Modeling and analysis of randomly cross-linked polymers and application to chromatin organization and dynamics

Ofir Shukron

THÉSE DE DOCTORAT
de l'Université de recherche
Paris Sciences Lettres –
PSL Research University

Préparée à

l'École normale supérieure

Modélisation et analyse de modèles de polymères aléatoirement réticulé et application à l'organisation et à la dynamique de la chromatine

Modeling and Analysis of Randomly Cross-Linked Polymers and Application To Chromatin Organization and Dynamics

École doctorale n°386
Spécialité: Mathématiques appliquées

Composition du Jury :

M Andrew Spakowitz
University of Stanford, Rapporteur

M Jean-Marc Victor
UPMC,Rapporteur

M David Holcman
ENS, Directeur de thèse

M Bernard Derrida
ENS, Membre du Jury

Mme Angela Taddei
Institute Curie, Membre du Jury

Mme Alessandra Carbone
UPMC, Présidente du Jury

M Marco Cosentino Lagomarsino
UPMC, Membre du Jury

M Jean-Phillippe Vert
ENS, Mine ParisTech, Membre du Jury

par Ofir Shukron

Dirigée par M David Holcman

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Acknowledgments

First, I would like to thank my family and my wife, Yonit. Their patience, love, and support are my true source of inspiration. I would also like to thank my lab mates: Jürgen Reingruber, Jérôme Cartailler, Pierre Parutto, Kanishka Basnayake, and Lou Zonca. Last but not least, I would like to thank my thesis advisor, David Holcman, for his support and direction throughout my years in the lab.

Resumé

Dans cette thèse nous étudions la relation entre la conformation et la dynamique de la chromatine en nous basant sur une classe de modèles de polymères aléatoirement réticulé (AR). Les modèles AR permettent de prendre en compte la variabilité de la conformation de la chromatine sur l'ensemble d'une population de cellules. Nous utilisons les outils tels que les statistiques, les processus stochastiques, les simulations numériques ainsi que la physique des polymères afin de déduire certaines propriétés des polymères AR a l'équilibre ainsi que pour des cas transitoires. Nous utilisons par la suite ces propriétés afin d'élucider l'organisation dynamique de la chromatine pour diverses échelles et conditions biologiques.

Au chapitre trois de ce travail, nous développons une méthode générale pour construire les polymères AR directement à partir des données expérimentales, c'est-à-dire des données de capture chromosomiques (CC). Nous montrons que des connections longue portées persistantes entre des domaines topologiquement associés (DTA) affectent le temps de rencontre transitoire entre les DTA dans le processus d'inactivation du chromosome X. Nous montrons de plus que la variabilité des exposants anormaux – mesurée en trajectoires de particules individuelles (TPI) – est une conséquence directe de l'hétérogénéité dans la position des réticulations.

Au chapitre quatre, nous utilisons les polymères AR afin d'étudier la réorganisation locale du génome au point de cassure des deux branches d'ADN (CDB). Le nombre de connecteurs dans le modèle de polymère AR est calibré à partir de TPI, mesurées avant et après la CDB. Nous avons trouvé que la perte modérée de connecteur autour des sites de la CDB affecte de façon significative le premier temps de rencontre des deux extrémités cassées lors du processus de réparation d'une CBD. Nous montrons comment un micro-environnement génomique réticulé peut confiner les extrémités d'une cassure, empêchant ainsi les deux brins de dériver l'un de l'autre.

Au chapitre cinq, nous déduisons une expression analytique des propriétés transitoires et a l' équilibre du modèle de polymère AR, représentant une unique région DTA. Les expressions ainsi obtenue sont ensuite utilisées afin d'extraire le nombre moyen de connexions dans les DTA provenant des données de CC, et ce à l'aide d'une simple procédure d'ajustement de courbe. Nous dérivons par la suite la formule pour le temps moyen de première rencontre (TMPR) entre deux monomères d'un polymère AR. Le TMPR est un temps clé pour des processus tels que la régulation de gènes et la réparation de dommages sur l'ADN.

Au chapitre six, nous généralisons le modèle AR analytique afin de prendre en compte plusieurs DTA de tailles différentes ainsi que les connectivités intra-DTA et extra-DTA. Nous étudions la dynamique de réorganisation de DTA lors des stages successifs de différentiations cellulaires à partir de données de CC. Nous trouvons un effet non-négligeable de la connectivité de l'inter-DTA sur les dynamiques de la chromatique. Par la suite nous trouvons une compactification et une décompactification synchrone des DTA à travers les différents stages.

Abstract

In this dissertation we study the relationship between chromatin conformation and dynamics using a class of randomly cross-linked (RCL) polymer models. The RCL models account for the variability in chromatin conformation over cell population. We use tools from statistics, stochastic process, numerical simulations and polymer physics, to derive the steady-state and transient properties of the RCL polymer, and use them to elucidate the dynamic reorganization of the chromatin for various scales and biological conditions. We introduce the biological background and polymer physics in chapter 1, followed by a list of results obtained in this thesis in chapter 2.

In chapter 3 of this dissertation work, we develop a general method to construct the RCL polymer directly from chromosomal capture (CC) data. We show that persistent long-range connection between topologically associating domain (TAD) affect transient encounter times within TADs, in the process of X chromosome inactivation. We further show that the variability in anomalous exponents, measured in single particle trajectories (SPT), is a direct consequence of the heterogeneity of cross-link positions.

In chapter 4 we use the RCL polymer to study local genome reorganization around double strand DNA breaks (DSBs). We calibrate the number of connectors in the RCL model using SPT data, acquired before and after DSB. We find that the conservative loss of connectors around DSB sites significantly affects first encounter times of the broken ends in the process of DSB repair. We show how a cross-linked genomic micro-environment can confine the two broken ends of a DSB from drifting apart.

In chapter 5 we derive analytical expressions for the steady-state and transient properties of the RCL model, representing a single TAD region. The derived expressions are then used to extract the mean number of cross-links in TADs of the CC data, by a simple curve fitting procedure. We further derive formula for the mean first encounter time (MFET) between any two monomers of the RCL polymer. The MFET is a key time in processes such as gene regulation.

In chapter 6 we generalize the analytical RCL model, to account for multiple TADs with variable sizes, intra, and inter-TAD connectivity. We study the dynamic reorganization of TADs, throughout successive stages of cell differentiation, from the CC data. We find non-negligible effect of inter-TAD connectivity on the dynamics of the chromatin. We further find a synchronous compaction and decompaction of TADs during differentiation.

Publications

Published

- Shukron Ofir, and David Holcman. "Transient chromatin properties revealed by polymer models and stochastic simulations constructed from Chromosomal Capture data." PLoS computational biology 13.4 (2017): e1005469.

- Shukron Ofir, and David Holcman. "Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data." Physical Review E 96.1 (2017): 012503.

- Shukron Ofir, Michael Hauer, and David Holcman. "Two loci single particle trajectories analysis: constructing a first passage time statistics of local chromatin exploration." Scientific Reports 7 (2017).

In preparation

- Shukron Ofir, and David Holcman. "Chromatin reorganization during cell differentiation captured by randomly cross-linked polymer models of multiple topologically associating domains."

- Shukron Ofir, and David Holcman. "A model for histone sliding and genome reorganization following UV-C induction."

Contents

Abbreviations and notations

EP - Encounter probability
EF - encounter frequency
TAD - topologically associating domain
SPT - single particle trajectory
RCL - randomly cross-linked
MSD - mean square displacement
MSRG - mean square radius of gyration
FET - first encounter time
MFET - mean first encounter time
FDT - first dissociation time
MFDT - mean first dissociation time
t - time
N - number of monomers
b - std of connectors' length
D - diffusion constant
r_n- monomer n of a polymer
ξ- connectivity fraction single TAD
Ξ- connectivity matrix, multiple TADs
$B(\xi)$- added connectivity matrix
$[B^{(i)}()]$ - i^{th} diagonal block matrix
$\langle \cdot \rangle$ - average operator

Introduction 1

1.1 The dynamic landscape of the mammalian chromatin

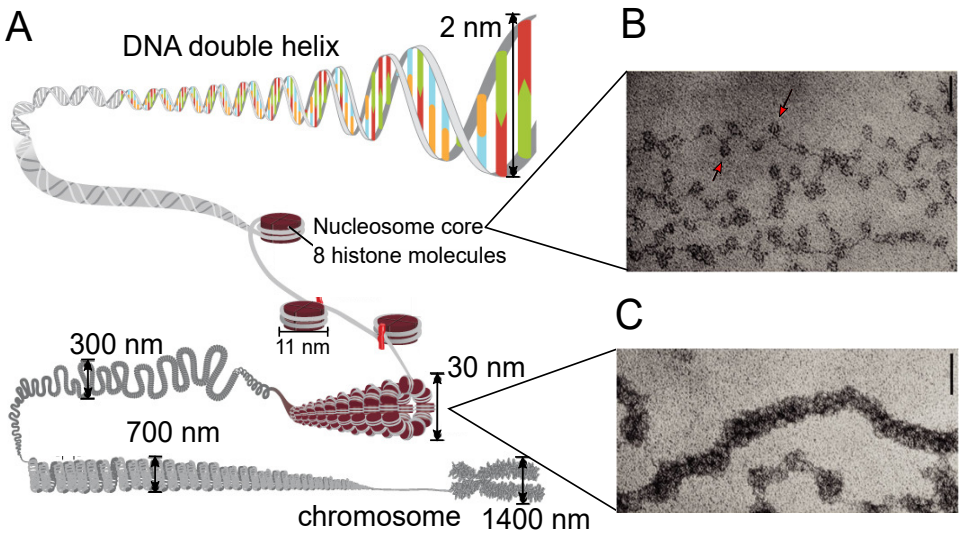


Figure 1.1 **DNA folding Hierarchy. A.** The double strand DNA is folded onto nucleosomes (bead-string model, 10-11 nm fiber) and further folded into the chromatin (30 nm fiber). Figure adapted from [10].**B.** electron microscopy image of the bead-string chromatin, with nucleosomes indicated by arrows. Adapted from [78]. **C.** electron microscopy image of the 30 nm fiber of the chromatin. Scale-bar=50 nm. Adapted from [78].

The DNA in mammalian nucleus is the focal point of sub-cellular activities. Continuous and rapid changes to three-dimensional organization of the DNA are due to process such as transcription, repair, and monitoring the integrity of the genomic content. The organization, function, and dynamics of the DNA are, thus, intimately linked. Polymer physics is the framework in which the complex relationship between DNA organization and dynamics can be studied in a rational manner. However, it is prohibitively difficult to construct a polymer model, which can capture the many complexities of genomic organization and dynamic.

The mammalian nucleus contains about 6×10^9 DNA base-pairs (bp), with each bp of size 0.3 nm, the total DNA in the cell reaches a length of roughly 2 m, which is compacted in a nucleus of about 10 μm in diameter. Such high level of compaction is achieved by an intricate mesh-work of proteins and DNA, which constitute the chromatin. The first level of chromatin compaction is due to nucleosomes (Fig. 1.1A), consisting of 150-170 DNA

bp wounded 1.6 times on histone protein octamers complexes, and forming the 10 nm nucleosomal fiber, also referred to as the bead-string model [60]. An additional level of compaction is due to higher-order folding of DNA-histone complexes into the 30 nm fiber. Much less is known about the folding principles of the genome beyond this level.

Experimental sub-cellular measurement techniques [36, 76, 29] revealed that protein-mediated kilo-bp (kbp) and mega-bp (mbp) genomic loops are main contributors to further compact the chromatin, and are key features of genome three-dimensional organization [85]. The juxtaposition of two distal genomic segments can arise due to the activity of transcription factors and other genome structural remodeling proteins, in processes such as gene regulation and DNA repair, but can also originate from random collisions and DNA entanglement [49, 76, 57]. As a result, genomic loops can take on either stable configuration, which persist throughout cell cycles [49], or transient, having a lifetime ranging from 2-35 minutes [43].

On a genome-wide scale, proximity ligation methods, known as the chromosome Conformation Capture (CC) techniques, provide a snapshot of the three-dimensional organization of the chromosome. Starting with the Chromatin Conformation Capture (3C) method [25] and its successors (the 4C [111], 5C [31], and Hi-C [85]). This CC family of techniques, developed in the last two decades, record genomic looping events simultaneously over a population of millions of nuclei. The common, principle steps of the CC methods are shown in Fig. 1.2A. The 3C method provides the contact frequency between two specific genomic loci, the 4C then extended this examination to a single known loci vs. all others, and the 5C and Hi-C methods generalized the formers by providing pairwise contact frequencies between all genomic segments (known and unknown) at a resolution of 1-3kpb.

1.1.1 Topologically Associating Domain

Shortly after its introduction, the 5C techniques was used to discover [65] that mammalian X chromosomes fold into discrete mega-base regions of enriched segment-segment interactions (Fig. 1.1B). These contact-enriched regions were termed Topologically Associating Domains (TADs), and were later found to be a common feature of all chromosomes[29, 76]. The mechanism driving the formation of TADs is unknown, and our understanding of internal TAD organization and its affect on cellular processes is largely incomplete. The internal TAD organization is highly variable between cells of similar type [87], whereas TAD boundaries are mostly conserved during cell cycle and

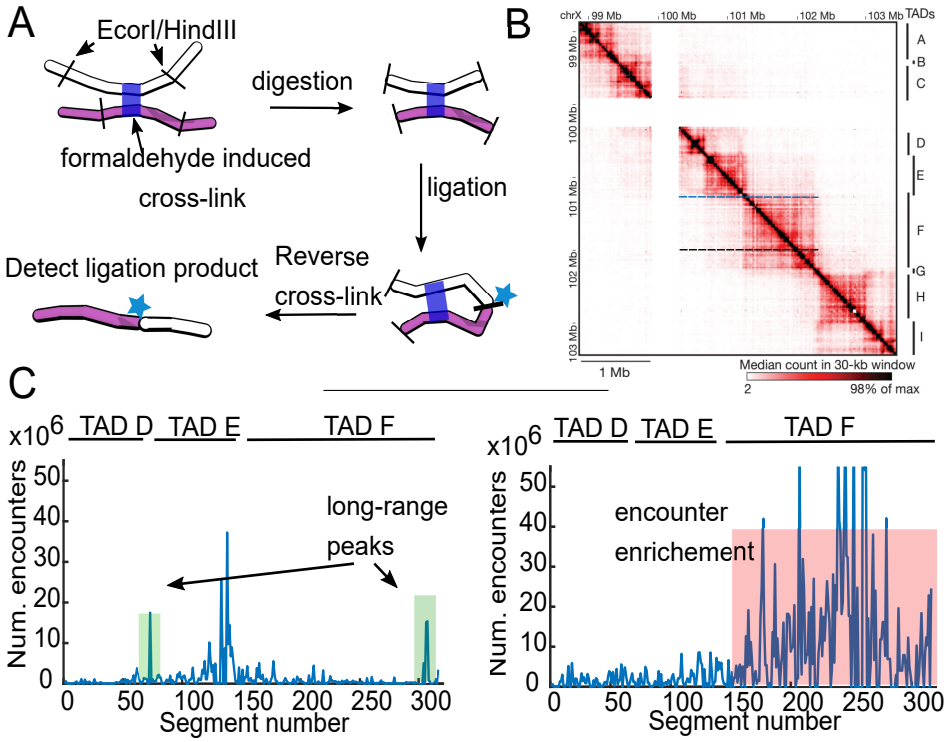


Figure 1.2 **The Chromosomal Capture experiments. A.** principle steps of the CC experiments include cross- linking by formaldehyde (blue, upper left) of genomic segment (white and purple) harboring recognition sequences (lines) of restriction enzymes, such as EcoRI and HindIII. Restriction enzyme then digest the cross-linked segments (upper right), and the digested DNA undergoes ligation in dilute solution (bottom right) to form a ligation product (star). The ligation products are detected, following the reversal of cross-links (bottom left). **B.** The encounter frequency matrix produced by the 5C experiments (extract from [76]). In this example, the genomic segment harbors 10 Topologically Associating Domains (TADs). **C.** Two sample cross sections of the EF matrix in panel B, showing long-range persistent peaks (left panel), corresponding to dashed blue line in panel B, and encounter enrichment (right), corresponding to dashed black line in panel B.

between cell of similar type[89, 76, 36]. Disruption of these boundaries resulted in merging of TADs [112, 89, 76] and mis-regulation of genes. TADs have also been found to be a unit of correlated gene regulation and DNA replication [84, 27]. Higher order folding hierarchy of TAD have also been discovered, where TADs cluster to form meta-TADs [36, 82].

Two distinct features of the CC maps are encounter enrichment at TAD position and isolated high amplitude local maxima (Fig. 1.2C). The exact origin of the peaks of CC maps are unknown but have been found to indicate a conserved genomic interactions between distal loci, which are involved in conserved and cell type-specific gene activity [36, 76, 70], or originating from conserved structural features of the chromatin, such as at TAD boundaries [96, 59].

1.1.2 Limitations of the CC data

Despite the deep insight into genome organization, brought about with the CC experimental results, interpretation of the CC encounter matrices remains difficult. An interpretation of the 3D genome organization in terms of the CC maps must take into account both probabilistic nature of cross-linking and ligation, and population averaging of CC data. The CC maps, thus, represent a large section of the looping configuration space of chromosomes of a particular cell type and state over a population. Indeed, TADs are only visible in population CC, and are not present in single-cell CC data [73, 99]. In addition, the CC matrices provide only static encounter information, completely deprived of any temporal data and dynamics, which leaves open the question of how TADs are formed and interact over time. Moreover, CC reads do not translate immediately into genomic distances between interacting loci, and cannot be used to infer the size of the folded chromatin in the nucleus.

The missing temporal dimension in the CC data calls for complementary methods to elucidate the origin of TAD formation, chromatin organization, and dynamics. Methods such as single particle tracking (SPT) [101, 39, 26, 44, 9, 63] can provide only limited information for a restrictive number of tagged loci over time. It is at this point, that polymer models come into play. The encounter probabilities, computed from the CC matrices, provide rich grounds on which the steady-state statistical properties of a given polymer model can be verified and, thus, act as a candidate mechanical model for representing chromatin. The trajectories from SPT, then, can provide the complementary temporal information to validate transient properties of a suggested polymer model, e.g., the first encounter times between monomers.

1.2 Polymer models as coarse-grained representation of the chromatin

The structural complexity of the chromatin and its interactions renders a study into the relationship between its structure and equilibrium and transient properties prohibitively difficult. Therefore, rigor in structural description is sacrificed in favor of mathematical tractability by studying mechanical polymer models, stripped of many details, as a coarse-grained representation of the chromatin.

A coarse-grained polymer model consists of sequence of connected units (monomers), where each monomer collectively represents a geneomic seg-

ment at a given coarse-graining scale (Fig. 1.3A). The choice of scale affects the amount of details captured by the model, and can often times be dictated by the resolution of the experimental measurements at hand.

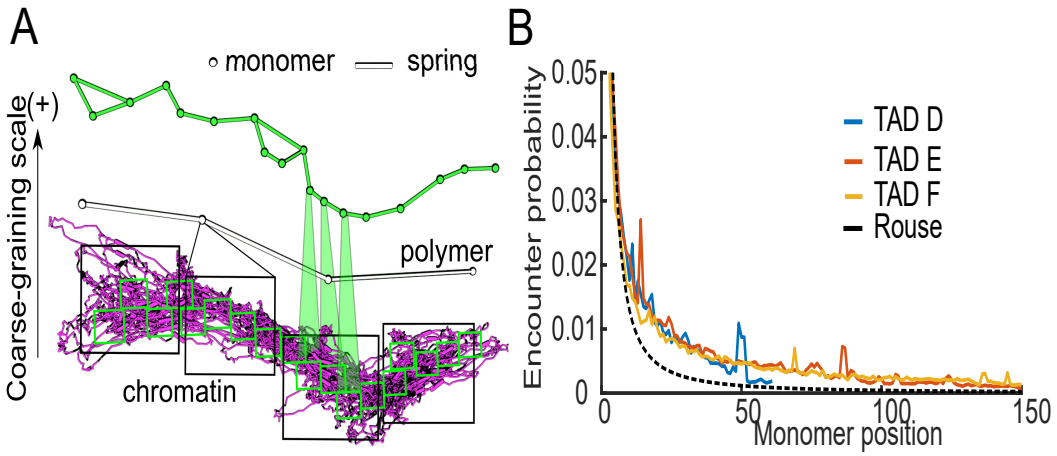


Figure 1.3 **Polymer models as coarse-grained representation of the chromatin A.** Two levels of coarse-graining of a sample chromatin segment (purple) to construct a polymer model consisting of monomers (spheres) connected by harmonic springs (bars). For the white polymer in this example, only nearest-neighboring monomers are connected to form the Rouse polymer. **B.** The encounter probability between genomic segments of TAD D (blue), TAD E (red) and TAD F of the mammalian X chromosome in Fig. 1.2B, decays with genomic distance and cannot be captures by the Rouse polymer (dashed), which does not account for long-range interactions.

The most tractable models for representing the dynamics of complex macromolecules are the bead-string polymer models [30, 68]. These models consist of N similar beads (monomers) of unit mass, located at position $\boldsymbol{R} = [r_1, r_2, .., r_N]^T$ in dimension d and connected arbitrarily by Hookean springs (Fig. 1.3A). According to the linear force law of Hookean springs, forces between connected monomers are derived from the quadratic potential

$$\phi(\boldsymbol{R}) = \frac{\kappa}{2}\boldsymbol{R}^T\boldsymbol{M}\boldsymbol{R}, \tag{1.1}$$

where $\kappa = \frac{dk_BT}{b^2}$ is the spring constant, related to the standard-deviation b of the spring between connected monomers, k_B is the Boltzmann's constant, and T is the temperature. The matrix $\boldsymbol{M} = \boldsymbol{C} - \boldsymbol{\Delta}$ is the Laplacian of the graph representation of the polymer [40], $\boldsymbol{C}$ is a diagonal $N \times N$ matrix containing the degree of connectivity (number of connectors) of each monomer, and $\boldsymbol{\Delta}$ is the $N \times N$ graph adjacency matrix, containing 1 at cell m, n when monomers

m and n are connected. The matrix $\boldsymbol{M}$ contains all the information about the internal connectivity of the polymer, where cells i, j of $\boldsymbol{M}$ are

$$\boldsymbol{M}_{ij} = \begin{cases} -1, & |i-j| = 1; \\ -1 & r_i, r_j \text{ are connected, } (|i-j| > 1); \\ -\sum_{j \neq i}^{N} \boldsymbol{M}_{ij}, & i = j. \end{cases} \tag{1.2}$$

Random and independent collisions of monomers with particles of their surrounding fluid (e.g., the nucleoplasm) result in an additional fluctuating force with no preferred direction. The distribution of this random force is a Gaussian with mean 0 and variance $2\zeta k_B T \delta(t-t')$ in a short time interval $t-t'$, and ζ is the friction coefficient, which describes the affect of the resistance of fluid particles to monomers' displacement (Stocks' law). Taken together, the dynamics of the polymer in the solvent is described by the over-dumped Langevin equation [30]

$$\begin{aligned} \frac{d\boldsymbol{R}(t)}{dt} &= -\frac{1}{\zeta}\nabla\phi(\boldsymbol{R}(t)) + \frac{\sqrt{2\zeta k_B T}}{\zeta}\frac{\mathrm{d}\boldsymbol{\omega}(t)}{\mathrm{d}t} \\ &= -d\frac{D}{b^2}\boldsymbol{M}\boldsymbol{R}(t) + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}(t)}{\mathrm{d}t}, \end{aligned} \tag{1.3}$$

where $D = \frac{k_B T}{\zeta}$ is the diffusion coefficient and $\boldsymbol{\omega}(t)$ are standard Brownian motion with mean 0 and standard-deviation 1.

When only nearest-neighbor monomer connectivity is considered, the resulting polymer is the Rouse polymer [30], defined by the tri-diagonal connectivity matrix $\boldsymbol{M}$

$$\boldsymbol{M}_{ij} = \begin{cases} -1, & |i-j| = 1; \\ -\sum_{j \neq i}^{N} \boldsymbol{M}_{ij}, & i = j, \end{cases} \tag{1.4}$$

which represent the linear polymer's backbone, and the potential energy of the Rouse polymer is

$$\phi_{Rouse}(\boldsymbol{R}) = -\frac{\kappa}{2}\sum_{n}^{N}(r_n - r_{n-1})^2. \tag{1.5}$$

In the Rouse polymer, the vector between any two monomers is Normally distributed, with probability density function given by [30]

$$f_{m,n} = \left(\frac{d}{2\pi b^2|m-n|}\right)^{d/2} \exp\left(-\frac{d(r_m(t)-r_n(t))^2}{2|m-n|}\right). \tag{1.6}$$

When $\|r_m(t) - r_n(t)\| \to 0$, where ϵ is the radius of an encounter sphere, the EP, $P_{m,n}$, between monomers m and n is then

$$P_{m,n} \propto \left(\frac{d}{2\pi b^2|m-n|}\right)^{d/2}. \tag{1.7}$$

The Rouse polymer describes well the statistics of long polymers [30, 45] (on a scale of Mbp) in solutions where hydrodynamics and exclusion forces are screened out. However, the lack of monomer connectivity beyond that of nearest neighboring monomers renders the Rouse polymer as inadequate in representing long-range genomic loops (Fig. 1.3B), such as in TADs.

1.2.1 Steady-state statistics of bead-string polymer models with general connectivity

The adequacy of a polymer model for representing CC data can be testes by comparing the steady-state EP of the polymer model to that of the empirical data. In system 1.3, the coordinates of all monomers are coupled, which poses difficulties in deriving steady-state properties e.g., the variance and encounter probability. Decoupling of system 1.3 is done by finding an orthonormal eigenbasis $\boldsymbol{V} = [v_0, v_1, .., v_{N-1}]$ and defining a new coordinate system $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}$ such that the potential 1.1 is transformed to

$$\phi(\boldsymbol{U}) = \frac{\kappa}{2}(\boldsymbol{V}^T\boldsymbol{U})^T\boldsymbol{M}(\boldsymbol{V}^T\boldsymbol{U}) = \frac{\kappa}{2}\boldsymbol{U}^T\Lambda\boldsymbol{U} \tag{1.8}$$

where $\Lambda = \boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^T = diag(\lambda_0, \lambda_1, .., \lambda_{N-1})$ are the eigenvalues of the connectivity matrix $\boldsymbol{M}$ (Eq. 1.2). We therefore obtain a convenient set of Ornstein-Uhlenbeck equations

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -\frac{1}{\zeta}\nabla\phi(\boldsymbol{U}) + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t} = -d\frac{D}{b^2}\Lambda\boldsymbol{U} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t}, \tag{1.9}$$

in the so called normal coordinates $\boldsymbol{U} = [u_0, u_1, .., u_{N-1}]^T$ [30], which represents the dynamics of the N modes of motion of the polymer [30]. In this coordinate system, the variance between monomers r_m and r_n is defined as

$$\sigma^2_{mn}(t) = \langle (r_m - r_n)^2 \rangle = \sum_{p=1}^{N-1} (v_p^m - v_p^n)^2 \langle u_p(t)^2 \rangle = \sum_{p=1}^{N-1} \frac{b^2 \left(1 - e^{-\frac{2dDt}{b^2}}\right)}{\lambda_p^2} \quad (1.10)$$

where v_p^m is the m-th row of v_p in the eigenbasis $\boldsymbol{V}$. In general monomer connectivity, the major difficulty is to obtain V and Λ analytically and therefore compute 1.10 for $t \to \infty$ to obtain the variance at steady-state. However, the eigenvalues can be computed for some particular polymer topologies such as rings, stars, cubes [40, 35] and other organized structure, but otherwise, one has to resort to procedures of numerical diagonalization [34].

1.2.2 Transient statistics of polymer with general monomer connectivity

The first encounter time between two genomic loci is a key time in processes of gene activation/silencing [39, 94, 66], RNA transcription and antibody coding, to name only a few. The first encounter time of two monomers r_1, r_N of a polymer chain is defined as

$$\tau_\epsilon = \inf\{t > 0; |r_1(t) - r_N(t)| < \epsilon\}, \quad (1.11)$$

and was first studied by Wilemski and Fixman [110] for the two end monomers of a Rouse chain, and later by many others [22, 23].

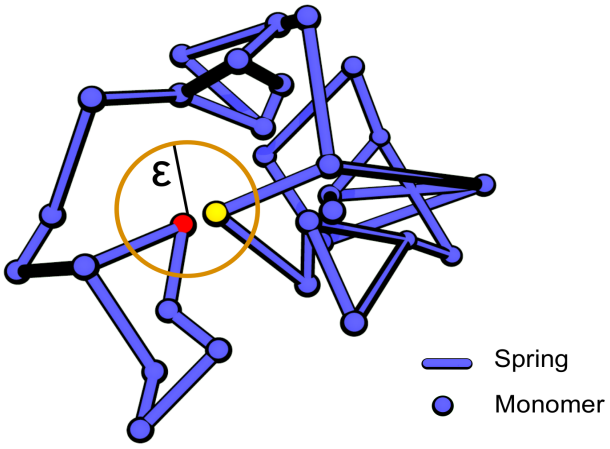


Figure 1.4 **The encounter between two monomers (red, yellow) at a distance ϵ in a polymer model with general connectivity.**

A general expression for the mean first encounter (MFET) was also derived in [8], and is given by

$$\langle \tau_\epsilon \rangle \approx \frac{1}{\lambda_0^\epsilon} = \frac{|\Omega|}{4\pi\epsilon \int_{C-P} e^{-\phi(\boldsymbol{U})} \mathrm{d}\boldsymbol{U}}, \tag{1.12}$$

where λ_0^ϵ is the first non-vanishing eigenvalue of the forward Fokker-Planck equation [91], associated with the stochastic system 1.9 in normal coordinates $\boldsymbol{U}$. The term $|\Omega|$ is the value of integral over the whole configuration space of the polymer and can be computed when the eigenvalues and eigenvectors of the connectivity matrix are known. For the Rouse polymer

$$|\Omega| = \int e^{-\phi(\boldsymbol{R})} d\boldsymbol{R} = \left(\frac{(2\pi)^{(N-1)}}{\prod_{p=1}^{N-1} \lambda_p} \right)^{d/2}. \tag{1.13}$$

with

$$\lambda_p = 4\kappa \sin^2(p\pi/2N), \quad p = 1..N-1. \tag{1.14}$$

In [8], the distribution of the first encounter time in expression 1.12 was shown to be well approximated by a sum of exponentials, where a single exponential is sufficient for short polymers in an open domain, and more terms are needed for long polymers or to account for boundary effects for polymers in confined domains. This result was validated by Brownian simulations in free and confined domains [3], and further for the β polymer [5], which included long-range monomer interactions.

In general polymer connectivity, where long-range monomer connectivity is permitted, the computation of the $C-P$ integral 1.12 is challenging, because the eigenvalues of the connectivity matrix must first be computed.

1.2.3 Polymer models with long-range monomer connectivity

Polymer models with long-range monomer connectivity were also considered as coarse-grained representation of the chromatin. In [16], a random looping model was studied numerically, where monomers m, n are connected by a fixed connector, based on monomers' distance along the chain. In that model, the authors provided analytical expression for the MSD, and their simulation results agreed with average loops lengths observed experimentally. On the same line, dynamic loop model [45] was studied numerically, where monomers can connect to form a loop only when they are located within the

encounter distance, and the loops are reversible. Dynamic loop formation was also considered in the strings and binders switch model [11], in which loop formation is mediated by diffusing molecules that must be present at binding sites at the moment of encounter. This model provides means of mapping a range of polymer folding phenomena based on the distribution of binding sites. However, loops are either stable or unstable collectively, and depend on binding site affinity and the concentration of binding molecules. Polymer melts with random connectivity were studied numerically on a discrete lattice in [98] with the aim of exploring the diffusion properties of the melt. The numerical study of the polymer models above obscures the dependence of dynamic behavior on model parameters, the dynamics in free domain, and require heavy numerical simulations. Nevertheless, the models mentioned above capture one of the key characteristics of an ensemble of chromatin, i.e., variability of polymer conformation in the ensemble as a result of random looping.

The β polymer [5] provides an analytical asymptotic expressions for the statistical properties of a polymer with long-range monomer interactions. The β polymer provides means of constructing a polymer, having a potential energy, which decays with monomer distance, and is based on a prescribed anomalous exponent $1 - 1/\beta$. In the polymer model by Giorgetti et al. [39], both attractive and repulsive potentials were placed between monomer pairs, and the strength of these potentials was deduced from the 5C EP matrices. In the two models above, the potential between monomers cannot, in general, be linked with any physical contact between genomic segments, or loop formation, and thus the structure of the polymer remains obscure.

Polymer models to predict 3D chromatin organization were also constructed from CC [105] and single cell Hi-C[99], and predict encounter probability which agreement with population Hi-C. However, it is unclear how to interpret the resulting structure, or to what extant does the predicted 3D structures reliably represent the dynamic chromatin conformation landscape or a characteristic chromosome conformation. Other notable polymer models include the well known fractal globule [72], random copolymer [67], and polymer with folding principles based on the epigenetic state of the chromatin [62].

1.2.4 The RCL polymer model

Before listing the results of this dissertation in the next Chapter, I now give a short description of the RCL polymer. The RCL polymer is a member of the general Gaussian models [40], represented by N monomers connected

sequentially by harmonic spring, and an additional N_c connectors between random monomer pairs (Fig. 1.5). In each realization of the polymer, the choice of monomer pairs to connect is randomized. This added level of random connectivity serves to capture the heterogeneity in configurations seen in an ensemble of chromatin, and the added connectors serve to mimic the effect of binding molecules (e.g., CTCF).

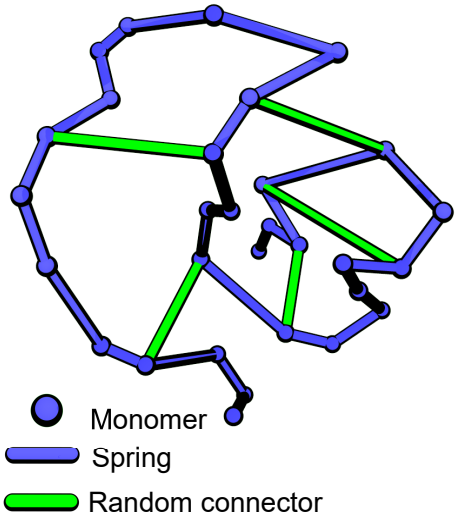


Figure 1.5 **The RCL polymer model.** A set of N monomers (blue sphere) connected sequentially by harmonic springs (blue bars) to form the backbone, and an additional set of connectors (green) is added between random monomer pairs in each realization of the polymer.

The potential energy of the RCL is a generalization of the potential in 1.1, defined as the sum of the potential energy of the deterministic linear backbone and random potential from added random connectors,

$$\phi(\boldsymbol{R}) = \frac{\kappa}{2}\boldsymbol{R}^T(\boldsymbol{M} + B(\xi))\boldsymbol{R}, \tag{1.15}$$

where $\boldsymbol{M}$ is the Rouse matrix (Eq. 1.4), $B(\xi)$ is the random added connectivity matrix, describing the position of N_c added random, non-nearest neighboring (NN) monomer pairs,

$$B_{m,n}(\xi) = \begin{cases} -1, & r_m, r_n \text{are connected, } |m-n| > 1; \\ -\sum_{j\neq m}^{N} B_{m,j}(\xi), & m = n, \end{cases} \tag{1.16}$$

and the choice of m and n is randomized for each realization of the polymer. The connectivity fraction $0 \leq \xi \leq 1$ is related to N_c by

$$N_c(\xi) = \lfloor \xi \frac{(N-1)(N-2)}{2} \rfloor, \tag{1.17}$$

and is defined as the fraction of the maximal possible choices of non NN monomers pairs to connect. Put differently, it is a fraction of the total number of cells in the upper triangular part of $N \times N$ matrix, excluding the super diagonal. The dynamics of monomers of the RCL polymer is given by the Langevin equation

$$\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} = -\frac{1}{\zeta}\nabla\phi(\boldsymbol{R}) + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t}, \tag{1.18}$$

where $\boldsymbol{\omega}$ are standard Brownian motion with mean 0 and standard-deviation 1. As we shall see in subsequent chapters, this simple construction of the RCL polymer can account for many of the features of the chromatin organization and dynamics.

Summary of results

2

In this chapter I give a short summary and a list of results for each of the chapters in this dissertation work.

2.1 Chapter 3: Transient chromatin properties revealed by the randomly cross-linked polymer model

In Chapter 3 I present a computational procedure to construct a bead-spring polymer model directly from the 5C data. The polymer captures two hallmarks of the 5C data: long-range persistent interactions and encounter enrichment, which accounts for TAD-like structures (see Fig. 1.2). I use this polymer model to elucidate the relationship between genome organization and transient properties, such as the mean first encounter time between genomic segments. We refrain from providing three-dimensional structure to interpret the data as in [62, 99, 104, 64], and focus on extracting the mean number of cross-links in a given genomic segment. This information is not contained in the 5C data. To demonstrate the construction process, I use a subset of the 5C data of the X chromosome of female mice embryonic stem cells [39, 76], harboring two TADs (TAD D and E), which span a genomic region of about 1 mbp.

2.1.1 Result 1: The chromatin is represented by a polymer model with random short-range and persistent long-range connectors

I present a method to construct a polymer model, with a combination of random and persistent connectors, directly from the empirical encounter probability (EP) of the 5C data [76]. I use connectors to resolve a reverse engineering problem, which is to recover the degree of connectivity from the EP-decay rate. To do so, I coarse-grain the 5C data at a resolution of 3 kb, in accordance with the median restriction fragment of the HindIII restriction enzyme used for generating the 5C data [39, 76]. The resulting polymer model is composed of $N = 307$ monomers, with TAD D (1-106) and TAD 107-307 (Fig. 2.1A). Then, I fitted a model of the form $c_0|m-n|^{-\beta}$, $c_0, \beta > 0$ (see Eq. 1.7) to the experimental EPs of TAD D and E, and obtained $\langle\beta_D\rangle = 0.74, \langle\beta_E\rangle = 0.8$, indicating that Rouse polymer ($\beta = 1.5$) is inadequate in

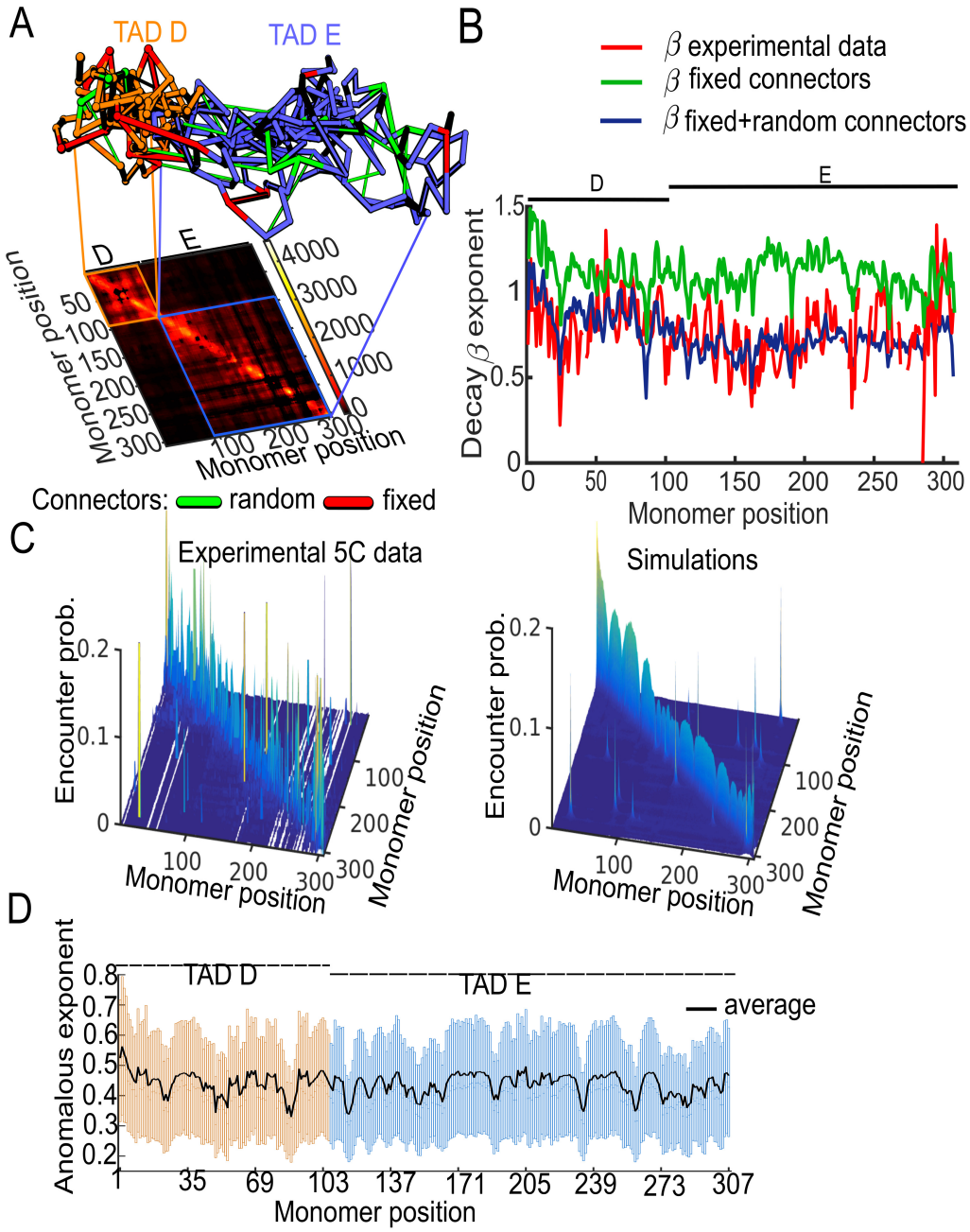


Figure 2.1 **Reconstruction of the chromatin by RCL polymer model A.** The 5C encounter data [76] of TAD D and E is represented by a polymer model of $N = 307$ monomers with monomer 1-106 (orange) representing TAD D and 107-207 representing TAD E (blue). The polymer model consist of random short-range connectors (green) within each TAD, and fixed long-range (red) within and across TADs. **B.** Decay β exponent obtained from fitting a model of the form $c_0 m^{-}\beta$, with m the monomer distance and c_0 a constant, to each one of the 307 monomers in panel A, for the experimental data (red), a polymer model with only fixed connectors (green), and a model with random and fixed connectors (blue). **C.** The encounter probability surface of the experimental 5C (left) versus that of the calibrated polymer model (right). **D.** Anomalous exponents α_m ($m = 1..307$) of the calibrated RCL model shows high variability around the mean value of 0.4.

representing the empirical EP (see also Fig. 1.3B). I generalize the potential 1.1 to include the linear backbone, random, and fixed long-range connectors

$$\phi(\boldsymbol{R}) = \frac{\kappa}{2}\sum_{m}(r_m - r_{m-1})^2 + \frac{\kappa}{2}\sum_{\mathcal{G}}(r_m - r_n)^2 + \frac{1}{2}\sum_{S_{max}}\kappa_{mn}(r_n - r_m)^2, \quad (2.1)$$

where κ is the spring constant of both the linear backbone and random connectors, $\mathcal{G}$ is a set of N_c randomly chosen non nearest-neighboring (NN) monomer pairs, S_{max} is a set of monomer indices pairs representing positions of persistent fixed connectors, and k_{mn} are spring constant for persistent long-range connectors, which I calibrated from the 5C data. In each realization of the polymer I randomize the choice of monomer pairs to connect (the set $\mathcal{G}$ in Eq. 2.1) to capture the effect of CTCF binding molecules and their heterogeneous positions in a large population of cells. I obtained the positions of persistent loops by thresholding the 5C data, using a threshold value computed from the NN monomers' EP (super and sub-diagonals of the 5C EP matrix). I find 24 persistent connectors within and between TADs.

