Scuola Internazionale Superiore di Studi Avanzati

Doctoral Thesis

Structure and dynamics of entangled biopolymers: from knotted DNA to chromosomes

Author:
Marco Di Stefano

Supervisor:
Cristian Micheletti
Angelo Rosa

A thesis submitted for the degree of
Philosophiae Doctor

in the

Statistical and Biological Physics sector
Ph.D course in Physics and Chemistry of Biological Systems

October 2014

Contents

To my family...

Introduction

A distinctive feature of DNA filaments that carry the genomic information is that, irrespective of the organisms of origin, they are typically confined in compartments with linear size that is much smaller than their contour length. For instance, bacteriophage DNA, which is typically $\sim 3\mu$m long is packed inside viral capsids that are about $50-80$nm in diameter [1]. The circular DNA of bacteria, on the other hand, spans about 1mm and is contained inside cells whose size is about $1-2\mu$m. Finally in eukaryotic cells, the DNA is partitioned in several chromosome each spanning a few cm, which are confined in a nucleus of about 10μm in diameter [2].

As any other polymer under such high packing densities, the structure and dynamics of genomic DNA are expected to be strongly affected by topological constraints, that is the entanglement of the DNA chains which cannot pass through each other unless aided by the action of specific enzymes [2]

These observations pose the need to understand more in detail the extent to which entanglement, and particularly self-entanglement in the form of physical knots, restricts the conformational plasticity of DNA filaments, which is arguably necessary for the biological viability and functionality of genomic DNA.

This general question has motivated the work presented in this thesis which reports on the theoretical and numerical investigation of the implication of self- and mutual entanglement of charged, semiflexible chains, such as DNA, in several contexts which vary both for the length of the filaments and their packing density.

These studies are presented in order of increasing complexity, starting from the case of individual polyelectrolyte filaments that are a few microns long and which are amenable to single-molecule manipulation techniques. Such systems are presented in chapters 2 and 3, where I consider the effect of external AC/DC electric fields on tensioned and untensioned polyelectrolytes, that accommodate physical knots along their contour.

Specifically, in chapter 2 I report on the motion of knots along mechanically-stretched DNA molecules, mimicking the typical setup of an optical tweezers experiment [3]. The

knotted region is shown to slide along the DNA in the direction of the DC electric force with a field dependent drift velocity. Interestingly, in AC fields the knot motion is found to follow the time-dependent external fields with a noticeable lag. This is one of the typical signatures of dynamical systems with memory and, since such effect was not pointed out before for DNA, we report in detail on how the dynamical memory, or hysteresis, changes with the frequency of the driving electric force.

In chapter 3, instead, I consider general polyelectrolyte chains, modelled after DNA, that are free in solution with trivalent counterions and study their response to DC and AC fields. I will focus on a particular aspect, which has not been considered before, namely the spontaneous formation of non-trivial entanglement in the form of physical knots in dependence of the strength and frequency of the external electric field. The impact of self-knotting on the chain internal dynamics is also discussed.

Finally, I turn to the more complex case of eukaryotic chromosomes, whose structural organization is increasingly studied with the aid of various models [11–13] thanks to the increasing availability of quantitative experimental data that has become available in recent years [7–10]. Our study focuses mostly on the extent to which the internal organization of chromosomes is tailored towards favouring specific functionally-oriented properties. In particular, one intriguing aspect of chromosome structure-function relationship is related to the so-called *gene-kissing hypothesis* [14, 15], according to which the efficient coregulation of genes should reflect in their close spatial proximity.

This hypothesis, has motivated the study reported in chapter 4 where a suitable numerical strategy is used to enforce the simultaneous colocalization of coregulated genes in model chromosomes. The study is aimed at ascertaining to what extent it would be at all physically feasible to bring together in space the large number of gene pairs that are known to be significantly coregulated.

This study is carried out for the gene-rich human chromosome 19 and its macrodomain organization resulting from colocalization of the genes will be compared with that based on experimental HiC data. The relationship between gene coregulation and colocalization will be further explored at a genome-wide level for *D. melanogaster* by using exclusively HiC and gene expression data.

Finally, in chapter 6, I report on a more direct chromosome modelling approach where the HiC data available for human cells [17] is used as knowledge-based constraints to carry out steered molecular dynamics simulations and identify putative three-dimensional chromosome conformations inside the nucleus.

The material presented in this thesis is largely based on the following published papers and ongoing investigations:

- Di Stefano M, Rosa A, Belcastro V, di Bernardo D, Micheletti C (2013)
  Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19.
  PLoS Comput Biol 9(3): e1003019. doi:10.1371/journal.pcbi.1003019
  Chosen for the monthly cover of the PLoS Computation Biology journal for March 2013.

- Di Stefano M, Tubiana L, Di Ventra M, Micheletti C (2014)
  Driving knots on DNA with AC/DC electric fields: topological friction and memory effects.
  Soft Matter 10(34), 6491-6498. doi:10.1039/C4SM00160E

- Nasica-Labouze J, Di Stefano M, di Bernardo D (2014)
  Gene coregulation and colocalization in Drosophila melanogaster: a critical assessment with HiC based on gene expression data and 3D models.
  Manuscript in preparation, to be submitted by the end of December 2014.

- In collaboration with: Hovig E, Paulsen J (University of Oslo), Micheletti C.
  Three-dimensional modelling of the human genome: a knowledge-based approach.
  Expected completion date: December 2014.

- In collaboration with: Di Ventra M (UCSD), Orland H and Micheletti C.
  Complex dynamics of polyelectrolytes in DC electric fields.
  Expected completion date: February 2015.

Chapter 1

Phenomenology of polymers and biopolymers: theory and numerical methods

In this thesis, we will employ state-of-the-art physical polymer models to treat the large-scale structure and dynamics of biomolecules, namely DNA and chromatin fibers (see Introduction).

From the physico-chemical point of view, a polymer is a macromolecular compound made of single, discrete units, generically called *monomers* [18, 19]. In this thesis we will treat mainly linear homopolymer chains, *i.e.* polymers with identical monomers arranged into a linear sequence. Although this might look as an oversimplification of the complex nature of the heteropolymer character of DNA and chromatin [20], it provides nonetheless a remarkably accurate description of their corresponding, generic large-scale spatial and temporal behaviors under different physical conditions [21].

1.1 Basics concepts in Polymer Physics

A (linear) polymer chain is schematically modeled as a set of $(N+1)$ *beads* at spatial positions $\vec{r}_i$, with $i = 0, 1, ..., N$ (Fig. 1.1). For practical purposes, it is also convenient to introduce the notation (see Fig. 1.1) for the corresponding:

- N bond vectors $\vec{b}_i \equiv \vec{r}_i - \vec{r}_{i-1}$, with $i = 1, 2, ..., N$;
- $N-1$ bond angles $\cos(\theta_i) \equiv \frac{\vec{b}_i}{|\vec{b}_i|} \cdot \frac{\vec{b}_{i+1}}{|\vec{b}_{i+1}|}$, with $i = 1, 2, ..., N-1$.

With this definition, the total polymer *curvilinear* or *contour* length, L_c, is given by $L_c = \sum_{i=1}^{N} \left|\vec{b}_i\right|$, which reduces to $L_c = bN$ in the most common case where all bond vectors have the same length, b.

