *Nature Nanotechnology*, 10(8):696–700, July 2015.

[227] Elena B. Krementsova, Kenya Furuta, Kazuhiro Oiwa, Kathleen M. Trybus, and M. Yusuf Ali. Small teams of myosin vc motors coordinate their stepping for efficient cargo transport on actin bundles. *Journal of Biological Chemistry*, 292(26):10998–11008, May 2017.

[228] Arthur R. Rogers, Jonathan W. Driver, Pamela E. Constantinou, D. Kenneth Jamison, and Michael R. Diehl. Negative interference dominates collective transport of kinesin motors in the absence of load. *Physical Chemistry Chemical Physics*, 11(24):4882, 2009.

[229] M. Yusuf Ali, Andrej Vilfan, Kathleen M. Trybus, and David M. Warshaw. Cargo transport by two coupled myosin va motors on actin filaments and bundles. *Biophysical Journal*, 111(10):2228–2240, November 2016.

[230] M. J. I. Muller, S. Klumpp, and R. Lipowsky. Tug-of-war as a cooperative mechanism for bidirectional cargo transport by molecular motors. *Proceedings of the National Academy of Sciences*, 105(12):4609–4614, March 2008.

[231] Florian Berger, Corina Keller, Stefan Klumpp, and Reinhard Lipowsky. External forces influence the elastic coupling effects during cargo transport by molecular motors. *Physical Review E*, 91(2), February 2015.

[232] Frank Jülicher and Jacques Prost. Cooperative molecular motors. *Physical Review Letters*, 75(13):2618–2621, September 1995.

[233] Akito Igarashi, Shinji Tsukamoto, and Hiromichi Goko. Transport properties and efficiency of elastically coupled brownian motors. *Physical Review E*, 64(5), October 2001.

[234] O. Campàs, Y. Kafri, K. B. Zeldovich, J. Casademunt, and J.-F. Joanny. Collective dynamics of interacting molecular motors. *Physical Review Letters*, 97(3), July 2006.

[235] Huong T. Vu, Shaon Chakrabarti, Michael Hinczewski, and D. Thirumalai. Discrete step sizes of molecular motors lead to bimodal non-gaussian velocity distributions under force. *Physical Review Letters*, 117(7), August 2016.

[236] M. E. Fisher and A. B. Kolomeisky. The force exerted by a molecular motor. *Proceedings of the National Academy of Sciences*, 96(12):6597–6602, June 1999.

[237] Bernard Derrida. Velocity and diffusion constant of a periodic one-dimensional hopping model. *Journal of Statistical Physics*, 31(3):433–450, June 1983.

[238] Hong Qian. Nonequilibrium steady-state circulation and heat dissipation functional. *Physical Review E*, 64(2), July 2001.

[239] Johan OL Andreasson, Bojan Milic, Geng-Yuan Chen, Nicholas R Guydosh, William O Hancock, and Steven M Block. Examining kinesin processivity within a general gating framework. *Elife*, 4:e07403, 2015.

[240] Jerome Irianto, Yuntao Xia, Charlotte R. Pfeifer, Roger A. Greenberg, and Dennis E. Discher. As a nucleus enters a small pore, chromatin stretches and maintains integrity, even with DNA breaks. *Biophysical Journal*, 112(3):446–449, February 2017.

[241] J. D. Pajerowski, K. N. Dahl, F. L. Zhong, P. J. Sammak, and D. E. Discher. Physical plasticity of the nucleus in stem cell differentiation. *Proceedings of the National Academy of Sciences*, 104(40):15619–15624, September 2007.

[242] T. G. Mason, Hu Gang, and D. A. Weitz. Diffusing-wave-spectroscopy measurements of viscoelasticity of complex fluids. *Journal of the Optical Society of America A*, 14(1):139, January 1997.

[243] Qingzhou Feng, Keith J. Mickolajczyk, Geng-Yuan Chen, and William O. Hancock. Motor reattachment kinetics play a dominant role in multimotor-driven cargo transport. *Biophysical Journal*, 114(2):400–409, January 2018.

[244] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[245] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC 98*. ACM Press, 1998.

[246] Moo K. Chung, Hyekyoung Lee, Victor Solo, Richard J. Davidson, and Seth D. Pollak. Topological distances between brain networks. In *Connectomics in NeuroImaging*, pages 161–170. Springer International Publishing, 2017.

[247] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. 2011. Preprint at https://arxiv.org/abs/1109.2378.

[248] J.A. Aronovitz and D.R. Nelson. Universal features of polymer shapes. *J. Phys.*, 47(9):1445–1456, 1986.

[249] Ruxandra I. Dima and D. Thirumalai. Asymmetry in the Shapes of Folded and Denatured States of Proteins. *J. Phys. Chem. B*, 108(21):6564–6570, may 2004.

[250] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. Siam, 1995.

[251] P. C. Hansen. The l-curve and its use in the numerical treatment of inverse problems. In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

[252] Isaac J Schoenberg. Remarks to maurice frechet's article"sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert. *Annals of Mathematics*, pages 724–732, 1935.

[253] John Clifford Gower. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.