By numerical simulations of system 1.3, I establish a relationship between the decay β exponent and N_c, the number of connectors. Using that relationship, I find that the experimental $\beta_D = 0.74, \beta_E = 0.8$ correspond to 6 and 10 connectors in TAD D and E, respectively. This result shows that the addition of only a few connectors is sufficient for reproducing the empirical EP of TADs. A comparison of the decay β exponents of the 5C data to a model with only long-range versus a model with a combination of random and fixed connectors, favors the latter (Fig. 2.1B).

2.1.2 Result 2: Long-range persistent connectors between TADs affect transient encounter times of monomers within TADs

Using stochastic simulations of the calibrated RCL model, I computed the histogram of the conditional first encounter time and probability of three key loci, corresponding to the genomic elements called Tsix/Xite, Chic1, and Linx. The pairwise conditional encounter of these loci initiates a cascade of events leading to production of proteins, which coat the X chromosome, and lead to its inactivation in female mammalian cells [76]. I show that the conditional EP is nearly 50% between loci pairs when both TAD D and E are present, whereas when TAD E is not considered and the connectors between TADs are removed, the probability shifts and the conditional encounter time increases

by 50%. This result is a consequence of removal of persistent long-range connectors between TADs and demonstrates how TADs can cross-regulate. I find the first encounter times to be Poissonian, which agree with theoretical results [8].

2.1.3 Result 3: The variability in anomalous exponents of monomers within TADs is caused by heterogeneous TAD organization

I computed the Mean-Squared-Displacement (MSD) of loci in both TAD D and E from simulations of the calibrated RCL model, where the position of connectors is randomized in each realization and persistent long-range connectors are included. I found large fluctuations in the values of anomalous exponents (Fig. 2.1D) around a mean value of 0.42, and around 0.4 when TAD E is removed. This result links the heterogeneity in internal chromatin organization in a population to the large variability in anomalous exponent measured experimentally. I attribute the sub-diffusive behavior of loci to the presence of persistent long-range connectors.

2.2 Chapter 4: Two loci single particle trajectories analysis, constructing a first passage time statistics of local chromatin exploration

A large body of work deals with the statistics extracted from SPTs, the characterization of locus motion, and inference of the characteristics of local locus environment [9, 14, 48, 108] to mention only a few. Less well dealt with is the analysis of the correlated motion of two tagged loci and its interpretation in terms of the local genome organization. In particular, when the two tagged loci on the same chromosomal arm are tracked, the connection between local chromatin architecture and transient first encounter times remains largely unexplored. However, several work do exists and are insightful [62].

In Chapter 4 of this dissertation I study chromatin reorganization using the transient first encounter time (FET) and first dissociation time (FDT) of two tagged loci (Fig. 2.2A). I study genome confinement, organization and reorganization based on the transient mean FET (MFDT) and mean

FDT (MFDT) from the experimental histograms of recurrent FET and FDT, reported in two data sets. The first dataset, generated in [26], provides simultaneous recording of two tagged loci on yeast genome, positioned at genomic separation ranging from 25-100 kbp. The second dataset, provided by Hauer et al. [44], contains simultaneous recording of two tagged loci 50 kbp apart, before and after the induction of double strand breaks (DSB) by the Zeocin drug. Zeocin is a radiomimetic substance, which induces DSB uniformly in the nucleus.

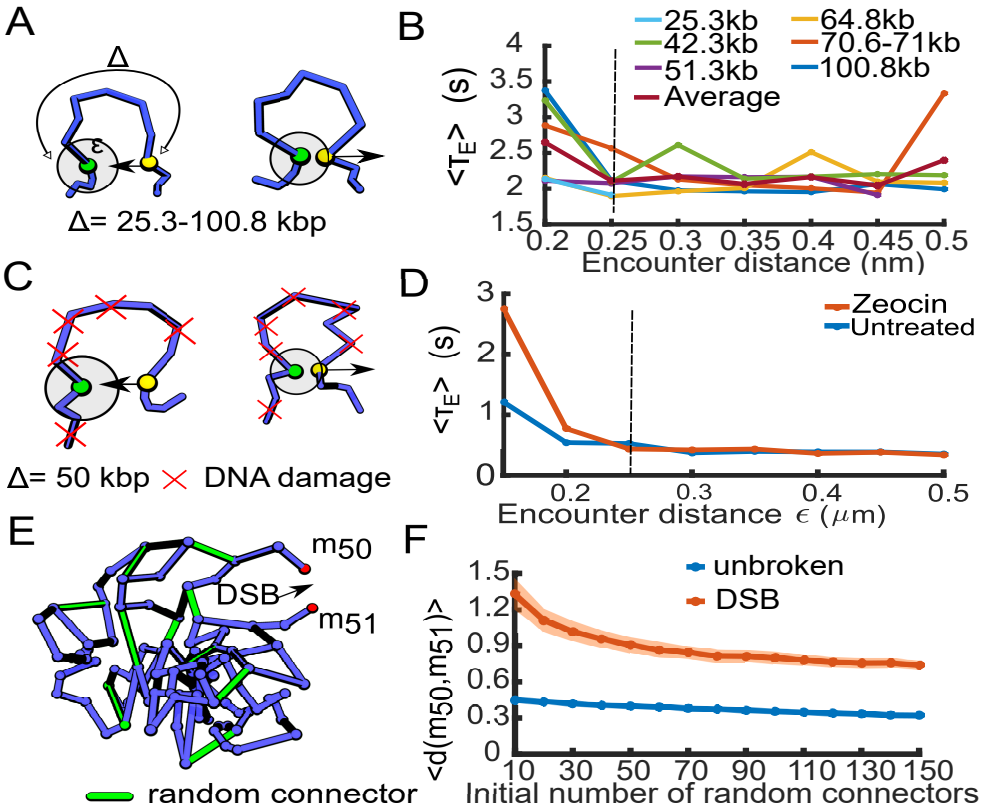


Figure 2.2 **First encounter and dissociation of two tagged loci reveal chromatin reorganization principles A.** First association (left) and dissociation (right) of two tagged loci (green and yellow spheres) at the encounter distance ϵ , recorded in [26] for genomic separation distance $\Delta \in [25, 100.8] kbp$. **B.** The mean first encounter time (MFET) $\langle \tau_E \rangle$ for all genomic strains in panel A, plotted against ϵ, shows that $\langle \tau_E \rangle$ is independent of ϵ above 250 nm. **C.** First association (left) and dissociation (right) of two tagged loci (green and yellow spheres), recorded in [44], before and after the induction of DSB (red crosses) by Zeocin drug. **D.** A plot of $\langle \tau_E \rangle$ before for untreated (blue) and zeocin treated (orange) cases, shows that $\langle \tau_E \rangle$ is independent of ϵ above 250 nm. **E.** A sketch of the randomly cross-linked (RCL) polymer of 100 monomers used for the study of genome reorganization following a single DSB between monomers m_{50} and m_{51} (red). **F.** The mean distance $\langle d(m_{50}, m_{51}) \rangle$ between monomers m_{50}, m_{51} computed from simulation of the RCL polymer in panel E for the unbroken (blue) and the DSB (orange) case.

2.2.1 Results 1: The two tagged loci are confined in a region of 250 nm in radius

I collected the histograms of the FET and FDT of trajectories of the data set in [26], and fitted an exponential distribution to it. The MFET and MFDT are the reciprocal of the decay rates of the fitted distributions [8]. By varying the encounter distance ϵ, I find that the MFDT and MFET are independent of ϵ above 250 nm (Fig. 2.2B). I find a similar confinement length scale in the data set [44] by repeating the FET and FDT analysis on the distribution (Fig. 2.2C) of the chromatin before and after the induction of DSB (Fig. 2.2D). This result implies that 250 nm is a characteristic length-scale for chromatin confinement. I use the transient statistics and analytical formulas derived from polymer model in confined domains [3] to compute the radius of confinement for the polymer itself to be 0.5 μm.

2.2.2 Result 2: Genome reorganization following DSB involves conservative loss of connectors around damaged sites

I use the RCL polymer to zoom in on a the process of local genome reorganization following a single DSB induction. The chromatin undergoes local expansion around DSB sites caused by recruitment of repair protein, push aside undamaged DNA, untangle the DNA, and evict histones to facilitate access to damaged sites. To calibrated the number of random connectors in the RCL polymer I use SPT data [9], in which the standard deviation of loci distance before and after DSBs were measures to be 0.14 μm and 0.23 μm respectively. I simulated the RCL model of 100 monomers and computed the variance of the distance of monomers m_{50} and m_{51} before and after the induction of DSB between them and the removal of all cross-links to these monomers (Fig. 2.2E). I find that 130 connectors are in good agreement with measurements in [9]. I find that only 4% of the total connectors are lost following a DSB. The MFET of the two broken ends in simulations were on a scale of 1-2 second, in line with experimental measurements [44, 26].

2.2.3 Result 3: The cross-linked micro-environment of DSB confine the two broken ends

I computed, $\langle d(m_{50}, m_{51})\rangle$, the mean distance between monomer m_{50}, m_{51} as a function of the initial number of connectors in the RCL polymer (Fig.

2.2F). The result curve shows how in the local cross-linked environment serve to confine the two broken ends from drifting apart. These results provides quantitative measure for the local reorganization of the genome following DNA breaks, in terms of loss of cross-links.

2.3 Chapter 5: Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data

In Chapters 3 and 4 I presented simulations of a RCL polymer, where either a combination of stable and random loops are added to it, or a sudden break in the chain occurs. These two situations render calculation of exact solutions to both steady-state and transient statistics difficult. In Chapter 5 I derive analytical expressions for the statistics of RCL polymers, representing the internal connectivity of a single TAD-like region, in which all monomers share a similar average level of connectivity (number of connectors). The resulting analytical expressions I derived are suitable for extracting the mean number of cross links directly from the CC data and construct a polymer, which reproduces the steady-state CC encounter data. This analytical approach allow to replace heavy numerical simulations to extract the polymer's connectivity, by a simple curve fitting of the expressions I derived for the EP to the empirical 5C EP data.

2.3.1 Result 1: The eigenvalues of the RCL random connectivity matrix are linear transformation of the Rouse eigenvalues

I adopt a mean-field approach to compute the spectrum of the random matrix $B(\xi)$, in which I replace the random matrix $B(\xi)$ in Eq. 1.18 by its structural average $\langle B(\xi)\rangle$, where averaging is performed over all possible choices of N_c non NN connected monomer pairs. I constructed the matrix $\langle B(\xi)\rangle$ based on the probability density of the monomer connectivity, which I show to be the hyper geometric. I further show that $\langle B(\xi)\rangle$ commutes with M and, therefore,

diagonalizable by the Rouse eigenbasis [30], such that $\chi = \boldsymbol{V}\langle B(\xi)\rangle \boldsymbol{V}^T$, and $\chi = \text{diag}[\chi_0(\xi), \chi_1(\xi), .., \chi_{N-1}(\xi)]$, are the new RCL eigenvalues, given by

$$\chi_p(\xi) = \begin{cases} 0, & p = 0; \\ N\xi + (1-\xi)4\sin^2\left(\frac{p\pi}{2N}\right), & p > 0. \end{cases} \tag{2.2}$$

The term $N\xi$ in 2.2 is the mean number of connectors for each monomer, which is approximately $\chi_1(\xi)$,for $N \gg 1$.

I, therefore, obtain a decoupled Ornstein-Uhlenbeck (OU) [91] system of equations describing the dynamics of the modes of the RCL polymer

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -d\frac{D}{b^2}\chi\boldsymbol{U} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}t}. \tag{2.3}$$

2.3.2 Result 2: The variance and encounter probability between monomers of the RCL polymer

Using the properties of the OU system 2.3 at steady-state I derived an expression for the variance $\sigma^2_{m,n}(\xi)$ between monomers m and n

$$\sigma^2_{m,n}(\xi) = \begin{cases} \frac{b^2((\zeta_0^{m-n}(N,\xi)-1)^2 - 2\zeta_0^{m+n-1}(N,\xi) + 2\zeta_0^{2m-1}(N,\xi))}{\zeta_0^{2m-1}(N,\xi)(\zeta_0(N,\xi)-\zeta_1(N,\xi))(1-\xi)}, m \geq n; \\ \frac{b^2((\zeta_0^{n-m}(N,\xi)-1)^2 - 2\zeta_0^{m+n-1}(N,\xi) + 2\zeta_0^{2n-1}(N,\xi))}{\zeta_0^{2n-1}(N,\xi)(\zeta_0(N,\xi)-\zeta_1(N,\xi))(1-\xi)}, m < n, \end{cases} \tag{2.4}$$

with

$$\begin{aligned} \zeta_0(N,\xi) &= 1 + \frac{N\xi}{2(1-\xi)} + \sqrt{\left(1 + \frac{N\xi}{2(1-\xi)}\right)^2 - 1}, \\ \zeta_1(N,\xi) &= 1 + \frac{N\xi}{2(1-\xi)} - \sqrt{\left(1 + \frac{N\xi}{2(1-\xi)}\right)^2 - 1}. \end{aligned} \tag{2.5}$$

I derive an approximation for $\sigma^2_{m,n}(\xi)$, when $\xi \ll 1$

$$\sigma^2_{m,n}(\xi) \approx \frac{b^2}{\sqrt{N\xi}}\left(1 - \exp(-|m-n|\sqrt{N\xi})\right). \tag{2.6}$$

The expression for the EP between monomers m and n is, therefore

$$P_{m,n}(\xi) = \left(\frac{d}{2\pi\sigma^2_{m,n}(\xi)}\right)^{\frac{d}{2}}, \tag{2.7}$$

with $\sigma^2_{m,n}(\xi)$ defined in 2.4.

2.3.3 Result 3: The mean-square radius of gyration (MSRG) of the RCL polymer

I further derived an analytical expression for $\langle R_G^2(\xi)\rangle$, the MSRG, which characterizes the size of the RCL polymer

$$\langle R_G^2(\xi)\rangle = \frac{b^2}{N^2(1-\xi)(\zeta_0-\zeta_1)}\Big[\frac{(1+2\zeta_0)N(1+N)}{2\zeta_0} + \frac{N(2(1+\zeta_0)^2-\zeta_0^3)}{1-\zeta_0^2} - \frac{\zeta_0^3(1-\frac{1}{\zeta_0^{2N}})}{(1-\zeta_0^2)^2} + \frac{2(1+\zeta_0)(1-\frac{1}{\zeta_0^N})}{(1-\zeta_0)^2}\Big], \tag{2.8}$$

where here we set $\zeta_0 = \zeta_0(N,\xi), \zeta_1 = \zeta_1(N,\xi)$, and the asymptotic approximation of 2.8 for $N_c(\xi) \ll \frac{N^2}{2}$ is

$$\langle R_G^2(\xi)\rangle \approx \frac{3b^2}{4(1-\xi)\sqrt{N\xi}}. \tag{2.9}$$

2.3.4 Result 4: The Mean square displacement of monomers of the RCL polymer

I obtain an expression for the MSD, which is approximated in three different time-scales based on the polymer relaxation times $\tau_p(\xi) = \frac{b^2}{D\chi_p(\xi)}$

$$\langle\langle r_m^2(t)\rangle\rangle \approx \begin{cases} 2dD_{cm}t + \frac{db^2\mathrm{Erf}[\sqrt{2DN\xi t/b^2}]}{2\sqrt{N\xi(1-\xi)}}, & \tau_{N-1}(\xi) \ll t \ll \tau_1(\xi); \\ \frac{db\sqrt{2dDt}}{\sqrt{\pi(1-\xi)}}\left(1-\frac{\exp(-2dDN\xi t/b^2)}{2}\right), & t \ll \tau_{N-1}(\xi); \\ 2dD_{cm}t + \frac{db^2}{2\sqrt{N\xi(1-\xi)}}, & t \gg \tau_1(\xi), \end{cases} \tag{2.10}$$

where $\mathrm{Erf}[t]$ is the error function. I conclude that the homogeneous behavior of MSD for the RCL polymer model gives an anomalous exponent $\alpha = 0.5$, similar to the Rouse model, and is a result of the mean-field approach used here.

2.3.5 Result 5: The mean first encounter time between monomers of the RCL polymer

I derive and expression for the MFET between any two monomers m and n of the RCL polymer

$$\langle \tau_{m,n}^{\epsilon}(\xi)\rangle = \frac{1}{4\pi D\epsilon}\left(\frac{2\pi\sigma_{m,n}^2(\xi)}{\kappa b^2}\right)^{\frac{d}{2}}, \tag{2.11}$$

and approximated it by

$$\langle \tau^{\epsilon}_{m,n}(\xi) \rangle \approx \frac{b^2 \left(1 - \exp(-|m-n|\sqrt{N\xi})\right)^{d/2}}{4\sqrt{N\xi}\pi D\epsilon(\kappa b^2)^{d/2}} + \mathcal{O}(N\xi), \tag{2.12}$$

where $|m-n| \ll N$, and $\xi \ll 1$.

Taken together, the expressions I derived in Results 1-5 provides rich set of tools to construct a RCL polymer directly from the CC data, where only one parameter ξ, the connectivity fraction remains free and is found by fitting expression 2.7 to the CC encounter data. Once ξ is obtained, computation of steady-state length scales (Eq. 2.8) and encounter times (Eq. 2.11) are straightforward. Results 1-5 were verified by Brownian simulations, which showed in excellent agreement with the theory.

2.4 Chapter 6: Chromatin reorganization during cell differentiation captured by randomly cross-linked polymer models of multiple topologically associating domains

The internal connectivity within TADs, as seen in the results of Parts I-III of this dissertation, affects the dynamic behavior of the polymer and its compaction. However, connectivity across TADs, sparse as it may be (Fig. 1.2A), must be taken into account in polymer models, to arrive at a more complete description of the folding pattern and dynamic of the chromosome. In the Chapter 6 of this thesis I present a generalization of the analytical RCL model of one TAD (Chapter 5) to multiple TADs of variable size, inter and intra-connectivities. I use the generalized model to study the affect of inter-TAD connectivity on the statistical properties of the chromatin throughout 3 stages of cell differentiation.

I generalize the construction of single TAD-like RCL polymer to N_T diagonal blocks, by chaining N_T RCL polymers of N_i monomers each. The construction of a single TAD in Section III is generalized into block matrices representations:

$$\boldsymbol{R} = \begin{bmatrix} [R^{(1)}] \\ [R^{(2)}] \\ . \\ . \\ . \\ [R^{(N_T)}] \end{bmatrix}, \boldsymbol{U} = \begin{bmatrix} [U^{(1)}] \\ [U^{(2)}] \\ . \\ . \\ . \\ [U^{(N_T)}] \end{bmatrix}, \boldsymbol{M} = \begin{bmatrix} [M_1] & 0 & 0 & ... & 0 \\ 0 & [M_2] & 0 & ... & 0 \\ 0 & 0 & [M_3] & ... & 0 \\ . & . & 0 & ... & . \\ . & . & & ... & . \\ 0 & . & . & ... & [M_{N_T}] \end{bmatrix}, \quad (2.13)$$

$$\boldsymbol{V} = \begin{bmatrix} [V_1] & 0 & 0 & ... & 0 \\ 0 & [V_2] & 0 & ... & 0 \\ 0 & 0 & [V_3] & ... & 0 \\ . & . & 0 & ... & . \\ . & . & & ... & . \\ 0 & . & . & ... & [V_{N_T}] \end{bmatrix}, \Lambda = \begin{bmatrix} [\Lambda_1] & 0 & 0 & ... & 0 \\ 0 & [\Lambda_2] & 0 & ... & 0 \\ 0 & 0 & [\Lambda_3] & ... & 0 \\ . & . & 0 & ... & . \\ . & . & & ... & . \\ 0 & . & . & ... & [\Lambda_{N_T}] \end{bmatrix}. \quad (2.14)$$

where each $[M_i]$ is a Rouse matrix 1.4 of a chain of N_i monomers, and $[\boldsymbol{U}^{(i)}] = [u_0^{(i)}, u_1^{(i)}, ..., u_{N_i}^{(i)}]$ and $[\Lambda^{(i)}]$ represent the normal coordinates and eigenvalues for chain i, respectively. Corresponding to the N_T blocks, the connectivity fraction $\Xi = \{\xi_{ij}\}$ is represented by a symmetric $N_T \times N_T$ matrix, from which we obtain the block matrix of added random connectivity $\langle B(\Xi) \rangle$.

2.4.1 Result 1: The distribution of the square radius of gyration in each TAD is approximately Normal

The distribution $P(R_g^2)$ of the square radius of gyration (SRG) for each TAD is given by [34]

$$P(R_G^2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(\mathtt{i_c} \beta R_g^2) \mathrm{Det}(\mathbf{1}_{N_k - 1} + \frac{\mathtt{i_c} d\beta b^2}{2N_k} [\Gamma_k]^{-1})^{-d/2} \mathrm{d}\beta, \quad (2.15)$$

where $\mathtt{i_c}$ is the complex unit, $\mathbf{1}_{N_k-1}$ is $N_k - 1 \times N_k - 1$ matrix of ones, $[\Gamma_k]$ is the block diagonal matrix of eigenvalues for TAD k except the first, and Det is the determinant operator. I approximate the integral in 2.15 to obtain the distribution of the SRG

$$P(R_G^2) = \sqrt{\frac{dN_k^2}{4\pi b^4 Tr([\Gamma_k]^{-2})}} \exp\left(- \frac{\left(R_g^2 - \frac{b^2 \mathrm{Tr}([\Gamma_k]^{-1})}{N_k} \right)^2}{4b^4 Tr([\Gamma_k]^{-2})/dN_k^2} \right), \quad (2.16)$$

where Tr is the trace operator. I then computed the mean SRG (MSRG) of TAD k and obtained

$$\langle R_g^2\rangle^{(k)} \approx \frac{b^2}{(1-\xi_{kk})(\zeta_0^{(k)}(\Xi)-\zeta_1^{(k)}(\Xi))}, \tag{2.17}$$

which is a Gaussian distribution, and the terms

$$\begin{aligned}
\zeta_0^{(k)}(N,\Xi) &= y^{(k)}(N,\Xi)+\sqrt{y^{(k)}(N,\Xi)^2-1},\\
\zeta_1^{(k)}(N,\Xi) &= y^{(k)}(N,\Xi)-\sqrt{y^{(k)}(N,\Xi)^2-1},\\
y^{(k)}(N,\Xi) &= 1+\frac{\sum_{k=1}^{N_T}\xi_{kj}N_j}{2(1-\xi_{kk})},
\end{aligned} \tag{2.18}$$

couple the connectivities ξ_{kj} of all TADs $j=1..N_T$ connected to TAD k.

2.4.2 Result 2: The encounter probability of monomers within and between TADs

I first show that the equation of motion in normal coordinates $u^{(k)}$ are decoupled for all internal modes

$$\frac{\mathrm{d}u_m^{(k)}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(\lambda_m^{(k)}(1-\xi_{kk})+\sum_{j=1}^{N_T}N_j\xi_{kj}\right)u_m^{(k)}+\sqrt{2D}\frac{\mathrm{d}\eta_m^{(k)}}{\mathrm{d}t}, \tag{2.19}$$

whereas the centers of masses $u_0^{(k)}$, remain coupled.

$$\frac{\mathrm{d}u_0^{(k)}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(N_j\xi_{kj}u_0^{(k)}-\sum_{j=1}^{N_T}\xi_{kj}\sqrt{N_kN_j}u_0^{(j)}\right)+\sqrt{2D}\frac{\mathrm{d}\eta_0^{(k)}}{\mathrm{d}t}, \tag{2.20}$$

where $\eta_0^{(k)}$ are standard Brownian motions.

I then obtained an expression for variance between monomers m and n of TAD k using the steady-state properties of Eq.2.19 and the MSRG (Eq. 2.17)

$$\sigma_{m,n}^2(\Xi) = \begin{cases}\langle R_G^2\rangle^{(k)}\left(\frac{(\zeta_0^{(k)}(N,\Xi)^{m-n}-1)^2-2\zeta_0^{(k)}(N,\Xi)^{m+n-1}}{\zeta_0^{(k)}(N,\Xi)^{2m-1}}+2\right), m\geq n;\\ \langle R_G^2\rangle^{(k)}\left(\frac{(\zeta_0^{(k)}(N,\Xi)^{n-m}-1)^2-2\zeta_0^{(k)}(N,\Xi)^{m+n-1}}{\zeta_0^{(k)}(N,\Xi)^{2n-1}}+2\right), m<n.\end{cases} \tag{2.21}$$

I further obtain an approximate expression for the variance between monomers $m^{(k)}$ and $n^{(j)}$of TADs k and j respectively

$$\begin{aligned}
\sigma^2_{m^{(k)}n^{(j)}}(\Xi) &= \langle R_G^2\rangle^{(k)}(1+\zeta_0^{(k)}(N,\Xi)^{1-2m})+\langle R_G^2\rangle^{(j)}(1+\zeta_0^{(j)}(N,\Xi)^{1-2n})\\
&+ b^2\left(\frac{1}{N_i\sum_{l\neq k}^{N_T}N_k\xi_{kl}}+\frac{1}{N_j\sum_{l\neq j}^{N_T}N_k\xi_{jl}}\right).
\end{aligned} \tag{2.22}$$

I obtain an expression for the EP within and between TADs by substituting 2.21 or 2.22 in

$$P^{(k)}(m,n) \propto \left(\frac{d}{2\pi\sigma^2_{m^{(k)},n^{(j)}}(\Xi)} \right)^{d/2}. \tag{2.23}$$

Expression 2.23 is confirmed by numerical simulations of a generalized RCL polymer harboring 3 TADs (Chap. 6, Fig. 6.3B).

2.4.3 Result 3: TADs of the X chromosome compact and de-compact synchronously throughout differentiation

I use the generalized RCL model to study the reorganization of TADs throughout cell lineage commitment in 3 subsequent stages of cell differentiation 1) embryonic stem cell 2) neuronal precursors, and 3) embryonic fibroblast (Chap. 6 Fig. 6.4). I fit expression 2.23 to the 5C data of TADs D, E, and F [76], coarse-grained at 6 kbp resolution, and obtain the matrix Ξ, the number of connectors within and between TADs, for all three stages of differentiation (Fig. 6.4B). I computed the radius of gyration for each TAD using the fitted Ξ, and found that the 3 TADs compact in the transition from embryonic stem cells to neuronal precursors, and then de-compact in the differentiation to embryonic fibroblasts (Fig. 6.4C). This compaction is associated with acquisition of connectors within and between TADs. This result shows the synchronous activity of TADs, where the inter-TAD affects their compaction. I thus demonstrate the applicability of the generalized RCL in modeling structural reorganization in the chromatin throughout differentiation, in which inter-TAD connectivity plays a significant role, and show how the RCL model can interpolate between 5C snapshots of chromosome organizations (Fig. 6.4D).

3

Transient chromatin properties revealed by polymer models and stochastic simulations constructed from Chromosomal Capture data

Published in Shukron Ofir, and David Holcman. "Transient chromatin properties revealed by polymer models and stochastic simulations constructed from Chromosomal Capture data." PLoS computational biology 13.4 (2017): e1005469.

Abstract

Chromatin organization can be probed by Chromosomal Capture (5C) data, from which the encounter probability (EP) between genomic sites is presented in a large matrix. This matrix is averaged over a large cell population, revealing diagonal blocks called Topological Associating Domains (TADs) that represent a sub-chromatin organization. To study the relation between chromatin organization and gene regulation, we introduce a computational procedure to construct a bead-spring polymer model based on the EP matrix. The model permits exploring transient properties constrained by the statistics of the 5C data. To construct the polymer model, we proceed in two steps: first, we introduce a minimal number of random connectors inside restricted regions to account for diagonal blocks. Second, we account for long-range frequent specific genomic interactions. Using the constructed polymer, we compute the first encounter time distribution and the conditional probability of three key genomic sites. By simulating single particle trajectories of loci located on the constructed polymers from 5C data, we found a large variability of the anomalous exponent, used to interpret live cell imaging trajectories. The present polymer construction provides a generic tool to study steady-state and transient properties of chromatin constrained by some physical properties embedded in 5C data.

3.1 Introduction

Chromatin is organized in heterogeneous sub-regions of various sizes, as recently revealed by Chromosome Capture (5C) data [24, 31]. This multiscale organization is generated by short and long-range genomic interactions between DNA segments, observed in the statistics of a large number of cells. Mammalian chromatin at a resolution of 3kb, [76, 29] contains an organization at 1Mbp scale, where several sub-structures are enriched with intra-connectivity, reflecting an increased encounter probability (EP) between genomic segments. This increased EP is described in the two-dimensional encounter frequency (EF) matrix, containing diagonal blocks called Topologically Associating Domains (TADs) [76, 39]. TADs are associated with gene regulation [76], DNA replication timing [84], DNA entanglement or cross-linking by molecules such as cohesin, CTCF [81] and condensin [76]. Cross-linking between chromatin sites are precisely the events sampled by Chromosome Capture data (3C, 4C, 5C, HiC) [76, 61], and single cell HiC confirms that positions of cross-links can vary between cell types and phases [73]. In that context, TADs represent average chromatin conformations, characterized by a higher numbers of binding molecules compared to non-TAD regions.

Over the past ten years, polymer models have been used to analyze statistics hidden in Chromosome Capture (CC) data and to characterize the decay of the encounter probability with the genomic distance s. For example, the EP between two monomers A, B for a (linear) Rouse polymer decays with $s^{-3/2}$, thus the exponent is 3/2 [30]. A range of decay exponents lower than $3/2$ can be produced by other polymer models, where stable transient loops are formed between segments [16]. By varying the number of loops, a large range of chromatin configurations can be generated and the associated polymer characteristics are reflected in the EP decay exponent. By including transcriptional information, dynamic-loop model [52, 16] can reproduce chromatin looping associated with transcriptional activity. Similar polymer models describe chromosomal territories [18], suggesting that inactive genes are located inside these territories. The strings-and-binders-switch polymer model consists of reversible binding between specific genomic segments located in close proximity, when an additional diffusing molecule is present at the binding site. In an extension of this model, genomic segments having similar epigenomic state can directly interact [55], revealing the phase diagram of polymer configurations. These models are used to interpret the contact probability decay in the Hi-C data [83, 75, 56, 77]. In a new class of polymer model, each monomer interacts with any other through a potential well [39, 102], where pairwise interaction between monomers is represented by either an attractive or repulsive potential. The parameters of the model are extracted from data by minimizing the chi-square norm between the EP empirical and Monte-Carlo simulation matrices. However, this highly connected model does not translate easily into molecular binding because the nature of these

potential wells does not have a direct physical interpretation. When this model is applied to the region containing the X inactivation center of mammalian embryonic stem cells, it predicted novel ensemble of polymer configurations, representing TAD structures present in the 5C data. Other minimization procedures were used to inter-chromosomal distances at a large-scale resolution of 1Mbp [104].

Polymer models have also been used to interpret single particle trajectories (SPTs) of tagged DNA locus [28, 38, 92, 20, 51, 50, 7], revealing that chromatin is constantly remodeled. SPTs are characterized by their anomalous behavior, which can deviate significantly from classical diffusion and are usually quantified by the mean-square displacement (MSD). In that context, a model with a minimal number of parameters is still needed to account for both types of data: 1) the EPs decay rate and 2) dynamical parameters such as the anomalous exponent extracted from SPTs. In the absence of a systematic procedures to convert the EP into a polymer model with a similar EP decay rate as the one presented in the 5C data, the connection between 1) and 2) was left open.

We present here a general computational and algorithmic procedure to estimate parameters of a randomly cross-linked polymer model following the 5C protocol. The procedure consists in constructing an ensemble of polymer models from the EP of 5C by randomly cross-linking monomers and in resolving the difficulty of assigning the minimal number of sparse interactions between monomer pairs. These interactions can be directly interpreted as binding molecules. The construction of the polymer model starts with the Rouse model [30], which consists of beads linearly connected by harmonic spring. We started with the coarse-grained Rouse polymer that describes accurately the statistics of the chromatin below a scale of few Mbp [88, 17]. To further constrain monomer interactions, we determine monomer connectivity from the 5C data of mammalian X chromosomes. The construction procedure is divided into two steps: first, to account for heterogeneity in the 5C data, we added a minimal required number of connectors (cross-links) between genomic sites chosen at random that can reproduce TAD blocks. We show that the number of random connectors to be added is uniquely determined from data. However, this step is insufficient to recover the EP decay peaks contained in the 5C data. Thus, in the second step, we account for consistent long-range interaction present in the EP matrix within and between TADs. We calibrate our model by requiring that the EP matrix, constructed from simulations of the polymer, has the same decay exponent (for each monomer) as that of the empirical data. These calibrated polymer models allow us to study transient properties and to estimate the conditional encounter probability and the first encounter time between three specific genomic sites.

Although the generalized Gaussian polymer model we are using here is well known, our reconstruction using minimal number of short and long-range connectors, derived from empirical EP, is new and permits generating novel statistics describing transient gene regulation. By exploring the statistics of simulated SPTs

(by computing the anomalous exponent [28, 38, 92, 20]), we further show that the large heterogeneity of the anomalous exponent present in live cell imaging can be explained by random binding locations on the chromatin that can vary from cell to cell.

3.2 Results

3.2.1 The encounter probability of coarse-grained 5C data

We previously described how we construct a polymer model from a symmetrized 5C matrix M (Fig 3.1A). By symmetrizing the EP matrix, we averaged-out asymmetrical fluctuations. The 5C data we used represent a sub-region of the X chromosome ($\approx 92kbp$), that was previously segmented into two regions called Topological Associating Domains (TADs) D and E [76]. The matrix M was further coarse-grained by binning the encounter frequencies into 307 monomers of $3kbp$ [39], where TAD D (resp. TAD E) is represented by the first 106 monomers (resp. 107-307), as shown in Fig 3.1A. We introduce a general polymer model (Fig 3.1B) with arbitrary configuration, the properties of which will be extracted from empirical 5C matrix M.

The encounter probabilities between monomer m and monomer n are computed from the experimental 5C matrix M (see Material and Methods) by

$$P_e(|m-n||n) = \frac{M_{n,n+|m-n|} + M_{n,n-|m-n|}}{\sum_{m=1}^{N} M_{n,m}}, \tag{3.1}$$

which depend on the genomic distance $|m-n|$ (Fig 3.1C). Although the average encounter probabilities decays with $|m-n|$ for each n, they contain peaks that reflect consistent long-range interactions between monomers. To quantify the decay of the EP, we fitted its average value $P_e(|m-n|) = \frac{1}{N}\sum_{n=1}^{N} P_e(|m-n||n)$ (black dotted line in Fig 3.1C) with the function

$$\tilde{P}(|m-n|) = \frac{C}{|m-n|^{\beta}}, \tag{3.2}$$

where C and $\beta > 0$ are two constants. For a Rouse polymer, the EP function $\tilde{P}$ is characterized by a decay exponent $\beta = 3/2$ [30]. Fitting 3.2 to data, revealed that $\beta = 0.77$, from which we concluded that the polymer model should be modified to account for higher compaction than allowed by a Rouse polymer [5].

To better account for the heterogeneity in the EP of each monomer, we plotted the distributions of the exponent β_n for $n = 1..307$ along the polymer (Fig 3.1D blue dots). The exponents β_n were extracted by fitting the function 3.2 to the empirical EPs 3.1. The large variability in β_n, $n = 1..307$ reflects the local heterogeneity of the chromatin architecture at the current scale (a monomer represents 3kbp).

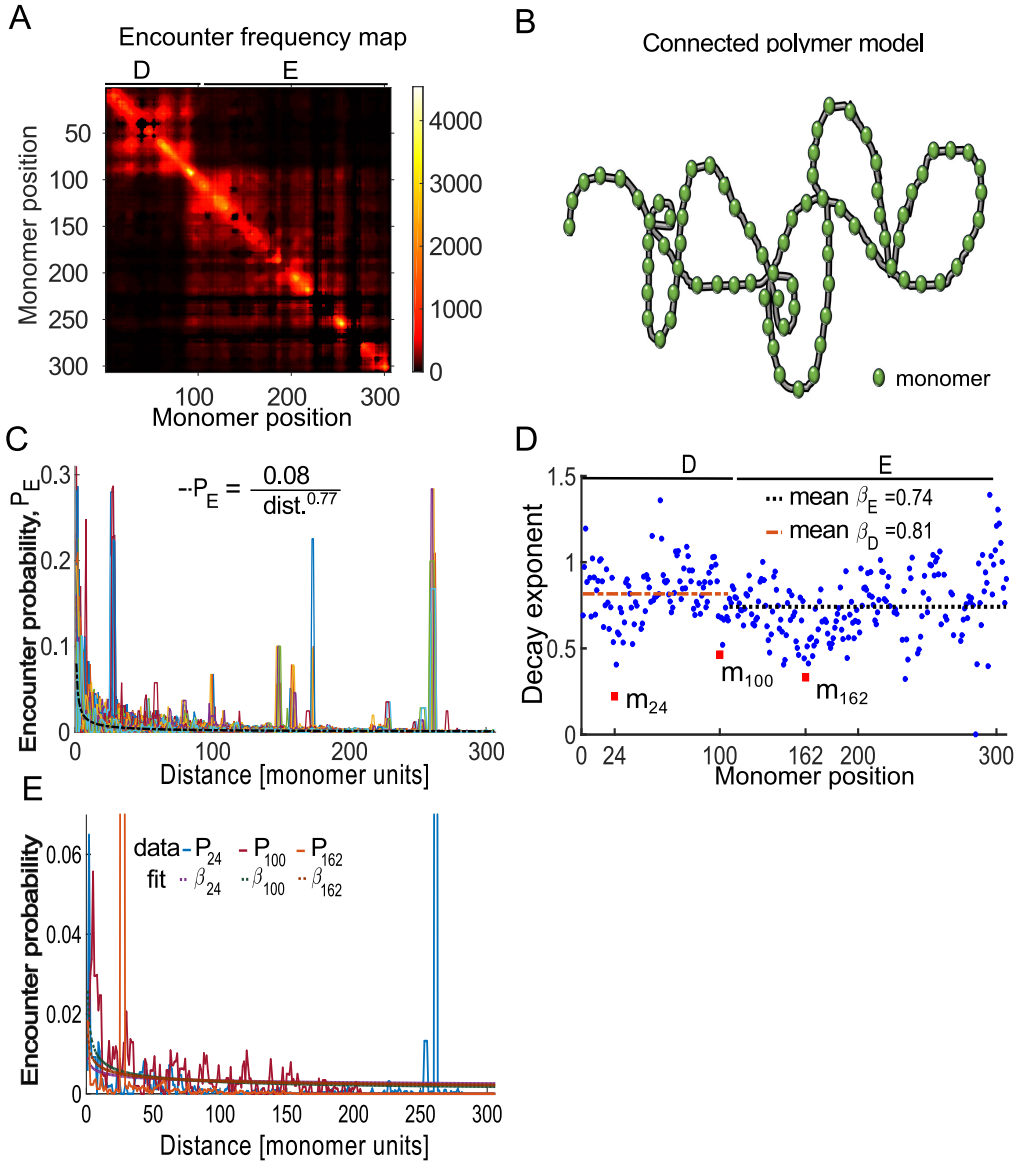


Figure 3.1 **Statistics of Conformation Capture data**. **A.** Average encounter frequency map of two 5C replica spanning ≈1Mbp genomic region containing two Topologically Associating Domain (TAD) D (monomers 1-106) and TAD E (monomers 107-307) [76], where the map was coarse-grained into 307 monomers of size 3kbp [39]. **B.** Schematic representation of a polymer model with randomly connected monomers. **C.** Empirical encounter probability P_n, for monomer n plotted with respect to the genomic distance d [monomer units], reveals long-range interactions (localized peaks). P_n are fitted with functions $Ad^{-\beta}$, where β is the decay exponent and A the normalization factor. For the mean encounter probability $\bar{P}$, the value of the parameters are $A = 0.08$ and $\beta = 0.77$ (thick red curve). **D.** Distribution of the β_n exponents $(n = 1..307)$ (blue dots): Monomers m_{24}, m_{100}, m_{162} (red square dots) with $\beta_{24} = 0.22$, $\beta_{100} = 0.46$, and $\beta_{162} = 0.33$, respectively, accounts for high peaks (first and last), while the middle one corresponds to the boundary between TADs. **E.** Encounter probability P_n for monomers n=24, 100, and 162, corresponding to local minima shown in box D.