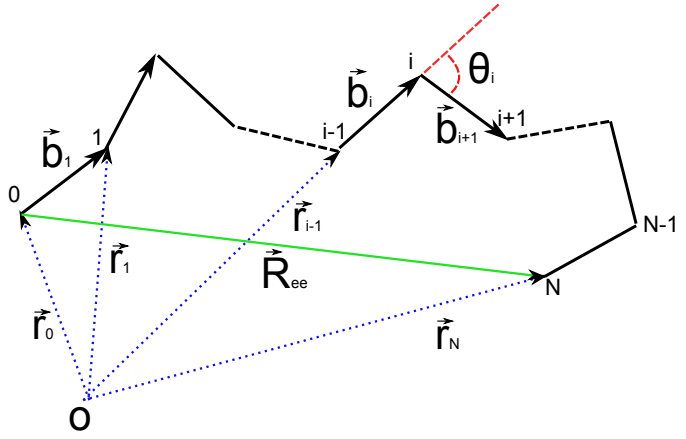


Figure 1.1: Representation of a polymer conformation.

The typical polymer size as a function of the number of its constituent monomers is a central quantity in Polymer Physics [18, 19]. It is aptly characterized by either

(1) the *mean square end-to-end distance* (Fig. 1.1):

$$R_{ee}^2 \equiv \left\langle |\vec{r}_N - \vec{r}_0|^2 \right\rangle = \left\langle \left| \sum_{i=1}^{N} \vec{b}_i \right|^2 \right\rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \vec{b}_i \cdot \vec{b}_j \right\rangle , \tag{1.1}$$

or

(2) the *mean square radius of gyration*:

$$R_g^2 \equiv \frac{1}{N+1} \sum_{i=0}^{N} \left\langle (\vec{r}_i - \vec{r}_{cm})^2 \right\rangle = \frac{2}{(N+1)^2} \sum_{i=0}^{N} \sum_{j=i+1}^{N} \left\langle (\vec{r}_i - \vec{r}_j)^2 \right\rangle , \tag{1.2}$$

where $\vec{r}_{cm} = \frac{1}{N+1} \sum_{i=0}^{N} \vec{r}_i$ is the center of mass of the polymer.

In previous equations, and in the rest of this thesis, quantities inside brackets "$\langle\,\rangle$" indicate their corresponding averages at thermal equilibrium.

We will now provide a simple application of Eq. 1.1 to the class of so-called semiflexible polymers, namely the same class which the polymers studied in this thesis (DNA and chromatin) belong to. Semiflexible polymers [18] are characterized by the fact that, due to thermal motion, the correlation function $\left\langle \vec{b}_i \cdot \vec{b}_j \right\rangle$ appearing in Eq. 1.1 decays as an exponential function of the contour length distance between corresponding monomers, $\left\langle \vec{b}_i \cdot \vec{b}_j \right\rangle = b^2 \exp\left(\frac{-b|i-j|}{l_p}\right)$. l_p is the so-called *persistence length*, whose physical meaning

is the following: after a few l_p's local directions of polymer chain become statistically uncorrelated. Upon insertion of the previous expression into Eq. 1.1, one has

$$\begin{aligned} R_{ee}^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N} \exp\left(\frac{-b\,|i-j|}{l_p}\right) \\ &= 2l_p^2\left[\frac{L_c}{l_p} + e^{-\frac{L_c}{l_p}} - 1\right] \end{aligned} \tag{1.3}$$

Eq. 1.3 has two interesting limits:

- Short-chain or rod-like limit, $L_c << l_p$. In this case the chain is sensibly shorter than its persistence length, thermal fluctuations are suppressed and Eq. 1.3 reads:

$$R_{ee}^2 \simeq L_c^2. \tag{1.4}$$

- Long- or flexible-chain limit $L_c >> l_p$. In this case the polymer is much longer than is persistence length. Therefore:

$$R_{ee}^2 \simeq 2L_c l_p. \tag{1.5}$$

  In this case, thermal fluctuations dominate chain statistics [18]. Moreover, from the ratio of R_{ee}^2 and L_c, we can define the *Kuhn length*, l_K, that is frequently used as another possible measure of chain flexibility [18]. l_K and l_p are simply linked by $l_K \equiv R_{ee}^2/L_c \simeq 2l_p$.

The specific value for l_p for DNA and chromatin depends on the experimental conditions of the solvent surrounding the polymer. Typical values are $l_p = 50$nm for DNA [22] and 150nm for chromatin fiber [2, 23].

1.2 Modelling of biopolymers

We have so far introduced phantom chain polymer models that have essentially two structural properties: the chain connectivity that ensures and a mechanical persistence length l_p which takes into account the flexibility (or rigidity) of the polymer.

Since the aim of the present thesis is to study the structural and topological properties of long biopolymers (both isolated or in solution), it is necessary to abandon the description of polymers as phantom chains and introduce the excluded volume effects. The latter should avoid unphysical overlap of beads and prevent the bond crossing.

Introduction of excluded volume effects in polymer models make analytical approaches technically very difficult if not impossible [18, 19]. For these reasons, one frequently resorts to numerical computer simulations of polymer models which have been proven to be extremely fruitful in the past [24–26]. In this thesis we are going to adopt the Kremer-Grest model for polymer dynamics [24], whose technical details we provide in the next Section.

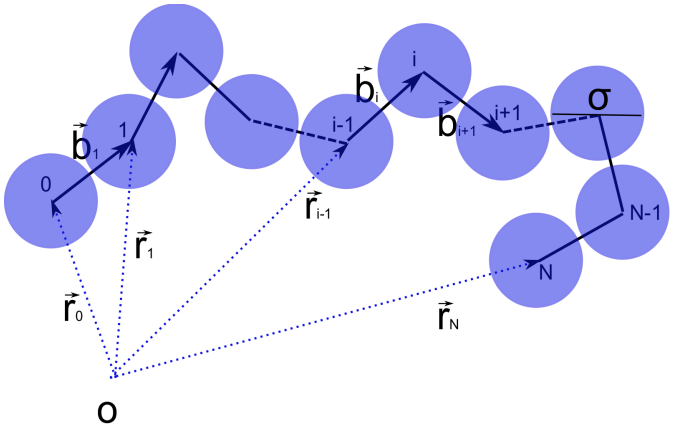


Figure 1.2: Representation of a Kremer and Grest polymer chain conformation.

1.2.1 Kremer and Grest polymer model

To model a single biomolecule we used the polymer model introduced by Kremer and Grest in 1990 [24], which describes a polymer as a chain of $N+1$ beads connected by N bonds with a suitable bending rigidity. Specifically, the beads interact with the following potential energy:

$$\mathcal{H}_{KG} = U_{LJ} + U_{FENE} + U_{KP} \tag{1.6}$$

The first term is a purely repulsive *Lennard-Jones* interaction of the *Weeks-Chandler-Andersen* form:

$$U_{LJ} = \begin{cases} \frac{\epsilon}{2} \sum_{(i,j),j\neq i}^{N} 4 \left[\left(\frac{\sigma}{d_{ij}}\right)^{12} - \left(\frac{\sigma}{d_{ij}}\right)^{6} + \frac{1}{4} \right] & \text{if } d_{ij} < 2^{1/6}\sigma \\ 0 & \text{otherwise} \end{cases} \tag{1.7}$$

where ϵ is the interaction strength and is taken as the unit energy of the system, $d_{ij} = |\vec{r}_i - \vec{r}_j|$ is the distance of the bead centers i and j and σ is the diameter of the beads, which is taken as the unit length. This energy contribution decays smoothly to zero at $2^{1/6}\sigma$ that is the distance at which the interaction has a minimum.

This energy term implements the excluded volume effect and avoids the unphysical overlapping of different beads of the chain, assigning to the latter a diameter σ.