The average value β for TAD D and E was found to be $\beta_D = 0.81$ and $\beta_E = 0.74$, respectively. The local minima of β deviated significantly from the mean values (Fig 3.1D red squares), where the deviation is 2-3 times the standard deviation computed from the β values in each TAD. The minima may represent chromatin features (Fig 3.1E) such as specific long-range interactions or boundary between chromatin sub-domains. Indeed, point m_{100} (around monomers 102-107 in Fig 3.1D red) is located at the boundary between TAD D and E, while m_{24} and m_{162} are characterized by strong long-range interactions (Fig 3.1E).

To conclude, the distribution of β values extracted from the EP is quite heterogeneous, which can disclose chromatin subregions and long-range strong interactions. We shall account in the next two sections for these characteristic features and include in the polymer model both random and persistent long-range connections between monomers.

3.2.2 Encounter probability in the random loop polymer model

To determine the level of connectivity of the generalized Rouse polymer which reproduce the EP-decay using a prescribed exponent β, we first studied the case of one TAD-like region using the simulation of a 307 monomer chain. For each realization, we added connectors between random non-nearest neighbor monomer-pairs in the subregion 103-203 (Fig 3.2A). The subregion 103-203 occupies the mid part of the 307 monomer in this synthetic example, and does not translates into biological meaningful subregion. The number of connectors, or the connectivity percentage ξ (fraction of the number of connected monomer-pairs to the maximum, described in Materials and Methods), was increased in the range $0 - 2\%$. By adding connectors, the EP between distant monomers has increased, as presented in the EP-matrix (Fig 3.2B). In contrast, outside the region 103-203 the EPs were similar to the case $\xi = 0$ (linear chain), showing that the connected region did not affect the EP in the non-connected ones. At this stage, we have shown that adding random connectors allows recovering the shape of TAD regions.

To find the minimal numbers of connectors necessary to recover a given TAD, we aim at elucidating the relationship between the connectivity percentage ξ and the decay exponent β. For that purpose, we simulated an ensemble of polymers to their relaxation time (Materials and Methods), and used the equilibrium configuration to estimate the EP of each monomer for $\xi \in [0, 2]\%$. We calculated the exponent β by fitting the function 3.2 to the simulated encounter probability data: the values of β_n for $n \in [103, 203]$ decreased with ξ. Indeed, for $\xi \approx 0.2\%$, the coefficients β_n decreases below the Rouse exponent (equals to $\beta_{Rouse} = 1.5$), indicating compact polymer configurations. For $\xi = 2\%$, the mean decay exponent β_n and $n = 103 - 203$ was $\bar{\beta} = 0.47$, with a minimal value 0.42 obtained for the boundary monomers

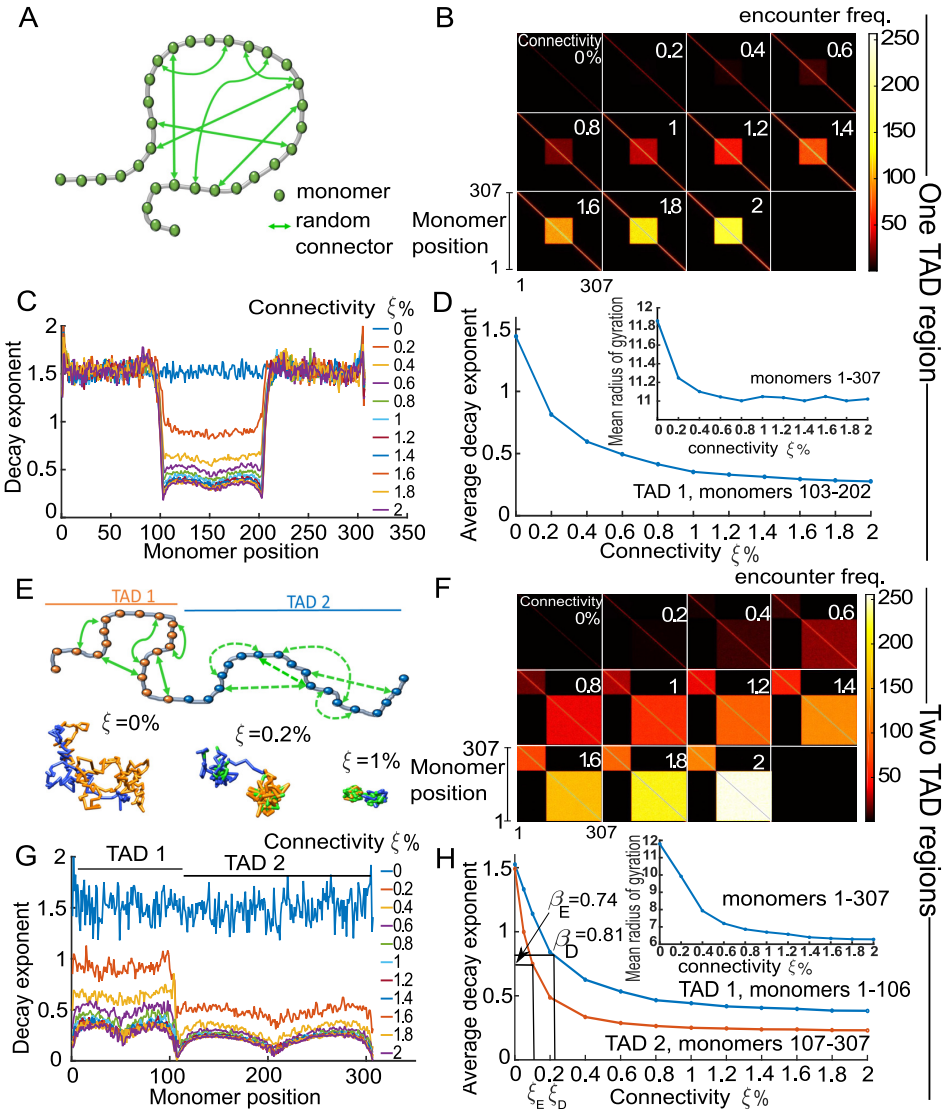


Figure 3.2 **Statistics of simulated generalized Rouse polymer chain for various connectivity in one and two sub-regions A.** Schematic bead-spring chain connected at random positions (two-sided green arrows) between non-nearest-neighbor monomers. **B.** Encounter frequency maps of a 307 monomers chain. Connectors are added randomly between monomers 103-202 for each realization. The connectivity ξ (number of connectors) increases from 0 to 2%. **C.** Distribution of β_n $(n = 1..307)$ fitted by the function $Ad^{-\beta}$, where d is the distance along the chain [monomer units], to the encounter probabilities of numerical simulation. **D.** Average value of β for monomers in the interval 103-202 with respect to the connectivity percentage ξ. **E.** Schematic polymer chain, where two defined regions: monomers 1-106 (TAD 1, orange circles) and monomers 107-307 (TAD 2, blue circles), are randomly connected (green arrows). No connections were added between the two TAD regions. Lower panel: three snapshot realizations of a random loop chain with TAD 1 (orange) and TAD 2 (blue) and random connectors (green) for three increasing values of connectivity $\xi = 0, 0.2, 1\%$. **F.** Encounter frequency maps showing two TAD regions for an increasing number of random connectors. **G.** Distribution of $\beta-$ exponent for $\xi \in [0, 2]\%$, showing the border (n=106) effect between TADs. **H.** Average β over TAD 1 (blue) and TAD 2 (orange) for $xi \in [0, 2]$. The curves decrease until plateau at 0.42 (0.24) for TAD 1 (resp. TAD 2). We use these curves to recover the connectivity percentage ξ of the experimental TAD D, with $\beta_D = 0.74$ (resp. TAD E with $\beta_E = 0.81$) for which $\xi_D = 0.23\%$ (resp. $\xi_E = 0.12\%$.)

103 and 203 (Fig 3.2C). These results confirm that polymer condenses when ξ increases, as confirmed by computing the mean radius of gyration R_g [30], showing a decay from $R_g = 11.9$ for $\xi = 0$ to $R_g = 11$ at $\xi = 2\%$ (measured for 307 monomers, Fig 3.2D). Values of β outside the TAD region (monomers 1 to 102 and 204-307) were mostly unchanged, fluctuating around $\beta = 1.5$, confirming that statistical properties of a Rouse chain are unaltered when connectors are added to the middle region ($n = 103 - 203$). Finally, the average value of β_n (computed over $n = 103 - 203$) versus ξ is shown in Fig 3.2D and, as we shall see, will serve to extract the connectivity percentage ξ from the empirical data.

To reproduce the two TADs D and E of the X-chromosome, we started with a polymer of 307 monomers and added connectors randomly (green arrows) between monomers 1-106 and between monomers 107-307, as described in Fig 3.2E upper panel). This partition follows the empirical TAD segmentation described in [76, 39] at the scale 3kbp of the polymer model. Three polymer realizations for $\xi = 0, 0.2, 1\%$ are shown in Fig 3.2E bottom panel, showing polymer condensation into two distinct regions. The EF matrix shows that two TAD-like regions, named TAD1 and TAD2 (Fig 3.2F), emerge as the connectivity ξ increases from 0 to 2%. To extract the exponent β (Fig 3.2G, colored curves) we fitted the function 3.2 to the EP matrix for each monomer inside TAD1 and 2. In both cases, the exponent β decreased below $\beta_{Rouse} = 1.5$ ($\xi = 0\%$ blue curve) and for $\xi = 0.2\%$, 11 and 39 random connectors were added for TAD1 and TAD2, respectively. The boundary between TADs is characterized by an abrupt decay of the β value, reflecting high long and short-range encounters. The average β exponent (averaged over each TAD), plotted with respect to the connectivity ξ (Fig 3.2H), was used to determine the number of connectors necessary to reconstruct the empirical data. Indeed, we extracted from the EP matrix (Fig 3.1A) that $\beta_D = 0.81, \beta_E = 0.78$ the associated connectivity percentages $\xi_D = 0.12, \xi_E = 0.23$, respectively (Fig 3.2H). To conclude, we obtain the minimal number of random connectors to be added on a generalized Rouse polymer such that the decay exponents of the reconstructed and empirical EP-matrix are as close as possible.

3.2.3 Incorporating long-range empirical interactions in the polymer model

A key feature present in the 5C EF-matrix (Fig 3.1A) is the ensemble of persistent long-range interactions between monomers (Fig 3.1C). To account for these interactions, we connected monomers corresponding to off-diagonal local maxima of the EF matrix, for which their EP exceeds that of nearest neighboring monomers threshold (Materials and Methods). We found 24 long-range connections: 7 (resp. 13) within TAD D (resp. E) and 4 across the two (see SI for the list of monomers pairs) as shown in Fig 3.3A. We adjusted the spring constant $\kappa_{m,n}$ between monomer m and

n corresponding to consistent long-range interactions, based on the values of their EPs (Materials and Methods). The scaled coefficient $\kappa_{m,n}$ are summarized in the SI, and were found to be 1.1 to 3 times higher than the spring constant assigned to connectors of the linear backbone ($\kappa = 0.97$). As revealed by simulations of a Rouse polymer with fixed long-range connectors, the polymer configuration are condensed, characterized by a radius of gyration of about $R_g = 9.1$ (compared to $R_g = 12$ for the Rouse chain). Three realizations of the polymer are shown in Fig 3.3A.

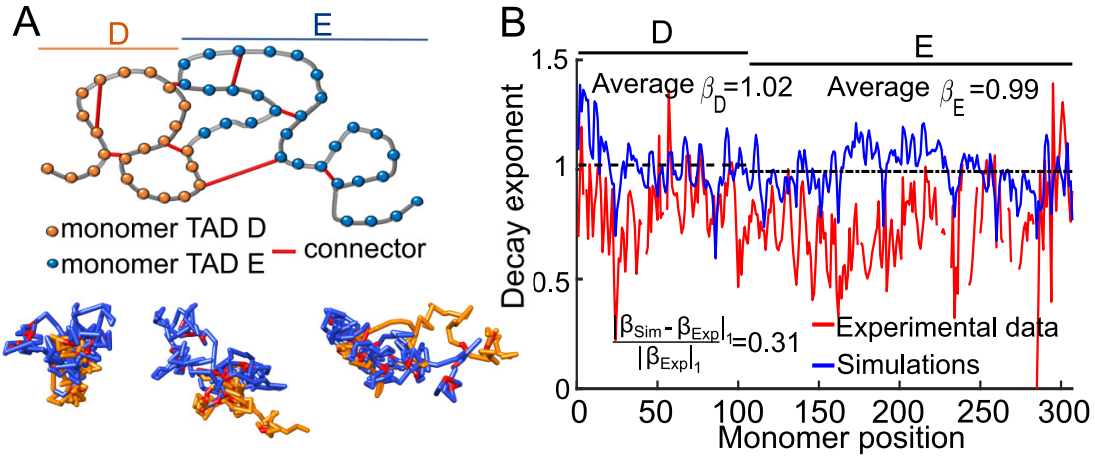


Figure 3.3 **Effect of persistent long-range connectors on polymer folding. A.** Upper panel: schematic representation of the bead-spring polymer model with added fixed connectors (red) representing specific long-range monomer interactions (peaks) shown in Fig. 3.1C. Lower panel: three different realizations of the same polymer, showing TAD D (orange), TAD E (blue), and fixed connectors (red). **B.** Simulated (blue) and experimental (red) β exponent of the fitted encounter probability. The polymer model contains only specific long-range interactions. Average β values for TAD D and E are $\beta_D = 1.02$ (resp. $\beta_E = 0.99$).

To quantify the effect of adding consistent long-range connections on the EP decay, we simulated a Rouse chain containing 307 monomers with the addition of fixed monomer connectivity extracted from the peaks of the empirical EP-matrix. We computed the decay exponents β_n of each monomer by fitting the function 3.2 to the EPs from simulations, and compared it to the ones computed from the experimental data (Fig 3.3B). To estimate the quality of the reconstruction, we use the L_1-norm, defined for a function f by $|f|_1 = \sum_k |f(k)|$, and computed the difference between the experimental β_{Exp} and simulated β_{Sim} curves normalized by the norm $|\beta_{Exp}\|_1$ and we find

$$\frac{\|\beta_{Exp} - \beta_{Sim}\|_1}{\|\beta_{Exp}\|_1} = 0.312. \tag{3.3}$$

The experimental values β_n (Fig 3.3B red) were generally lower than the ones obtained from simulations (blue), indicating that the reconstructed chromatin polymer is less condensed for both TADs. The mean β values for TAD D and E were quite

similar with $\beta_D = 1.02$ and $\beta_E = 0.99$. Therefore, we conclude that long-range connectors are insufficient to reproduce the statistics of the 5C data.

3.2.4 Combination of random loops and long-range interactions to construct a polymer model of a TAD

We previously evaluated separately the effect of adding random connectors and fixed long-range interactions on the EPs. We computed the decay exponent β and compared it with coarse-grained 5C data. We now combine these two constraints, such that specific and non-specific connectors are added to each realization of a generalized Rouse polymer (Fig 3.4). We first find the connectivity percentage matching that of the experimental data in each TAD $\xi_D = 0.23\%$ and $\xi_E = 0.12\%$. These values summarize the contribution from the two types of connectors (Fig 3.2H). For long-range specific interactions (Fig 3.3), we previously obtained $\beta_D = 1.02, \beta_E = 0.99$, corresponding to $\xi_D = 0.12\%, \xi_E = 0.07\%$(Fig 3.2H) for TAD D and E, respectively. Therefore, we attributed the remaining percentages to the addition of random connectors, that is $\xi_D = 0.11, \xi_E = 0.05$. The number of added random connectors corresponding to $\xi_D = 0.11\%, \xi_E = 0.05\%$ are 6 and 10 in TAD D and E, respectively.

To reconstruct the polymer, we started with a Rouse chain (Fig 3.4A (gray)) and added 24 connectors between monomer pairs corresponding to peaks of the EP matrix (red). Three polymer realizations, simulated with the two types of connectors, are shown in Fig 3.4B, characterized by a radius of Gyration $R_g = 6.4$. We computed the EF-matrix (Fig 3.4C) that showed similarity with the experimental data (compare Fig 3.1A with Fig 3.4B), for which two TAD-like structures are visible. We find a satisfactory agreement between simulations and experimental data, measured by the Kolmogorov-Smirnov distance $D_{Max} = 0.06$ computed on the cumulative density function between the reconstructed and experimental data, as shown in S1 Fig).

We further quantified the similarity between the two matrices by comparing the decay exponent for each monomer β_n,$(n = 1..307)$ from numerical simulations to those of the experimental data. We used the function 3.2 to fit the EP of monomers $1 - 307$ after long time polymer simulations (Materials and Methods). The fitted value for β shows an good agreement with the experimental β values (Fig 3.3C). Comparing the normalized difference between β signals we find $\|\beta_{Exp} - \beta_{Sim}\|_1 / \|\beta_{Exp}\|_1 = 0.2$ (compared with 0.312 Fig 3.3B). To conclude, accounting for long-range deterministic and short-range stochastic interactions leads to a more accurate polymer model for chromatin reconstruction. The quality of this approximation is measured by the decay norm of the β exponent between the data and the simulations, also confirmed by the Kolmogorov-Smirnov distance as shown in S1 Fig C and to be compared with S1 Fig D.

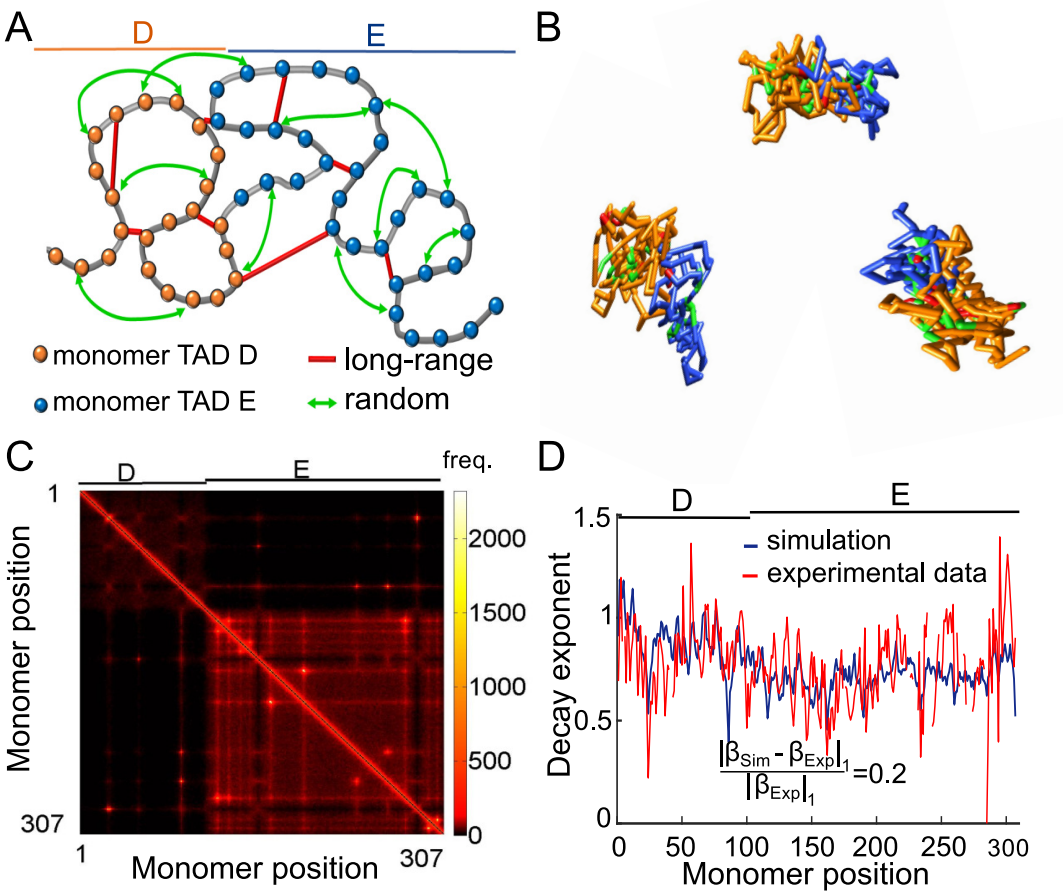


Figure 3.4 **Coarse-grained reconstruction of chromatin using extracted random loops and connectors corresponding to peaks of the 5C data.** **A.** Schematic polymer model, where TAD D (orange, monomers 1-106), and TAD E (blue, monomers 107-307) are recovered by random loops (green arrows) using the connectivity ξ and persistent long-range connectors (red bars). **B.** Three realizations of the polymer model. **C.** Encounter frequency matrix of the simulated polymer model, showing two TADs where off-diagonal points indicate fixed connectors. **D.** Comparison between β computed from experiments and simulations data.

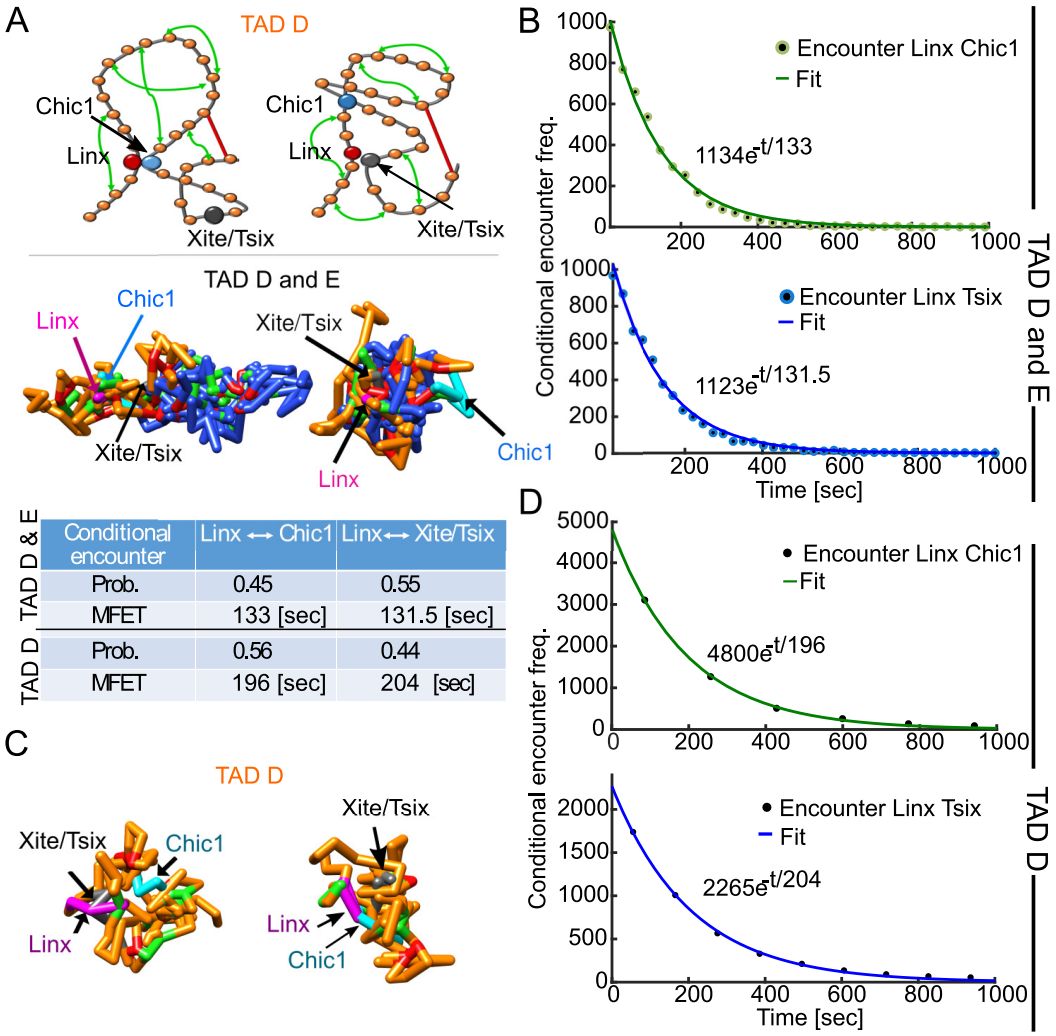
3.2.5 Encounter probabilities and distribution of search times of three genomic sites

We showed previously how to construct a polymer, the statistical properties of which match the ones extracted from 5C data. However, the 5C data cannot be used to study transient properties of the chromatin, as they represent a static genomic encounter interactions averaged over cell population. We shall now use the calibrated polymer described above, constrained by the steady-state properties of the 5C data (Fig 3.5A), to evaluate transient properties of the chromatin.

We focus here on the 5C data harboring the Xist loci, which is the master regulator of X chromosome inactivation (XCI), and its antisense transcript, Tsix, which plays a key role in modulating Xist expression during mouse development [39, 76]. Tsix is believed to play a key role in the choice of the Xist allele that will be expressed during X inactivation. Thus, we decide to estimate how chromatin conformation within TADs can contribute to this transcriptional variability using the present polymer model. We estimated the first encounter time distribution and the probability that monomer 26 (position of the Linx) meets monomer 87 (Xite) before monomer 62 (Chic1). These monomers represent three key sites on the X chromosome [39, 76], located in TAD D. We show three realizations in Fig 3.5A and indicate the location of the three sites inside TAD D (yellow) and E (blue).

We started the polymer simulations from the steady-state distribution and performed around 10,000 runs. As predicted by the narrow escape theory [48, 8], the encounter time between two of the three monomers is Poissonian, and we confirmed this result by simulating the distributions (Fig 3.5B). The reciprocal of the mean encounter time is by definition [90] the encounter rate that we extract in Fig 3.5B-D. We found also that the encounter probability, computed from the encounter rates (rate divided by the sum of the rates) as between Linx and Chic1 is $P = 0.55$, while the mean encounter times between each pair (Linx-Chic1) and (Xist-Linx) are comparable of the order of 131s (table C in Fig 3.5).

Finally, to check the impact of TAD E on the encounter time inside TAD D, we ran another set of stochastic simulations, after removing TAD E (Fig 3.5D). Surprisingly, the encounter probability was inverted compared to the case of no deletion, while the mean time was increased by almost 50% to 195s (Linx to Chic1) and 205s (Linx to Xite). This result suggests that specific long-range interactions between TAD D and E (table in S1 Text) serve to modulate the probability and the encounter time between the three key genomic sites. This result further indicates that the search time inside a TAD can be influenced by neighboring chromatin configurations.



TAD D & E	Conditional encounter	Linx ↔ Chic1	Linx ↔ Xite/Tsix
	Prob.	0.45	0.55
	MFET	133 [sec]	131.5 [sec]
TAD D	Prob.	0.56	0.44
	MFET	196 [sec]	204 [sec]

Figure 3.5 **Transient properties of the chromatin: Conditional mean time and probability for three sites to meet. A.** (upper panel) Representation of the polymer model for TAD D (orange, monomers 1-106), where the locus Linx (monomer 26, red) meets Chic1 (monomer 62, cyan) and Xite/Tsix (monomer 87, gray), respectively. Random connectors (green arrows) and specific long range-connectors (red bar) are added, following the connectivity ξ. Fixed connectors (red bars) correspond to specific peaks of the 5C data. Two realizations (bottom panel) of the polymer model containing TAD D and E, show the encounter of Linx (magenta) with Chic1 (cyan), and Xite/Tsix (gray), respectively. The color code is from the upper panel. **B.** Histogram of the conditional encounter times between Linx and Chic1 (upper panel, green),and Linx and Xite/Tsix (bottom panel, blue) with TAD D and E. **C.** Two polymer realizations with a single TAD D (monomers 1-106, orange), showing the encounter between Linx (magenta) and Xite/Tsix (gray, left panel), and the encounter between Linx and Chic1 (cyan, right panel). **D.** Histogram of the conditional encounter times for a polymer with only TAD D, showing an exponential decay as in sub-figure B.

3.2.6 Statistics of single loci trajectories in the reconstructed polymer model

To further study the statistical properties of loci trajectories, we simulated the three loci monomer 26 (Linx), monomer 87 (Xite) and 62 (Chic1), following classical single particle tracking experiments (Fig 3.6A). We used the calibrated polymer model reconstructed in section Encounter probabilities and distribution of search times of three genomic sites. Starting from an equilibrium configuration following a relaxation time, we estimated the mean-square displacement (MSD) and computed the anomalous exponent over all realizations. The MSD of a stochastic process $X(t)$ is computed by averaging over trajectory $X_i(t)$ realizations [90],

$$\langle |X(t) - X(0)|^2 \rangle = \lim_{N_p \to \infty} \frac{1}{N_p} \sum_{i=1}^{N_p} |X_i(t) - X_i(0)|^2 \tag{3.4}$$

and for short time, we use the asymptotic behavior

$$\langle |X(t) - X(0)|^2 \rangle \sim A t^{\alpha}, \tag{3.5}$$

where A is a constant and α is called the anomalous exponent. In practice, we computed the MSD from the estimator $\frac{1}{N_p} \sum_{i=1}^{N_p} |X_i(t) - X_i(0)|^2$, where the number of trajectories N_p is of the order 1000. We fitted the function At^{α} to the MSD 3.5, computed on the simulated trajectories and then extracted α.

We explore now the consequence of computing the MSD, when averaging over a cell population. To reproduce this situation, we consider a class of polymer configurations $C_{6,10}$ obtained by fixing the number of random connectors 6, and 10 for TAD D and E respectively. However, their positions in each realization is allowed to vary, generating a variability. We shall now estimate the anomalous exponent α_w of a locus for a given configuration $w \in C_{6,10}$ and then compute the average $< \alpha_w >_{w \in C_{6,10}}$ over the ensemble $C_{6,10}$. We shall emphasize that long-range specific connectors (table in S1 Text) are also accounted for in the polymer ensemble $C_{6,10}$.

For two specific configurations from $C_{6,10}$, one with TADs $E{+}D$ (Fig 3.6A Upper) and the other TAD D alone (Fig 3.6A Lower), we show SPTs for each monomer $m = 26, 62, 87$, projected in two dimensions. We then computed the MSD curves by averaging over all realization (Fig 3.6B). Further, the distribution of the anomalous exponents is shown in Fig 3.6C. Interestingly, we found that each locus $m = 26, 62, 87$ had a different mean anomalous exponent $\alpha_{26}^{D+E} = 0.39, \alpha_{62}^{D+E} = 0.4, \alpha_{87}^{D+E} = 0.31$, revealing the intrinsic heterogeneity present in the chromatin. The α values we have computed for all monomers are indeed heterogeneous, as can be seen in histograms of Fig 3.6B (right column), 6C and S2 Fig A-B left columns. Values are spreading in the range 0.3 to 0.5, however there is a peak in the histogram between values 0.4 to 0.45.

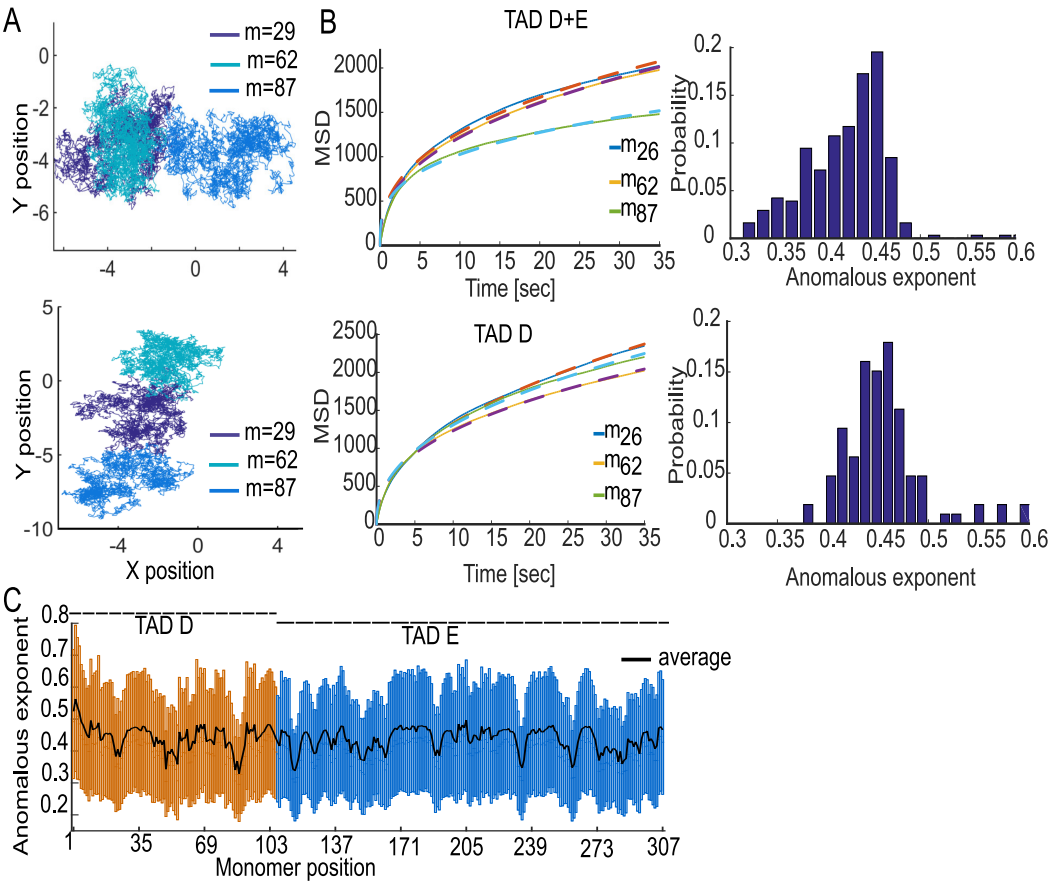


Figure 3.6 **Statistics of SPTs simulated from the reconstructed polymer model**. **A.** Representation of the polymer model described in Fig 3.5 for TAD D (orange, monomers 1-106). Trajectories for the three loci Linx (monomer 26, red) Chic1 (monomer 62, cyan) and Xite/Tsix (monomer 87, gray). **B.** MSD of trajectories shown in **A** for TAD D+E (left,upper) and TAD D alone (left, lower). The anomalous exponents for the three loci are $\alpha_{26} = 0.39, \alpha_{62} = 0.4$, and $\alpha_{87} = 0.31$ (TAD D+E), while the anomalous exponent become $\alpha_{26}^{-E} = 0.46, \alpha_{62}^{-E} = 0.4, \alpha_{87}^{-E} = 0.43$, when TAD E is removed. The histograms of the anomalous exponent (right column) is computed by averaging over 500 realizations with different random connectors positions. **C.** Box plot of the anomalous exponents (25-75%) computed over 500 realizations by changing the random connectors locations.

To evaluate how the anomalous exponent is influenced by persistent long-range interactions present in the EP data, we repeated polymer simulations and MSD analysis by removing TAD E. The distribution of the anomalous exponent α shows that removing TAD E results in the loss of low values, which characterize high chromatin connectivity (Fig 3.6B, right). We found the following changes in the anomalous exponents: $\alpha^D_{26} = 0.46, \alpha^D_{62} = 0.4, \alpha^D_{87} = 0.43$. This result shows how specific long-range interactions affect the local chromatin dynamics and locus motion. Finally, the anomalous exponent averaged over all monomers is $\alpha^{D+E} = 0.425$, whereas for TAD D and E we found $\alpha^D = 0.433, \alpha^E = 0.42$. However, when TAD E is removed, the average anomalous exponent for all monomers in TAD D is $\alpha^D = 0.458$.

To represent the contribution of a single cell experiments in a population, we simulated SPTs for a fixed polymer configuration (we chose an ensemble of connectors in $C_{6,10}$). We computed the MSD of all loci $n = 1..307$ that can be divided into 3 classes: low medium and high anomalous exponent (SI). The distribution of the anomalous exponents of all sites is quite uniform (S2 Fig). We then varied the connector positions and computed the spread of the anomalous exponents (Fig 3.6C). This result shows that changing the connectors position account for the variability of the anomalous exponents. This result clarifies how the local chromatin organization affects the MSD when computed across cell population [28]. Indeed, random connectors model molecular binding that can vary from cell-to-cell.

To conclude, the construction of polymer models from 5C data can now be used to simulate SPTs of any loci of interest and thus to explore the inherent statistical variability found experimentally (Fig 3.6C). The present results can be used to interpret the variability of the MSD found in SPTs of live cell imaging, where the same locus in different cells can exhibit a different anomalous exponent, depending not only on the locus position, but also on the intrinsic variability due to the random arrangements of binding molecules between cells.

3.3 Discussion

We presented here a general method to construct a coarse-grained polymer model from the 5C encounter probability (EP) matrix. This construction preserves the statistical properties of the 5C data, such as the decay rate of the EP of each monomer. The present approach is not used to study the configuration space of chromatin geometry, because it is too large to be fully sampled by elementary polymer models. However, we used this approach to generate statistics of passage times and radius of gyration, which characterize more accurately chromatin dynamics and are not contained in the 5C data. We built here a coarse-grained polymer model of the chromatin based on Rouse and we disregarded the repulsion forces between monomers and possible cross over of bonds that certainly influences the dynamics of the chromatin and statistical results. Future models should clearly

examine the effect of repulsion forces on the present reconstruction, especially to study refined spatial scales below few kbp. We added connectors between random monomer-pairs to characterize sub-configurations present in 5C data. Connectors are represented by springs between monomer-pairs and account for the chromatin architectures. By adding connectors between random monomers, we were able to recover TAD sub-regions. By randomizing connectors positions, we could reproduce the inherent variability in chromatin architecture of nuclei population captured in 5C experiments.

Using our methodology, the characteristics of the reconstructed polymer are derived directly from the empirical data and do not require any minimization procedure [39]. One of the key result here is to determine the number of connectors directly from the experimental EP decay (Fig 3.2). Connectors can directly be interpreted as molecular interactions mediated by proteins such as cohesin, condensing and CTCF bindings. For example, cohesin could bind at random places scattered along the chromatin at 5C data acquisition time. These bounds could generate TADs, as shown here using simulations (Fig 3.2). 5C Contact maps represent steady-state distribution obtained from looping events in large ensemble of millions of cells, where TAD structures appears. The present approach differs from classical reconstruction methods, where 3D structures of a genome are inferred from 5C contact frequency data [105, 32, 42, 54]. Previous models explored the effect of connectors between regions of the chromatin [11, 77] and examined the consequences on the EP-decay rate, but the positions and the number of these connectors were not derived from data. Here, we use connectors to resolve a reverse engineering problem, which is to recover the degree of connectivity from the EP-decay rate (Fig 3.2). The relation between the mean number of connectors and the decay exponent β of the EP (Eq 3.2) is found using simulations in section Encounter probability in the random loop polymer model. The decay exponent of the EP characterizes the polymer scaling statistics [11] and for this reason, the present model extends the key switch-and-binders model developed in [11].

It was surprising to find that a TAD subregion could influence the encounter distribution between monomers located in different TADs (Fig 3.5). Indeed, the distribution of looping time in free space depends only on the distance between monomers, while in confined domain, the nuclear boundary has an effect [53, 8, 5]. We found here that this modulation of loop regulation is due to long-range inter-TAD interactions, that are present in the 5C data and accounted for here in the construction of the polymer model. Indeed, we reported (SI) significant inter-TAD interactions between three monomer pairs: 86-234, 86-260, 24-285, where spring constants, representing long-range interactions are at least twice larger than other interactions. The threshold procedure described in Material and Methods (Eq 3.11) disregards the peaks of the 5C data for which the EP falls below the nearest-neighbor threshold T_{th}. The peaks in the 5C below T_{th} were ignored because we interpret

them as transient events, or a statistic that was not shared by the majority of the chromatin events or it could simply be due to some random fluctuations in the data. Long-range stable interactions can signify cohesin-CTCF mediated genomic loops [49] and have also been found in other mammalian chromosomes [29].

Finally, the present polymer reconstruction allows probing the dynamics of single particle trajectories (SPTs). We use our reconstructed polymer model to explore transient properties and the encounter time distribution between any two sites. Additionally, we simulated single particle trajectories from a reconstructed polymer model (Fig 3.4) and estimated the anomalous exponents, following the routine procedure in experimental SPTs. The present analysis reveals that the variability of the anomalous exponent of a given loci is due to the heterogeneity of the local polymer configurations constructed from the 5C data. This heterogeneity is simulated here as the random cross-links between monomer pairs. Indeed, for a Rouse polymer or any uniformly connected polymers, the anomalous exponent is constant [5, 109, 108].

We suggest that the inherent fluctuations of the chromatin interactions can be due to the random positions of binding molecules. Indeed, even for a fixed number of connectors, there still remains an intrinsic variability of monomer binding positions, leading to the same EP decay rate. This structural heterogeneity originating from connector positions certainly influences the anomalous exponents and can be used to re-interpret experimental SPTs [41, 2, 28, 102]. The present method and algorithms are generic and can be used to reconstruct a polymer model at a given scale (number of monomers and number of bps coarse-grained in a monomer) in a limit of few Mbp. Analyzing chromatin condensation and its transient properties can now integrate chromosomal capture data and SPTs statistics.

3.4 Materials and Methods

Presentation of a generalized Rouse polymer model with long and short-range connections

The Rouse model [30] describes a polymer as a collection of beads $R_n(n = 1...N)$ linearly connected by harmonic springs and driven by Brownian motion. The energy of the polymer is given by [30]

$$\phi_{Rouse}(R_1, .., R_N) = \frac{1}{2}\sum_{j=1}^{N-1} \kappa(R_j - R_{j+1})^2, \tag{3.6}$$

where $\kappa = \frac{3k_BT}{\gamma b^2}$ and b is the standard deviation of the distance between adjacent monomers, γ is the friction coefficient, k_B the Boltzmann coefficient, and T the temperature.

To account for a sub-chromatin region $\mathcal{C}_\mathcal{N}$, characterized by a higher EP than the rest, we will add connectors between monomer-pairs chosen randomly (with uniform distribution) inside this subregion such that an additional potential

$$\phi_{Rand}(R_1, .., R_N) = \frac{1}{2} \sum_{j,k \in \mathcal{C}_\mathcal{N}} \kappa (R_j - R_k)^2, \tag{3.7}$$

is added to ϕ_{Rouse}, where $\mathcal{C}_\mathcal{N}$ is the ensemble of indices defining the sub-region. The number of connectors is a free parameter, that will be determined from experimental 5C data.

In addition, to account for consistent long-range interactions, reflected by peaks in EP matrix (Fig 3.1C), we will fix connectors between monomer-pairs by adding a spring constant $\kappa_{m,n}$ between monomer m and n, so that the associated energy related to the peak interactions is described by

$$\phi_{Peaks}(R_1, ..., R_N) = \frac{1}{2} \sum_{n,m \in S_{Max}} \kappa_{n,m} (R_n - R_m)^2. \tag{3.8}$$

We will discuss below how the spring constant $\kappa_{m,n}$ is computed from empirical data. In summary, the total energy of a polymer containing random connectors and prescribed peaks, is the sum of three energies 3.6-3.7 and 3.8:

$$\Phi(R_1, .., R_N) = \phi_{Rand}(R_1, .., R_N) + \phi_{Peaks}(R_1, ..., R_N) + \phi_{Rouse}(R_1, .., R_N) \tag{3.9}$$

and the stochastic equation of motion for $n = 1, .., N$ is

$$\frac{dR_n}{dt} = -\nabla_{R_n} \Phi(R_1, .., R_N) + \sqrt{2D} \frac{d\omega_n}{dt}, \tag{3.10}$$

where $D = \frac{k_B T}{\gamma}$ is the diffusion constant, γ is the friction coefficient, and ω_n are independent 3-dimensional Brownian motion with mean 0 and standard deviation 1.

3.4.1 Polymer parameter calibration from 5C data

To account for the 5C-data, comprised of a subsection of the X-chromosome from female mice embryonic stem cells reported in [76], showing TAD D and E as two diagonal blocks (Fig 3.1A), we use the coarse-graining procedure of [39], with a polymer of length $N = 307$. Each monomer represents a genomic segment of$3kbp$ and is connected to its 2 nearest neighbors by a harmonic spring (see subsection above). TAD D (resp. E) is represented by the range of monomers from 1 to $N_D = 106$ (resp. 107 to 307), and $N_E = 201$ for TAD E.

To reproduce the empirical EP extracted from data (see formula 3.1), we connected non-nearest neighbor pairs of monomers chosen randomly with uniform

probability $\frac{1}{N_1}$ (resp. $\frac{1}{N_2}$)) where $N_1 = (N_D - 2)(N_D - 1)/2$ (resp. $N_2 = (N_E - 2)(N_E - 1)/2$). The number of connectors in each TAD is a percentage ξ of the total possible number of non nearest neighbor pairs, for TAD D (resp. E), we have $C_D = \xi_D \frac{N_1}{100}$ (resp. E $C_E = \xi_E \frac{N_2}{100}$) and $\xi_D, \xi_E \in [0, 100]$, which will be extracted from data. Random connectors were not added between monomers belonging to different TADs.

The procedure of adding random loops to a Rouse polymer is implemented using the energy of random loops, as described in the previous subsection. Finally, 24 connectors were added (SI) in all polymers, corresponding to the selected peaks present in the EP matrix, obtained in the following procedure: we located the positions of peaks that form a subset S_{Max} of the ensemble of the off-diagonal local maxima in the EP-matrix, such that their EP is higher than a threshold T_{th}. In practice, we assumed that the encounter probability between neighboring monomers is almost not affected by the global chromatin structure. Therefore, any EP value in the matrix $M_{i,j}$ above the threshold T_{th} (equals to the the EP of the nearest neighbor monomers) is considered to be a stable loop. The threshold T_{th} is defined as follows:

$$T_{th} = \frac{\sum_i M_{ii-1} + M_{ii+1}}{\sum_{i,j} M_{ij}}. \tag{3.11}$$

In a second step, we determine the spring constants $\kappa_{m,n}$ between monomer m and n in the ensemble S_{Max} from the empirical EP $P_{m,n}$ at each peak position. We recall that for a Rouse chain the joint probability density function of beads R_m and R_n is given by [30, p.15]

$$\Phi(R_m, R_n, \Delta_{m,n}) = \left(\frac{3}{2\pi b^2 \Delta_{m,n}}\right)^{3/2} \exp\left(-\frac{3(R_m - R_n)^2)}{2b^2\Delta_{m,n}}\right) \tag{3.12}$$

where $\Delta_{m,n} = |m - n|$. We assume that the encounter probability between neighboring monomers is not affected by global polymer structure, thus for the nearest neighbors $\Delta_{m,n} = 1$ the EP occurs at small distances such that the exponential term is almost 1, that is

$$P_{m,n} = \left(\frac{3}{2\pi b^2}\right)^{3/2} \approx \left(\frac{\kappa_{m,m+1}}{2\pi}\right)^{3/2}, \tag{3.13}$$

We approximate the chromatin structure as a polymer chain with a variance b^2 between adjacent monomers. Thus, the constant $\bar{b}$ is estimated as the mean EP over the sub- and super-diagonals:

$$\left(\frac{3}{2\pi\bar{b}^2}\right)^{3/2} \approx \sum_i (P_{ii-1} + P_{ii+1}) = \frac{\sum_i M_{ii-1} + M_{ii+1}}{\sum_{i,j} M_{ij}} = T_{th} \tag{3.14}$$

To account for long-range interactions, we applied formula 3.13 to estimate the effective spring constant from the empirical EP $\tilde{P}$, so that

$$\kappa_{m,n} = 2\pi \tilde{P}_{m,n}^{2/3}. \tag{3.15}$$

Finally, the energy related to peak interactions is described by

$$\phi_{Peaks}(R_1, ..., R_N) = \frac{1}{2} \sum_{n,m \in S_{Max}} \kappa_{n,m}(R_n - R_m)^2. \tag{3.16}$$

3.4.2 Numerical simulations of the reconstructed polymer model

Using the reconstruction method described in the two previous subsections, we generate polymer realizations, each differs in the position of random connectors inside a TAD. To generate statistics from the EP matrix, we started from an initial random walk configuration and simulated the polymer until its equilibrium after a relaxation time τ_R. The time τ_R is determined for each realization from the formula $\tau_R = 1/(\kappa_{min}\lambda_1)$, where κ_{min} is the minimal positive spring constant, and λ_1 is the smallest non-vanishing eigenvalue of the polymer's connectivity matrix [40], which we computed numerically (in practice it is of the order of thousands of simulation steps with $\Delta t = 0.01s$).

For the numerical simulations, we divide Eq 3.10 by $\sqrt{D}$ and the spring constants are scaled by the friction coefficient such that $\kappa = \frac{dk_BT}{\gamma b^2} = \frac{dD}{b^2}$, in dimension $d = 3$. The encounter frequency matrix of the 307 monomers is computed at time τ_R, where two monomers are considered to have encountered if their distance is less than ϵ. The time step for all simulations is $\Delta t = 10^{-2}s$. The value of the parameter b is computed using formula 3.14. Using the b value in the relation $\kappa = 3k_BT/b^2$ [30], we computed the spring constant for random connectors and the linear backbone of the polymer. The spring constants for persistent long-range connectors were estimated using Eq. 3.15 and are summarized in the table in S1 Text, other simulation parameters are summarized in Table 3.1.

3.5 Supporting Information

This supplementary information contains three sections. In the first section, we summarize in a table the spring constants associated to long-range connections used in Fig 3.3-3.6 of the main text. In the second section, we present the procedure for comparing the experimental encounter probability (EP) matrix with simulations. In the third section, we discuss the distribution of anomalous exponents and the Mean-Square-Displacement (MSD) computed over the simulated single particle trajectories

Parameter	Value	Description
d	3	Dimension
b	0.6 $[\mu m]$	STD of distance between adjacent monomers
D	4×10^{-2} $[\mu m^2/s]$	Monomer diffusion coefficient [3]
ϵ	0.03 $[\mu m]$	Encounter distance
γ	3.1×10^{-5} Ns/m	Friction coefficient [3]
κ	3×10^{-5} N/m	Spring constant
Δt	$10^{-2}s$	Time step

Table 3.1 Values of simulation parameters

(SPTs). These trajectories are generated by the polymer model reconstructed from the 5C data. This part links chromosomal capture data to SPTs.

3.5.1 Values for the spring constants of long-range monomer interaction

Table 3.2 summarizes the values of the spring constants $\kappa_{m,n}$ that we computed from the EP matrix (see Materials and methods) for simulating long-range connectors between monomer m and n. The right column indicates which TADs are connected by the monomer-pairs. The values are 1.1 to 3 times higher than the spring constant of randomly added connectors, and for adjacent monomers in the linear backbone (nearest-neighbor connection) of the polymer ($\kappa = 3 \times 10^{-5} Nm^{-1}$).

3.5.2 Comparison of the experimental and simulation encounter data

We now compare the EP matrices about the steady-state of our models and the experimental data. We compare the experimental EP matrix (S3.7A Fig) with the simulation of the model with persistent and random connectors (S3.7 Fig B). The simulated EP matrix is much smoother than the experimental 5C EP matrix.

To estimate the similarity between the EP matrices, we computed the cumulative distribution function (CDF) of the monomer encounters. The CDF is computed by averaging the EP between all monomers m and n from an encounter frequency matrix M. For each row of M, we start from the diagonal entry and average counts at symmetrical positions. The EP is given by

$$P(|m-n||n) = \frac{M_{n,n+|m-n|} + M_{n,n-|m-n|}}{\sum_{m=1}^{N} M_{n,m}}, \tag{3.17}$$

Spring const.	value$\times 10^{-5}[Nm^{-1}]$	connecting TADs
$\kappa_{14,10}$	3.9555	$D \leftrightarrow D$
$\kappa_{48,44}$	4.1396	$D \leftrightarrow D$
$\kappa_{54,50}$	3.6777	$D \leftrightarrow D$
$\kappa_{60,56}$	5.9841	$D \leftrightarrow D$
$\kappa_{69,65}$	3.1722	$D \leftrightarrow D$
$\kappa_{79,75}$	3.0864	$D \leftrightarrow D$
$\kappa_{95,91}$	4.6673	$D \leftrightarrow D$
$\kappa_{113,109}$	3.2280	$E \leftrightarrow E$
$\kappa_{126,117}$	3.4386	$E \leftrightarrow E$
$\kappa_{144,136}$	3.6859	$E \leftrightarrow E$
$\kappa_{151,50}$	3.2399	$E \leftrightarrow D$
$\kappa_{161,157}$	3.4420	$E \leftrightarrow E$
$\kappa_{190,162}$	8.3386	$E \leftrightarrow E$
$\kappa_{205,201}$	3.5671	$E \leftrightarrow E$
$\kappa_{217,213}$	3.1621	$E \leftrightarrow E$
$\kappa_{234,86}$	4.2015	$D \leftrightarrow E$
$\kappa_{259,255}$	3.4765	$E \leftrightarrow E$
$\kappa_{260,86}$	7.2143	$D \leftrightarrow E$
$\kappa_{275,116}$	3.5335	$E \leftrightarrow E$
$\kappa_{279,275}$	3.2174	$E \leftrightarrow E$
$\kappa_{285,24}$	8.4070	$D \leftrightarrow E$
$\kappa_{293,289}$	4.7508	$E \leftrightarrow E$
$\kappa_{302,294}$	7.1703	$E \leftrightarrow E$
$\kappa_{305,301}$	3.1879	$E \leftrightarrow E$

Table 3.2 **Spring constants for long-range interactions.** Long-range fixed connectors added between monomers pairs indices (left column) and their computed spring constant (middle column, see Methods in main text), form connections within and between TADs as indicated in the right column

We average over $N = 307$ rows of the matrix, leading the CDF defined by

$$F(k) = \frac{1}{N} \sum_{m=1}^{k} \sum_{n=1}^{N} P(|m-n|n). \tag{3.18}$$

We use Eq 3.18) to compare the EP matrices, by computing the CDF for the experimental $F(k)_{Exp}$ and the simulation $F(k)_{Sim}$ data respectively. Finally, we shall use the Kolmogorov-Smirnov (KS) distance, defined by

$$D_{max} = \max_{k} |F(k)_{Exp} - F(k)_{Sim}|. \tag{3.19}$$

The KS distance is computed for the simulation of the model with persistent long-range connectors (subsection 3.2.3 of the main text) that we compare with the model with persistent long-range and random connectors (subsection 3.2.4 of the main

text). In S3.7 Fig C, we show the CDF of the model with only persistent connectors gives $D_{max} = 0.15$. The CDF of the model that contain both persistent and random connectors (S3.7 Fig D), leads to an improved agreement with experimental data indicated by the value $D_{max} = 0.06$ at a level of 0.001 (P-Value= 0.06).

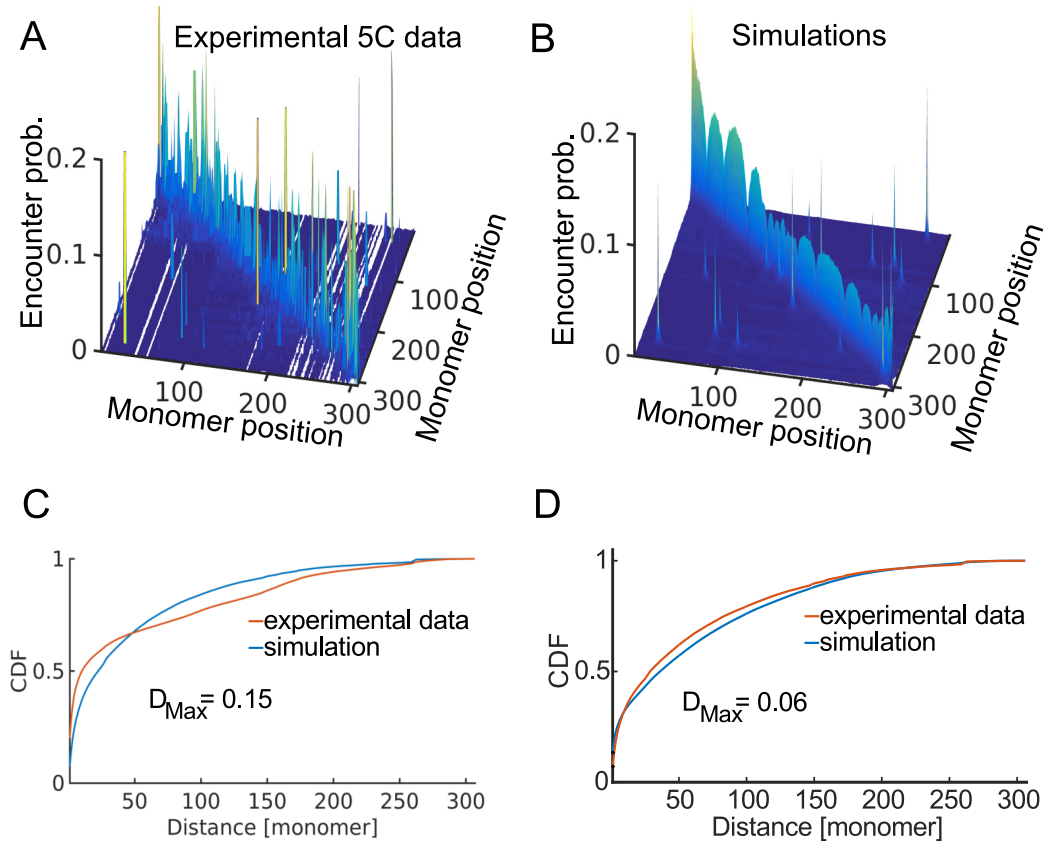


Figure 3.7 **Simulations and experimental 5C encounter matrices**. **A.** Three-dimensional representation of the empirical encounter probability matrix **B.** Three-dimensional representation of the simulation encounter probability matrix for polymer with persistent long-range and random connectors. **C.** Cumulative distribution function of monomer encounters for a polymer model with only persistent long-range connectors, simulations (blue) vs. experimental (orange) data. The Kolomogorov-Smirnov distance is $D_{max} = 0.15$. **D.** Cumulative distribution function of monomer encounter for a polymer model with persistent long-range and random connectors, simulations (blue) vs. experimental (orange) data. The Kolmogorov-Smirnov distance is $D_{max} = 0.06$.

3.5.3 MSD and anomalous exponent statistics for single polymer realization

We describe in this section the MSD along single monomer trajectories for polymer realizations having connectivity matching the one calibrated from the 5C data based

on 6 and 10 randomly connectors for TAD D and E, respectively (this construction is already described in Fig 3.4 of the main text). Long-range connectors are included, as listed in Table 3.2. The anomalous exponent values for each single monomer trajectory is estimated by fitting a power law to the MSD:

$$f(t) = At^{\alpha}. \tag{3.20}$$

To each MSD curve, we computed the two parameters A and α (anomalous exponent). The values of the anomalous exponents are distributed in [0,1]. However, if we average over realizations of random connector positions, the anomalous exponents are concentrated around the mean 0.42 (Fig 3.6 of the main text), as shown for three realization in S3.8 Fig.

We classify the curves into low, medium and high MSD trajectories, by dividing equally the range of values at time $35s$ into 3 regions (S3.8A Fig right column) for each realization. The majority of the MSD curves (92.3%) saturate for long-time simulations (blue in S3.8A Fig right column), while a fraction 7.3% had a mix saturated an linear behavior. Finally, only a small fraction of 0.5% curves (red) are characterized by a rapid increase. S3.8A Fig right column shows an example of the MSD curves with three realizations.

We repeated the MSD analysis and the computation of the anomalous exponent when TAD E was removed. We found that the values of the anomalous exponent remained distributed in the range [0,1] (S3.8B Fig left column), similar to the case where TAD E is included. The MSD curves were divided into 3 classes as described above (S3.8B Fig right column), and we found that the majority of them (87%) saturate (blue), the medium class included 12% of the curves (green), which displayed mixed saturated and rapid MSD increase. Only 1% of the MSD curves were classified into high class (red curves), displaying rapid MSD increase, similar to the case with TAD E included.

We conclude that the distribution of anomalous exponents extracted from SPTs depends on the polymer configuration generated by random connectors. In addition, the distribution of MSDs in TAD D was sensitive to the influence of TAD E (as shown in comparing S3.8A Fig with B). The exact scaling law that connect the anomalous exponent measured in SPTs with the EP decay exponent of the Chromosomal Capture data remains unknown, although it is now clear that increasing the chromatin connectivity, leading to a higher decay exponent is characterized by a smaller anomalous exponent.

3.5.4 Computational tools

The data analysis and stochastic simulations were performed using our codes on Matlab 2015. The source codes and description are now available on our website

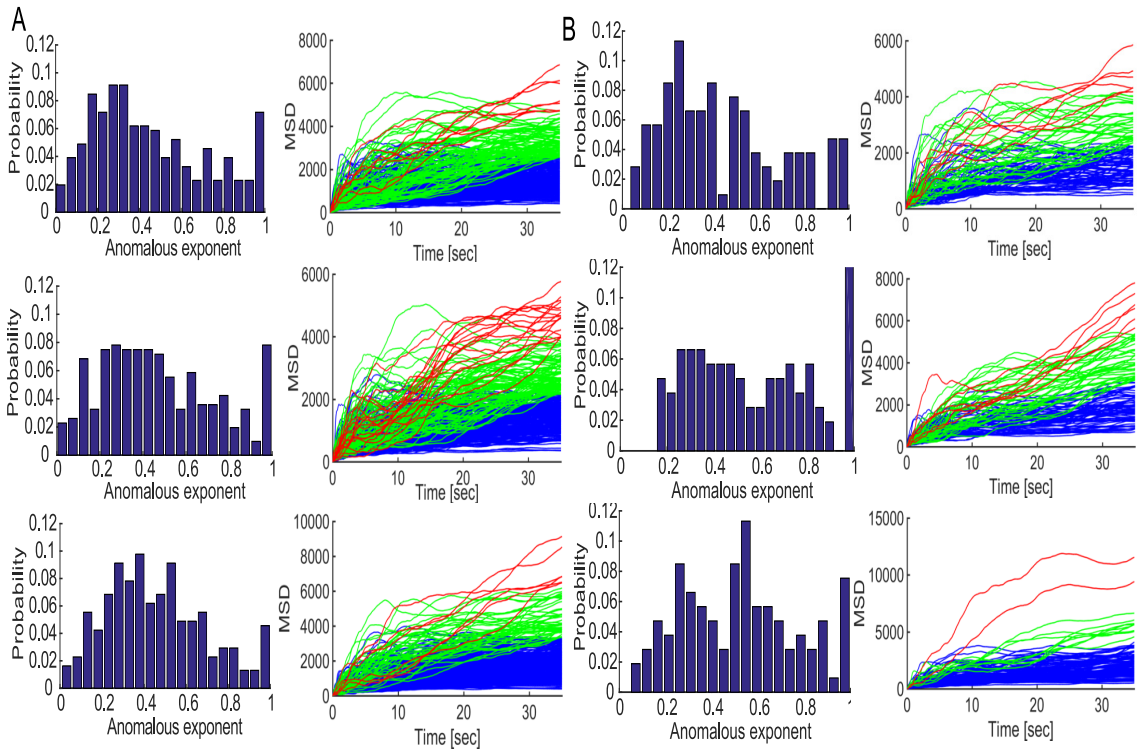


Figure 3.8 **MSD and anomalous exponent of polymer realizations**. **A.** Distribution of the anomalous exponent for 3 polymer realizations (left column). The MSD curves for each realization is shown in the right column. MSD curves are classified into 3 classes: low (blue), medium (green) and high (red) based on the MSD value at time 35 sec. **B** Distribution of anomalous exponent (right column) for TAD D, when TAD E is removed. The distribution in each class is given by low (blue, 85%) medium (green, 12% of curves) and high (red, 1% of curves).

http://bionewmetrics.org/. For the chromatin visualizations in Figs 3.2-3.5 of the main text, we use the UCSF Chimera software version 1.11.

4 Two loci single particle trajectories analysis: constructing a first passage time statistics of local chromatin exploration

Accepted in Shukron Ofir, Michael Hauer, and David Holcman. "Two loci single particle trajectories analysis: constructing a first passage time statistics of local chromatin exploration." Scientific Reports.

Abstract

Stochastic single particle trajectories are used to explore the local chromatin organization. We present here a statistical analysis of the first contact time distributions between two tagged loci recorded experimentally. First, we extract the association and dissociation times from data for various genomic distances between loci and we show that the looping time occurs in confined nanometer regions. Second, we characterize the looping time distribution for two loci in the presence of multiple DNA damages. Finally, we construct a polymer model that accounts for the local chromatin organization before and after a double-stranded DNA break (DSB), to estimate the level of chromatin decompaction. This novel passage time statistics method allows extracting transient dynamic at scales from one to few hundreds of nanometers, predicts the local changes in the number of binding molecules following DSB and can be used to better characterize the local dynamic of the chromatin.

4.1 Introduction

Analysis of single particle trajectories (SPTs) of a tagged single locus revealed that chromatin dynamics is mostly driven by stochastic forces [44, 9]. The statistic of a locus motion has been characterized as sub-diffusive [1, 58, 28, 107, 71] and confined into nano-domains. The confinement is probably due to an ensemble of local tethering forces, generated either at the nuclear periphery [106], or internally [7], where binding molecules such as CTCF or cohesin play a key role [62, 69]. Chromatin dynamics is involved in short-range loop formation in the sub-Mbp scale, and contributes to processes such as gene regulation. However, the analysis of the chromatin dynamics in the sub-Mbp scale is insufficient to describe processes involving long-range chromatin looping (above Mps scale), such as in homologous

dsDNA repair. When two neighboring loci located on the same chromosome arm are tracked simultaneously over time, their correlated position can be used to explore the local chromatin organization in the range of tens to hundreds of nanometers (of the order of the genomic distance between the loci).

Statistical parameters, characterizing short-range chromatin motion, have been studied in stochastic polymer models, starting with the Rouse polymer [30], copolymers [55], the beta polymer [4], and polymer models with additional diffusing or fixed binding molecules [17, 16, 45, 94]. The extracted statistical parameters are the diffusion coefficient, local tethering forces, the radius of gyration, radius of confinement [44, 9], and the distribution of anomalous exponents of tagged loci along the chromatin, which characterizes the deviation from pure diffusion [94, 58].

Here we analyze the transient statistics of two loci SPTs and use it to explore the local chromatin reorganization, following DSB and its confining geometry. Thus, we further contribute to the global chromatin reorganization explored in [44]. We adopt here the formalism of Brownian polymer dynamics, as we have already shown [7] that the auto-correlation function of a single locus decays exponentially, but not as power laws, as would be predicted by the fractional Brownian motion description [19]. Specifically, we explore the chromatin state from the transient statistics of recurrent visits of two tagged loci. This approach is new and is not contained in other work involving two spots trajectories, which use equilibrium thermodynamic models for steady-state encounter frequency [46], or specific chromatin arrangement [66]. We study the distributions of 1) the first encounter time (FET) and 2) the first dissociation time (FDT) of two tagged loci. The FET is defined as the first arrival time of one locus to the neighborhood of the second, while the FDT is the first time the two loci are separated by a given distance. The statistics of FDT and FET is not contained in moments associated with each locus separately, but revealed by their correlated motion.

This article is organized as follow: in the first part, we introduce and estimate the FET and FDT distribution from SPTs of two loci (data from [26]). In the second part, we analyze empirical data of loci motion before and after the induction of DNA damages by Zeocin (data from [44]). The local effects of DSBs on the loci motion was not the goal in [44], but multiple DSBs and single stand breaks (caused by Zeocin), together with a strong DNA damage checkpoint response can trigger global chromatin changes. We shall study here the consequences of multiple tether losses on the chromatin not just around the break site, on the local loci motion. In the third part, we use a randomly cross-linked (RCL) polymer model [16, 94] to simulate the trajectories of two loci following a DSB on the DNA strand between them and evaluate the number of binding molecules required to restrict their motion. We thus use the RCL polymer to explore the chromatin reorganization on the scale of a single DSB. In the last section, we estimate the number of binding molecules required to obtain SPTs with the same statistics as the measured ones. We conclude that

the statistics of two correlated loci provide complementary information about local chromatin organization, not contained in the statistics of individual non-correlated loci. The present method is general and can be applied to any SPT of any number of loci. It can further be used to reveal characteristic lengths, local chromatin dynamics, remodeling following DSB and estimate the changes in the number of molecular interactions.

4.2 Results

4.2.1 First passage time analysis

The construction of the present statistical method is based on the first passage time for two loci entering and exiting a small ball of radius ϵ (that can vary continuously). We will thus estimate the FET and FDT (introduced above). The statistics of these times contain information about the local chromatin organization at a scale of one to few hundreds of nanometers, because the fluctuations in loci distance depend not only on their stochastic dynamics but also on the restricted geometry. We now briefly recall the published data we will used to construct the analysis. In the data of [26], two fluorescently tagged loci are tracked over a course of 60-120 s. We only used recording for which the time interval did not exceed 1 s. The experiment is repeated for seven DNA strains of genomic length between the tagged loci between 25-100 kbp. We also use the dataset reported in [44], which tracks two tagged loci located on yeast chromosome III, at a genomic distance of 50 kbp, at time intervals of 300 ms for a total of 60 s. The trajectories of two loci are tracked after the induction of DSB breaks uniformly over the genome by Zeocin 500 $\mu g/ml$.

We first analyze trajectories of two tagged loci of [26], when they are separated by various genomic distances: $\Delta = 25.3; 42.3; 51.3; 71$, and 100.8 kbp. The distance $d(t) = dist(X(t), Y(t))$ between the two trajectories $X(t)$ and $Y(t)$ fluctuates in time, thus we estimate the distribution of the FET τ_E and the FDT τ_D (Fig. 4.1A). The FET is the first time the distance between the two loci becomes less than ϵ, when the initial distance is larger. The FDT is defined as the first time that the distance between the two loci reaches ϵ, when they are initially inside a ball of radius ϵ. The FET (FDT) are collected between successive dissociation (association) events, after which we reset the time to $t = 0$, by definition:

$$\tau_E = \inf\{t > 0; d(t) \leq \epsilon | d(0) > \epsilon\}, \tag{4.1}$$

and

$$\tau_D = \inf\{t > 0; d(t) \geq \epsilon | d(0) < \epsilon\}. \tag{4.2}$$

In practice, we constructed the distributions of τ_E, τ_D for a continuum of encounter distances ϵ that varies in the range 150-500 nm.

We gathered the distributions of the FET and FDT for seven various DNA strains of genomic length $\Delta = 25 - 108$ kbp [26] (Fig. 4.1B). As predicted by the polymer looping theory in confined domains [8, 4] (formula 4.5 of the Method), the distribution (Fig. 4.1C, red curves) follows a single exponential decay, with rate λ, which is the reciprocal of the mean FET (MFET) between the two loci. Using an exponential fit to the data for all strains of length Δ, we find that the MFET slightly decreases from 3.2 s for $\Delta = 25$ kbp to 2 s for $\Delta = 108$ kbp (Fig. 4.2A blue circles).

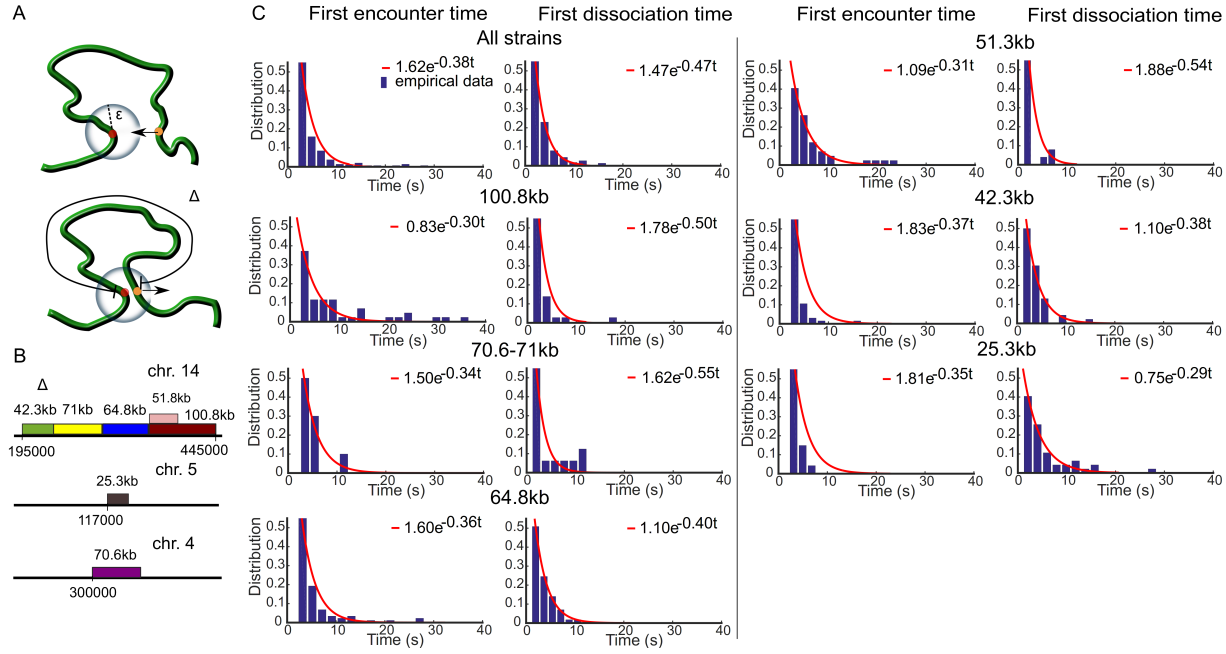


Figure 4.1 **Statistics of two loci trajectories**. **A.** Schematic representation of the first encounter time (FET) τ_E (upper) and the first dissociation time (FDT) τ_D (lower). The FET is computed when the two loci are within an encounter distance ϵ when they are initially apart. The FDT is computed when the distance between two loci is larger than ϵ when they initially encountered. The genomic distances are $\Delta \in [25.3, 100.8]$ kbp between tagged loci **B.** Experimental setting for tagging seven chromatin strains by inserting lac and tet flanking operators at their ends on chromosome 4,5 and 14 [26] **C.** Distribution of the FET (left column) and the FDT (right column) with respect to Δ, fitted with $a\exp(-\lambda t)$, with a a constant.

To estimate the effect of chromatin confinement on transient properties, we used the formula 4.6, derived for confined polymer, to fit the MFET of the two loci for all Δ. Because the two loci are located along the same chromosome arm [26], we model them as the two end monomers of a polymer chains with N monomers. To fit the MFET data using 4.6, we use $\Delta \in [25, 108]$ kbp, $b = 0.2$ μm, the length of 1 bp to be 3×10^{-4} μm, the number of monomers $N = (3 \times 10^{-4}\Delta/b)$ and the parameters $D = 8 \times 10^{-3}$ $\mu m^2/s$, $\kappa = 1.75 \times 10^{-2}$ N/m, and $\epsilon = 0.2$ μm (see Table 4.1). We find the value for the confined parameters $\beta = 2.4$ μm^{-2}, and substituting in 4.7, we finally obtain the radius of confinement of $A = 0.5$ μm in agreement with data presented in [26]. Furthermore, the MFET in a confined environment does not

exceed a limit of 2.65 s for all genomic distances between tagged loci (Fig. 4.2A dashed blue), suggesting that the dynamics has already reached the asymptotic limit and , thus, the loci are confined at all scales. We conclude that the motion of two loci located in the range 25-108 kbp is largely influenced by the local chromatin confinement.

The stochastic model for the FDT is the escape problem from a parabolic potential well [91]. The mean escape time is given by formula 4.9 (Methods), which shows that the dissociation time increases with the genomic length Δ. We thus fitted the mean FDT (MFET) data points using formula 4.9 (Fig. 4.2A, purple squares), confirming that the MFDT increases from 1.5 s to 2.5 s when Δ increases from 25 to 108 kbps (Fig. 4.2A) (Using a Matlab fit, we estimated the parameters of relation 4.9 to be $a_2 = 0.01, b_2 = 40.36$).

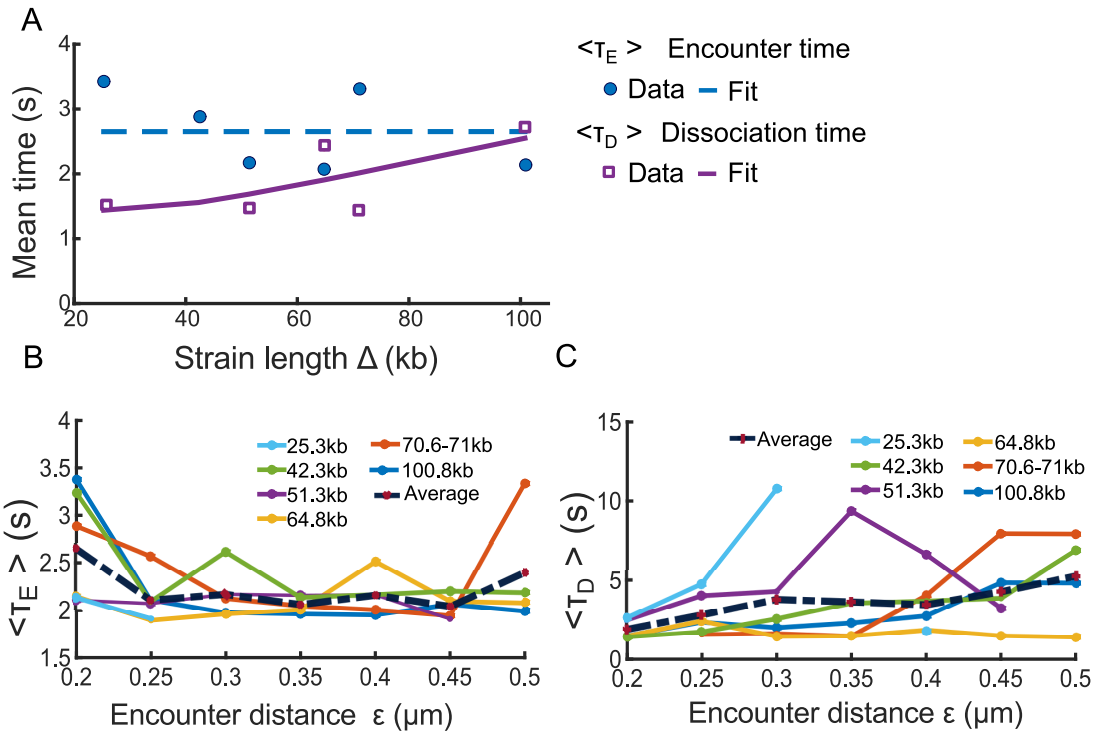


Figure 4.2 **Effect of the genomic separation distance Δ and the encounter distance ϵ. A.** The mean first encounter time (MFET) data (blue circles) are fitted using eq. 4.6 (blue dashed). The mean first dissociation time (MFDT) data (purple squares) is fitted using eq. 4.9 $a_2\Delta\exp(b_2/\Delta)$ (purple curve), where $a_2 = 0.01, b_2 = 40.36$ **B.** The MFET $\langle\tau_E\rangle$ for 6 strain lengths Δ kbp, shown in Fig. 4.1B, where the encounter distance ϵ varies in 0.2 and 0.5 μm. **C.** MFDT $\langle\tau_D\rangle$ extracted from Fig. 4.1B and plotted for all Δ with respect to the encounter distance ϵ.