The other two terms are meant to recover the properties of the polymer chain that we already mentioned in Section 1.1: the chain connectivity and the chain flexibility. The former is ensured by the *Finite-extensible non-linear elastic* (FENE) potential acting between consecutive beads:

$$U_{\text{FENE}} = -\frac{\epsilon}{2}\sum_{i=1}^{N} K\left(\frac{R_0}{\sigma}\right)^2 \ln\left[1-\left(\frac{\left|\vec{b}_i\right|}{R_0}\right)^2\right] \tag{1.8}$$

where $K = 30.0$ is the potential strength, $R_0 = 1.5\sigma$ is the maximum bond length and $\vec{b}_i$ is the i_{th} bond vector and is the distance of the bead centers i and $i+1$ that are consecutive along the chain. The length of the bonds is not fixed (*extensible bonds*), but is kept under strict control by the logarithmic form of the potential, which diverges in R_0.

Next, triplets of consecutive beads are involved in the *Kratky-Porod* (KP) interaction:

$$U_{\text{KP}} = \epsilon\sum_{i=1}^{N}\left(\frac{K_b}{\sigma}\right)\left(1-\frac{\vec{b}_i\cdot\vec{b}_{i+1}}{\left|\vec{b}_i\right|\left|\vec{b}_{i+1}\right|}\right) = \epsilon\sum_{i=1}^{N-1}\left(\frac{K_b}{\sigma}\right)(1-\cos(\theta_i)) \tag{1.9}$$

where K_p is the bending rigidity with the physical dimension of a length. As it as been discussed in the previous Section, this term describes the flexibility (or rigidity properties) of the polymer chain. In particular, the bending rigidity K_b is related to the persistence length of the chain by the following relation: $K_b = l_p$.

1.2.2 Chromatin fiber and DNA models

The Kremer and Grest model in Eq. 1.6 provides a basis to describe various biopolymers starting from their salient physical properties. The model is able, in fact, to take into account the thickness, the connectivity and the bending rigidity of the chain.

For instance, in the case of the chromatin fiber, which we study in Chapters 4 and 6, we will use $\sigma = 30nm$ and $l_p = 150\,nm$ [11].

In Chapters 2 and 3, we shall consider the properties of polyelectrolytes, e.g. dsDNA. In this case in addition to using proper parameters for the thickness ($\sigma = 2.5nm$) and the bending rigidity ($l_p = 50nm$), we have to take into account also the the electrostatic charge of the molecule. This will be done by adding to the potential energy in Eq. 1.6 the electrostatic one U_{EL}. The resulting Hamiltonian of the polyelectrolyte chain is:

$$\mathcal{H}_{\text{PE}} = U_{\text{FENE}} + U_{\text{KP}} + U_{\text{LJ}} + U_{\text{EL}} \tag{1.10}$$

We treated the electrostatic part in two ways. The first is the Coulombic interaction:

$$U_{\mathrm{C}} = \frac{1}{2} \sum_{(i,j),j\neq i}^{N} \frac{q_i q_j}{4\pi\epsilon_0\, \epsilon_r d_{ij}} \tag{1.11}$$

where q_i (q_j) is the charge of the i-th (j-th) bead and $\epsilon_0\epsilon_r$ is the permittivity of the medium. An important length-scale associated to the electrostatics is the *Bjerrum length*, $l_B = e^2/(4\pi\epsilon_0\epsilon_r\kappa_B T)$, that is the spatial separation at which the interaction between two electronic charges, e, is comparable in magnitude to the thermal energy scale $k_B T$, where k_B is the Boltzmann constant and T the temperature.

When describing isolated or dilute solutions of polyelectrolytes, the electrostatic forces between charged beads are considerable at short distances, but at sufficiently large spatial separations can be considered as completely screened by the surrounding ions (counterions). Following the Debye-Hückel mean-field approach [27], the screening effect of the counterions can be described, first, by correcting the thickness to account for the ions condensed on the charged chain, for instance in chapter 2 we increase the bare thickness of DNA (2nm) to 2.5nm. Next, it is assigned to each bead an effective (smaller) charge q_i' that usually is related to its intrinsic charge q_i by a factor one half, $q_i' = 0.5\, q_i$ [28]. Finally, the electrostatic interaction is depleted with an exponential decay as in the following:

$$U_{\mathrm{DH}} = \frac{1}{2} \sum_{(i,j),j\neq i}^{N} \frac{q_i' q_j'}{4\pi\epsilon_0\, \epsilon_r d_{ij}} e^{-\frac{d_{ij}}{\lambda_{\mathrm{DH}}}} \tag{1.12}$$

In the latter expression, we introduced the *Debye-Hückel screening length* that is the characteristic length scale at which the complete electrostatic interaction is depleted to 30% of its unscreened value:

$$\lambda_{\mathrm{DH}} = (8\pi N_A I l_B)^{-1/2} \tag{1.13}$$

where N_A is the Avogadro number, I is the ionic strength of the solution and l_B is the Bjerrum length.

1.3 Dynamics of polymers

1.3.1 Langevin Equation

Time evolution of polymer dynamics under the force field described in previous Section is obtained by numerical integrations of the corresponding Newton equations of motion,

by using Langevin dynamics [24].

Specifically, the coordinate $r_{i\alpha}$ of particle i evolves according to equation:

$$m\ddot{r}_{i\alpha} = -\partial_{i\alpha}\mathcal{H} - \gamma\dot{r}_{i\alpha} + \eta_{i\alpha}(t) \tag{1.14}$$

where $i = 0, 1, 2, \ldots, N$ is the bead index and $\alpha = x, y, z$ says that the equation applies to all the Cartesian coordinates. In Eq. 1.14:

- The first term on the right is the usual expression $-\partial_{i\alpha}\mathcal{H}$ that relate the Hamiltonian $\mathcal{H}$ to the components of the deterministic force acting on the bead i.
- $-\gamma\dot{r}_{i\alpha}$ is a deterministic damping Stokes term, in which $\gamma > 0$ is the dumping coefficient. This describes the fact that a particle moving with a certain velocity $\vec{\dot{r}}_i$ with respect to the fluid sees more fluid particles coming from the direction in which it is moving with respect to the other directions.
- $\eta_{i\alpha}(t)$ is a Gaussian term, describing the continuous stochastic collisions of the fluid particles with the beads. The average time lapse between two of this collapses the shortest characteristic time scale of the system and it is usually named the collision time τ_c. This time can be seen as the time scale at which the small particles of the fluid decorrelate their velocities.

The energetic properties of the system polymer+surrounding fluid, in particular the temperature at equilibrium, are, then, contained in the last term.

1.3.2 Properties of the noise term and their physical interpretation.

The statistics of the Gaussian term $\vec{\eta}(t)$ is aptly defined by the average and the variance:

$$\begin{cases} \langle \eta_{i\alpha}(t) \rangle & = 0 \\ \langle \eta_{i\alpha}(t)\, \eta_{j\beta}(t') \rangle & = 2\kappa_B T \gamma\, \delta_{ij}\, \delta_{\alpha\beta}\, \delta(t - t') \end{cases} \tag{1.15}$$

where δ_{ij} is the Kronecher delta that is equal to 1 for $i = j$ and zero otherwise and $\delta(t - t')$ is the Dirac delta. To understand the reason of the latter choices, we shall see that we are really describing particles in thermal equilibrium at temperature T.