To evaluate the sensitivity of our approach to the choice of encounter distance ϵ, we estimated the FET and FDT when ϵ varied in the range $0.2 - 0.5$ μm. For $\epsilon \in 0.2 - 0.25$ μm, the MFET decreased from 2.7 s for $\epsilon = 0.2$ μm to a 2.1 s when $\epsilon = 0.25$ μm (Fig. 4.2B dashed). In the range $\epsilon \in [0.25, 0.5]$ μm, we find that the

MFET is almost constant, independent of ϵ with an average of 2.1 s (Fig. 4.2B left). This result shows that it takes around 2 seconds for the two loci to meet and thus to explore a ball of radius of 0.25 μm. Indeed, the MFET is the time to meet when almost all points of the domain have been visited [47].

We conclude that any loci explore constantly and recurrently the neighborhood of the chromatin with a time constant of 2 seconds in a tubular neighborhood of 0.25 μm and this spatial constraint does not depend on the position of the locus. We note that this result about the exploration and the recurrence exploration is not contained in the statistics SPT of a single loci, because a reference point is needed. Finally, we find that the MFDT increases with ϵ for all DNA strain of length Δ (Fig. 4.2C), from an average of 2 s for $\epsilon = 0.2\ \mu m$ to 5 s when $\epsilon = 0.5\ \mu$m.

4.2.2 Loci dynamics in the presence of double-strand DNA break

To continue exploring how two loci trajectories provide information about local chromatin organization, we focus now on the consequences of double-strand DNA breaks (DSBs) on chromatin dynamic. For that purpose, we analyzed the transient statistics of two loci, before and after treatment with the radiomimetic drug Zeocin (data presented in [44]). The Zeocin drug induces uniformly distributed DSBs on the chromatin, leading to chromatin expansion and enhanced chromatin flexibility [44]. Thus, we repeated the FET and FDT statistical analysis (Fig. 4.3A) as described in the previous subsection. As predicted by the polymer model theory, the FET and FDT follow a Poisson distribution, and we fitted a single exponential (formula 4.5) to the empirical distributions (Fig. 4.3B). We then computed the MFET and MFDT for encounter distances ϵ in the range $0.1 - 0.5\ \mu$m.

For both the untreated and Zeocin treated data, the MFET graphs in Fig. 4.3C show two phases: in the first, when $\epsilon \in [0, 0.2]\ \mu$m, the MFET decays with the radius ϵ, while in the second phase ($\epsilon \in [0.2, 0.5]\ \mu$m), it is independent (Fig. 4.3C). The boundary between the two phases at $\epsilon = 0.25\ \mu$m indicates that this length is a characteristic of local chromatin folding and crowding. Interestingly, following Zeocin treatment, the MFET increases at a scale lower than 0.25 μm, compared to the untreated case, probably due to the local chromatin expansion around DSBs. Furthermore, the increase of the confinement length L_c, when the chromatin is decompacted [44] and the restriction of the loci dynamics can be due to repair molecules.

To further investigate how DSB affect the separation of two loci, we computed the MFDT for untreated and Zeocin treated cells (Fig. 4.3C), that shows an increase pattern with ϵ. The MFDT for the untreated case increased from 0.2 to 0.9 s, whereas the MFDT for Zeocin treated increased from 0.2 to 0.5 s as ϵ increased from 0.15 to 0.5 μm. We conclude that it takes less time for the two loci to dissociate following

DSB, probably due to the chromatin decompaction, enhanced mobility and the activity of repair proteins. This result suggests that repair molecules do not impair the local chromatin motion.

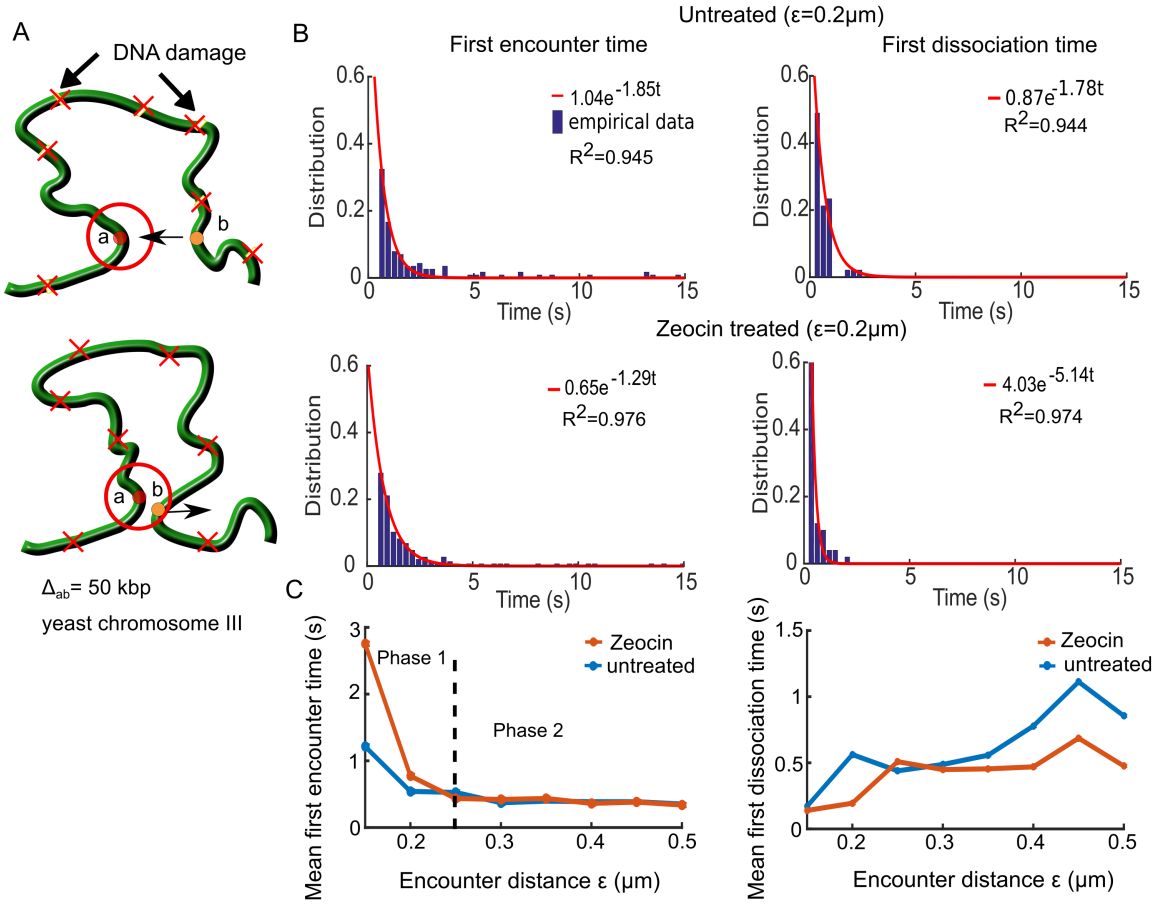


Figure 4.3 **Two loci dynamics before and after Zeocin treatment. A.** Two tagged loci (a and b, circles) separated by a genomic distance $\Delta_{ab} = 50$ kbp. When the loci are within a distance ϵ μm (red circle), they are considered to encounter for computing the FET, and above ϵ (lower) for the FDT. We analyzed the untreated and Zeocin treated cases, where Zeocin induces DNA damages (red X) at random positions along the DNA. **B** the FET (left column) and FDT (right) empirical distributions in the untreated (upper) and Zeocin treated (lower) cases, fitted by $a\exp(-\lambda t)$ (red curves), where the reciprocal of λ is the MFET and MFDT in their respective cases. R^2 values from the fit are reported in each box **C** The MFET (left) is plotted with respect to the encounter distance ϵ for the untreated (blue) and Zeocin treated (orange) cases. For the MFET (left), both curves are at a plateau at 0.5 s (phase 2) above $\epsilon > 0.25$ μm. The MFDT (right) increase with ϵ from 0.2 s at $\epsilon = 0.15$ μm to 0.5 s at $\epsilon = 0.5$ μm for the untreated (blue) and Zeocin treated (orange) case.

We conclude that uniformly distributed DSBs impair the MFET only at a scale below 0.25 μm, suggesting that this scale characterizes the local chromatin organization, in which undamaged loci can freely move, but become restricted above it. These finding confirms the confinement found in [44] (Supplementary Fig.5b),

which reported that a radius of confinement of 0.23 μm for Zeocin 500 $\mu g/ml$ treatment. However, we show here that following DSBs, the local exploration of the chromatin remain characterized by recurrent motion, and if repair molecules do affect the encounter time at a spatial scale below 0.25 μm, they do not prevent the dissociation time of the two loci.

4.2.3 Stochastic simulations of a DSB in randomly cross-linked (RCL) polymer

To further investigate the difference between chromatin reorganization before and after DSBs, reported above for the two loci dynamics, we now use a Randomly Cross-Linked (RCL) polymer, generalizing the Rouse polymer model (Methods), to evaluate the changes in the constraints of SPTs statistics following DSBs. The parameters of the RCL model are calibrated from experimental data [44], in which two tagged monomers are tracked before and after DSB induction between the tagged loci. The tagged loci were tracked over 60 s at a time interval of 300 ms. The confinement length [6] is defined as

$$L_c^2 = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} (R_c(k\Delta t) - \langle R_c \rangle)^2, \tag{4.3}$$

and we computed it before and after DSB induction, where $R_c(t)$ is the vector position between the two tagged loci at time t. As reported in [9], for an unbroken locus $L_c = 0.13$ μm, and $L_c = 0.23$ μm after DSB induction.

To identify the possible local chromatin reorganization underlying this difference in L_c, we simulate a RCL polymer (Methods) with $N = 100$ monomers, containing an additional N_c connectors between randomly chosen non-nearest neighboring monomers [30, 45, 94] (Fig. 4.4A). The added connectors reflects the compaction in the coarse-grained representation of the chromatin by molecules such as cohesin CTCF and condensin [81]. Randomly positioning connectors reflects the heterogeneity in chromatin architecture in a population of cells.

We first find the minimal number of random added connectors N_c by varying N_c in the range 10-150, for both unbroken loci and after DSB induction. We then computed L_c from simulations (1000 realizations for each N_c) and adjusted N_c to match the measured one. For each realization, we randomized the choice of monomer pairs to connect. We simulated each realization until the relaxation time $\tau_R = b^2 D/(3\lambda_1)$, where b is the standard-deviation of the vector between adjacent monomers and λ_1 is the smallest non-vanishing eigenvalue of the polymer's connectivity matrix [40], which we calculated numerically. We empirically found λ_1 to vary between 0.15 when $N_c = 10$ to 0.8 for $N_c = 150$, resulting in τ_R in the range of 20 minutes to 23 s until polymer relaxation. We then continued the simulations for an additional 200 steps at $\Delta t = 300$ ms for a total of 60 s, to match

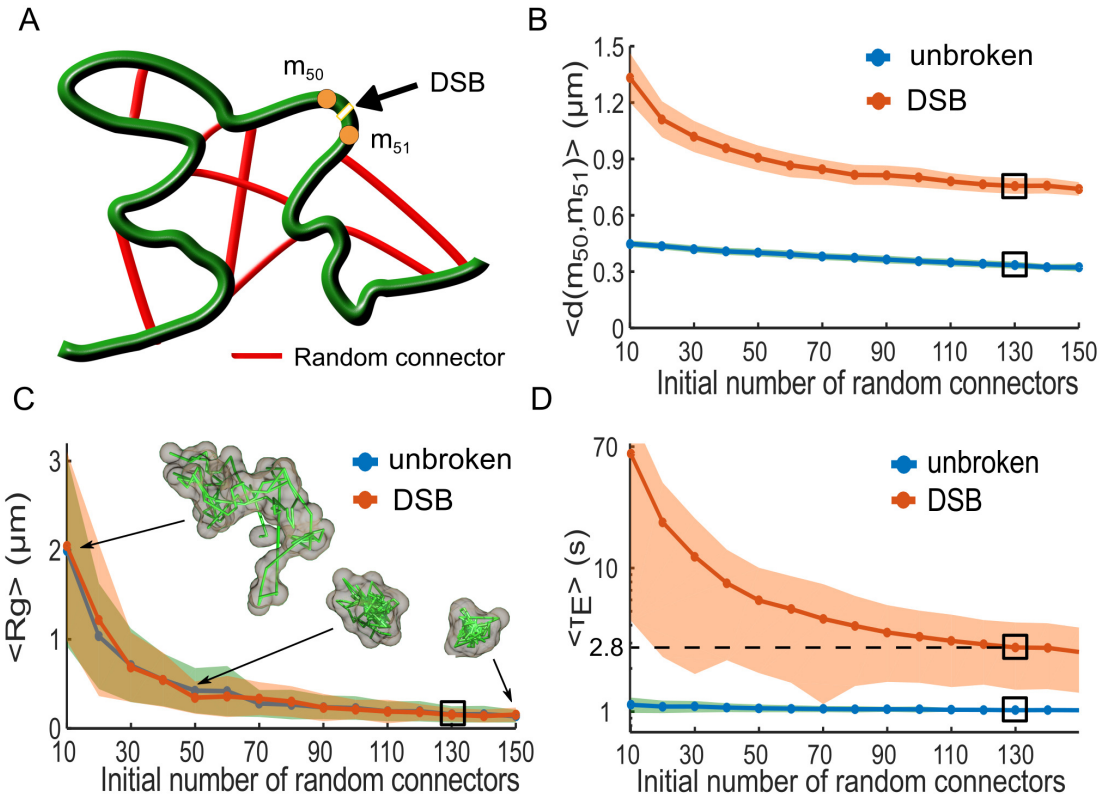


Figure 4.4 **Local force destabilization following a double-strand DNA break (DSB). A.** Schematic representation of a randomly cross-linked (RCL) polymer, where N_c random connectors (red) are initially added to the linear backbone (green) of a Rouse chain. A DSB is induced between monomers m_{50} and m_{51}, modeled by removing the spring connectors between them and all random connectors to these monomers. **B.** Mean maximal distance $\langle Max(d(m_{50}, m_{51}))\rangle$ for both the unbroken loci (blue) and DSB (orange) simulations, where the shaded are the STD. The black rectangle indicates the value obtained for $N_c = 130$ matching L_c (eq. 4.3) measurements reported in [9], where we obtain 0.37 μm for the unbroken and 0.86 μm for DSB simulation **C.** The mean radius of gyration (MRG), $\langle R_g\rangle$, obtained from simulations of 100 monomer RCL polymer (blue) and after DSB between monomers 50 and 51 (orange). Three sample polymer realizations are shown for $N_c = 10, 50,$ and 150. For $N_c = 130$ we obtain $\langle R_g\rangle = 0.15$ μm for both cases **D.** The mean first encounter time (MFET) $\langle \tau_E\rangle$ for m_{50} and m_{51} plotted with respected to N_c, for both the unbroken (blue) and DSB (orange) simulations. The MFET is displayed on a semi-log axes, where before DSB we obtained $\langle \tau_E\rangle = 1$ s and 2.8 s following DSB and the removal of 5 random connectors on average.

the experimental recorded time [44]. For DSB simulations, we induced a DSB between monomer 50 and 51 after the relaxation time τ_R and we then removed the spring connector between them. To account for the local chromatin decompaction, we further removed all random connectors to monomers 50 and 51. We discarded polymer configurations where the polymer chain was divided into two separated chains after the induction of DSB and removal random connector. Simulation parameters are summarized in Table 4.1, where N_c remains a free parameter.

We computed the average values of L_c for monomers 50 and 51 over each realization and found a good agreement between simulations and experimental data when $N_c = 130$ for both the unbroken ($L_c = 0.13$ μm) and broken loci ($L_c = 0.23$ μm). After DSB induction, we recover the value $L_c = 0.23$ μm for $N_c = 125$, where 5 connectors, on average are removed. Furthermore, the mean maximal distance between monomers 50 and 51 (Fig. 4.4B) decayed from 0.4 to 0.3 μm when N_c varied between 10 and 150 for the unbroken loci simulation, while it changes from 1.3 μm for $N_c = 10$ to 0.75 μm for $N_c = 150$ in the DSB simulations. When $N_c = 130$, the mean maximum monomer distance was 0.33 μm and increased to 0.75 μm after DSB induction.

Using the mean radius of gyration, $\langle R_g \rangle$ (Fig. 4.4C), computed for RCL polymer configuration, we show that compaction increases with N_c. Thus, $\langle R_g \rangle$ decreased from 2 μm for $N_c = 10$ to 0.15 μm at $N_c = 130$. Note that a single DSB does not affect the radius of gyration, $\langle R_g \rangle$, for all $N_c \in [10, 150]$ (Fig. 4.4C). In that figure, we further represented three polymer realizations for $N_c = 10, 50, 150$. For $N_c = 130$, the value of the gyration radius is $\langle R_g \rangle = 150$ nm for both the unbroken and DSB (numerical simulations). The average length $\langle L \rangle$ of loops of the RCL polymer can be computed using Eq. 4.17, with $b = 200$ nm, $N = 100$, and $Nc = 130$ (Table 4.1), we found (see Methods) that $\langle L \rangle = 4.8$ μm. This length should be compared to the total contour length L_{RCL} of the RCL polymer (expression 4.18), and obtained $L_{RCL} = 14$ μm. Thus, the average length of loops is roughly a third of the total polymer length. Looping occur in a confined cross-linked micro-environment, where the polymer is compacted in a ball of radius $R_g = 0.15$ μm (Fig. 4.4C) for $N_c = 130$. A length of 4.8 μm is converted with a compaction ratio of 50 bp/nm to 240 kbp, falling in the middle part of the range 2-500 kbp of loops reported between enhancers and promoters. It would be interesting to compare this size with the one generated by the interaction of enhancer-promoter in high resolution simulations with $b < 200$ nm.

To further examine the relationship between the chromatin local architecture and transient properties of the chromatin, we computed the first passage time (FET) between monomer m_{50} and m_{51}, before and after DSB induction (Fig. 4.4A). Each simulation realization was terminated when m_{50} and m_{51} enter for the first time within a distance less than ϵ, where we recorded the first encounter time, τ_E. Terminating each simulation after the first encounter, allowed us to randomize the

position of connectors for any other simulation and thus better account for chromatin structural heterogeneity. The MFET for the unbroken loci simulations decreases from 1.1 s to 1 s when N_c increases from 10 to 150, while it decreases from 62 s for $N_c = 10$ to 2.6 s for $N_c = 150$ (Fig. 4.4C) following DSB. Interestingly, when $N_c = 130$, only 5 connectors were removed on average to account for a DSB, but before induction $\langle \tau_E \rangle$ between m_{50} and m_{51} was 1 s and increased to 2.8 s after DSB induction. These time scales are consistent with data used in figures 4.1 and 4.2.

We conclude that the empirical confinement length can be accounted for in the RCL polymer using a $N_c = 130$ connectors. Following a DSB, the average number of removed connectors was 5, which represents 4% of the total number of connectors, N_c. Interestingly, the mean radius of gyration, $\langle R_g \rangle \approx 150$ nm, is mostly unchanged between the unbroken and DSB. However, the encounter time τ_E (Fig. 4.4D changed from 1 to 2.8 s, showing that removing the key connector could affect the local encounter time. The drastic effect of changing the number of connectors appears in the mean maximal distance between the two monomers, increasing from 0.33 μm in the unbroken case to 0.75 μm after DSB, leading to high increase in the local search time. This search in local restricted environment could be at the basis of the Non-Homologous end joining (see Discussion).

4.3 Discussion

We introduced here a transient analysis of loci trajectories based on computing the first encounter times between two simultaneously tagged chromatin loci to a small distance. Because the positions of the loci fluctuate in time but return recurrently into close proximity, this dynamics generates enough statistical events. We showed here that this statistics revealed a characteristics length around 250 nm, where the chromatin constrains the two loci dynamics. This analysis cannot be obtained from the traditional parameters, extracted from SPTs such as the mean square displacement (MSD) or the anomalous exponent [46], which characterize the dynamics of individual locus separately. Such parameters were used in the past to study the deviation from Brownian motion [66]. Further information about chromatin organization is obtained from individual single loci trajectories [26], such as the length of constraint characterizing confinement or the tethering force to account for the first statistical moment and the mean force responsible for confinement [28, 106, 7, 9, 44, 38].

The statistics of the FET and FDT account for the correlated properties of two loci and are directly related to the transient properties of the chromatin: the FET reveals that the recurrent visit time between the loci, depending on the genomic distance (Fig. 4.1), varies from 1.5 to 2.5 s. We further confirm that altering the chromatin integrity by generating DNA damages using the Zeocin drug [44] affects the MFET at a distance lower than 250 nm (Fig. 4.2B), showing that this scale

is certainly critical in chromatin remodeling [44]. Above this distance, the MFET was constant and we interpret this result as a consequence of the local crowding effect. These results further show that the recurrent visits between two loci can be modulated by chromatin remodeling. The confinement length of few hundreds of nanometers estimated here is compatible with the one extracted from Hi-C data of the order of 220 nm, using polymer looping in confined microdomains [4].

To further explore how chromatin re-organization affects recurrent loci encounters, we use the RCL polymer model [94], which is a Rouse polymer with added random connectors. This approach consists in adding random connectors is more accurate in representing tethering force than the average computation that we introduced in [7]. Random cross-linking in the RCL model serves to simulate the confined environment [103] and the heterogeneity in chromatin architecture in cell population. We used the changes in the length of constraint reported in [44] to calibrate the number of added random connectors and simulated trajectories of the RCL before and after the induction of DSB. Interestingly, the consequence of DSB damages on chromatin reorganization is equivalent to removing 4% of the connectors in the vicinity of the DSB, leading to an increase of distance between the two broken part from 0.4 to 0.9 μm, while the mean radius of gyration, $\langle R_g \rangle$, was almost unchanged at 0.15 μm. However, the MFET increased from 1 to 2.8 s. The random cross-links in the RCL model thus play the role of the confining environment, which prevents the two ends from drifting apart (Fig 4.4B and C), similarly to the crowding effect seen in [103] for self-avoiding polymers. Bending elasticity and self avoidance could be accounted for by altering the number of random connectors.

The present model reveals that the local confined decompaction following DSB prevents the two ends to drift apart, which could have drastic consequences in dsDNA break repair processes, such as during non-homologous-end-joining (NHEJ), where the two ends should be re-ligated together. The possible role of stabilizing the broken ends by maintaining a large number of connectors is probably to avoid inappropriate NHEJ religations that can lead to translocations or telomere fusion. We remark that the MFET that we computed here cannot be used to study the other repair process called homologous recombination, which is based on a long-range spatio-temporal search for a homologous template [12, 28]. We conclude that the present first passage time statistics, derived from polymer simulations, can be used to analyze any temporal correlation between loci pairs. It would certainly be interesting to record three loci simultaneously at different distances and apply our method to it to obtain refine properties of chromatin reorganization.

Acknowledgments

We thank Tom Owen-Hughes lab for sending us their SPTs data. Data sets were previously published in [44, 26]. DH thanks the Simons foundation for support. This research was supported by a Marie Curie Award to DH.

4.4 Theory and Methods

4.4.1 Looping times in chromatin polymer models

To analyze the statistics of two loci located on the same chromatin arm, we use the classical Rouse polymer model that describes a collection of beads $R_n(n = 1...N)$ connected by harmonic springs and driven by Brownian motion [30]. The energy of the polymer is given by

$$\phi_{Rouse}(R_1, .., R_N) = \frac{1}{2}\sum_{j=1}^{N-1} \kappa(R_j - R_{j+1})^2, \tag{4.4}$$

where $\kappa = \frac{3k_BT}{\gamma b^2}$ is the spring constant, b is the standard deviation of the connector between adjacent monomers, γ is the friction coefficient, k_B the Boltzmann constant, and T the temperature.

The first encounter time (FET) between two loci is defined as the first time the two loci are positioned within a ball of radius ϵ. The distribution of FET between the two ends of a polymer chain is well approximated by a Poisson process in free and confined domains [4, 8]. In both cases, the distribution of the decay rate constant, λ_E, is the reciprocal of the mean first encounter time (MFET) $\langle \tau_E \rangle = \frac{1}{\lambda_E}$, and the probability density function is

$$p(t) \approx e^{-\lambda_E t}, \tag{4.5}$$

In a confined domain, the expression for the MFET is

$$\begin{aligned} \langle \tau_E^c \rangle &\approx \frac{2^{1/2}}{4\pi\epsilon D}\left[\frac{4\pi/N}{\beta + \kappa(\pi/N)^2} + \frac{4}{\sqrt{\kappa\beta}}\left[\frac{\pi}{2}\right.\right. \\ &- \left.\left.\tan^{-1}\left(2\sqrt{\kappa/\beta}\tan\left(\pi/2N\right)\right)\right]\right]^{3/2} + \mathcal{O}(1), \end{aligned} \tag{4.6}$$

where

$$\beta = \frac{12}{A^4/b^2 + 2A^2}, \tag{4.7}$$

and A is the radius of a sphere confining the polymer [4].

4.4.2 Dissociation times in a parabolic potential

To characterize the dissociation time of two loci, we adopt the Kramer's escape over a potential barrier [91]. The potential can be due to the average forces between local monomers. We model it as an effective parabolic well truncated at a hight H.

In the deep circular well approximation of size a_0 [91], the escape time for a process $\dot{X} = -\nabla U + \sqrt{2D}\dot{w}$ is (in two dimensions)

$$\langle \tau_D \rangle = \frac{Da^2}{4\alpha} e^{\frac{\alpha}{D}}, \tag{4.8}$$

where $U(r) = \alpha \frac{r^2}{a_0^2}$ and the energy is $E = U(a) = \alpha$ and $U(0) = 0$. The distribution of escape time is Poissonian with rate $\frac{1}{\langle \tau_D \rangle}$.

For the effective problem of unlooping to a certain distance, we consider that this problem is equivalent to the escape of a particle from a well with diffusion coefficient ND, where N is the number of monomers. In the present case, N is proportional to Δ and we have used the empirical formula:

$$\langle \tau_D \rangle = a_2 T e^{\frac{b_2}{T}}, \tag{4.9}$$

where a_2 and b_2 are two constants.

4.4.3 Computing the average loop size from the randomly cross-linked (RCL) polymer model

We summarize here our computations for the average length of loops in the RCL polymer. We define the length of the loop between any two connected monomers, m and n, as their linear distance $|m - n|$ along the backbone of the polymer. We do not compute here the average shortest possible loop length between monomer m and n, which might result from configurations of other connected monomers of the polymer.

Each realization of the RCL polymer is a (uniformly random) choice of N_c non-neighboring monomer pairs to connect from the ensemble of possible N_L pairs, given by

$$N_L = \frac{(N-1)(N-2)}{2}, \tag{4.10}$$

The ensemble of N_L possible choices of monomer pairs contains the disjoint subsets $\{L_k\} = \{(m,n); |m-n| = k\}$ of loops with length $2 \leq k \leq N-2$, where the size of each subset $\{L_k\}$ is

$$|L_k| = N - k. \tag{4.11}$$

The fraction p_k of each subset $\{L_k\}$ out of the total N_L possibilities is

$$p_k = \frac{|L_k|}{N_L}. \tag{4.12}$$

Thus, the number of loops of length k monomers is

$$E(k) = N_c p_k = N_c \frac{|L_k|}{N_L}. \tag{4.13}$$

The expected length (in non physical units) of a loop is obtained by averaging over all loops of size k of the RCL polymer,

$$\begin{aligned} E(L) &= \sum_{k=2}^{N-1} p_k k = \sum_{k=2}^{N-1} \frac{k(N-k)}{N_L} = \frac{1}{N_L}\left(N\sum_{k=1}^{N} k - \sum_{k=1}^{N} k^2 - (N-1)\right) \\ &= \frac{1}{N_L}\left(\frac{N^2}{1+N} - (N-1) - \frac{N^3}{3} - \frac{N^2}{2} - \frac{N}{6}\right) = \frac{N^3 - 7N + 6}{6N_L}. \end{aligned} \tag{4.14}$$

To obtain the physical length of the average loop size in μm units, we multiply the mean length (non-dimensional units, equation 4.14) by standard-deviation (STD) of the distance between adjacent monomers for the RCL polymer. An analytical expression is available (Eq. 30 [93]), and can be approximated by

$$\sigma(N, \xi, b) = \left(\frac{b^2(1 - \exp(-\sqrt{N\xi}))}{\sqrt{N\xi}}\right)^{\frac{1}{2}}, \tag{4.15}$$

where

$$\xi = \frac{N_c}{N_L} \tag{4.16}$$

is the connectivity fraction and b has units of μm. By multiplying Eq. 4.14 by 4.15, we obtain an approximation for the average loop length

$$\langle L \rangle = \sigma(N, \xi, b) E(L) = \left(\frac{b^2(1 - \exp(-\sqrt{N\xi}))}{\sqrt{N\xi}}\right)^{\frac{1}{2}} \left(\frac{N^3 - 7N + 6}{6N_L}\right). \tag{4.17}$$

We note that the total contour length L_{RCL} of the RCL polymer is computed by multiplying the number of monomers, N, by expression 4.15

$$L_{RCL} = N \left(\frac{b^2(1 - \exp(-\sqrt{N\xi}))}{\sqrt{N\xi}}\right)^{\frac{1}{2}}. \tag{4.18}$$

4.4.4 Construction of the randomly cross-linked (RCL) polymer model

The Rouse polymer [30] describes chromatin below a scale of few Mbp [88, 17]. Starting from a Rouse model [30], the RCL is constructed by adding sparse connected

pairs (Fig. 4.4A red), chosen with a uniform probability such that the potential for the polymer is the sum of the ϕ_{Rouse} plus the potential

$$\phi_{Rand}(R_1,..,R_N) = \frac{1}{2}\sum_{j,k\in\mathcal{C}_\mathcal{N}} \kappa(R_j - R_k)^2, \tag{4.19}$$

where $\mathcal{C}_\mathcal{N}$ is an ensemble of indices from 1 to N. The chromatin is modeled as a polymer chain with a uniform variance b^2 between adjacent monomers. The total energy of a polymer containing random connectors is the sum of two energies 4.4, and 4.19

$$\Phi(R_1,..,R_N) = \phi_{Rand}(R_1,..,R_N) + \phi_{Rouse}(R_1,..,R_N), \tag{4.20}$$

and the stochastic equation of motion for $n = 1,..,N$ is

$$\frac{\mathrm{d}R_n}{\mathrm{d}t} = -\nabla_{R_n}\Phi(R_1,..,R_N) + \sqrt{2D}\frac{\mathrm{d}\omega_n}{\mathrm{d}t} \tag{4.21}$$

where $D = \frac{k_B T}{\gamma}$ is the diffusion constant, γ is the friction coefficient, and ω_n are independent 3-dimensional Brownian motion with mean 0 and standard deviation 1. We use this construction to estimate the minimal number of connectors before and after a dsDNA-breaks.

Simulations of the RCL polymer were performed using codes written in Julia v0.5.1 [15]. Codes are available on the Bionewmetric website http://bionewmetrics.org/. We summarize in Table 4.1 the values of parameters used in simulations

Parameter	Value	Description
N	100	number of monomers
b	200 nm	STD of adjacent monomers distance
D	$8 \times 10^{-3}\ \mu m^2/s$	Diffusion coefficient [4]
ϵ	0.2 μm	Encounter distance
κ	1.75×10^{-2} N/m	Spring constant[3]
γ	3.1×10^{-4} Ns/m	friction coefficient [3]

Table 4.1 values of simulation parameters

5 Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data

Published in Shukron Ofir, and David Holcman. "Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data.", Physical Review E 96.1 (2017): 012503.

Abstract

Polymer models are used to describe chromatin, which can be folded at different spatial scales by binding molecules. By folding, chromatin generates loops of various sizes. We present here a statistical analysis of the randomly cross-linked (RCL) polymer model, where monomer pairs are connected randomly, generating a heterogeneous ensemble of chromatin conformations. We obtain asymptotic formulas for the steady-state variance, encounter probability, the radius of gyration, instantaneous displacement and the mean first encounter time between any two monomers. The analytical results are confirmed by Brownian simulations. Finally, the present results are used to extract the mean number of cross-links in a chromatin region from conformation capture data.

5.1 Introduction

DNA in the nucleus is constantly remodeled by regulatory factors, and compacted genomic regions form transient and stable loops [76, 101]. Looping is thus a key event in chromatin regulation: although rare for a single polymer, it is frequent in a population of hierarchy folded genome. Genome organization is probed by chromatin Conformation Capture (CC) techniques [25, 95, 65], which give access to simultaneous looping events in an ensemble of millions of chromatin segments. This experimental approach provides contact frequency matrices at various scale from few kilo- to Mega-base-pairs. Analysis of these matrices remains difficult, but already revealed that mammalian genomes contain Mbp "blocks" of enriched connectivity, called Topologically Associating Domains (TADs) [76, 29]. The role of TADs and their organization remains unclear, although they are involved in gene regulation [76, 95] and replication. TADs appear by averaging encounters over an ensemble of

millions of samples [65] and represents steady-state looping frequencies, but does not contain neither information about the size of the folded genomic section nor any transient genomic encounter times.

To reconstruct chromatin at a given scale and explore its transient properties, polymer models are used as a coarse-grained representation. The Rouse model [30], characterized by nearest neighbors interactions, predicts an encounter probability (EP) that decays with $|m-n|^{-3/2}$ between monomer m and n, but cannot account for long-range interactions, observed inside TADs [94, 76]. Other polymer models include attractive and repulsive forces between monomers [97, 17, 16, 45, 5, 61], to account for long-range interactions, and have been used to probe the heterogeneous steady-state organization of the chromatin [55, 11].

We study here a randomly cross-linked (RCL) polymer model used in [94] to describe the ensemble of steady-state chromatin conformations, present in CC data [76, 25, 29]. Cross-links could be generated by either binding molecules (CTCF [76]) or by a hypothetical loop extrusion mechanism [37], but this exact formation mechanism is not the focus of the present model. Randomly cross-linked polymers were previously studied on a scale of a single protein molecule [100] and for cross-linked networks [86], where monomer connectivity modulate the energy landscape. The steady-state statistical properties of cross-linked polymers is similar to other physical areas, such as resistor networks, where analytical formula were derived for the mean-square distance between resistor with prescribed connectivity, such as rings and stars [40, 34]. Other applications came from the dynamic of random loop models in polymer physics [17, 21, 53] or fractal networks [74].

The RCL polymer configuration space was so far mostly explored numerically [45, 53, 11, 33]. However, computing the encounter probability and the mean first encounter time, which are key quantities of interest to extract chromatin dynamics from the CC, was left open. We derive here formulas for the EP, the variance, and the radius of gyration of the RCL polymer, that we use in a key step of chromatin reconstruction using polymer models. The present model can be used to determine from CC empirical EP, the average number of cross-links, a quantity inaccessible from CC experiments. We further derive an asymptotic formula for the mean first encounter time between any two monomers, which plays a key role in gene regulation [57]. Our asymptotic derivations are further confirmed by Brownian simulations.

A general procedure to extract the average number of cross-links in a genomic section based on the EP decay of the 5C data [94] is available, but it requires to perform heavy iterative simulations. Using the present analysis, we derive an analytical formula that allow us to determine the number of loops or connectors directly from the EP of CC data.

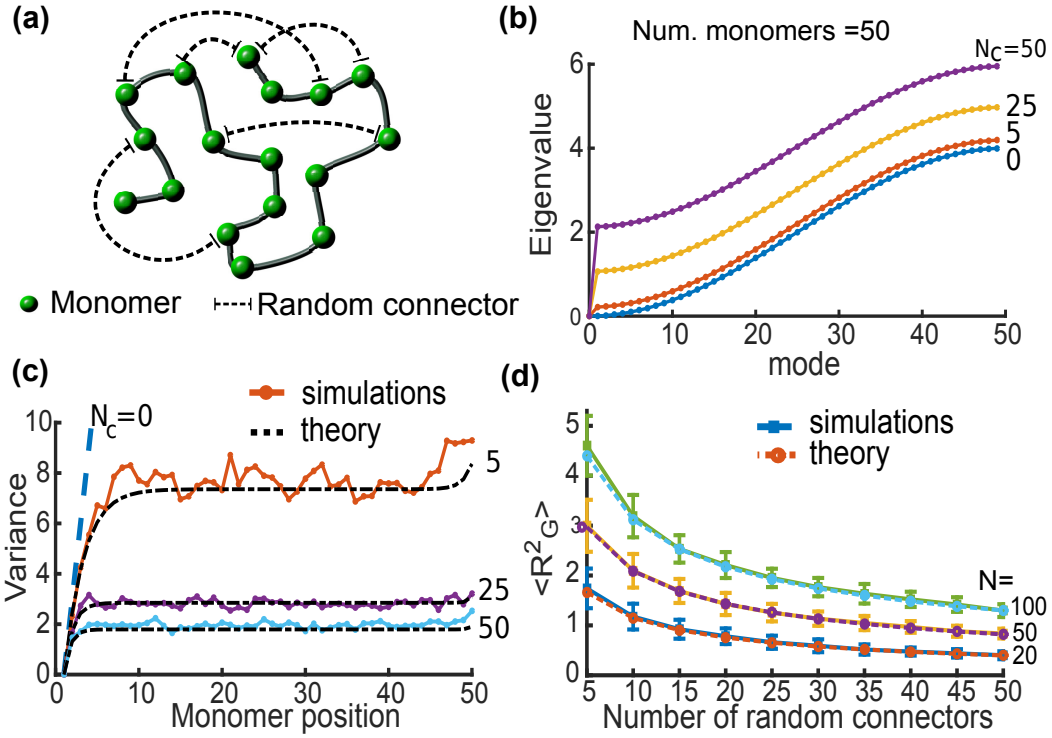


Figure 5.1 Properties of a Randomly-Cross-Linked (RCL) polymer. **(a)** RCL polymers are composed of a linear backbone of N monomers (spheres), and $N_c(\xi)$ random connectors (dashed) between non-nearest neighboring monomers pairs. **(b)** Eigenvalues of the RCL polymer (Eq. 5.16) with $N = 50$ monomers, and $N_c(\xi) = 5$ (blue), 25 (red), and 50 (yellow) connectors. **(c)** Variance of the monomers distance: analytical (dashed, Eq.5.28) versus simulations (Eq.5.13), between monomer 1 and monomers 2-50 of the RCL polymer, 500 realizations with $N = 50, b = \sqrt{d}, D = 1, \Delta t = 0.01s$, for $N_c(\xi) = 5$ (blue), 25 (yellow) and 50 (green) added random connectors, computed after 10^4 steps, corresponding to the slowest relaxation time τ_0 (see Eq.5.22). **(d)** Mean square radius of gyration, $\langle R_G^2(\xi)\rangle$, with $N = 20$ (blue), 50 (yellow), and 100 (green) monomers: analytical (dashed, Eq. 5.32), where $N_c(\xi) \in [5, 50]$, versus stochastic simulations of 5.13 (continuous).

5.2 Results

5.2.1 The RCL polymer model

A linear polymer in dimension d $(d = 3)$ consists of N monomers at positions $\boldsymbol{R} = [r_1, r_2, ..., r_N]^T$, connected sequentially by harmonic springs [30], and we added spring connectors between random non-nearest neighboring (NN) monomer pairs (Fig. 5.1a). The energy of the RCL polymer, introduced in [21, 17], is the sum of the spring potential of linear backbone and that of random connectors

$$\phi_{\mathcal{G}}(\boldsymbol{R}) = \frac{\kappa}{2}\sum_{n=2}^{N}(r_n - r_{n-1})^2 + \frac{\kappa}{2}\sum_{\mathcal{G}}(r_m - r_n)^2, \tag{5.1}$$

where $\kappa = dk_BT/b^2$ is the spring constant, b the standard-deviation of the connector between connected monomers, k_B is the Boltzmann's constant, and T the temperature. The ensemble $\mathcal{G}$ is composed of N_c randomly chosen indices m, n among the non-NN monomers and this set is re-computed for each polymer realization to account for the large polymer conformational space. The connectivity fraction $0 \leq \xi \leq 1$, is the fraction of the total connector numbers $N_L = \frac{(N-1)(N-2)}{2}$, defined by

$$N_c(\xi) = \lfloor \xi N_L \rfloor. \tag{5.2}$$

For each polymer realization, we choose $N_c(\xi)$ pairs from N_L possible NN monomers, thus, leading each time to a new ensemble of indices in $\mathcal{G}$ (equation 5.1). The dynamics of the resulting polymer model (vector $\boldsymbol{R}$) is given by Smoluchowski limit of the Langevin equation, which is the sum of Brownian motion and the gradient force induced by the potential energy (equation 5.1),

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} &= -\frac{1}{\zeta}\nabla\phi_{\mathcal{G}}(\boldsymbol{R}) + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t} \\ &= -\frac{d}{b^2}D\left(M + B^{\mathcal{G}}(\xi)\right)\boldsymbol{R} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t},\end{aligned} \tag{5.3}$$

where $D = \frac{k_BT}{\zeta}$ is the diffusion constant, ζ is the friction coefficient, $\boldsymbol{\omega}$ are independent Brownian motion with mean 0 and variance 1, and M is the $N \times N$ Rouse matrix [30]

$$M_{m,n} = \begin{cases} -\sum_{j\neq m} M_{m,j}, & m = n; \\ -1 & |m-n| = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

For a given connectivity fraction, ξ, the square symmetric matrix $B^{\mathcal{G}}(\xi)$ with random connectivity is given by

$$B^{\mathcal{G}}_{mn}(\xi) = \begin{cases} -1, & |m-n| > 1, \text{and connected in } \mathcal{G}; \\ -\sum_{i\neq j}^{N} B_{mj}(\xi), & m = n; \\ 0, & \text{otherwise.} \end{cases}$$

The steady-state properties of an ensemble of RCL polymers are contained in the mean-field model, where we replace the matrix $B^{\mathcal{G}}(\xi)$ in Eq. 5.3 by its average $\langle B^{\mathcal{G}}(\xi)\rangle$ (averaging over all realizations $\mathcal{G}$ of non NN monomer pairs when the number of connectors $N_c(\xi)$ is fixed). We construct $\langle B^{\mathcal{G}}(\xi)\rangle$ using the probability density of the monomer connectivity.

For a fixed number of connector $N_c(\xi)$, the probability that monomer m has $k \leq (N-2)$ non-NN connections is obtained by choosing k position in row m of the matrix $B^{\mathcal{G}}(\xi)$ (excluding the super- and sub- and the diagonal), and the remaining $Nc-k$ connectors in any row or column $n \neq m$, thus:

$$Pr_m(k) = \begin{cases} \frac{C^{N_c(\xi)-k}_{N_L-(N-3)} C^k_{N-3}}{C^{N_c(\xi)}_{N_L}}, & 1 < m < N; \\ \frac{C^{N_c(\xi)-k}_{N_L-(N-2)} C^k_{N-2}}{C^{N_c(\xi)}_{N_L}}, & m = 1, N, \end{cases} \tag{5.5}$$

where the binomial coefficient is $C^j_i = \frac{i!}{(i-j)!j!}$. This probability is the hyper-geometric distribution for the number of connections for monomer m. The mean number of connectors for each monomer is, therefore

$$\beta_m(\xi) = \begin{cases} \frac{(N-3)N_c(\xi)}{N_L} \approx (N-3)\xi, & 1 < m < N; \\ \frac{(N-2)N_c(\xi)}{N_L} \approx (N-2)\xi, & m = 1, N. \end{cases} \tag{5.6}$$

Using the mean values in 5.6, we obtain the expression for the matrix $\langle B^{\mathcal{G}}(\xi)\rangle$

$$\langle B^{\mathcal{G}}_{mn}(\xi)\rangle = \begin{cases} -\xi, & |m-n| > 1; \\ \beta_m(\xi), & m = n; \\ 0, & \text{otherwise,} \end{cases} \tag{5.7}$$

which can be decomposed as the sum

$$\langle B^{\mathcal{G}}(\xi)\rangle = \xi(N\boldsymbol{I_d} - \boldsymbol{M} - \mathbf{1}_N), \tag{5.8}$$

where $\boldsymbol{I_d}$ is the $N \times N$ identity matrix, and $\mathbf{1}_N$ is a $N \times N$ matrix of ones. To study the mean properties of the RCL polymer, we study the stochastic process 5.3 using the average matrix $\langle B(\xi)\rangle$ in relation 5.8.

5.2.2 Eigenvalues of the RCL polymer.

To study the steady-state properties of system 5.3, we diagonalize the averaged connectivity matrix $M+\langle B(\xi)\rangle$. Using Rouse normal coordinates $\boldsymbol{U} = [u_0, u_1, ..u_{N-1}]$ [30], defined as

$$\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}, \tag{5.9}$$

where

$$\boldsymbol{V} = (\alpha_p^n) = \begin{cases} \sqrt{\frac{1}{N}}, & p = 0; \\ \sqrt{\frac{2}{N}} \cos\left((n-\frac{1}{2})\frac{p\pi}{N}\right), & \text{otherwise.} \end{cases} \tag{5.10}$$

The Rouse orthonormal basis $\boldsymbol{V}$ [30] diagonalizes M to

$$VMV^T = \Lambda = diag(\lambda_0, \lambda_1, ..., \lambda_{N-1}), \tag{5.11}$$

where

$$\lambda_p = 4\sin^2\left(\frac{p\pi}{2N}\right), \quad p = 0, .., N-1, \tag{5.12}$$

are the eigenvalues of the Rouse matrix (relation 5.4). Substituting $\langle B^{\mathcal{G}}(\xi)\rangle$ for $B^{\mathcal{G}}(\xi)$ in system 5.3 and multiplying it from the left by $\boldsymbol{V}$ in 5.10, we obtain the mean-field equations

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -\frac{d}{b^2} D\left[\Lambda + \boldsymbol{V}\langle B^{\mathcal{G}}(\xi)\rangle \boldsymbol{V}^T\right]\boldsymbol{U} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}t}, \tag{5.13}$$

where $\boldsymbol{\eta} = \boldsymbol{V}\omega$ are independent Brownian motion with mean 0 and variance 1. From identity 5.8, the matrix $\langle B^{\mathcal{G}}(\xi)\rangle$ commutes with M and, therefore, is diagonalizable using the same orthonormal basis $\boldsymbol{V}$:

$$\boldsymbol{V}\langle B^{\mathcal{G}}(\xi)\rangle \boldsymbol{V}^T = diag(\gamma_0(\xi), ..., \gamma_{N-1}(\xi)). \tag{5.14}$$

Using 5.8 and 5.14, we obtain the eigenvalues

$$\gamma_p(\xi) = \begin{cases} 0, & p = 0; \\ \xi(N - \lambda_p), & 1 \le p \le N-1. \end{cases} \tag{5.15}$$

To conclude, the eigenvalues of system 5.13 are the sum of eigenvalues of the Rouse matrix M and $\langle B^{\mathcal{G}}(\xi)\rangle$:

$$\chi_p(\xi) = \gamma_p(\xi) + \lambda_p = N\xi + 4(1-\xi)\sin^2\left(\frac{p\pi}{2N}\right). \tag{5.16}$$

The stochastic system 5.13 consists of $N-$independent equations. For $\xi = 0$, we recover the Rouse polymer [30], whereas for $\xi = 1$, we obtain a fully connected polymer, with a circular matrix $\boldsymbol{M} + \langle B^{\mathcal{G}}(\xi)\rangle$, for which all eigenvalues equal to N except for the first vanishing one. Using relations 5.9, 5.16 and 5.10 in 5.1, the energy of the RCL polymer is written as

$$\phi_{\mathcal{G}}(\boldsymbol{U}) = \frac{\kappa}{2}\sum_{p=1}^{N-1}\chi_p(\xi)u_p^2. \tag{5.17}$$

The statistics of the RCL system (relation 5.3), can be recovered from 5.13 in the diagonalized form (expression 5.17), by scaling ξ with the ratio of mean number of random connectors to the mean of total number of connectors:

$$\xi^* = \xi\frac{N_c(\xi)}{N + N_c(\xi)}. \tag{5.18}$$

The ensemble of eigenvalues eigenvalues 5.16 for RCL polymers, for $N = 50$ monomers, and $N_c(\xi)$=5, 25, and 50 added random connectors is shown in Fig.5.1b.

5.2.3 Encounter probability (EP) between monomers of the RCL polymer.

The RCL polymer belongs to the class of generalized Gaussian chain models, studied in [97, 40, 53, 34], for which the EP between any two monomers m and n at equilibrium is given by

$$P_{m,n}(\xi) = \left(\frac{d}{2\pi\sigma^2_{m,n}(\xi)}\right)^{\frac{d}{2}}. \tag{5.19}$$

To compute expression 5.19 explicitly, we estimate now the variance $\sigma^2_{m,n}(\xi) = \langle(r_m - r_n)^2\rangle$ in normal coordinates (Eq. 5.9):

$$\sigma^2_{m,n}(\xi) = \sum_{p=0}^{N-1}\left(\alpha_p^m - \alpha_p^n\right)^2\langle u_p^2(\xi)\rangle. \tag{5.20}$$

Although computational methods to study the steady-sate variance of Gaussian models were introduced already in [34], we provide here a computation of the variance using the normal coordinates (Eq. 5.9) and the eigenvalues (Eq. 5.16), which we will use below to compute time-dependent polymer properties.

The time-dependent variance is computed from the decoupled Ornstein-Uhlenbeck equations 5.13 [91], we obtain

$$\langle u_p^2(\xi)\rangle = \frac{b^2}{\chi_p(\xi)}\left(1 - \exp\left(-\frac{2D\chi_p(\xi)t}{b^2}\right)\right). \tag{5.21}$$

The relaxation times $\tau_0 \geq \tau_1(\xi) \geq ..\tau_{N-1}(\xi)$ are

$$\tau_p(\xi) = \frac{b^2}{2D\chi_p(\xi)}, \tag{5.22}$$

and the slowest, $\tau_0(\xi)$, corresponds to the diffusion of the center of mass. At steady-state,

$$\langle u_p^2(\xi)\rangle = \frac{b^2}{2(1-\xi)\left(y(N,\xi) - \cos\left(\frac{p\pi}{N}\right)\right)}, \tag{5.23}$$

where

$$y(N,\xi) = 1 + \frac{N\xi}{2(1-\xi)}. \tag{5.24}$$

Substituting relations 5.10 and 5.23 into 5.20, we get

$$\sigma^2_{m,n}(\xi) = \sum_{p=0}^{N-1} \frac{b^2\left(\cos\left(\frac{p(m-\frac{1}{2})\pi}{N}\right) - \cos\left(\frac{p(n-\frac{1}{2})\pi}{N}\right)\right)^2}{N(1-\xi)(y(N,\xi) - \cos(\frac{p\pi}{N}))}. \tag{5.25}$$

For $N \gg 1$, the sum 5.25 is approximated by an integral (Euler Mac-Lauren formula),

$$\begin{aligned}\sigma^2_{m,n}(\xi) &= \int_{-\pi}^{\pi} \frac{b^2\left(\cos(x(m-\frac{1}{2})) - \cos(x(n-\frac{1}{2}))\right)^2 dx}{2\pi(1-\xi)(y(N,\xi) - \cos(x))} \\ &= \oint_{|z|=1} \frac{-b^2(z-z^{m+n})^2(z^m - z^n)^2 dz}{4\pi i(1-\xi)(z-\zeta_0(N,\xi))(z-\zeta_1(N,\xi))z^{2(m+n)+1}},\end{aligned} \tag{5.26}$$

where the boundaries of integration $[0,\pi]$ are transformed in the complex plane using the contour of the unit disk parameterized by $z = e^{ix}$, and we define

$$\begin{aligned}\zeta_0(N,\xi) &= y(N,\xi) + \sqrt{y^2(N,\xi) - 1}, \\ \zeta_1(N,\xi) &= y(N,\xi) - \sqrt{y^2(N,\xi) - 1}.\end{aligned} \tag{5.27}$$

When $\zeta_0(N,0) = 1$, we recover from expression 5.25 the variance, $\sigma^2_{m,n}(0) = b^2|m-n|$, of the Rouse chain ($N_c(\xi) = 0$) [30]. The integrand in 5.26 is symmetric in m and n and has a pole of order $2(m+n)+1$ at $z = 0$ and simple poles at $z = \zeta_0(N,\xi), z = \zeta_1(N,\xi)$. Because $y(N,\xi) \geq 1$, we have $\zeta_0(N,\xi) \geq 1$, which is outside the contour $|z| = 1$, and $\zeta_1(N,\xi) \leq 1$, for all N, $\xi \geq 0$. The pole $\zeta_0(N,\xi)$ is not in the disk and does not contribute to the residues of 5.26. For $\xi > 0$, we solve

the integral 5.26 to obtain an exact expression for the variance. With the notations $\zeta_0 = \zeta_0(N,\xi), \quad \zeta_1 = \zeta_1(N,\xi)$, we have

$$\sigma^2_{m,n}(\xi) = \begin{cases} \dfrac{b^2\left((\zeta_0^{m-n}(N,\xi)-1)^2 - 2\zeta_0^{m+n-1} + 2\zeta_0^{2m-1}\right)}{(1-\xi)(\zeta_0-\zeta_1)\zeta_0^{2m-1}(N,\xi)}, m \geq n; \\ \dfrac{b^2\left((\zeta_0^{n-m}-1)^2 - 2\zeta_0^{m+n-1} + 2\zeta_0^{2n-1}\right)}{(1-\xi)(\zeta_0-\zeta_1)\zeta_0^{2n-1}}, m < n. \end{cases} \tag{5.28}$$

For $0 < \xi \ll 1$, $k > 1$, we approximate the terms 5.27 by

$$\begin{aligned} \zeta_0^k(N,\xi) &\approx \exp(k\sqrt{N\xi}); \\ \zeta_1^k(N,\xi) &\approx \exp(-k\sqrt{N\xi}); \end{aligned} \tag{5.29}$$

and use 5.29 in expression 5.28, to obtain the asymptotic expression for the variance

$$\sigma^2_{m,n}(\xi) \approx \frac{b^2}{\sqrt{N\xi}}\left(1 - \exp(-|m-n|\sqrt{N\xi})\right). \tag{5.30}$$

To check the range of validity of formula 5.28, we use Brownian simulations (Fig. 5.1c), computed after a relaxation time τ_0 (10^4 numerical steps) for $N = 50, N_c = 5, 25$ and 50. Substituting relation 5.28 in 5.19, we obtain a novel expression for the steady-state EP $P_{m,n}(\xi)$ between any two monomers. We then compare the EP obtained from Brownian simulations of RCL polymer for $N = 20, 50$ with the analytical formula 5.19 for $N_c(\xi) = 25, 50$ connectors (Fig. 5.2a), which shows a very good agreement.

5.2.4 Mean square radius of gyration (MSRG) of the RCL polymer.

The MSRG, $\langle R_G^2(\xi)\rangle$, characterizes the size of the RCL polymer, and can be computed from the expression of the variance 5.28 by the following formula [30]

$$\langle R_G^2(\xi)\rangle = \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{m}\sigma^2_{m,n}(\xi). \tag{5.31}$$

To compute the sum 5.31, we use elementary formula for the sum of geometric series. Using relation 5.28 in 5.31 and with the notations $\zeta_0 = \zeta_0(N,\xi), \quad \zeta_1 = \zeta_1(N,\xi)$, we obtain

$$\begin{aligned} \langle R_G^2(\xi)\rangle \;=\; & \frac{b^2}{N^2(1-\xi)(\zeta_0-\zeta_1)}\Big[\frac{(1+2\zeta_0)N(1+N)}{2\zeta_0} + \frac{N(2(1+\zeta_0)^2-\zeta_0^3)}{1-\zeta_0^2} \\ & - \frac{\zeta_0^3(1-\frac{1}{\zeta_0^{2N}})}{(1-\zeta_0^2)^2} + \frac{2(1+\zeta_0)(1-\frac{1}{\zeta_0^N})}{(1-\zeta_0)^2}\Big]. \end{aligned} \tag{5.32}$$

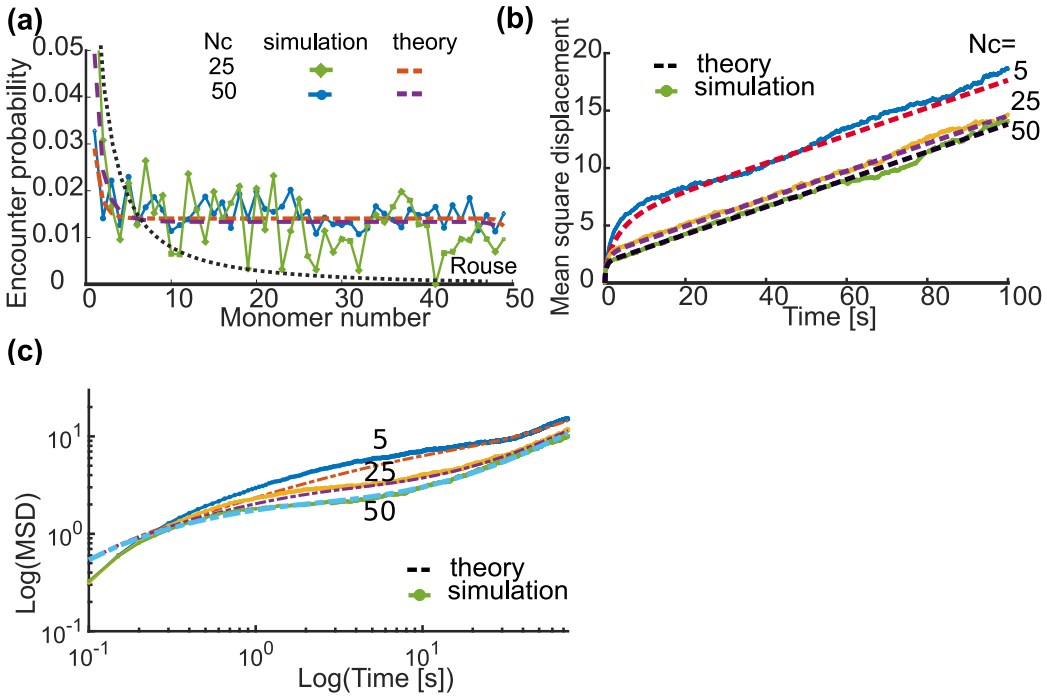


Figure 5.2 Properties of the RCL polymer **(a)** Encounter probability between monomers 1 and 2-50, simulated from (Eq. 5.3). The statistics of the simulations is recorded after the slowest relaxation time (Eq. 5.22). Parameters are $N = 50$ monomers, $N_c(\xi) = 25$ (green diamonds) and 50 (blue circles) connectors. We average over 500 realizations (changing each time the ensemble $\mathcal{G}$) and compare with the analytical formula Eq. 5.19 (dashed curves) with $D = 1, d = 3, b = \sqrt{d}, \epsilon = b/10, \Delta t = 0.01s$. The encounter probability of the Rouse polymer where $N_c(\xi) = 0$ (dotted black), which cannot account for long-range connectivity. **(b)** Mean square displacement (MSD) simulations for a RCL polymer, where $N = 50$ monomers and $N_c(\xi) = 5$ (blue), 25 (yellow), and 50 (green) added connectors. The formulas (Eq. 5.35) is shown in dashed. **(c)** The MSD in panel (b) for short time-scales (0-1 s) in log-log form , shows slight deviation between simulations of the polymer as in (a) and formula Eq. 5.35, for $N_c = 5$ (blue), 25 (yellow), and 50 (cyan) random connectors, where the deviation is more prominent for low connectivity.

In the low connectivity case, $N_c(\xi) \ll \frac{N^2}{2}$, we use 5.29 in 5.32 and discarding terms of higher order in $O(N^{-1})$, we obtain the asymptotic expansion

$$\langle R_G^2(\xi) \rangle \approx \frac{3b^2}{4(1-\xi)\sqrt{N\xi}}. \tag{5.33}$$

In Fig. 5.1d, we compare formula 5.32 with $\langle R_G^2(\xi) \rangle$ computed from Brownian simulations for $N = 20, 50$, and 100 monomers and $N_c(\xi) \in [5, 50]$ added random connectors and both agree.

5.2.5 Mean Square Displacement (MSD) of a single monomer of the RCL polymer.

Using the normal coordinate system 5.9 in dimension d, the MSD of monomers in the RCL polymer is given by

$$\langle r_m^2(t)\rangle = \langle \left(\sum_{p=0}^{N-1} \alpha_p^m u_p(t)\right)^2 \rangle = \frac{2dDt}{N} + \sum_{p=1}^{N-1} (\alpha_p^m)^2 \langle u_p^2\rangle = 2dD_{cm}t + \frac{2db^2}{N}\sum_{p=1}^{N-1} \frac{\cos^2\left(\frac{p\pi(m-\frac{1}{2})}{N}\right)\left(1-\exp(-\frac{2D\chi_p(\xi)t}{b^2})\right)}{\chi_p(\xi)}, \tag{5.34}$$

where we used $\langle u_p, u_q\rangle = 0$, $\forall p \neq q$ and $D_{cm} = \frac{D}{N}$. Averaging over all monomers and approximating the sum in 5.34 by an integral (Eurler Mac-Lauren formula) for $N \gg 1$, we obtain

$$\begin{aligned}
\langle\langle r_m^2(t)\rangle\rangle &= 2dD_{cm}t + \frac{2db^2}{N^2}\sum_{p=1}^{N-1} \frac{\left(1-\exp(-\frac{2dD\chi_p(\xi)t}{b^2})\right)}{\chi_p(\xi)} \sum_{m=1}^{N} \cos^2\left(\frac{p\pi(m-1/2)}{N}\right) \\
&= 2dD_{cm}t + \frac{db^2}{\pi}\int_0^\pi \frac{1-e^{\left(-\frac{2dD\chi_x(\xi)t}{b^2}\right)}}{\chi_x(\xi)}\mathrm{d}x \\
&= 2dD_{cm}t + \frac{db^2}{\sqrt{\pi N\xi(1-\xi)}}\int_0^{\sqrt{2dDN\xi t/b^2}} \exp(-g^2)\mathrm{d}g \\
&= 2dD_{cm}t + \frac{db^2 Erf[\sqrt{2DN\xi t/b^2}]}{2\sqrt{N\xi(1-\xi)}},
\end{aligned} \tag{5.35}$$

where $Erf[t]$ is the error function. Equation 5.35 characterizes the MSD for intermediate time scale $\tau_{N-1}(\xi) \ll t \ll \tau_1(\xi)$. For short time scale $t \ll \tau_{N-1}(\xi)$, the MSD is approximated by

$$\begin{aligned}
\langle\langle r_m^2(t)\rangle\rangle &= \frac{b^2\int_0^{\sqrt{2dDN\xi t/b^2}} \exp(-g^2)\mathrm{d}g}{\sqrt{\pi N\xi(1-\xi)}} \\
&\approx \frac{b\sqrt{2dDt}}{\sqrt{\pi(1-\xi)}}\left(1-\frac{\exp(-2dDN\xi t/b^2)}{2}\right).
\end{aligned} \tag{5.36}$$

Thus, for $N\xi \gg 1$, the MSD behaves like

$$\langle\langle r_m^2(t)\rangle\rangle \propto \frac{db\sqrt{2dDt}}{\sqrt{\pi(1-\xi)}}. \tag{5.37}$$

We conclude that the homogeneous behavior of MSD for the RCL polymer model gives an anomalous exponent $\alpha = 0.5$, similar to the Rouse model due to the mean-field approximation. This result is in contrast with the spectrum of anomalous exponents obtained for each configuration [94]. Finally, for long time scales ($t \gg$

$\tau_1(\xi)$), (slow diffusion of the polymer's center of mass), the error function in 5.35 is almost constant and, therefore,

$$\langle\langle r_m(t)^2\rangle\rangle \approx 2dD_{cm}t + \frac{db^2}{2\sqrt{N\xi(1-\xi)}}. \tag{5.38}$$

5.2.6 Mean First Encounter Time (MFET) $\langle\tau^\epsilon(\xi)\rangle$ between monomers of the RCL polymer.

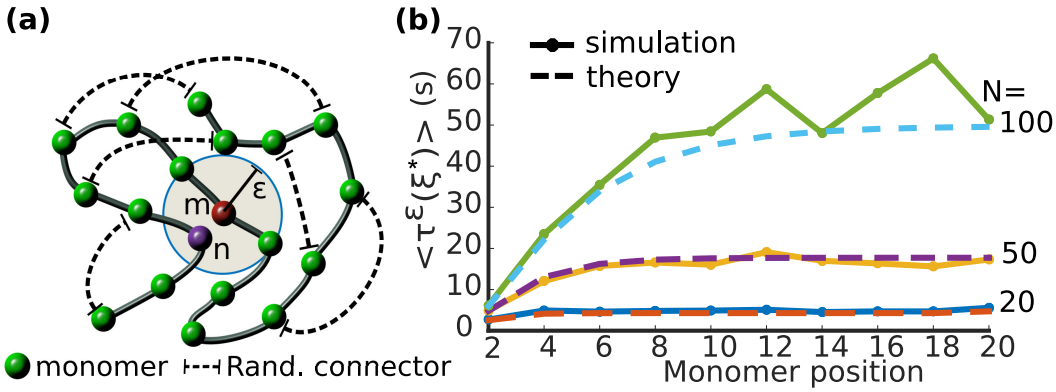


Figure 5.3 Transient RCL polymer properties: **(a)** Two monomers m (red) and n (purple) meet when they enter a ball of radius ϵ. Random connectors (dashed arrows) are added to a linear Rouse chain. **(b)** Stochastic simulations (dots) of the MFET between monomer 1 and monomers 2-20 of RCL polymers with $N = 20$ (blue), 50 (yellow) and 100 (green) monomers, with $N_c(\xi) = 25$ random connectors, agree with the formula in Eq. 5.43 (dashed). Parameters: $\epsilon = b/10, D = 1, b = \sqrt{3}, \Delta t = 0.01s$, the RCL system is 5.3 (we used Eq. 5.43 with ξ^*, Eq. 5.18).

We compute here the mean time for two monomers of the RCL polymer to enter for the first time in a ball of radius $\epsilon > 0$, at which they can possibly interact (Fig. 5.3a). The MFET for both the Rouse and β-polymer [5, 79] were computed [8] using the regular expansion with respect to $\epsilon > 0$ of the first eigenvalue λ_0^ϵ of the Fokker-Planck operator associated to the stochastic equation 5.13, so that

$$\langle\tau^\epsilon(\xi)\rangle \approx \frac{1}{D\lambda_0^\epsilon(\xi)}. \tag{5.39}$$

The first order approximation in ϵ is given by [8]

$$\lambda_0^\epsilon(\xi) = \frac{4\pi\epsilon\int_{C-P} e^{-\phi_G(U)}dU}{|\tilde{\Omega}(\xi)|} + O(\epsilon^2), \tag{5.40}$$

where $\phi_{\mathcal{G}}(U)$ is the diagonalized potential 5.17, $|\tilde{\Omega}(\xi)|$ is the integral over the entire RCL configuration space, computed using Gaussian integrals

$$\begin{aligned} |\tilde{\Omega}(\xi)| &= \int e^{-\phi_{\mathcal{G}}(U)} \mathrm{d}\boldsymbol{U} = \int \prod_{p=1}^{N} e^{-\frac{\kappa}{2}\chi_p(\xi)u_p^2(\xi)} \mathrm{d}\boldsymbol{U} \\ &= \left(\frac{(2\pi)^{N-1}}{\prod_{p=1}^{N-1} \kappa\chi_p(\xi)} \right)^{\frac{d}{2}}. \end{aligned} \tag{5.41}$$

The integral over the space $C-P$ of closed RCL polymer ensemble with fixed connector between monomers m and n and additional $N_c(\xi)$ random connectors in relation 5.40, is computed directly and gives [9]

$$\begin{aligned} &\int_{C-P} e^{-\phi_{\mathcal{G}}(U)} \mathrm{d}\boldsymbol{U} = \int e^{-\phi_{\mathcal{G}}(U)} \delta \left(\sum_{p=1}^{N} (\alpha_p^m - \alpha_p^n) u_p^2 \right) \mathrm{d}\boldsymbol{U} \\ &= (2\pi)^{\frac{(N-2)d}{2}} \left(\frac{\kappa}{2} b^2 \prod_{p=1}^{N-1} e^{-(\kappa\chi_p(\xi))} \sigma_{m,n}^2(\xi) \right)^{\frac{d}{2}}, \end{aligned} \tag{5.42}$$

where δ is the delta function. Using relations 5.41 and 5.42 in 5.39, we obtain the MFET between any two monomers m and n of the RCL polymer for a given connectivity fraction, ξ, in dimension $d = 3$:

$$\langle \tau_{m,n}^{\epsilon}(\xi) \rangle = \frac{1}{4\pi D\epsilon} \left(\frac{2\pi\sigma_{m,n}^2(\xi)}{\kappa b^2} \right)^{\frac{3}{2}}, \tag{5.43}$$

Using 5.30 into 5.43, we obtain the new looping formula

$$\langle \tau_{m,n}^{\epsilon}(\xi) \rangle \approx \frac{b^2 \left(1 - \exp(-|m-n|\sqrt{N\xi})\right)^{d/2}}{4\sqrt{N\xi}\pi D\epsilon(\kappa b^2)^{d/2}} + \mathcal{O}(N\xi),$$

where $|m-n| \ll N$, and $\xi \ll 1$. The analytical formula 5.43 agrees with Brownian simulations of the MFET for the RCL polymer (Eq. 5.3) with $N = 20, 50,$ and 100 monomers, and $N_c(\xi) = 25$ added random connectors (Fig. 5.3b).

5.2.7 Applications of the RCL polymer model to chromatin reconstruction.

We derived here several analytical formulas for the variance, encounter probability, radius of gyration, mean-square displacement and the mean first encounter time of RCL polymer models. These formulas can be used to extract parameters from CC experiments [29, 76]. Formula 5.19 can be used to fit the empirical encounter probability to extract the connectivity fraction ξ. This parameter has a direct interpretation and represents the mean number of cross-links, that can be mediated by CTCF molecules present in a genomic region. The parameter ξ depends on the

coarse-grained scale [94] and is used directly to estimate the radius of gyration (Eq. 5.32) of any region of interest. This radius characterizes the size of the folded genomic region relative to other genomic segments. It also provides insight into the relative compaction of TADs and local organization of the chromatin in the cell nucleus.

To demonstrate the applicability of the present method, we coarse-grained the 5C data reported in [76] of male neuronal progenitors NPC-E14 cells, replicate 1. Coarse-graining was performed at a scale of 3kbp according to the method presented in [39]. The assumptions of the RCL model, require that monomers share similar average connectivity, and thus we took only a subset of the 5C data containing TAD H. The TAD did not contain any long-range persistent loops (peaks) and we decided to test the present model. The length of the genomic section in TAD H is 679 kbp, which after coarse-graining resulted in a polymer of $N = 226$ monomers. We fit the EP (Eq. 5.19) of each of the 226 monomers as explained in [94] (Materials and Methods) and obtained an average connectivity of $\xi = 0.0022$, corresponding $N_c(\xi) = 56$ added connectors, that could be interpreted as the number of binding molecules. Fitting the EP of TAD H with a power law $a|m-n|^{-\beta}$ led to $\beta = 0.77$, showing that the Rouse (β=1.5) model is inadequate to represent the empirical EP. With persistence length of $b = 0.05\ \mu m$ and $N = 226, \xi = 0.0022$ in Eq. 5.32, we predicted that the radius of gyration is 43 nm for TAD H. Thus, the 679 kbp TAD H is compacted into a ball of volume $3.4 \times 10^5\ nm^3$ (2 bp per nm^3).

Finally, a possible test for the robustness of the RCL to coarse-graining at any scale, is that the value of the MSRG should persist. By coarse graining, we change the number N of monomers and the variance b, which should be known experimentally for each scale. In the absence of such knowledge, from Eq. 5.30 or 5.33, we see that to keep the MSRG constant for all scales, b^2 needs to be proportional to $\sqrt{N\xi}$, the square root of the mean number of connectors. Here the coarse-graining is imposed by the 5C protocol at resolution 3kb. We change the coarse-graining from 3kb to 10kb resolution of TAD H and we find that the number of connectors decreases from 56 at 3kb to 7 at 10 kb and, thus, b^2 should increases from 0.025 to 0.075 μm at 10 kb resolution.

Another application of the present analysis is the fitting of the MSD (Eq. 5.35) to single particle trajectories data: by fitting the experimental MSD curves using Eqs. 5.35-5.38, we obtain the degree of connectivity ξ. We can then interpret the mean deviation of loci dynamic from pure diffusion as the confinement due to cross-linked genomic environment [94, 109, 108].

To conclude, the main goal of this paper was to derive asymptotic formulas to extract the connectivity ξ or the mean number of connectors of a polymer model to account for the block matrices (Topological Associated Domains) present in CC data (3C,5C and Hi-C). The procedure consists in fitting the EP of the RCL model (Eq. 5.19) to CC data to extract the value of connectivity, that can later be used in

formula 5.43 to compute the mean first encounter time between any two monomers and, thus, for any two genes of interest. Encounter times are key for understanding processes, such as mammalian X chromosome inactivation [76], or non-homologous-end joining after DNA double-strand break [6, 8].

6 Chromatin reorganization during cell differentiation captured by randomly cross-linked polymer models of multiple topologically associating domains

Abstract

The chromatin in mammalian nucleus folds into discrete, contact enriched regions called Topologically Associating domains (TADs). The folding hierarchy of the TADs and their internal organization is highly dynamic throughout cellular differentiation, where structural changes within and between TADs are correlated with gene activation and silencing. To elucidate the relationship between chromatin conformation and gene regulation, polymer models are used. Here we introduce a heterogeneous randomly cross-linked (RCL) polymer model, accounting for multiple interacting TADs, to study the affect of connectivity within and between TADs on chromatin organization and dynamics. We derive analytical formula for the steady-state encounter probability within and between TADs and show the non-negligible affect of inter-TAD connectivity on the statistical and dynamical properties of the chromatin. We further show that the RCL model can capture high-order TAD organization resulting from inter-TAD connectivity. We demonstrate the applicability of the heterogeneous RCL model to the study the dynamic reorganization of three TADs of the mammalian X chromosome during three successive stages of differentiation from chromosomal capture data.

6.1 Introduction

The mammalian chromosome folds into discrete Mega-bp (Mbp) contact enriched regions termed Topologically associating domain (TADs). Our understanding of role of TADs remain incomplete and, to date, TADs have been found to participate in gene regulation [76, 95] and as replication timing units [84]. The regulation of genes within TADs is affected by the dynamics of genomic loop formation at sub Mbp scale [25, 29, 76], where DNA loops within TADs are actively formed by

regulatory factors [101]. However, sparse connectors between TADs (at scale >Mbp) can significantly affect the chromatin dynamics within TADs and act as a mechanism for TAD cross-regulation [94, 39]. The mutual affect of TAD on the dynamics of the chromatin remains largely unexplored.

The genome organization is now routinely probed by chromosome conformation capture (CC) methods [25, 65, 95], which simultaneously record genomic contacts (loops) at scales of kilo bp (kbp) to Mbp. TADs appear by averaging encounters over an ensemble of millions of samples [65, 76] and represents steady-state looping frequencies, but does not contain neither directly information about the size of the folded genomic section, nor any transient genomic encounter times. The CC methods were used to elucidate the dynamic organization of the chromatin throughout cell differentiation stages [76, 36], where the boundaries of TADs remain stable, but their internal looping pattern is highly variable. Moreover, the TADs have been found to form hierarchy into meta TADs, formed by inter TAD connectivity, which was found to be correlated with transcription state of the chromatin [36].

To study the steady-state and transient properties of the chromatin at a given scale, polymer models are used as a coarse-grained representation. Starting with the Rouse polymer [30], composed of a of N monomers connected sequentially by harmonic springs. The linear connectivity in the Rouse model neglects complexity in molecular representation and therefore cannot account for contact enriched TAD regions. Other polymer models include short and long-range looping [16, 97, 17, 45, 5, 61], self avoiding interactions and random loops [39, 16, 53, 21], and epigenomic state [55] to demonstrated the formation of TADs.

In this work we study the mutual affect of multiple TAD on the dynamics of the chromatin. We generalize the construction of a randomly cross-linked (RCL) polymer model (Chapter 5), to account for multiple connected TADs of variable size and internal connectivity. The random cross-links in the RCL model can be due to binding molecules such as CTCF [76] or by loop extrusion mechanism [37], although the exact mechanism by which they form is not the focus of the present work. The random position of the cross-links serve to capture the heterogeneity in chromatin structure in an ensemble of cells.

We derive novel analytical formulas for the encounter probability (EP), variance, and the radius of gyration of the RCL polymer, that we use to study the polymer dynamics. We show that TADs and higher order genome organization are affected by connectivity between TADs, which also affect both steady-state and time-dependent statistical properties of monomers within each TADs. We further demonstrate the applicability of the RCL model to study the reorganization of three neighboring TADs on the mammalian X chromosome throughout successive stages of cellular differentiation. We derive the average number of connectors within and between TADs by fitting the novel expression we obtained for the EP, to the empirical 5C data [76] and find a correlated compaction and decompaction of TADs, which is linked to

gene silencing and activation, respectively. We provide here an analytical framework, which enables to study a near full representation of the 5C data, derive statistical properties of the chromatin, and account for the influence of multiple connected TADs on each other.

6.2 Results

6.2.1 The RCL polymer for multiple associating domains (AD)

In Chapter 5 we have constructed a RCL polymer model for one TAD-like region, where all monomers share similar average level of connectivity. We now construct a heterogeneous RCL polymer model, comprised of N_T ADs of variable sizes and connected randomly within and between ADs. We construct the polymer by stitching N_T successive RCL chains of $N = [N_1, N_2, .., N_{N_T}]$ monomers, respectively (Fig. 6.1B).

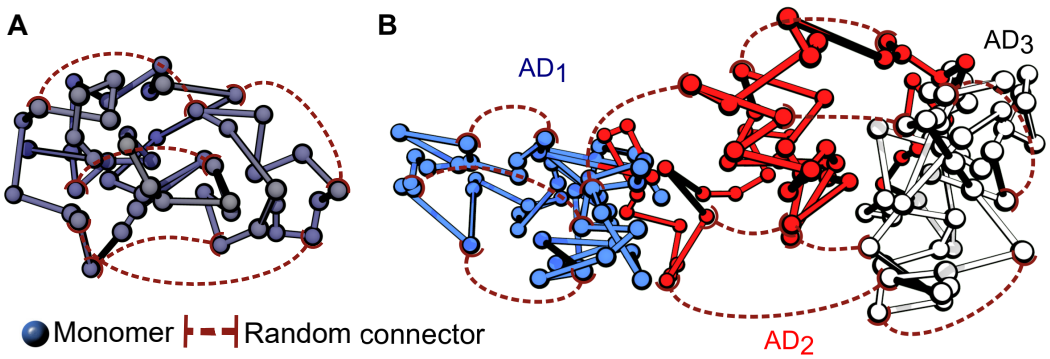


Figure 6.1 **The RCL polymer for single and multiple associating domains. A.** A schematic representation of the randomly cross-linked (RCL polymer) for a single Associating domain (AD). Monomers (circles) are connected linearly by harmonic springs to form the polymer's backbone (gray). Spring connectors (dashed red) are then added between randomly chosen non nearest-neighbor monomers. The choice of monomer pairs to connect is randomized in each realization of the polymer. **B.** A schematic representation of the RCL polymer model for three connected ADs (blue, red, white), where monomers (circles) are randomly connected (dashed red) within and between ADs by harmonic springs.