To do this, let's consider the Langevin equation for a single isolated bead in absence of external forces ($\partial_{i\alpha}\mathcal{H} = 0$) with an initial velocity $\dot{r}_{0x}$ performing a 1D motion along the $\hat{x}$ direction. In this case the Langevin equation, simplifies as follow:

$$m\ddot{r}_x = -\gamma\dot{r}_x + \eta_x(t)\,. \tag{1.16}$$

First, integrating the Langevin equation from the initial time $t = 0$ to a time t, we can write down a formal solution for the bead velocity:

$$\dot{r}_x(t) = \dot{r}_{0x} e^{-(\gamma/m)t} + \frac{1}{m}\int_0^t dt' \eta_x(t') e^{-(\gamma/m)(t-t')} \tag{1.17}$$

Next, by using the properties of the Gaussian term, we can easily compute the average of the formal solution:

$$\begin{aligned} \langle \dot{r}_x(t) \rangle &= \dot{r}_{0x} e^{-(\gamma/m)t} + \frac{1}{m}\int_0^t dt' \left\langle \eta_x(t') \right\rangle e^{-(\gamma/m)(t-t')} \\ &= \dot{r}_{0x} e^{-(\gamma/m)t} \end{aligned} \tag{1.18}$$

where the last equality follows from the property of the average of the random force $\langle \eta_x(t) \rangle = 0$ for every t. The latter is, *a posteriori*, a right choice to ensure that, as expected, a particle starting at rest $\dot{r}_{0x} = 0$ remains (on average) motionless.
The expression of the velocity in Eq. 1.18 reveals the existence of a characteristic relaxation time of the velocity of the bead: $\tau_v = m/\gamma$. This time is much larger than the previously mentioned collision time $\tau_v >> \tau_c$ because the velocity of the bead decorrelates much later than the ones of the fluid small particles.

The physical meaning of the property 1.15 of the random term can be recognized by looking at the variance of the velocity:

$$\begin{aligned} \left\langle [\dot{r}_x(t) - \langle \dot{r}_x(t) \rangle]^2 \right\rangle &= \frac{1}{m^2}\int_0^t \int_0^t dt' dt'' \left\langle \eta_x(t') \eta_x(t'') \right\rangle e^{-(t-t')/\tau_v} e^{-(t-t'')/\tau_v} \\ &= \frac{1}{m^2}\int_0^t \int_0^t dt' dt'' 2 k_B T \gamma \delta_{t't''} e^{-(t-t')/\tau_v} e^{-(t-t'')/\tau_v} \\ &= \frac{k_B T}{m}\left(1 - e^{-2t/\tau_v}\right) \end{aligned} \tag{1.19}$$

For $t >> \tau_v >> \tau_c$, we recover, at equilibrium, the more general result of the *equipartition theorem*:

$$\left\langle [\dot{r}_x(t) - \langle \dot{r}_x(t) \rangle]^2 \right\rangle = \frac{k_B T}{m} \tag{1.20}$$

1.4 Time-scales of the simulations

The Langevin equation is integrated numerically with the LAMMPS simulation package [29], by using a standard velocity Verlet algorithm [30].

The natural time-scale of the simulated dynamics is the *Lennard-Jones time*, τ_{LJ}, which is obtained as the simplest combination of the unit scales of the model σ (*length*), m

(*mass*) and ϵ (*energy*):

$$\tau_{\mathrm{LJ}} = \sigma\sqrt{\frac{m}{\epsilon}}. \tag{1.21}$$

The integration time step is, in fact, a fraction of τ_{LJ} and, in particular, we usually set in this thesis $\Delta t = 0.012\tau_{\mathrm{LJ}}$, accordingly to the choice in ref. [24].

1.4.1 Mapping simulation time onto real time

We shall discuss how it is possible to map the simulation time of the numerical integration onto a realistic time scale. To do so, we rely on simple considerations about the diffusive motion of an isolated spherical bead as described with the Langevin equation 1.14 of motion. We point out that in the latter the hydrodynamic effects, which might have a large impact in the following consideration, are neglected.

We recall that, from Stokes' law applied to a spherical particle of radius $\sigma/2$:

$$\gamma = 3\pi\eta\sigma \tag{1.22}$$

where η is the viscosity of the fluid surrounding the particle. In ref. [24], Kremer and Grest have set:

$$\tau_v = \frac{m}{\gamma} = 2\tau_{\mathrm{LJ}} = 2\sigma\sqrt{\frac{m}{\epsilon}} \Rightarrow \gamma = \frac{\sqrt{m\epsilon}}{2\sigma} \tag{1.23}$$

By equating the two expressions for γ, we obtain the square root of the mass of one bead:

$$\sqrt{m} = \frac{6\pi\eta\sigma^2}{\sqrt{\epsilon}}\ . \tag{1.24}$$

In conclusion, we can provide an estimate of the mapping of the integration time onto real time:

$$\tau_{LJ} = \sigma\sqrt{\frac{m}{\epsilon}} = \frac{6\pi\eta\sigma^3}{\epsilon}. \tag{1.25}$$

As an example, we can consider a spherical bead in water ($\eta = 1cP$) of diameter $\sigma = 2.5\,nm$ (the typical thickness of dsDNA). Since at the temperature $T = 300\,K$, $\epsilon = k_BT = 4.2\,pN\,nm$, we have that the simulation time is of the order of $\tau_{LJ} \simeq 0.75ns$.

Chapter 2

Driving knots on DNA with AC/DC electric fields: topological friction and memory effects

Long and densely packed polymers are inevitably entangled and, in particular, are prone to develop physical knots. In fact, from general arguments based on topology and polymer physics [31], one has that the probability that an equilibrated unconstrained chain is unknotted decreases exponentially with chain length and the same holds for increasing confinement [32].

The impact of knots, on the physical properties of biopolymers which can be both long, and densely packed (as it is usually the case for DNA that is subject to spatial confinement for any type of organism) [33], has been an actively researched topic both for its general connection with polymer physics [34–42] and for its biological [1, 43] and technological ramifications [31, 44].

In particular, recent experimental and theoretical studies have dealt with fundamental issues such as clarifying the mechanisms leading to the spontaneous formation and untying of knots in biopolymers [45–49], identifying the length and time scales of relaxation introduced by the presence of knots in chains and rings [32, 50–52] or the topological friction hindering translocation of polymers through pores [44, 53, 54], their motion in a channel [55] or the DNA ejection from viral capsids [56, 57].

A classic experimental setup, in which several of these aspects can be simultaneously observed, is offered by DNA filaments that can be both knotted and stretched using optical tweezers [3, 58, 59]. The technique has provided considerable insight into the complex interplay of characteristic time-scales [60, 61] and length-scales [39] that generally control the diffusive dynamics of the knotted region along the chain.

In this chapter we shall report on a computational study on knotted tensioned chain that has recently appeared in Soft Matter [62]. Specifically, we performed extensive molecular dynamics simulations of model knotted DNA chains under mechanical tension and analyzed how the knotted region evolves under the action of external DC or AC electric fields.

2.1 The model

We considered a knotted polyelectrolyte (PE) chain that is pulled at both termini by a mechanical force $\vec{F}_s$ directed along the z Cartesian axis, whose unit vector we shall indicate with $\hat{z}$. After equilibrating the stretched knotted chain, the two termini were pinned and the DNA filament was subjected to the action of a spatially-uniform longitudinal external electric field, both constant (DC) and time-dependent (AC).

As described in Section 1.2, the polyelectrolyte was modeled as a chain of N charged beads. To accommodate the elongated chain, the periodic parallelepiped simulation box had the z-side of length $N\sigma$ while each of the other two sides had length $100\,\sigma$. Each bead had mass m, electric charge q and diameter σ.