The position vector $\boldsymbol{R}$ of monomers is defined by $N_i \times d$ position vectors given in block matrix form

$$\boldsymbol{R} = \left[[R^{(1)}], [R^{(2)}], .., [R^{(N_T)}]\right]^T . \tag{6.1}$$

where the superscript in brackets indicates membership to AD $1,..N_T$, and the square brackets indicate a block matrix. The linear polymer backbone is composed of N_T Rouse matrices, defined in a matrix block form by

$$\boldsymbol{M} = diag([M_1],[M_2],..,[M_{N_T}]) = \begin{bmatrix} [M_1] & 0 & 0 & ... & 0 \\ 0 & [M_2] & 0 & ... & 0 \\ 0 & 0 & [M_3] & ... & 0 \\ . & . & 0 & ... & . \\ . & . & & ... & . \\ 0 & . & . & ... & [M_{N_T}] \end{bmatrix}, \tag{6.2}$$

where each $[M_j]$ is a Rouse matrix of the form 5.4 with N_j monomers. The total number of monomers in the heterogeneous RCL polymer model is thus $\sum_{j=1}^{N_T} N_j$. We further define $\Xi = \{\xi_{ij}\}$, $1 \leq i,j \leq N_T$ to be the square symmetric connectivity fraction matrix within and between ADs. For a given ξ, the random connectivity matrix $B(\xi)$ is defined by

$$B_{mn}(\Xi) = \begin{cases} -1, & |m-n| > 1, \text{and connected;} \\ -\sum_{i \neq j}^{N} B_{mj}(\Xi), & m = n; \\ 0, & \text{otherwise,} \end{cases} \tag{6.3}$$

To construct the average, random connectivity matrix, $\langle B(\Xi) \rangle$, we first define the indices b_i, to indicate the boundary between successive blocks i and $i+1$, $1 \leq i \leq N_T$ (Fig. 6.2A).

$$b_i = \sum_{m=1}^{i} N_m, \tag{6.4}$$

with $b_0 = 0$, which define the indices of a grid $T_{ij} = (b_i, b_j)$, which represent the bottom right index in each block of $\langle B(\Xi) \rangle$ (Fig. 6.2A). The block i,j stretches between $T_{i-1,j-1} + (1,1)$ to T_{ij}, and we denote it with square brackets as $[T_{ij}]$. The number of added connectors in block $[T_{ij}]$ is

$$N_c(\xi_{ij}) = \lfloor \xi_{ij} N_L \rfloor, \tag{6.5}$$

where N_L is the maximal possible number of connected pairs in block $[T_{ij}]$, given by

$$N_L = \begin{cases} (N_i - 1)(N_i - 2)/2, & |i-j| = 0; \\ N_i N_j - 1, & |i-j| = 1; \\ N_i N_j, & \text{otherwise.} \end{cases} \tag{6.6}$$

We assume here, that the random choice of $N_c(\xi_{ij})$ connected pairs is independent between blocks $[T_{ij}], (i \neq j)$, and that all monomer pairs are equally likely to

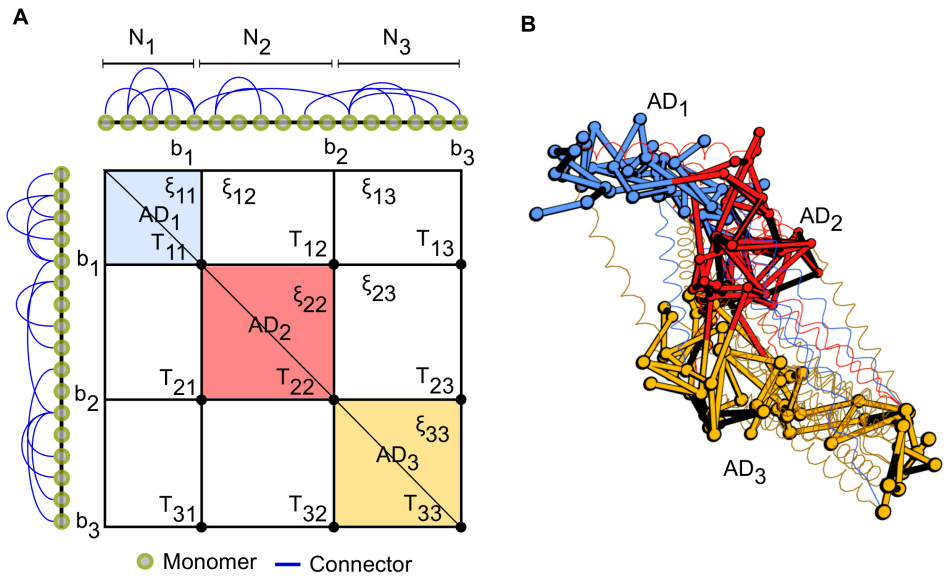


Figure 6.2 **Construction of the heterogeneous RCL polymer A.** Schematic description of 3 Associating domains (ADs), AD_1 (blue) AD_2 (red) and AD_3 (orange) composed of N_1, N_2 and N_3 monomers connected linearly, and additional random connectors within and between ADs, determined by ξ_{ij} the connectivity fraction (Eq. 6.5) . The indices b_i represent the bottom right end of each AD block and give rise to the grid points $T_{ij} = (b_i, b_j)$, which define blocks $[T_{ij}]$. **(B)** Sample configuration of the heterogeneous RCL polymer with 3 ADs corresponding to the construction in panel A. Monomers (spheres) are connected linearly by harmonic springs to form the backbone and additional connectors (springs) are added randomly between non-nearest neighboring monomers within and between ADs.

be chosen within each block. Due to symmetry, we shall only consider here choices of monomer pairs in the upper triangular part of $B(\Xi)$. The probability of monomer $b_i < m < b_{i+1}$ to have $k \geq 0$ connectors in $[T_{ii}]$ is computed by choosing k positions in row m and the remaining $N_L - k$ connectors in any row and column $n \neq m$, leading to the hyper-geometric probability

$$Pr_{ii}(k) = \begin{cases} \frac{C^k_{N_c(\xi_{ii})} C^{N_c(\xi_{ii})-k}_{N_L-(N_i-3)}}{C^{N_c(\xi_{ij})}_{N_L}}, & b_i < m < b_{i+1}; \\ \frac{C^{N_c(\xi_{ii})}_k C^{N_c(\xi_{ii})-k}_{N_L-(N_i-2)}}{C^{N_c(\xi_{ii})}_{N_L}}, & m = b_i, b_{i+1}, \end{cases} \tag{6.7}$$

and for $[T_{ij}]$, $j > i$

$$Pr_{ij}(k) = \frac{C^{N_j}_k C^{N_c(\xi_{ij})-k}_{N_L-N_j}}{C^{N_c(\xi_{ij})}_{N_L}}, \tag{6.8}$$

where $C^j_i = \frac{i!}{(i-j)!j!}$ is the binomial coefficient. Thus, the average number of connectors, $\bar{\beta}^{(i,j)}_m(\Xi)$, for monomer $b_i < m < b_{i+1}$ in each block $[T_{ij}]$ follows the average of the hyper-geometric distribution:

$$\bar{\beta}^{(i,j)}_m = \begin{cases} \xi_{ij} N_j, & |i-j| > 1; \\ \xi_{ii}(N_i - 2), & |i-j| = 0, m = b_i, m = b_{i+1}; \\ \xi_{ii}(N_i - 3), & |i-j| = 0, b_i < m < b_{i+1}. \end{cases} \tag{6.9}$$

We consider the average number of non nearest neighbor (NN) connectors $\beta_m(\Xi)$ for monomer $b_i \leq m \leq b_{i+1}$ to be the sum of average number of connectors in each block $[T_{ij}]$, $j = 1, .., N_T$

$$\beta_m(\Xi) = \sum_{k=1}^{N_T} \bar{\beta}^{(i,k)}_m = \langle B_{mm}(\Xi) \rangle = \begin{cases} -3\xi_{ii} + \sum_{j=1}^{N_T} N_j \xi_{ij}, & b_i \leq m \leq b_{i+1}; \\ -2\xi_{ii} + \sum_{j=1}^{N_T} N_j \xi_{ij}, & m = b_i, b_{i+1}. \end{cases} \tag{6.10}$$

Therefore, the average connectivity matrix, $\langle B(\Xi) \rangle$, is constructed as

$$\langle B_{mn}(\Xi) \rangle = \begin{cases} \beta_m(\Xi), & m = n; \\ 0, & |m-n| = 1; \\ -\xi_{ij}, & b_i \leq m \leq b_{i+1}, b_j \leq n \leq b_{j+1}. \end{cases} \tag{6.11}$$

The matrix $\langle B(\Xi) \rangle$ can be defined by blocks $[T_{ij}]$, such that

$$[T_{ij}] = \begin{cases} -\xi_{ii} \left(\mathbf{1}_{ii} + M_i \right) + I_i \sum_{k=1}^{N_T} N_k \xi_{ik}, & i = j; \\ -\xi_{ij} \mathbf{1}_{ij}, & i \neq j, \end{cases} \tag{6.12}$$

where $\mathbf{1}_{ij}$ is a $N_i \times N_j$ matrix of ones, M_i is a Rouse matrix of N_i monomers (Eq. 5.4), and I_i is an $N_i \times N_i$ identity matrix.

The mean-field stochastic system of equations describing the dynamics of monomers $\boldsymbol{R}$ is

$$\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(\boldsymbol{M} + \langle B(\Xi)\rangle\right)\boldsymbol{R} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t}, \tag{6.13}$$

where we replaced $B(\Xi)$ by its average $\langle B(\Xi)\rangle$ (average over choices of connected monomers [93, 17]), $\boldsymbol{\omega}$ are Brownian motions with mean 0 and standard-deviation 1.

To study the steady-state properties of system 6.13, we first decouple the system 6.13 into independent modes using the normal coordinate transform, defined by

$$\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R} = [u_0, u_1, ...], \tag{6.14}$$

where

$$\boldsymbol{V} = diag([V_1], [V_2], .., [V_{N_T}]), \tag{6.15}$$

is a diagonal block matrix, and each block $[V_i]$ is an $N_i \times N_i$ matrix of the form:

$$[V_i]_{m,n} = \begin{cases} \sqrt{\frac{1}{N_i}}, & m = 1; \\ \sqrt{\frac{2}{N_i}}\cos\left(\frac{m\pi}{N_i}(n-\frac{1}{2})\right), & 1 < m \leq N_i. \end{cases} \tag{6.16}$$

We multiply 6.13 from the left by 6.15 to obtain system of equations in the normal form

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(\Lambda + \boldsymbol{V}\langle B(\Xi)\rangle\boldsymbol{V}^T\right)\boldsymbol{U} + \sqrt{2D}\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}t}, \tag{6.17}$$

where $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\omega}$ are Brownian motion with mean 0 and standard-deviation 1, and

$$\Lambda = diag([\Lambda_1], [\Lambda_2], .., [\Lambda_{N_T}]), \tag{6.18}$$

is a $N_T \times N_T$ block diagonal matrix of the eigenvalues of Rouse chains [30] of $N_1, N_2, .., N_T$ monomers, defined by

$$[\Lambda_i] = diag(\lambda_0, \lambda_1, ..\lambda_{N_i-1}), \tag{6.19}$$

and

$$\lambda_p = 4\sin^2\left(\frac{p\pi}{2N_i}\right)), \quad p = 0, .., N_i - 1, \tag{6.20}$$

are the Rouse eigenvalues.

From 6.12, each $[T_{ii}]$ commutes with $[M_i]$ (Eq. 5.4) and therefore the multiplication $\boldsymbol{V}\langle B(\Xi)\rangle\boldsymbol{V}^T$ in the right hand side of 6.17, can be carried out for each block separately, and we obtained

$$[V_i][T_{ij}][V_j]^T = \begin{cases} -\xi_{ii}\left(N_i G_{ii} + [\Lambda_i]\right) + I_{N_i}\sum_{k=1}^{N_T} N_k\xi_{ik}, & i = j; \\ -\xi_{ij}\sqrt{N_i N_j}G_{ij}, & i \neq j, \end{cases} \tag{6.21}$$

where G_{ij} is a $N_i \times N_j$ matrix of zeros with 1 in the top left cell. The stochastic differential equation system that describes the dynamic of the normal coordinate $u_m^{(i)}$, $1 < m \leq N_i - 1$ in AD i is then

$$\frac{\mathrm{d}u_m^{(i)}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(\lambda_m^{(i)}(1-\xi_{ii}) + \sum_{k=1}^{N_T} N_k\xi_{ik}\right)u_m^{(i)} + \sqrt{2D}\frac{\mathrm{d}\eta_m^{(i)}}{\mathrm{d}t}, \tag{6.22}$$

and for the centers of masses $u_0^{(i)}$

$$\frac{\mathrm{d}u_0^{(i)}}{\mathrm{d}t} = -d\frac{D}{b^2}\left(\sum_{k=1}^{N_T} N_k\xi_{ik}u_0^{(i)} - \sum_{k=1}^{N_T}\xi_{ik}\sqrt{N_i N_k}u_0^{(k)}\right) + \sqrt{2D}\frac{\mathrm{d}\eta_0^{(i)}}{\mathrm{d}t}, \tag{6.23}$$

where $\eta_m^{(i)}$ are d-dimensional standard Brownian motion.

6.2.2 The MSRG for each AD

We now give an expression for the square radius of gyration in each AD for the case of dominant intra-AD connectivity, i.e., $N_i\xi_{ii} \gg N_j\xi_{ij}, j \neq i$. The distribution, $P(R_g^2)$, of the square radius of gyration is given by [34]

$$P(R_g^2) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(\mathtt{i_c}\beta R_g^2)Det(\mathbf{1}_{N_i-1} + \frac{\mathtt{i_c}d\beta b^2}{2N_i}[\Gamma_i]^{-1})^{-d/2}\mathrm{d}\beta, \tag{6.24}$$

where $[\Gamma_i]$ is the block diagonal matrix of eigenvalues for chain i except the first, Det is the determinant operator, and $\mathtt{i_c}$ is the complex unit. We approximate $[\Gamma_i]$ by $[V_i][T_{ii}][V_i]^T$ from 6.21 (removing the first vanishing row), which allows us to evaluate the determinant in the integral 6.24 by [34]

$$Det(\mathbf{1}_{N_i-1} + \frac{\mathtt{i_c}d\beta b^2}{2N_i}[\Gamma_i]^{-1})^{-d/2} = \exp\left(\frac{d}{2}\sum_{p=1}^{\infty}\left(-\frac{2\mathtt{i_c}\beta b^2}{dN_i}\right)^p\frac{Tr([\Gamma_i]^{-p})}{p}\right), \tag{6.25}$$

where Tr is the trace operator. We truncate the series in 6.25 at $p = 2$ and substitute the resulting expressing in 6.24 to obtain

$$\begin{aligned} P(R_g^2) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{b^4\beta^2}{dN_i^2}Tr([\Gamma]^{-2}) + \mathtt{i_c}\beta(R_g^2 - \frac{b^2}{N_i}Tr([\Gamma_i]^{-1}))\right)\mathrm{d}\beta, \\ &= \sqrt{\frac{dN_i^2}{4\pi b^4 Tr([\Gamma_i]^{-2})}}\exp\left(-\frac{\left(R_g^2 - \frac{b^2 Tr([\Gamma_i]^{-1})}{N_i}\right)^2}{4b^4Tr([\Gamma_i]^{-2})/dN_i^2}\right). \end{aligned} \tag{6.26}$$

The distribution of the square radius of gyration of each AD i is therefore approximately Normal, with mean and variance given, respectively, by

$$\langle R_g^2\rangle^{(i)} = \frac{b^2 Tr([\Gamma_i]^{-1})}{N_i}, \quad \sigma^2(R_g^2) = \frac{2b^4 Tr([\Gamma_i]^{-2})}{dN_i^2}. \tag{6.27}$$

For $N_i \gg 1$, we approximate the trace $Tr([\Gamma_i]^{-1})$ by an integral, where $x = p\pi/N_i$, and compute it along the unit contour in the complex plane, parametrized by $z = e^x$

$$\begin{aligned} Tr([\Gamma_i]^{-1}) &= \sum_{p=1}^{N_i-1}\frac{1}{\lambda_p(1-\xi_{ii}) + \sum_{k=1}^{N_T} N_k\xi_{ik}} \\ &\approx \frac{N_i}{2\pi(1-\xi_{ii})}\int_{-\pi}^{\pi}\frac{\mathrm{d}x}{y^{(i)}(N,\Xi) - \cos(x)} \\ &= -\frac{N_i}{2\pi\mathtt{i_c}(1-\xi_{ii})}\oint_{|z|=1}\frac{\mathrm{d}z}{(z-\zeta_0^{(i)}(\Xi))(z-\zeta_1^{(i)}(\Xi))} \\ &= \frac{N_i}{(1-\xi_{ii})(\zeta_0^{(i)}(\Xi) - \zeta_1^{(i)}(\Xi)}, \end{aligned} \tag{6.28}$$

where

$$y^{(i)}(N,\Xi) = 1 + \frac{\sum_{k=1}^{N_T}\xi_{ik}N_j}{2(1-\xi_{ii})}, \tag{6.29}$$

which couples the connectivities of all ADs connected to AD i, and

$$\begin{aligned} \zeta_0^{(i)}(N,\Xi) &= y^{(i)}(N,\Xi) + \sqrt{y^{(i)}(N,\Xi)^2 - 1}, \\ \zeta_1^{(i)}(N,\Xi) &= y^{(i)}(N,\Xi) - \sqrt{y^{(i)}(N,\Xi)^2 - 1}. \end{aligned} \tag{6.30}$$

Substituting 6.30 and 6.29 into 6.28 and then into 6.27, we obtain the expression for the MSRG of AD i

$$\langle R_g^2\rangle^{(i)} \approx \frac{b^2}{(1-\xi_{ii})(\zeta_0^{(i)}(\Xi) - \zeta_1^{(i)}(\Xi))} \tag{6.31}$$

6.2.3 Encounter probability of monomers of the heterogeneous RCL chain

To compute the encounter probability (EP) in the heterogeneous RCL polymer we distinguish between two cases: encounters of monomers of the same AD (intra AD), and monomers between ADs (inter-AD). We now derive formulas for the EP for those two cases under the assumption of non-vanishing connectivity, i.e., $\xi_{ij} > 0, \forall 1 \leq i, j \leq N_T$.

Intra-AD encounter probabilities

We first derive an expression for the variance between monomers of the same AD. For monomers $1 \leq m, n \leq N_i$ of AD i, we use the normal coordinates 6.14 to compute

$$\begin{aligned}\langle (r_m^{(i)} - r_n^{(i)})^2 \rangle &= \frac{2}{N_i} \langle \left(\sum_{p=1}^{N_i-1} \cos\left(\frac{p(m-\frac{1}{2})\pi}{N_i}\right) u_p - \cos\left(\frac{p(n-\frac{1}{2})\pi}{N_i}\right) u_p \right)^2 \rangle \\ &= \frac{2}{N_i} \sum_{p=1}^{N_i-1} \left(\cos\left(\frac{p(m-\frac{1}{2})\pi}{N_i}\right) - \cos\left(\frac{p(n-\frac{1}{2})\pi}{N_i}\right) \right)^2 \langle u_p^2(\Xi) \rangle. \end{aligned} \tag{6.32}$$

The time-dependent variance of the internal modes $p > 0$ of the Ornstein-Uhlenbeck system 6.22 is defined as

$$\langle u_p^2(\Xi) \rangle = \frac{b^2 \left(1 - \exp(-2\kappa(\lambda_p(1-\xi_{ii}) + \sum_{j=1}^{N_T} N_j \xi_{pj})t)\right)}{\lambda_p(1-\xi_{ii}) + \sum_{j=1}^{N_T} N_j \xi_{ij}}, \tag{6.33}$$

where the Rouse eigenvalues λ_p, $p = 1..N_i-1$, are defined in 6.19, and $\langle u_p(\Xi) u_q(\Xi) \rangle = 0, \quad \forall p \neq q$. Taking the limit in 6.33 as $t \to \infty$, we obtain the value of $\langle u_p^2(\Xi) \rangle$ at steady-state

$$\langle u_p^2(\Xi) \rangle = \frac{b^2}{\lambda_p(1-\xi_{ii}) + \sum_{j=1}^{N_T} N_j \xi_{ij}}. \tag{6.34}$$

Substituting 6.34 in 6.32, we obtain

$$\langle (r_m^{(i)} - r_n^{(i)})^2 \rangle = \frac{2b^2}{N_i} \sum_{p=1}^{N_i-1} \frac{\left(\cos\left(\frac{p(m-1/2)\pi}{N_i}\right) - \cos\left(\frac{p(n-1/2)\pi}{N_i}\right)\right)^2}{\lambda_p(1-\xi_{ii}) + \sum_{j=1}^{N_T} N_j \xi_{pj}}. \tag{6.35}$$

For $N_i \gg 1$ we approximate the sum 6.35 by an integral (Euler-Mclaurin formula), making a change of variable $x = p\pi/N_i$, $dx = \frac{N_i}{\pi}dp$ to obtain

$$\begin{aligned}\langle (r_m^{(i)} - r_n^{(i)})^2 \rangle &= \frac{2b^2}{N_i} \sum_{p=1}^{N_i-1} \frac{\left(\cos(\frac{p\pi(m-\frac{1}{2})}{N_i}) - \cos(\frac{p\pi(n-\frac{1}{2})}{N_i})\right)^2}{\sum_{j=1}^{N_T} \xi_{ij} N_j + 4(1-\xi_{ii})\sin^2(\frac{p\pi}{2N_i})} \\ &\approx \int_{-\pi}^{\pi} \frac{b^2 \left(\cos\left(x(m-\frac{1}{2})\right) - \cos\left(x(n-\frac{1}{2})\right)\right)^2}{\pi(1-\xi_{ii})(y^{(i)}(N,\Xi) - \cos(x))} \mathrm{d}x, \quad (6.36)\end{aligned}$$

with $y^{(i)}(N,\Xi)$ defined in 6.29. The integral in 6.36 can be computed in the complex plane along the contour of the unit disk [93], and we obtain

$$\sigma^2_{m,n}(\Xi) = \begin{cases} \langle R_g^2 \rangle^{(i)} \left(\frac{(\zeta_0^{(i)}(N,\Xi)^{m-n} - 1)^2 - 2\zeta_0^{(i)}(N,\Xi)^{m+n-1}}{\zeta_0^{(i)}(N,\Xi)^{2m-1}} + 2 \right), m \geq n; \\ \langle R_g^2 \rangle^{(i)} \left(\frac{(\zeta_0^{(i)}(N,\Xi)^{n-m} - 1)^2 - 2\zeta_0^{(i)}(N,\Xi)^{m+n-1}}{\zeta_0^{(i)}(N,\Xi)^{2n-1}} + 2 \right), m < n, \end{cases} \quad (6.37)$$

where $\langle R_g^2 \rangle^{(i)}$ is the MSRG defined in 6.31. The EP between monomer m and n of chain i is, thus

$$P^{(i)}(m,n) \propto \left(\frac{d}{2\pi\sigma^2_{m,n}(\Xi)} \right)^{d/2}, \quad (6.38)$$

Inter-AD encounter probabilities

We now compute the EP between monomers of AD i and j ($i \neq j$). We start by computing the variance of the vector between $r_m^{(i)}$ of AD i and $r_n^{(j)}$ of AD j, using the vectors $r_{cm}^{(i)}, r_{cm}^{(j)}$, the center of masses for AD i and j, respectively, as

$$\begin{aligned}\sigma^2_{m^{(i)}n^{(j)}}(\Xi) &= \langle (r_m^{(i)}) - r_n^{(j)})^2 \rangle = \langle (r_m^{(i)} - r_{cm}^{(i)} + r_{cm}^{(i)} - r_{cm}^{(j)} + r_{cm}^{(j)} - r_n^{(j)})^2 \rangle \\ &= \langle (r_m^{(i)} - r_{cm}^{(i)})^2 \rangle + \langle (r_n^{(j)} - r_{cm}^{(j)})^2 \rangle + \langle (r_{cm}^{(i)} - r_{cm}^{(j)})^2 \rangle. \quad (6.39)\end{aligned}$$

We compute the variance of the vector between a monomer m of AD i and its center of mass of using the normal coordinates (Eq. 6.14) at steady-state

$$\begin{aligned}\langle (r_m^{(i)} - r_{cm}^{(i)})^2 \rangle &= \frac{b^2}{2N_i(1-\xi_{ii})} \sum_{p=1}^{N_i-1} \frac{\cos^2\left(\frac{p\pi}{N_i}(m-\frac{1}{2})\right)}{y^{(i)}(\Xi) - \cos(\frac{p\pi}{N_i})} \\ &\approx \frac{-b^2}{2\pi(1-\xi_{ii})\mathtt{i_c}} \oint_{|z|=1} \frac{(z^{2m-1}+1)^2}{z^{2m-1}(z-\zeta_0^{(i)}(N,\Xi))(z-\zeta_1^{(i)}(N,\xi))} \mathrm{d}z \\ &= \frac{b^2(1+\zeta_0^{(i)}(N,\Xi)^{1-2m})}{(\zeta_0^{(i)}(N,\Xi) - \zeta_1^{(i)}(N,\Xi))(1-\xi_{ii})} = \langle R_g^2 \rangle^{(i)}(1+\zeta_0^{(i)}(N,\Xi)^{1-2m})), \quad (6.40)\end{aligned}$$

and similarly for AD j

$$\langle (r_n^{(j)} - r_{cm}^{(j)})^2 \rangle = \langle R_g^2 \rangle^{(j)} (1 + \zeta_0^{(j)}(N, \Xi)^{1-2n})). \tag{6.41}$$

We compute the variance of the vector between center of masses of ADs i and j by

$$\langle (r_{cm}^{(i)} - r_{cm}^{(j)})^2 \rangle = \frac{\langle (u_0^{(i)})^2 \rangle}{N_i} + \frac{\langle (u_0^{(j)})^2 \rangle}{N_j}. \tag{6.42}$$

Under the assumption $\sum_{k \neq i}^{N_T} N_k \xi_{ik} \gg \sum_{k=1}^{N_T} \xi_{ik} \sqrt{N_i N_k}$, we compute the variance $\langle (u_0^{(i)})^2 \rangle$ from the Ornstein-Uhlenbeck equation 6.23 at steady-state to be

$$\langle (u_0^{(i)})^2 \rangle = \frac{b^2}{\sum_{k \neq i}^{N_T} N_k \xi_{ik}}. \tag{6.43}$$

Substituting expression 6.43 for AD i and j into 6.42, we obtain

$$\langle (r_{cm}^{(i)} - r_{cm}^{(j)})^2 \rangle = b^2 \left(\frac{1}{N_i \sum_{k \neq i}^{N_T} N_k \xi_{ik}} + \frac{1}{N_j \sum_{k \neq j}^{N_T} N_k \xi_{jk}} \right). \tag{6.44}$$

Substituting expressions 6.40, 6.41 and 6.44 into 6.39, we obtain the final expression for the inter-AD variance

$$\begin{aligned}
\sigma^2_{m^{(i)} n^{(j)}}(\Xi) &= \frac{b^2 (1 + \zeta_0^{(i)}(N, \Xi)^{1-2m})}{(\zeta_0^{(i)}(N, \Xi) - \zeta_1^{(i)}(N, \Xi))(1 - \xi_{ii})} \\
&+ \frac{b^2 (1 + \zeta_0^{(j)}(N, \Xi)^{1-2n})}{(\zeta_0^{(j)}(N, \Xi) - \zeta_1^{(j)}(N, \Xi))(1 - \xi_{jj})} \\
&+ b^2 \left(\frac{1}{N_i \sum_{k \neq i}^{N_T} N_k \xi_{ik}} + \frac{1}{N_j \sum_{k \neq j}^{N_T} N_k \xi_{jk}} \right) \\
&= \langle R_g^2 \rangle^{(i)} (1 + \zeta_0^{(i)}(N, \Xi)^{1-2m}) + \langle R_g^2 \rangle^{(j)} (1 + \zeta_0^{(i)}(N, \Xi)^{1-2n}) \\
&+ b^2 \left(\frac{1}{N_i \sum_{k \neq i}^{N_T} N_k \xi_{ik}} + \frac{1}{N_j \sum_{k \neq j}^{N_T} N_k \xi_{jk}} \right).
\end{aligned} \tag{6.45}$$

The EP between monomer m of AD i and n of AD j is then given by

$$P_{m^{(i)}, n^{(j)}}(\Xi) \propto \left(\frac{d}{2\pi \sigma^2_{m^{(i)} n^{(j)}}(\Xi)} \right)^{d/2}. \tag{6.46}$$

Asymptotic approximation to the heterogeneous RCL variance

To obtain an approximated term for the variance Eq.6.45 we first approximate the terms

$$\zeta_0^{(i)}(N,\Xi)^m \approx \exp\left(m\sqrt{\sum_{k=1}^{N_T} N_k\xi_{ik}}\right),$$
$$\langle R_g^2\rangle^{(i)} \approx \frac{b^2}{2(1-\xi_{ii})\sqrt{\sum_{k=1}^{N_T} N_k\xi_{ik}}}, \tag{6.47}$$

under the assumption of $\boldsymbol{N}_i \gg 1$, and substituting in 6.45, to obtain the approximations to the inter-AD variance

$$\sigma^2_{m^{(i)}n^{(j)}}(\Xi) \approx \frac{b^2\left(1+\exp\left((1-2m)\sqrt{\sum_{k=1}^{N_T} N_k\xi_{ik}}\right)\right)}{2(1-\xi_{ii})\sqrt{\sum_{k=1}^{N_T} N_k\xi_{ik}}}$$
$$+\frac{b^2\left(1+\exp\left((1-2n)\sqrt{\sum_{k=1}^{N_T} N_k\xi_{jk}}\right)\right)}{2(1-\xi_{jj})\sqrt{\sum_{k=1}^{N_T} N_k\xi_{jk}}}$$
$$+b^2\left(\frac{1}{N_i\sum_{k\neq i}^{N_T} N_k\xi_{ik}}+\frac{1}{N_j\sum_{k\neq j}^{N_T} N_k\xi_{jk}}\right) \tag{6.48}$$

6.2.4 Mean-Square Displacement of monomers of the heterogeneous RCL polymer

Calculation of the MSD of monomers in AD i follows similar lines as in 5.2.5, where we replaced $N\xi$ for a single AD by $\sum_{k=1}^{N_T} N_k\xi_{ik}$ for multiple ADs. For intermediate times $\tau_{N-1}(\xi) \leq t \leq \tau_1(\xi)$ (Eq. 5.22)

$$\langle\langle r_m^{(i)}(t)^2\rangle\rangle \approx 2dD_{cm}t + \frac{db^2 Erf[\sqrt{2dDt\sum_{k=1}^{N_T} N_k\xi_{ik}/b^2}]}{2\sqrt{(1-\xi_{ii})\sum_{k=1}^{N_T} N_k\xi_{ik}}}, \tag{6.49}$$

where $D_{cm} = \frac{D}{\sum_{k=1}^{N_T} N_k}$. Expression 6.49 can be further approximated using the trapezoidal rule as

$$\langle\langle r_m^{(i)}(t)^2\rangle\rangle \approx 2dD_{cm}t + \frac{db\sqrt{dDt}\left(1+\exp(-\frac{2dDt}{b^2}\sum_{k=1}^{N_T} N_k\xi_{ik})\right)}{\sqrt{2\pi(1-\xi_{ii})}} \tag{6.50}$$

Expression 6.50 scales as $\sqrt{t}$ and highlights the dependence of the MSD curve on the connectivity between AD i and all other ADs.

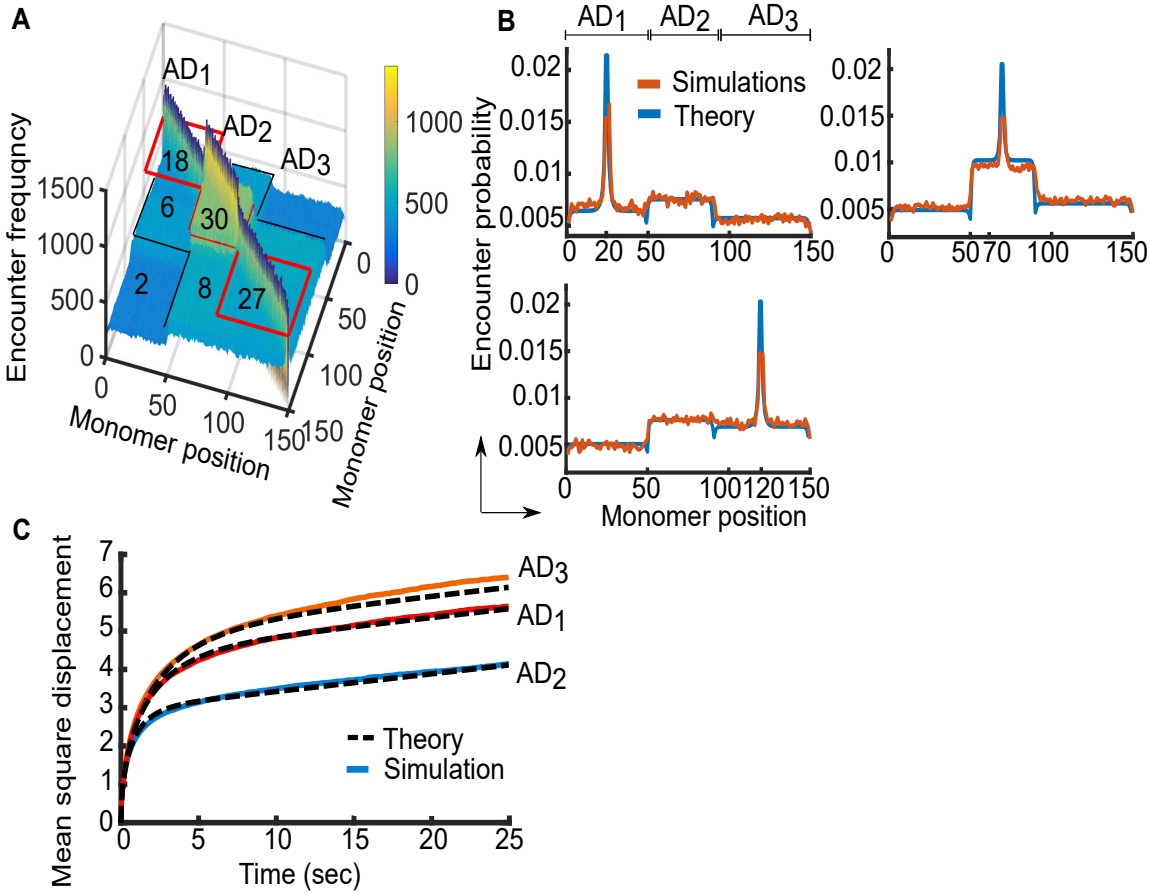


Figure 6.3 **Statistical properties of the heterogeneous RCL polymer.** **A.** Encounter frequency matrix of a polymer with 3 AD blocks (AD_1, AD_2, AD_3) of $N_1 = 50, N_2 = 40, N_3 = 60$ monomers each, result of 5000 simulations of the system 6.17 with $\Delta t = 0.1s, D = 1, d = 3, b^2 = \sqrt{d}$. The number of added connectors appears in each block. Three distinct diagonal ADs are visible (red boxes), where higher order structures appear (black lines) due to weak inter-AD connectivity. **B.** Encounter probability (EP) of the heterogeneous RCL described in panel A, where the simulation EP (orange) is in agreement with theoretical EP (blue, Eqs. 6.38, 6.46), plotted for the middle monomer in each AD: monomer 20 (top left), monomer 70 (top right) and monomer 120 (bottom left). The decreased EP at boundaries of ADs is an artifact of the use of the center of masses in Eqs. 6.40 and 6.41. **C.** The mean square displacement of monomers in each AD of the heterogeneous RCL polymer, simulated as described in panel A, simulation (full) versus theory (dashed Eq. 6.49) are in good agreement for time up to 25 s. The MSD for AD 1 overshots at $t > 15$ because the center of masses of the three AD cannot be fully decoupled (see Eq. 6.23).

6.2.5 Validation of the analytical expression of the heterogeneous RCL model

We validate the asymptotic expressions for the steady-state EP within (Eq. 6.38) and between ADs (Eq. 6.46) by Brownian simulations. We constructed an RCL polymer with 3 ADs of $N_1 = 50, N_2 = 40, N_3 = 60$ monomers, and set a dominant number of connectors in each diagonal AD block, such that the diagonal ADs contain 200% more connectors than between ADs (Fig. 6.3A). We ran 5000 simulations of system 6.13, with $d = 3, b = \sqrt{d}, D = 1, \Delta t = 0.1s$ until relaxation. The relaxation time $\tau(\Xi)$ was determined as the maximal relaxation time of the first mode over all ADs, defined as

$$\tau(\Xi) = \max_{1\leq i\leq N_T} \left(\frac{b^2}{2dD \sum_{k=1}^{N_T} N_k \xi_{ik} + 4(1-\xi_{ii}) \sin^2(\frac{\pi}{2N_i})} \right),$$

which resulted in number of steps of the order of tens of thousands. After time $\tau(\Xi)$ we constructed the encounter frequency matrix for 5000 realizations, where we set the encounter distance to $\epsilon = b/10$. The encounter frequency matrix showed three distinct diagonal blocks resulting from high internal AD connectivity but also higher order organization (meta ADs), which resulted from the inter-AD connectivity (Fig. 6.3A). To validate the expression for the theoretical steady-state EP (Eqs. 6.38 and 6.46), we computed simulation EP by dividing each row of the encounter frequency matrix by its sum and plotted it against the theoretical EP. In Fig. 6.3 B we show three sample fitted curves of the theoretical EP versus simulation EP for monomer 20 (upper left), 70 (upper right) and 120 (lower), which are in good agreement. We further computed the $\langle R_g^2 \rangle^{(i)}$ the MSRG of AD $i = 1, 2, 3$ from simulation and compared it to expression Eq. 6.31. We find the simulation MSRG is $\langle R_g^2 \rangle^{(i)} = 2.42, 1.5, 2.15$ and theoretical $\langle R_g^2 \rangle^{(i)} = 2.38, 1.31, 2.09$ for $i = 1, 2, 3$, respectively, which are in good agreement.

To validate the expression for the theoretical MSD (Eq. 6.49), we simulated system 6.13 for 500 step past relaxation time $\tau(\Xi)$ (Eq. 6.51 In Chapter 5) and computed the MSD for each monomers in each AD. In Fig. 6.3C we plotted the mean MSD in each AD against the theoretical expression Eq. 6.49 (dashed), which are in good agreement. The overshoot of the MSD of AD 1, is the result of the fact that center of masses of ADs cannot be fully decoupled (see Eq. 6.23). The height of the MSD curve is inversely proportional to the total connectivity of each AD, with AD 1 (Fig. 6.3C, orange, 26 connectors), AD 2 (blue, 46 connectors) and AD 3 (red, 37 connectors).

6.2.6 Application of the heterogeneous RCL to the study of genome reorganization from the 5C data

We now demonstrate the use of the heterogeneous RCL model with 5C data and study X chromosome reorganization during three stages of differentiation in cell lineage commitment: undifferentiated mouse embryonic stem cells (MESC), Neuronal precursor cells (NPC) and mouse Embryonic fibroblasts (MEF). The differentiation from MESC through NPC to MEF took place over the course of 84 hours [76]. We used the average of two replica of a subset of the 5C data [76], harboring 3 topologically Associating domains (TADs): TAD D, E, and F, which span a genomic section of about 1.9 Mbp. We coarse-grained the 5C encounter frequency data at a scale of 6 kb, which is twice the median length of the restriction segment of the HindII enzyme used in producing the 5C data [39, 94]. At this scale we found empirically that long-range persistent peaks of the 5C encounter data are smoothed-out sufficiently to enable us to use expressions 6.38 and 6.46 for fitting the 5C EP. The resulting coarse-grained encounter frequency matrix includes pair-wise encounter data of 302 genomic segments of equal size. To determine the position of TAD boundaries, we mapped the TAD boundaries reported in bp ([76] Supplementary Information, 'Analysis of 5C data, section) to genomic segments after coarse-graining. Accordingly, we constructed an RCL polymer with $N_D = 62, N_E = 88, N_F = 152$ monomers for TAD D, E and F, respectively. We fitted the EP of each monomer in the coarse-grained empirical EP matrix using formulas 6.38 and 6.46. In Fig. 6.4A we present the fitted EP matrices for MESC (left) NPC (middle) and MEF (right).

We computed N_c, the average number of connectors, within and between each TAD by averaging the connectivity values obtained from fitting the EP over all monomers to obtain the connectivity matrix Ξ and used relations 6.5 and 6.6. The mean number of connectors in the differentiation from MESC to NPC showed an increase by 145-200% (Fig. 6.4B) within and between TADs. The number of connectors within TAD F increases by 145% from 22 for MESC to 32 in NPC, the inter-connectivity between TAD F and E increases by 150% from 9 at MESC to 14 for NPC and the connectivity between TAD F and D doubled from 6 connectors for MESC to 12 for NPC, whereas the number of connectors within TAD E remained constant of 14. In the differentiation stage between NPC to MEF, the number of connectors within TADs D, E, and F returned to values comparable to MESC, whereas the inter-TAD connectivity between TAD F and E decreased from 14 for NPC to 9 for MEF.

The mean square radius of gyration of the three TADs, throughout differentiation stages, correlated with the acquisition and lose of connectors in all TADs. We Computed the MSRG (Eq. 6.31) for each TAD using the calibrated connectivity

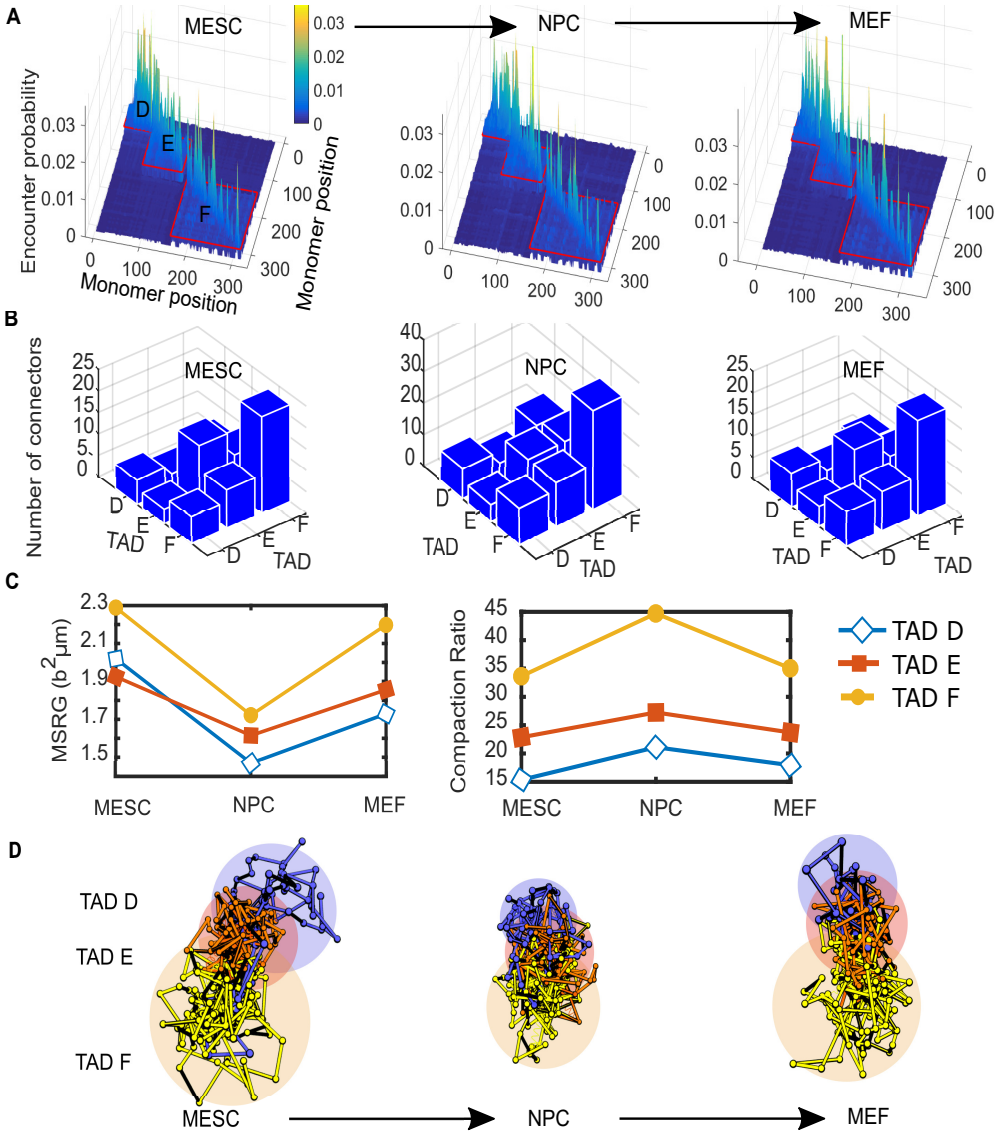


Figure 6.4 **Dynamic reorganization of the X Chromosome during differentiation. A.** Result of the fit of the expressions 6.38 and 6.46 to the empirical 5C encounter probability data at a scale of 6kb for 3 TADs (red rectangles) at 3 successive stages of cell differentiation: Embryonic stem cell (MESC, left), Neuronal precursor cells (NPC, middle) and Embryonic Fibroblasts (MEF, right). **B.** The mean number of connectors found by fitting of the EP as described in panel A, within and between TAD D, E and F for MESC (left), NPC (middle) and MEF (right). The number of connectors within TAD F grows from 22 for MSEC to 32 for NPC and drops back to 22 for MEF cells. The inter-TAD F and D connectivity increases from 9 for MSEC to 14 for NPC and decreases to 9 for MEF. The number of connectors in TAD E remains 14 throughout the 3 stages, whereas,the number of connectors for TAD D increases from 6 for MSEC to 9 for NPC and decreases to 7 for MEF. **C.** The mean square radius of gyration (left) at units of b^2 decreases for all TADs in NPC stage and increases back to the levels of MESC for MEF. The compaction ratio (right) computed as the reciprocal of the MSRG of each TAD to the MSRG of the Rouse chain ($Nb^2/6$) with the same number of monomers, shows that the compaction in TAD E is higher than in D at NPC stage despite having a higher MSRG (1.6 and 1.3 for TAD D and E, respectively). **D.** Three sample realization of the RCL polymer showing the compaction of TAD D (blue) , E (red) and F (orange) in the transition from MESC to NPC and decompaction in the transition to MEF. The shaded areas represents spheres of radius of gyration as in panel C for each TAD in each stage.

matrix Ξ, obtained from fitting the experimental 5C EP (Fig. 6.4C, left) in the units of b^2. The MSRG of all TADs decreased from average of 2 $b^2\mu m$ at MESC stage to 1.7 $b^2\mu m$ for NPC and increased back to 2 $b^2\mu m$ for MEF cells. In the transition from MESC to NPC, the MSRG of TAD D surpassed that of TAD E (Fig. 6.4C, red squares), due to the increased in inter-TAD connectivity between TAD E and F at NPC stage.

The value of the MSRG does not disclose the level of compaction in each TAD, which can indicate physical quantities, such as number of bp per nm^3. We, therefore, computed the compaction ratio (Fig. 6.4C, right), defined as the reciprocal of the ratio between the MSRG of each TAD (Eq. 6.31) and the MSRG of the linear Rouse chain ($Nb^2/6$, [30]) of the same size. We find that TAD F ($N = 154$ monomers) has the highest compaction between the TADs, throughout the 3 stages of differentiation (Fig. 6.4C, yellow circles), where we find it to be 33, 45, and 35 times more compact than the linear Rouse chain for MESC, NPC, and MEF stages, respectively. TAD E ($N = 88$ monomers) was found to be 22, 27, and 23 times more compact than the linear Rouse chain (Fig. 6.4C right, red squares), despite retaining similar levels of intra-TAD of 14 connectors throughout stages of differentiation. This effect is due to the increase in inter-TAD connectivity between TAD E and F at NPC stage. TAD D ($N = 62$ monomers) was found to be 15, 21, and 18 times more compact than the linear Rouse chain (blue squares) for MESC, NPC and MEF stages, respectively. To conclude, inter-TAD connectivity affects the compaction of TADs and therefore cannot be neglected in the study of genome reorganization from the 5C data.

6.3 Discussion

In this work we presented a generalization of the analytical RCL model [94, 93] from one AD region to multiple interacting regions of variable size, intra and inter-AD connectivities. We, thus, provided an analytical framework for studying a near full representation of conformation capture (5C and Hi-C) encounter frequency data, where we account for the affect of both intra and inter-TAD connectivity in the heterogeneous RCL model. The RCL polymer provides mean of deriving the average number of cross-links within and between each TAD, length scales, such as the mean square radius of gyration, which characterizes the size of the folded AD, and the mean square displacement in multiple ADs. These quantities cannot be derived from the empirical conformation capture data and requires complementary method, such as single particle tracking [38, 7].

We demonstrate the applicability of the generalized RCL model to the study of multiple TAD reorganization, throughout three successive stages of mammalian cell differentiation: mouse embryonic stem cell (ESC), neuronal precursor cells (NPC) and mouse embryonic fibroblasts (MEF). We fitted expressions 6.38 and 6.46 to the empirical 5C encounter matrices of three differentiation stages (Fig. 6.4A) and showed that the RCL model can capture contact enriched TADs and the variability in

monomer connectivity within each TAD (Fig. 6.4A, super and sub diagonals of EP matrices). We use the average connectivity we obtained from the fitting procedure to derive the average number of connectors between and within TADs (Fig. 6.4B). In the coarse graining scale of 6 kb, our results show that the X chromosome acquires connectors within and between TADs in the transition between MESC to NPC and later returns to values comparable to those of MESC at MEF stage. The increasing connectivity for NPC cells can be associated with the acquisition of LaminB1 ([76], Figure 3). The increase in the number of connectors in each TAD at NPC stage further correlates with TAD compaction. Indeed, the MSRG curves (Fig. 6.4C left) decrease at NPC stage, which indicate higher chromatin compaction for all TADs. The compaction ratio (in relation to linear Rouse polymer) showed high compaction at NPC stages (Fig. 6.4C, right) for all TADs, which can be associated with TADs being in heterochromatin state, suppressed gene expression and lamina associating domains [80]. The mutual affect of inter-TAD connectivity can clearly be seen, when the number of connectors for TAD E remain constant at 14, however, the MSRG and compaction ratio decreased at NPC stage. This result indicates that the a full description of the 5C data by polymer models must take into account inter-TAD connectivity. Overall, the RCL polymer captures well the correlated reorganization of TAD D, E, and F during differentiation. The correlation in TAD reorganization is in accordance with transcription co-regulation in these TADs [76].

Our construction of the RCL polymer accounts for multiple connected regions with weak inter-TAD connectivity, in line with experimental CC data [29, 25, 36, 76]. Despite such weak inter-AD connectivity the affect of multiple ADs on the dynamics of monomers cannot be fully neglected. The analytical expression we derived for the asymptotic steady-state encounter probability (Eqs. 6.38 and 6.46) are suited for modeling the 5C empirical encounter probability matrices. We validated the asymptotic expression for the EP by Brownian simulations, where we showed that we can capture both TADs and higher order structures (metaTADs [36]), resulting from inter-TAD connectivity (Fig. 6.3A and B). We further demonstrated how the dynamics of monomers is affected by the local connectivity between and within ADs, as can be seen in the MSD curves (Fig. 6.3C). This results provides an explains the variability in MSD behavior seen experimentally in live cell single particle tracking experiments [94, 38, 28] in terms of heterogeneous local chromatin organization in cell population.

To conclude, the analytical framework presented in this work provides means of studying chromatin structural reorganization on a genome-wide scale. The generalized RCL model can be used to interpolate between chromatin organization of cells at successive differentiation stages, directly from experimental chromatin capture data. The generalized RCL model we presented here can be used with any ligation proximity experimental data (Hi-C, 5C 4C).

7 Discussion and perspectives

Polymer models are instrumental in advancing our understanding of the relationships between chromatin conformation, dynamics, and function. The growing sophistication of polymer models over the past century went hand in hand with advancement in microscopy and sub-cellular measurement techniques. These techniques, in turn, exposed the complex relationship between chromatin folding, dynamics, and function, in all stages of the cell cycle. Polymer models, then, provided the framework in which the vast amount of experimental data could be reconciled, interpolated, and explained in a rational manner.

In this dissertation work, I developed and studied a randomly cross-linked (RCL) polymer models, to elucidate the relationship between chromatin organization and dynamics, at the sub-chromosome scale and across cell population. Experimental proximity ligation procedures (the conformation capture) provide rich grounds on which the prediction of steady-state statistical properties of the polymer could be validated. However, the missing dimension in these data is time, which hinders study of the chromatin dynamics directly from these data. Between the well defined scale of the DNA double helix to that of a full chromosome, the chromatin is constantly remodeled, and the principles governing its folding remain unclear. It is unclear whether the chromatin ever reaches a steady-state conformation, rather experimental results suggest otherwise. However, some genomic structures do remain invariant across populations, but with no unbiased technique to study chromosome-wide dynamics, the use of polymer models is, thus, necessary. It is at this stage the RCL polymer comes into play for bridging such gaps.

In Chapter 3 I presented a method to construct a polymer model directly from the 5C data, and use it to study the relationship between steady-state chromatin organization and transient encounter times. The method we've proposed has an advantage over existing ones, because we extract both random short-range and persistent long-range connectors directly from the data, which allows us to model any genomic section. Moreover, we extracted the average number of cross-links in a Topologically associating domain (TAD), which can be directly related to measurable biological quantities (binding molecules). The presence of long-range persistent connectors between TADs proved to be non-negligible on the encounter times of loci within TADs, and revealed the manner by which two neighboring TADs can cross-regulate. Because sparse long-range connectors strongly affect monomer dynamics and, in-turn, gene regulation, a natural continuation of this project would be to study the effect of loss and re-acquisition of these connectors on chromatin dynamics, following DNA damage by e.g. UV irradiation. Application of the presented

methodology to reconstruct a polymer from single cell Hi-C data is currently under development.

In Chapter 4 I presented a study into the dynamical reorganization of the chromatin, by computing transient first encounter and dissociation times of two tagged genomic loci. I show how to extend the possibilities of calibrating the number of random connectors of the RCL model, from static CC maps (Chapter 3) to dynamic SPT data. From simulations of the RCL model after DSB, I find a conservative loss of connectors (4%) around DSB sites. A continuation of this study would include modeling the dynamic reorganization of the chromatin following multiple DSBs. This reorganization involves many structural changes imposed by repair protein, which apply pushing forces on the chromatin to facilitate access to damaged sites. A model for re-acquisition of connectors and compaction of the genome following repair of damaged site is currently under development. Specifically, we would like to unveil the manner by which looping patterns are restored following large genomic reorganization in process such as damage repair and cell division (retention of epigenomic memory [24]).

In chapter 5 I take on an analytical approach for deriving expressions for steady-state and transient statistics of the RCL polymer, representing a single TAD-like region. I derived a new expression for the encounter probability (EP), which is particularly suited for fitting the EP of CC data and to extract the connectivity fraction. This provides a way to replace heavy numerical simulations by a simple curve fitting to the EP of the CC, with ξ, the connectivity fraction, as the only variable. However, this analytical approach cannot fully replace the one presented in Chapter 3, because persistent connectors are not accounted for in the analysis. Attempts to derive the statistics of the RCL with as little as 3 persistent connectors, led to prohibitive complex expressions. Nevertheless, all analytical expressions I derived can still be applied to TAD regions with little or no persistent connectors, or by coarse-graining the CC data at resolution of 10kb and above to smooth out the long-range peaks. The derivation of analytical expression for steady-state and transient statistics was made possible due to the mean-field approach I took. However, this approach came with the price of a mismatch between the anomalous exponent of 0.5, found analytically, and about 0.4-0.45, found by simulations in Chapter 3. Nevertheless, the analytical model remains insightful. The expression I derived for the mean first encounter time between any two monomers is a generalization of the mean first encounter times computed, by many authors, only between the two end monomers of a polymer chain. The first encounter times are directly related to the rate of gene expression, time of DNA repair. in the RCL model, the average added random connectivity matrix is uniform, where the genomic distance between monomers does not affect the choice of monomers to connect. It is, therefore, that the RCL model captures well the tails of the EP (long-range encounter enrichment in TADs) as apposed to the method in Chapter 3, from which the initial slope of the EP is well

captured. The analysis of RCL models, with non-uniform (in monomer distance) probability density functions, is currently under development. Preliminary study I conducted showed that the eigenvalues of a non-uniform random connectivity matrix $\langle B(\xi)\rangle$ can be computed, if one constructs it as a Toeplitz matrix embedded in a circular matrix, where each row is circular shift of the chosen pdf of monomer connectivity. Retrieving the eigenvalues of the connectivity matrix is the key for deriving steady-state and transient expressions in Chapter 5. Further development of the RCL analysis will include derivation of the simultaneous first encounters time of 3 monomers, which describes the encounter of multi-enhancer, as recently shown in [13].

In Chapter 6 I construct and analyze a generalized RCL polymer model, to account for several TAD-like regions, having variable intra and inter-TAD connectivity. As seen in Chapter 3, the organization and dynamics of TADs cannot be studied in isolation, and sparse connections between TADs must be included in the analysis. In Chapter 6 the goal was to provide steady-state expressions for the encounter probability, which enables to simultaneously fit the 5C encounter data in a genome-wide manner. I achieved this under the assumption of low inter-TAD connectivity, in which case the center of masses of TADs were regarded as independent. Low inter-TAD connectivity is in-line with experimental findings. The generalized RCL polymer model was then used to interpolate between snapshot 5C data of three successive stages of cell differentiation, and provided insight into the dynamic reorganization in the chromatin (compaction and de-compaction) throughout these stages. Limitation of time and insufficient data prevented me from continuing to explore whether the patterns of synchronous compaction and de-compaction of TADs occurs throughout other stages of differentiation, or whether there exists a recurring such pattern. These question are waiting to be answered in a genome-wide study I plan to undertake, retrospectively analyzing 5C and Hi-C data upon data availability. Finally, the multi-TAD model in this Chapter 6 provides a way to deal with the shortcoming of the models in Chapter 5 and 3, such that persistent, isolated, and strong interactions could be accounted for analytically in the model. By treating isolated peaks of the CC data as TADs of very small size and high internal connectivity, this problem can be partially resolved. I performed a preliminary exploration into this approach and results were encouraging.

To summarize, the analysis and methods presented in this dissertation can be used to study genome dynamics, where the complex organization of the chromatin is accounted for, thus, making a step towards resolving some of the difficulties presented in the Introduction. This dissertation was written at a time when new and existing sub-cellular measurement techniques have emerged and opened up new avenues for exploring chromatin dynamics and organization. Each such development enabled to elucidates an aspect of chromatin organization, but unveiled further unknowns. It was, therefore, the goal of this dissertation to provide computational and

analytical tools in polymer physics to illuminate aspects of chromatin organization and dynamics, and resolve such unknowns.

Bibliography

8

[1] Benjamin Albert, Julien Mathon, Ashutosh Shukla, et al. „Systematic characterization of the conformation and dynamics of budding yeast chromosome XII“. In: *The Journal of Cell Biology* 202.2 (2013), pp. 201–210.

[2] A. Amitai, A. Seeber, S. Gasser, and D Holcman. „Chromatin interaction removal and local reorganization during homologous search in the cell nucleus.“ In: *pre-print* (2016).

[3] A Amitai, Carlo Amoruso, Avi Ziskind, and D Holcman. „Encounter dynamics of a small target by a polymer diffusing in a confined domain“. In: *The Journal of Chemical Physics* 137.24 (2012), p. 244906.

[4] Assaf Amitai and David Holcman. „Diffusing polymers in confined microdomains and estimation of chromosomal territory sizes from chromosome capture data“. In: *Physical review letters* 110.24 (2013), p. 248105.

[5] Assaf Amitai and David Holcman. „Polymer model with long-range interactions: Analysis and applications to the chromatin structure“. In: *Physical Review E* 88.5 (2013), p. 052604.

[6] Assaf Amitai and David Holcman. „Polymer physics of nuclear organization and function“. In: *bioRxiv* (2016), p. 076661.

[7] Assaf Amitai, Mathias Toulouze, Karine Dubrana, and David Holcman. „Analysis of single locus trajectories for extracting in vivo chromatin tethering interactions“. In: *PLoS Comput Biol* 11.8 (2015), e1004433.

[8] Assaf Amitai, Ivan Kupka, and David Holcman. „Computation of the mean first-encounter time between the ends of a polymer chain“. In: *Physical Review Letters* 109.10 (2012), p. 108302.

[9] 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“. In: *Cell Reports* 18.5 (2017), pp. 1200–1214.

[10] A Annunziato. „DNA packaging: nucleosomes and chromatin“. In: *Nature Education* 1.1 (2008), p. 26.

[11] Mariano Barbieri, Mita Chotalia, James Fraser, et al. „Complexity of chromatin folding is captured by the strings and binders switch model“. In: *Proceedings of the National Academy of Sciences* 109.40 (2012), pp. 16173–16178.

[12] Adi Barzel and Martin Kupiec. „Finding a match: how do homologous sequences get together for recombination?“ In: *Nature Reviews Genetics* 9.1 (2008), pp. 27–37.

[13] Robert A Beagrie, Antonio Scialdone, Markus Schueler, et al. „Complex multi-enhancer contacts captured by genome architecture mapping“. In: *Nature* 543.7646 (2017), pp. 519–524.

[14] Irena Bronshtein Berger, Eldad Kepten, and Yuval Garini. „Single-particle tracking for studying the dynamic properties of genomic regions in live cells“. In: *Imaging Gene Expression: Methods and Protocols* (2013), pp. 139–151.

[15] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. „Julia: A fresh approach to numerical computing“. In: *SIAM Review* 59.1 (2017), pp. 65–98.

[16] Manfred Bohn and Dieter W Heermann. „Diffusion-driven looping provides a consistent framework for chromatin organization“. In: *PloS one* 5.8 (2010), e12218.

[17] Manfred Bohn, Dieter W Heermann, and Roel van Driel. „Random loop model for long polymers“. In: *Physical Review E* 76.5 (2007), p. 051805.

[18] Miguel R Branco and Ana Pombo. „Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations“. In: *PLoS Biol* 4.5 (2006), e138.

[19] I Bronshtein, E Kepten, I Kanter, et al. „Loss of lamin A function increases chromatin dynamics in the nuclear interior“. In: *Nature communications* 6 (2015).

[20] I Bronstein, Y Israel, E Kepten, et al. „Transient anomalous diffusion of telomeres in the nucleus of mammalian cells“. In: *Physical review letters* 103.1 (2009), p. 018102.

[21] JD Bryngelson and D Thirumalai. „Internal constraints induce localization in an isolated polymer molecule“. In: *Physical review letters* 76.3 (1996), p. 542.

[22] Jeff ZY Chen, Heng-Kwong Tsao, and Yu-Jane Sheng. „Diffusion-controlled first contact of the ends of a polymer: crossover between two scaling regimes“. In: *Physical Review E* 72.3 (2005), p. 031804.

[23] Ting Cui, Jiandong Ding, and Jeff ZY Chen. „Mean first-passage times of looping of polymers with intrachain reactive monomers: Lattice Monte Carlo simulations“. In: *Macromolecules* 39.16 (2006), pp. 5540–5545.

[24] Job Dekker. „Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture“. In: *Epigenetics & chromatin* 7.1 (2014), p. 1.

[25] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. „Capturing chromosome conformation“. In: *science* 295.5558 (2002), pp. 1306–1311.

[26] David Dickerson, Marek Gierliński, Vijender Singh, et al. „High resolution imaging reveals heterogeneity in chromatin states between cells that is not inherited through cell division“. In: *BMC Cell Biology* 17.1 (2016), p. 33.

[27] Vishnu Dileep, Ferhat Ay, Jiao Sima, et al. „Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program“. In: *Genome research* (2015).

[28] Vincent Dion and Susan M Gasser. „Chromatin movement in the maintenance of genome stability“. In: *Cell* 152.6 (2013), pp. 1355–1364.

[29] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, et al. „Topological domains in mammalian genomes identified by analysis of chromatin interactions“. In: *Nature* 485.7398 (2012), pp. 376–380.

[30] M Doi and SF Edwards. *The Theory of Polymer Dynamics Clarendon*. Oxford, 1986.

[31] Josée Dostie, Todd A Richmond, Ramy A Arnaout, et al. „Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements“. In: *Genome research* 16.10 (2006), pp. 1299–1309.

[32] Zhijun Duan, Mirela Andronescu, Kevin Schutz, et al. „A three-dimensional model of the yeast genome“. In: *Nature* 465.7296 (2010), pp. 363–367.

[33] Edgardo R Duering, Kurt Kremer, and Gary S Grest. „Relaxation of randomly cross-linked polymer melts“. In: *Physical review letters* 67.25 (1991), p. 3531.

[34] BE Eichinger. „Configuration statistics of Gaussian molecules“. In: *Macromolecules* 13.1 (1980), pp. 1–11.

[35] Bruce E Eichinger and JE Martin. „Distribution functions for Gaussian molecules. II. Reduction of the Kirchhoff matrix for large molecules“. In: *The Journal of Chemical Physics* 69.10 (1978), pp. 4595–4599.

[36] James Fraser, Carmelo Ferrai, Andrea M Chiariello, et al. „Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation“. In: *Molecular systems biology* 11.12 (2015), p. 852.

[37] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, et al. „Formation of chromosomal domains by loop extrusion“. In: *Cell reports* 15.9 (2016), pp. 2038–2049.

[38] Susan M Gasser. „Nuclear Architecture: Past and Future Tense“. In: *Trends in Cell Biology* (2016).

[39] Luca Giorgetti, Rafael Galupa, Elphège P Nora, et al. „Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription“. In: *Cell* 157.4 (2014), pp. 950–963.

[40] Andrey A Gurtovenko and Alexander Blumen. „Generalized Gaussian Structures: Models for Polymer Systems with ComplexTopologies“. In: *Polymer Analysis Polymer Theory*. Springer, 2005, pp. 171–282.

[41] H. Hajjoul, S. Kocanova, I. Lassadi, K. Bystricky, and A. Bancaud. „Lab-on-Chip for fast 3D particle tracking in living cells“. In: *Lab Chip.* 9 (2009), pp. 3054–3058.

[42] Houssam Hajjoul, Julien Mathon, Hubert Ranchon, et al. „High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome“. In: *Genome research* 23.11 (2013), pp. 1829–1838.

[43] Anders S Hansen, Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. „CTCF and cohesin regulate chromatin loop stability with distinct dynamics“. In: *Elife* 6 (2017).

[44] Michael H Hauer, Andrew Seeber, Vijender Singh, et al. „Histone degradation in response to DNA damage enhances chromatin dynamics and recombination rates“. In: *Nature Structural & Molecular Biology* (2017).

[45] Dieter W Heermann. „Physical nuclear organization: loops and entropy“. In: *Current Opinion in Cell Biology* 23.3 (2011), pp. 332–337.

[46] Zach Hensel, Xiaoli Weng, Arvin Cesar Lagda, and Jie Xiao. „Transcription-factor-mediated DNA looping probed by high-resolution, single-molecule imaging in live E. coli cells“. In: *PLoS Biol* 11.6 (2013), e1001591.

[47] D Holcman and Z Schuss. „Control of flux by narrow passages and hidden targets in cellular biology“. In: *Reports on Progress in Physics* 76.7 (2013), p. 074601.

[48] D Holcman, N Hoze, and Z Schuss. „Analysis and interpretation of superresolution single-particle trajectories“. In: *Biophysical journal* 109.9 (2015), pp. 1761–1771.

[49] Chunhui Hou, Ryan Dale, and Ann Dean. „Cell type specificity of chromatin organization mediated by CTCF and cohesin“. In: *Proceedings of the National Academy of Sciences* 107.8 (2010), pp. 3651–3656.

[50] Avelino Javer, Nathan J Kuwada, Zhicheng Long, et al. „Persistent super-diffusive motion of Escherichia coli chromosomal loci“. In: *Nature communications* 5 (2014).

[51] Avelino Javer, Zhicheng Long, Eileen Nugent, et al. „Short-time movement of E. coli chromosomal loci depends on coordinate and subcellular localization“. In: *Nature communications* 4 (2013).

[52] Hansjoerg Jerabek and Dieter W Heermann. „Expression-dependent folding of interphase chromatin“. In: *PloS one* 7.5 (2012), e37525.

[53] S Jespersen, IM Sokolov, and A Blumen. „Small-world Rouse networks as models of cross-linked polymers“. In: *The Journal of Chemical Physics* 113.17 (2000), pp. 7652–7655.

[54] Fulai Jin, Yan Li, Jesse R Dixon, et al. „A high-resolution map of the three-dimensional chromatin interactome in human cells“. In: *Nature* 503.7475 (2013), pp. 290–294.

[55] Daniel Jost, Pascal Carrivain, Giacomo Cavalli, and Cédric Vaillant. „Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains“. In: *Nucleic acids research* (2014), gku698.

[56] Ivan Junier, Olivier Martin, and François Képès. „Spatial and topological organization of DNA chains induced by gene co-localization“. In: *PLoS Comput Biol* 6(2).2 (2010), e1000678.

[57] Stephan Kadauke and Gerd A Blobel. „Chromatin loops in gene regulation“. In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1789.1 (2009), pp. 17–25.

[58] Eldad Kepten, Irena Bronshtein, and Yuval Garini. „Improved estimation of anomalous diffusion exponents in single-particle tracking experiments“. In: *Physical Review E* 87.5 (2013), p. 052713.

[59] Tae Hoon Kim, Ziedulla K Abdullaev, Andrew D Smith, et al. „Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome“. In: *Cell* 128.6 (2007), pp. 1231–1245.

[60] Roger D Kornberg. „Chromatin structure: a repeating unit of histones and DNA“. In: *Science* 184.4139 (1974), pp. 868–871.

[61] Jörg Langowski and Dieter W Heermann. „Computational modeling of the chromatin fiber“. In: *Seminars in cell & developmental biology*. Vol. 18(5). 5. Elsevier. 2007, pp. 659–667.

[62] Imen Lassadi, Alain Kamgoué, Isabelle Goiffon, Nicolas Tanguy-le Gac, and Kerstin Bystricky. „Differential chromosome conformations as hallmarks of cellular identity revealed by mathematical polymer modeling“. In: *PLoS Comput Biol* 11.6 (2015), e1004306.

[63] François Le Dily, François Serra, and Marc A Marti-Renom. „3D modeling of chromatin structure: is there a way to integrate and reconcile single cell and population experimental data?“ In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* (2017).

[64] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. „3D genome reconstruction from chromosomal contacts“. In: *Nature methods* 11.11 (2014), pp. 1141–1143.

[65] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, et al. „Comprehensive mapping of long-range interactions reveals folding principles of the human genome“. In: *science* 326.5950 (2009), pp. 289–293.

[66] 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“. In: *Cell* 158.2 (2014), pp. 339–352.

[67] Shifan Mao, Quinn MacPherson, Jian Qin, and Andrew J Spakowitz. „Field-theoretic simulations of random copolymers with structural rigidity“. In: *Soft Matter* 13.15 (2017), pp. 2760–2772.

[68] G Marrucci, RB Bird, CF Curtiss, RC Armstrong, and O Hassager. *Dynamics of polymeric liquids. Volume 2: Kinetic Theory By R. Ryron Bird, Charles F. Curtis, Robert C. Armstrong, and Ole Hassager, John Wiley & Sons, Inc., New York, 1987, 437+ xxi pp.* 1989.

[69] Wallace F Marshall. „Order and disorder in the nucleus“. In: *Current Biology* 12.5 (2002), R185–R192.

[70] David Martin, Cristina Pantoja, Ana Fernández Miñán, et al. „Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes“. In: *Nature structural & molecular biology* 18.6 (2011), pp. 708–714.

[71] Judith Miné-Hattab and Rodney Rothstein. „Increased chromosome mobility facilitates homology search during recombination“. In: *Nature Cell Biology* 14.5 (2012), pp. 510–517.

[72] Leonid A Mirny. „The fractal globule as a model of chromatin architecture in the cell“. In: *Chromosome research* 19.1 (2011), pp. 37–51.

[73] Takashi Nagano, Yaniv Lubling, Tim J Stevens, et al. „Single-cell Hi-C reveals cell-to-cell variability in chromosome structure“. In: *Nature* 502.7469 (2013), pp. 59–64.

[74] Tsuneyoshi Nakayama, Kousuke Yakubo, and Raymond L Orbach. „Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations“. In: *Reviews of modern physics* 66.2 (1994), p. 381.

[75] Mario Nicodemi and Ana Pombo. „Models of chromosome structure“. In: *Current opinion in cell biology* 28 (2014), pp. 90–95.

[76] Elphege P Nora, Bryan R Lajoie, Edda G Schulz, et al. „Spatial partitioning of the regulatory landscape of the X-inactivation centre“. In: *Nature* 485.7398 (2012), pp. 381–385.

[77] Juan D Olarte-Plata, Noelle Haddad, Cédric Vaillant, and Daniel Jost. „The folding landscape of the epigenome“. In: *Physical Biology* 13.2 (2016), p. 026001.

[78] Donald E Olins and Ada L Olins. „Chromatin history: our view from the bridge“. In: *Nature reviews Molecular cell biology* 4.10 (2003), pp. 809–814.

[79] Richard W Pastor, Robert Zwanzig, and Attila Szabo. „Diffusion limited first contact of the ends of a polymer: comparison of theory with simulation“. In: *The Journal of Chemical Physics* 105.9 (1996), pp. 3878–3882.

[80] Daan Peric-Hupkes, Wouter Meuleman, Ludo Pagie, et al. „Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation“. In: *Molecular cell* 38.4 (2010), pp. 603–613.

[81] Jennifer E Phillips and Victor G Corces. „CTCF: master weaver of the genome“. In: *Cell* 137.7 (2009), pp. 1194–1211.

[82] Jennifer E Phillips-Cremins, Michael EG Sauria, Amartya Sanyal, et al. „Architectural protein subclasses shape 3D organization of genomes during lineage commitment“. In: *Cell* 153.6 (2013), pp. 1281–1295.

[83] Ana Pombo and Miguel R Branco. „Functional organisation of the genome during interphase“. In: *Current opinion in genetics & development* 17.5 (2007), pp. 451–455.

[84] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, et al. „Topologically associating domains are stable units of replication-timing regulation“. In: *Nature* 515.7527 (2014), pp. 402–405.

[85] Suhas SP Rao, Miriam H Huntley, Neva C Durand, et al. „A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping“. In: *Cell* 159.7 (2014), pp. 1665–1680.

[86] Christian Roos, Annette Zippelius, and Paul M Goldbart. „Random networks of crosslinked manifolds“. In: *Journal of Physics A: Mathematical and General* 30.6 (1997), p. 1967.

[87] Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, et al. „Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture“. In: *Cell reports* 10.8 (2015), pp. 1297–1309.

[88] RK Sachs, G Van Den Engh, B Trask, H Yokota, and JE Hearst. „A random-walk/giant-loop model for interphase chromosomes.“ In: *Proceedings of the National Academy of Sciences* 92.7 (1995), pp. 2710–2714.

[89] Adrian L Sanborn, Suhas SP Rao, Su-Chen Huang, et al. „Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes“. In: *Proceedings of the National Academy of Sciences* 112.47 (2015), E6456–E6465.

[90] Zeev Schuss. *Nonlinear filtering and optimal phase tracking*. Vol. 180. Springer Science & Business Media, 2011.

[91] Zeev Schuss. *Theory and applications of stochastic processes: an analytical approach*. Vol. 170. Springer Science & Business Media, 2009.

[92] Andrew Seeber, Vincent Dion, and Susan M Gasser. „Checkpoint kinases and the INO80 nucleosome remodeling complex enhance global chromatin mobility in response to DNA damage“. In: *Genes & development* 27.18 (2013), pp. 1999–2008.

[93] O Shukron and D Holcman. „Statistics of randomly cross-linked polymer models to interpret chromatin conformation capture data“. In: *Physical Review E* 96.1 (2017), p. 012503.

[94] Ofir Shukron and David Holcman. „Transient chromatin properties revealed by polymer models and stochastic simulations constructed from Chromosomal Capture data“. In: *PLOS Computational Biology* 13.4 (2017), e1005469.

[95] Marieke Simonis, Petra Klous, Erik Splinter, et al. „Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C)“. In: *Nature genetics* 38.11 (2006), pp. 1348–1354.

[96] Sevil Sofueva, Eitan Yaffe, Wen-Ching Chan, et al. „Cohesin-mediated interactions organize chromosomal domain architecture“. In: *The EMBO journal* 32.24 (2013), pp. 3119–3129.

[97] IM Sokolov. „Cyclization of a polymer: first-passage problem for a non-Markovian process“. In: *Physical Review Letters* 90.8 (2003), p. 080601.

[98] Jens-Uwe Sommer, Michael Schulz, and Hans L Trautenberg. „Dynamical properties of randomly cross-linked polymer melts: A Monte Carlo study. I. Diffusion dynamics“. In: *The Journal of chemical physics* 98.9 (1993), pp. 7515–7520.

[99] Tim J Stevens, David Lando, Srinjan Basu, et al. „3D structures of individual mammalian genomes studied by single-cell Hi-C“. In: *Nature* 544.7648 (2017), pp. 59–64.

[100] Yoko Suzuki, Jeffrey K Noel, and José N Onuchic. „A semi-analytical description of protein folding that incorporates detailed geometrical information“. In: *The Journal of chemical physics* 134.24 (2011), 06B610.

[101] Mariliis Tark-Dame, Hansjoerg Jerabek, Erik MM Manders, Dieter W Heermann, and Roel van Driel. „Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling“. In: *PLoS Comput Biol* 10.10 (2014), e1003877.

[102] Guido Tiana, Assaf Amitai, Tim Pollex, et al. „Structural fluctuations of the chromatin fiber within topologically associating domains“. In: *Biophysical journal* 110.6 (2016), pp. 1234–1245.

[103] Ngo Minh Toan, Davide Marenduzzo, Peter R Cook, and Cristian Micheletti. „Depletion effects and loop formation in self-avoiding polymers“. In: *Physical review letters* 97.17 (2006), p. 178302.

[104] Tuan Trieu and Jianlin Cheng. „Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data“. In: *Nucleic acids research* 42.7 (2014), e52–e52.

[105] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. „A statistical approach for inferring the 3D structure of the genome“. In: *Bioinformatics* 30.12 (2014), pp. i26–i33.

[106] Jolien Suzanne Verdaasdonk, Paula Andrea Vasquez, Raymond Mario Barry, et al. „Centromere tethering confines chromosome domains“. In: *Molecular Cell* 52.6 (2013), pp. 819–831.

[107] Stephanie C Weber, Michael A Thompson, WE Moerner, Andrew J Spakowitz, and Julie A Theriot. „Analytical tools to distinguish the effects of localization error, confinement, and medium elasticity on the velocity autocorrelation function“. In: *Biophysical Journal* 102.11 (2012), pp. 2443–2450.

[108] Stephanie C Weber, Andrew J Spakowitz, and Julie A Theriot. „Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm“. In: *Physical review letters* 104.23 (2010), p. 238102.

[109] Stephanie C Weber, Julie A Theriot, and Andrew J Spakowitz. „Subdiffusive motion of a polymer composed of subdiffusive monomers“. In: *Physical Review E* 82.1 (2010), p. 011913.

[110] Gerald Wilemski and Marshall Fixman. „Diffusion-controlled intrachain reactions of polymers. I theory“. In: *The Journal of Chemical Physics* 60.3 (1974), pp. 866–877.

[111] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, et al. „Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions“. In: *Nature genetics* 38.11 (2006), pp. 1341–1347.

[112] Jessica Zuin, Jesse R Dixon, Michael IJA van der Reijden, et al. „Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells“. In: *Proceedings of the National Academy of Sciences* 111.3 (2014), pp. 996–1001.

Résumé

Dans cette thèse nous étudions la relation entre la conformation et la dynamique de la chromatine en nous basant sur une classe de modèles de polymères aléatoirement réticulé (AR). Nous utilisons les outils tels que les statistiques, les processus stochastiques, les simulations numériques ainsi que la physique des polymères afin de déduire certaines propriétés des polymères AR a l'équilibre ainsi que pour des cas transitoires. Nous utilisons par la suite ces propriétés afin d'élucider l'organisation dynamique de la chromatine pour diverses échelles et conditions biologiques.

Au chapitre trois nous développons une méthode générale pour construire les polymères AR directement à partir des données expérimentales, c'est-à-dire des données de capture chromosomiques (CC). Nous montrons que des connections longue portées persistantes entre des domaines topologiquement associés (DTA) affectent le temps de rencontre transitoire entre les DTA dans le processus d'inactivation du chromosome X. Nous montrons de plus que la variabilité des exposants anormaux – mesurée en trajectoires de particules individuelles (TPI) – est une conséquence directe de l'hétérogénéité dans la position des réticulations.

Au chapitre quatre, nous utilisons les polymères AR afin d'étudier la réorganisation locale du génome au point de cassure des deux branches d'ADN (CDB). Nous avons trouvé que la perte modérée de connecteur autour des sites de la CDB affecte de façon significative le premier temps de rencontre des deux extrémités cassées lors du processus de réparation d'une CBD. Nous montrons comment un microenvironnement génomique réticulé peut confiner les extrémités d'une cassure, empêchant ainsi les deux brins de dériver l'un de l'autre.

Au chapitre cinq, nous déduisons une expression analytique des propriétés transitoires et a l' équilibre du modèle de polymère AR, représentant une unique région DTA. Nous dérivons par la suite la formule pour le temps moyen de première rencontre (TMPR) entre deux monomères d'un polymère AR. Le TMPR est un temps clé pour des processus tels que la régulation de gènes et la réparation de dommages sur l'ADN.

Au chapitre six, nous généralisons le modèle AR analytique afin de prendre en compte plusieurs DTA de tailles différentes ainsi que les connectivités intra-DTA et extra-DTA. Nous étudions la dynamique de réorganisation de DTA lors des stages successifs de différentiations cellulaires à partir de données de CC. Par la suite nous trouvons une compactification et une décompactification synchrone des DTA à travers les différents stages.

Mots Clés

processus stochastique, modèles de polymères, dynamique de la chromatine, mathéématiques appliquées.

Abstract

In this dissertation we study the relationship between chromatin conformation and dynamics using a class of randomly cross-linked (RCL) polymer models. We use tools from statistics, stochastic process, numerical simulations and polymer physics, to study the properties of the chromatin in processes of DNA breaks.

In the third chapter of this dissertation work, we develop a general method to construct the RCL polymer directly from chromosomal capture (CC) data. We show that persistent long-range connection between topologically associating domain (TAD) affect transient encounter times within TADs, in the process of X chromosome inactivation. We further show that the variability in anomalous exponents, measured in single particle trajectories (SPT), is a direct consequence of the heterogeneity of cross-link positions.

In the forth chapter, we use the RCL polymer to study local genome reorganization around double strand DNA breaks (DSBs). We find that the conservative loss of connectors around DSB sites significantly affects first encounter times of the broken ends in the process of DSB repair. We show how a cross-linked genomic microenvironment can confine the two broken ends of a DSB from drifting apart.

In the fifth chapter, we derive analytical expressions for the steady-state and transient properties of the RCL model representing a single TAD region. We further derive formula for the mean first encounter time (MFET) between any two monomers of the RCL polymer. The MFET is a key time in processes such as gene regulation.

In the sixth chapter, we generalize the analytical RCL model, to account for multiple TADs with variable sizes, intra, and inter-TAD connectivity. We study the dynamic reorganization of TADs, throughout successive stages of cell differentiation, from the CC data. We further find a synchronous compaction and decompaction of TADs during differentiation.

Keywords

stochastic processes, polymer models, chromatin dynamics, applied mathematics.

Numéro national de thèse:
xxxxx