The beads interacted with the following potential energy:

$$\mathcal{H} = \mathcal{H}_{\rm PE} + U_{\rm ext} \tag{2.1}$$

The first term accounted for the potential energies between the beads of the PE chain and has been introduced in Eq. 1.10 of Chapter 1.

Specifically, the parameters of the model PE were tuned so as to match the nominal properties of a dsDNA filament in water solution (dielectric constant $\epsilon_r = 80$) with $0.01\,M$ NaCl at the temperature $T = 300$K. The effective charge of the beads, each spanning ~ 8 basepairs, was set equal to $q = -8$e, to account for the $\sim 50\%$ reduction of the nominal phosphate charge due to monovalent counterions [28]. The Bjerrum length was accordingly set to $l_B = e^2/(4\pi\epsilon_0\epsilon_r\kappa_B T) = 0.7$nm. The interaction strength ϵ was set equal to $\kappa_B T = 4 \cdot 10^{-21}\,J$ at $T = 300\,K$. The bead diameter was equal to the nominal hydrated DNA diameter, $\sigma = 2.5$nm, and the persistence length was set equal to $l_p = 50$nm.

The last term in eqn. 2.1, $U_{\rm ext}$, described the interactions of the PE chain with external forces.

When studying the free diffusion of knots along the stretched DNA chains, U_{ext} is simply the potential energy associated to the stretching force F_s pulling the two termini in the $-\hat{z}$ and $+\hat{z}$ directions, respectively:

$$U_{\rm ext} \;=\; F_s \hat{z} \cdot (\vec{r}_1 - \vec{r}_N) \tag{2.2}$$

When considering the interaction with a spatially-uniform, and possibly time-dependent, external field $E(t)$ along $\hat{z}$, the stretching force was replaced by a constraint which kept

the terminal beads fixed and apart in space. This change of boundary conditions was made necessary to prevent the chain to drift as a whole. The field potential energy was in this case:

$$U_{\text{ext}} \quad = \quad \sum_{i=1}^{N} qE(t)\,\hat{z}\cdot\vec{r}_i \ . \tag{2.3}$$

As mentioned in Section 1.14, the chain dynamics was described with an underdamped Langevin equation:

$$m\ddot{r}_{i\alpha} = -\partial_{i\alpha}\mathcal{H} - \gamma r_{i\alpha} + \eta_{i\alpha}(t) \tag{2.4}$$

where $i = 1,\ldots,N$ runs over the particles of the system and α over the Cartesian coordinates. The Gaussian white noise $\eta_{i\alpha}(t)$ has the usual statistical properties $\langle\eta_{i\alpha}\rangle = 0$ and $\langle\eta_{i\alpha}(t)\,\eta_{j\beta}(t')\rangle = 2\kappa_B T\gamma\,\delta_{ij}\delta_{\alpha\beta}\delta(t-t')$, where δ_{ij} is the Kronecher delta, $\delta(t-t')$ is the Dirac delta, k_B the Boltzmann constant and T the temperature of the system. The Langevin equation was solved numerically with the LAMMPS simulation package [29], with an integration time step $\Delta t = 0.012\tau_{LJ}$, where $\tau_{LJ} = \sigma\sqrt{m/\epsilon}$ and with $m/\gamma = 2\tau_{LJ}$ (see Section 1.4 and ref. [24]). Consistently with the time mapping in Section 1.4.1, for water solutions, where the viscosity $\eta = 1$ cP, one has $\tau_{LJ} \sim 75$ns so that each integration time step had a duration of $\Delta t \sim 1$ns.

2.1.1 Simulation details

The initial configuration of the knotted chains were generated by first discretizing in about 300 beads a parametric closed knotted configuration obtained from the Wolfram *MathWorld* web resource, next opening it by removing $15-20$ beads and finally prolonging the termini with straight segments running in opposite directions. The segments are equally long and their length was chosen so to yield a total of $N = 2000$ beads in the chain. This chain length was sufficiently long to study the knot motion avoiding border effects and short enough to be treated computationally.

The configurations, once knotted, were equilibrated for $1\cdot 10^8$ integration steps under the action of a longitudinal stretching force of $F_s = 10.0$pN applied at both termini. After this equilibration time, various observables were recorded at fixed stretching force.

As mentioned before, when investigating the action of an external longitudinal electric field, E, we replaced the action of the stretching force with the constraint of keeping

the termini at fixed positions. This pinning of the ends was introduced on chain configurations previously equilibrated at $F_s = 10.0$pN, so that for $E = 0$ the nominal average chain tension was equal to the stretching force in zero field, 10.0pN.

For each considered condition (tensile force in no external field, fixed termini at various DC and AC field) we gathered from a minimum of 3 to a maximum of 10 different production runs each typically spanning 10^9 integration steps. The reported values of the average knot length, $\langle l_{knot} \rangle$ and its estimated error were obtained from the block averages of l_{knot} calculated over non-overlapping intervals of about 10^8 integration steps. The knot diffusion coefficient along the chain contour was calculated from the linear fit of the knot mean-square displacements on the chain versus time cumulated over all production runs. Its error was estimated from the dispersion of the diffusion coefficients calculated separately for the various runs. An analogous procedure was used to compute the average and standard deviations of the drift velocity in DC and AC fields. The standard expression for error propagation was used to compute the error on derived quantities.

2.1.2 Identifying and locating knots on open DNA chains

As it is shown in Figure 2.1a, we considered five different knot topologies, corresponding to the simplest types of prime knots, namely: 3_1 (also called trefoil knot), 4_1, 5_1, 5_2 and 7_1. Each knot was inserted in the middle region of the polymer and allowed to equilibrate on the chain pulled at both termini with a stretching force of $F_s = 10.0$pN. Close-ups of the equilibrated knots are shown in Figure 2.1b.

To track the position of the knot in the course of the simulation we adopted the method described in ref. [64]. The strategy involves a bottom-up search of the knotted region. Specifically, we starts by considering all possible chain segments involving as few as 5 beads and checks whether whether any of them accommodates a physical knot. If no knot is found, the search is expanded to segments of six beads and so on, until one identifies the shortest knotted region. We note that, in order to detect whether a linear portion of the chain accommodates a physical knot, it is necessary to close it into a circle, so to trap its entanglement in proper topological state. For this purpose we used the minimally-interfering closure scheme of ref. [64].

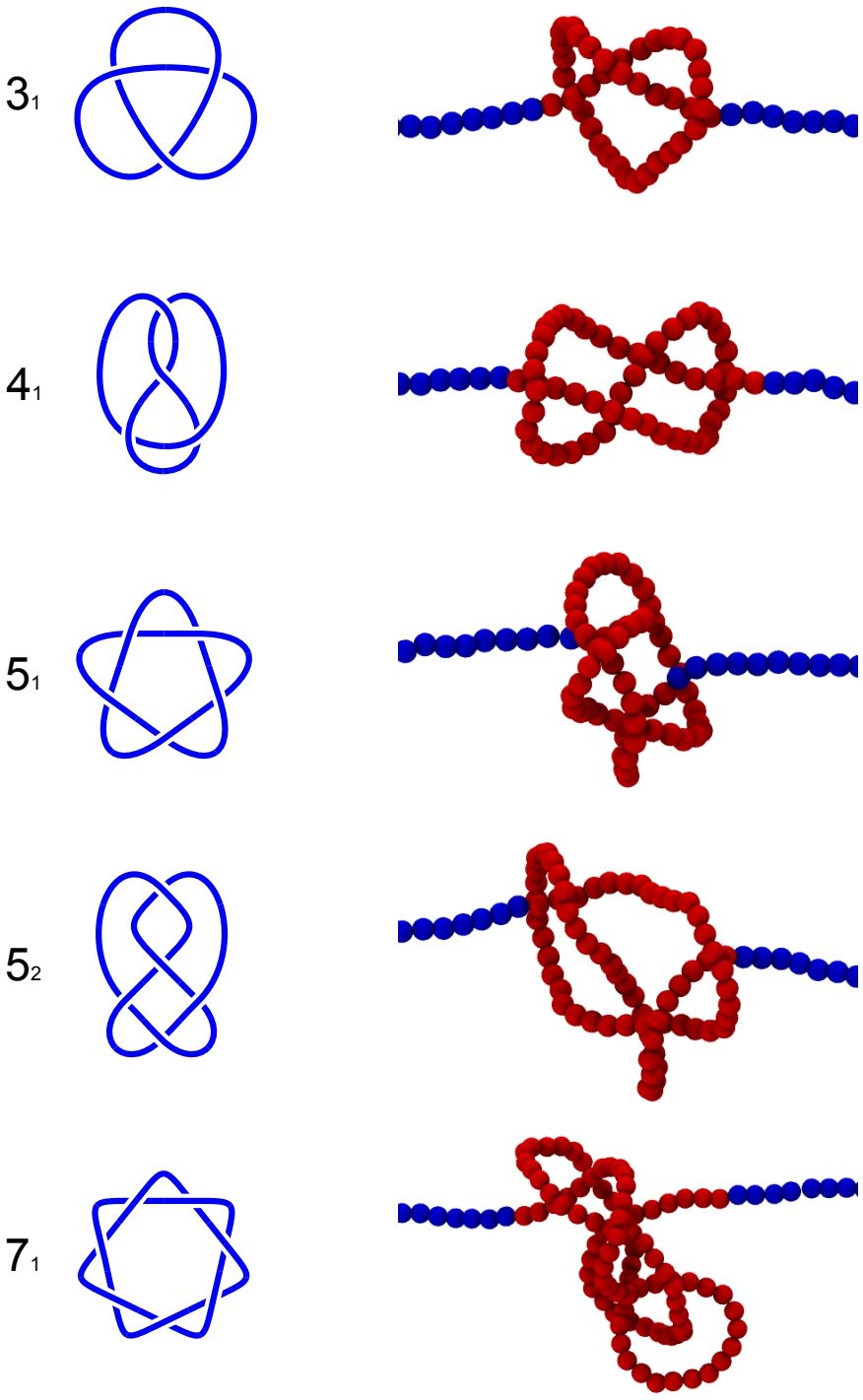


Figure 2.1: Knot diagrams and their counterpart tied in open dsDNA model chains - *Left:* Diagrammatic representation of the knot types used for the present investigation. Notice that the diagrams are for closed curves, as required for a proper mathematical definition of knots. The corresponding physical knots tied in open chains are shown on the right. The minimal knotted portions are highlighted in red. These and other graphical representations of model chains were rendered with the VMD graphical package [63]. On the leftmost column, we indicate the type of knot in the standard notation C_s, where C is the minimum number of crossings the knot can have in a planar projection. Since for any given C there can be several different knots, an additional number s labels the standard sequential ordering of the knots. For $C = 3$ and $C = 4$ there is only one knot type, while for $C = 5$ there are two and for knots with 7 crossings there are 7 distinct topologies.

2.2 Simulations of tensioned DNA chains

We started our investigation by characterizing the motion of knots along a polyelectrolyte chain that was stretched by pulling both ends in opposite directions with a constant force, $F_s = 10.0$pN. This force value falls in the typical range used in current DNA manipulation experiments with optical tweezers.

A typical configuration of the stretched DNA chain, with the time evolution of the knotted region location is provided in Fig. 2.2a,b for a trefoil (3_1) knot.

A relevant parameter to characterize the internal dynamics, mechanics and equilibrium properties of the knotted chains is the contour length of the knotted region, l_{knot} [58, 61, 65].

The distribution of l_{knot} of the different knot types is shown in Fig. 2.2c along with their average values, given also in Table 2.1. It was seen that $\langle l_{knot} \rangle$ increased with the nominal complexity of the knot. Such trend is, in fact, known to hold more generally because it was observed in untensioned knotted rings as well as in tensioned knotted chains without electrostatic self-repulsion [61].

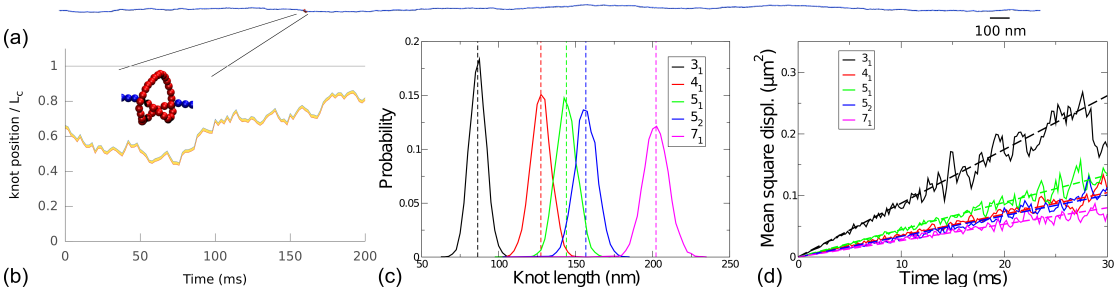


Figure 2.2: Knot diffusion along a stretched DNA chain at zero external electric field - All panels refer to model DNA chains with $L_c = 5\,\mu$m subjected to a stretching force $F_s = 10.0$pN. (a) Typical configuration of a trefoil-knotted (3_1) DNA chain. The knotted region is highlighted in red. (b) Typical time evolution of the boundaries of the knotted region. (c) Knot length distributions for 3_1, 4_1, 5_1, 5_2, 7_1 knots; the dashed lines mark the average values, see Table 2.1. (d) Time dependence of the mean-square displacement of the knot center along the chain. The dashed lines indicate the linear fitting curves.

2.2.1 Effect of screened electrostatic on the knot length

In all cases, it is seen that the knotted region spanned only a small fraction of the chain. This is an interesting point which highlights the competition between several effects: the applied mechanical tension, the chain intrinsic bending rigidity and its electrostatic self-repulsion.

It is, in fact, important to realize that electrostatic interactions affect the chain on scales

much larger than the Debye-Hückel screening length, as they affect the effective rigidity of the chain [50].

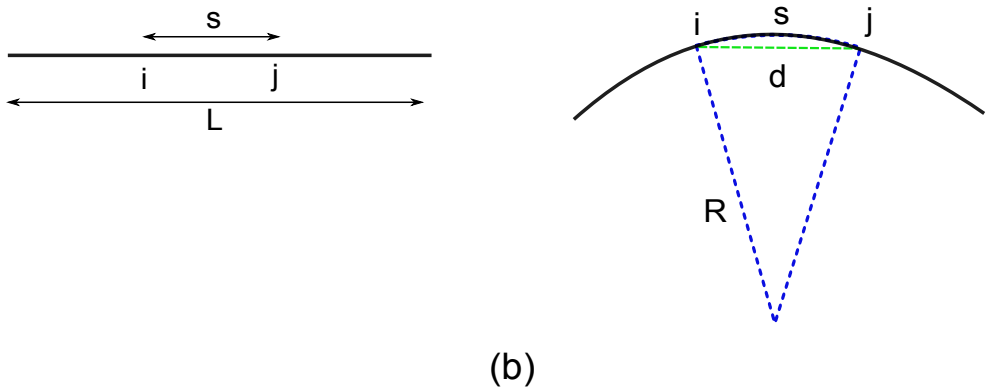


Figure 2.3: Straight (a) and bent (b) conformations of a polymer chain.

To illustrate this point, we follow, here, the spirit of ref. [50] and consider two beads i and j of at arc length distance s on the chain, see Figure 2.3a. We also assume that the chain contour length L is much larger than the Debye-Hückel screening length λ_{DH}. Accordingly, the electrostatic screened energy between the two points is:

$$E^s_{i,j} = \frac{q^2}{4\pi\epsilon_0\epsilon_r s}\, e^{-s/\lambda_{DH}} \tag{2.5}$$

where the superscript s stands for *straight*.

Next, we bend the chain to a curvature radius $R >> L >> s$. Thus, we bring the two charged beads closer to each other and, hence, increase their mutual electrostatic energy:

$$E^b_{i,j} = \frac{q^2}{4\pi\epsilon_0\epsilon_r d}\, e^{-d/\lambda_{DH}} \tag{2.6}$$

where the superscript b stands for *bent* and d is the spatial separation of point i and j, as sketched in Figure 2.3b:

$$d \;=\; 2R\sin\left(\frac{s}{2R}\right). \tag{2.7}$$

By expanding d in powers of s/R one has:

$$d \simeq s\left[1 - \frac{1}{24}\left(\frac{s}{R}\right)^2\right] \tag{2.8}$$

which, after substitution in Eq. 2.6, yields:

$$\begin{aligned} E^b_{i,j} &\simeq \frac{q^2}{4\pi\epsilon_0\epsilon_r s}\frac{1}{1-\frac{1}{24}\left(\frac{s}{R}\right)^2}e^{-\frac{s}{\lambda_{DH}}\left[1-\frac{1}{24}\left(\frac{s}{R}\right)^2\right]} \\ &= E^s_{i,j}+\frac{1}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2}s\left(1+\frac{s}{\lambda_{DH}}\right)e^{-s/\lambda_{DH}} \end{aligned} \tag{2.9}$$

Hence, the electrostatic energy difference between the two beads in the straight and bent conformation is approximated by:

$$\begin{aligned} \Delta E_{i,j}(s) &= E^b_{i,j}-E^s_{i,j} \\ &= \frac{1}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2}s\left(1+\frac{s}{\lambda_{DH}}\right)e^{-s/\lambda_{DH}} \end{aligned} \tag{2.10}$$

Next, to compute the electrostatic energy difference between the bent and the straight chain conformations we shall integrate over the possible values of i from 0 to L and the possible values of the arc length s at fixed i from 0 to $L-i$. When integrating we adopt a continuous description of the polymer chain and, for this reason, the pointwise charge q is substituted with a linear charge density q/σ.

We obtain:

$$\begin{aligned} \Delta E &\simeq \frac{1}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2\sigma^2}\int_0^L di\int_0^{L-i}s\left(1+\frac{s}{\lambda_{DH}}\right)e^{-s/\lambda_{DH}}ds \\ &= \frac{1}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2\sigma^2}\int_0^L di\left[-(L-i)^2e^{-\frac{L-i}{\lambda_{DH}}}-3\lambda_{DH}\left((L-i)e^{-\frac{L-i}{\lambda_{DH}}}+\lambda_{DH}e^{-\frac{L-i}{\lambda_{DH}}}-\lambda_{DH}\right)\right] \end{aligned}$$

Finally, to perform the integration over i, we substitute $k=L-i$ thus obtaining:

$$\begin{aligned} \Delta E &= \frac{1}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2\sigma^2}\int_0^L di\left[-(L-i)^2e^{-\frac{L-i}{\lambda_{DH}}}-3\lambda_{DH}\left((L-i)e^{-\frac{L-i}{\lambda_{DH}}}+\lambda_{DH}e^{-\frac{L-i}{\lambda_{DH}}}-\lambda_{DH}\right)\right] \\ &= \frac{L}{24}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2}\left(\frac{\lambda_{DH}}{\sigma}\right)^2\left[3-8\frac{\lambda_{DH}}{L}+\left(\frac{L}{\lambda_{DH}}+5+8\frac{\lambda_{DH}}{L}\right)e^{-\frac{L}{\lambda_{DH}}}\right] \end{aligned} \tag{2.11}$$

In the limit $L >> \lambda_{DH}$ the expression simplifies to:

$$\Delta E = \frac{L}{8}\frac{q^2}{4\pi\epsilon_0\epsilon_r R^2}\left(\frac{\lambda_{DH}}{\sigma}\right)^2 \tag{2.12}$$

It should be noted that ΔE depends linearly on L and quadratically on $1/R$ and, as such, it can be assimilated to an effective bending rigidity energy. Clearly, in this case, the origin of the bending rigidity is the electrostatic self-repulsion of the chain, not its

intrinsic mechanical resistance. By analogy with the definition of the mechanical persistence length (see section 1.1), one can thus define an effective electrostatic persistence length, as first done by Odijk, Skolnick and Fixman [66, 67]:

$$l_c = \frac{1}{4}\frac{q^2}{4\pi\epsilon_0\epsilon_r\,\kappa_B T}\left(\frac{\lambda_{DH}}{\sigma}\right)^2 \tag{2.13}$$

The effect of the electrostatic persistence length on the size of knotted regions has been investigated in ref. [50] for untensioned rings. In particular, in the regime $4l_c > l_p$,[1] the effective chain rigidity properties are dominated and controlled by the electrostatic persistence length, l_c, over the mechanical one, l_p. As a result, the unknotted portions of the chains tend to become swollen and to shape the ring in an open conformation. This behaviour reverberates on the knotted region, which ends up spanning only a small portion of the chain. As a consequence knots are found to be tight at least as metastable states.

In the case of our dsDNA model, $4l_c$ was equal to 64nm and, as expected, was much larger than the screening length $\lambda_{DH} = 8$nm. Moreover, $4l_c$ was also larger, though comparable, than the intrinsic persistence length $l_p = 50$nm of the chain.
Both this effect and the mechanical tension applied at the termini (which in our study replaced the chain closure condition of ref. [50]) are arguably the causes for the tightness of the knots shown in Fig. 2.2a.

2.2.2 Free knot diffusion

The spread of the knot length distributions in Fig. 2.2c indicates that, although being relatively tight, the knots are still capable of fluctuating substantially in size. In fact, in thermal equilibrium, they can also sustain stochastic longitudinal displacements along the chain contour as it is illustrated in Fig. 2.2d which portrays time-evolution of the the mean square displacement of the center of the knotted region. The fitting lines in Fig. 2.2d show that the mean square displacement has an overall linear dependence on time. Therefore, although knots fluctuate in size and were subject to the three-dimensional fluctuations of the chain, their motion along the contour is compatible with a standard one-dimensional diffusive process.

The corresponding diffusion coefficients, D, are provided in Table 2.1 along with the average knot length, and self-diffusion time.

[1] In the paper in ref. [50] the calculation of l_c was done by dropping the constant term 1/4 that is present in equation 2.13 as well as in the the original papers by Odijk and Skolnick-Fixman [66, 67]. As a consequence, the condition to determinate the regime in which the electrostatic persistence length prevails over the mechanical one involves $4l_c$ and not l_c itself.

Knot type	$\langle l_{knot} \rangle$ nm	D μm^2/s	Self diffusion time ms
3_1	86.415 ± 0.055	4.19 ± 0.81	0.89 ± 0.17
4_1	127.415 ± 0.094	1.77 ± 0.31	4.57 ± 0.80
5_1	143.970 ± 0.059	2.21 ± 0.27	4.68 ± 0.58
5_2	156.60 ± 0.15	1.70 ± 0.37	7.2 ± 1.6
7_1	202.139 ± 0.058	1.32 ± 0.10	15.5 ± 1.1

Table 2.1: Properties of knots in tensioned DNA chains at zero external electric field - Average knot length, diffusion coefficient and self diffusion time for various knot types in the model DNA chains tensioned by a stretching force of 10.0pN and without an external field. The self-diffusion time is the time required by the knot to diffuse along the chain by a distance equal to the average knot length.

One observes that D decreases with the increasing complexity of torus knots (3_1, 5_1 and 7_1) and that the diffusion coefficients of 5_1 and 5_2 knots are comparable. These properties are consistent with earlier results of Makarov [61] on knotted semi-flexible tensioned chains (without the electrostatic self-repulsion considered here).

To compare this results with previous investigations, we point out that the values of D reported in Table 2.1 are between 2 and 3 times larger than those reported in experiments on much longer dsDNA filaments (from 40 to 100μm) in a buffer with polyethylene glycol and stretched by smaller forces (from 0.1 to 2.0pN) [3]. Given the coarse-grained description of the system this is a fairly reasonable agreement with the experimental findings. Also, the diffusion coefficients are between a third and a half of those computed numerically for longer DNA chains treated with a different model (chains of cylinders) which further accounted for hydrodynamic effects [59].

2.3 DC field

We proceeded by studying the behavior of 3_1, 4_1, 5_1, 5_2 and 7_1 knots in the stretched model DNA chain after introducing a uniform DC electric field of magnitude E. In this situation, the stretching of the chain cannot be achieved by simply applying a mechanical force to the termini, as the chain would drift following the electric force. To prevent the global drift, we instead fixed the termini at a z separation corresponding to a nominal tension of $\sim$ 10.0pN in zero field.

The field was collinear to the stretching direction, and its magnitude was set equal to $E \simeq 20$, 40 and 80V/cm. The corresponding total electrostatic forces acting on the chain were 5.0, 10.0 and 20.0pN, i.e. comparable with the chain tension in zero field.

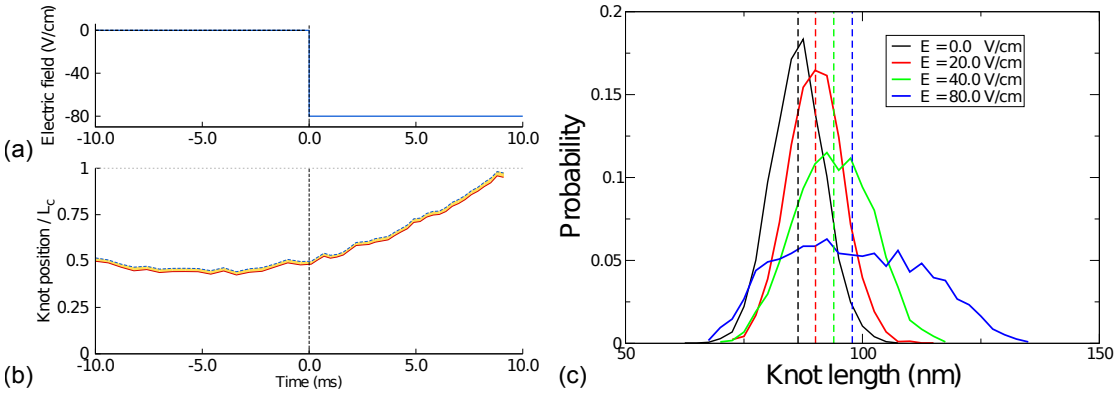


Figure 2.4: Effect of a DC electric field on a trefoil knotted tensioned DNA chain - Upon switching an external field of magnitude 80V/cm at time $t = 0$, see panel (a), the knotted region slided along the chain contour in the direction of the electric force with a definite average drift velocity, see panel (b). As it is shown in the knot length distributions in panel (c), the introduction of the field mildly affected the average knot lengths, which are indicated by the dashed lines.

We found that knots were only mildly affected in their length by the electric field action and acquired a systematic drift in the direction of the electric force. Both aspects are illustrated in Fig. 2.4 for trefoil knots.

The dependence of the knot drift velocity, v_{drift}, on the field strength and knot type is reported in Fig. 2.5a. The results are noteworthy in two respects.

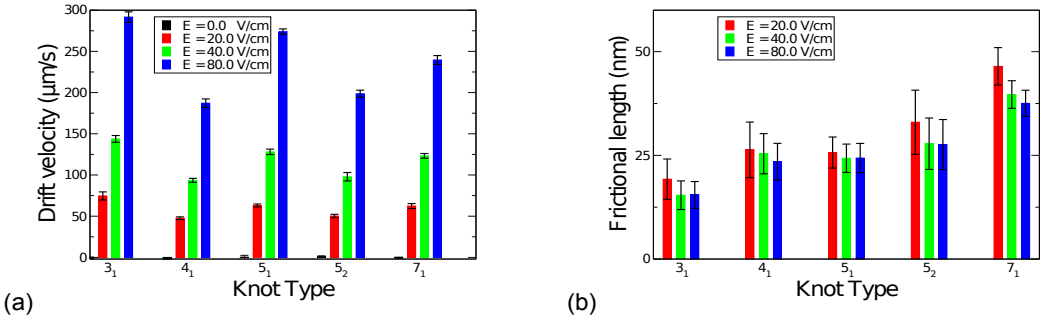


Figure 2.5: Drift velocity (a) and frictional length (b) for various knot types and DC field strengths - The drift velocity was computed from the displacements of the knot center along the oriented contour at time lags of 0.1ms. This time lag was chosen because it is much smaller than the self-diffusion time of any considered knot type.

First, one observes that, for any given value of the field, the drift velocities were similar across the various knot types, although torus knots (3_1, 5_1 and 7_1) velocities were systematically larger than those of twist knots (4_1 and 5_2). This effect was reminiscent of what is commonly observed for macroscopic threads and ropes where twist knots are employed as stoppers because of their enhanced hindrance to sliding compared to torus knots.

The second interesting aspect regards the possibility to rely on the drift velocity and the diffusion coefficient in zero field, D, to infer an effective "frictional length", l_{friction} associated with the various knot types.

In fact, by modelling the stochastic motion of the knot along the chain as a one-dimensional Langevin process, one can relate the drift velocity to the effective electrostatic force acting on the knot and the underlying friction coefficient of the process:

$$v_{\text{drift}} = qEl_{\text{friction}} \frac{D}{\kappa_B T} \frac{1}{\sigma} \,. \tag{2.14}$$

It should be noted that *a priori*, the effective frictional length needs not to coincide with the knot length, l_{knot}, which measures the extension of the minimal region that (upon closure) has a definite topological state.

The value of l_{friction} obtained by inverting eq. 2.14 are shown in Fig. 2.5b. For each knot type, the frictional length was practically independent of the field strength. This is consistent with the fact that the average knot length l_{knot} was also rather constant across the explored values of E and hence the self-contact effects which were arguably responsible for the topological friction were expected to be largely independent of E too. The notable quantitative fact emerging from Fig. 2.5b is that l_{friction} was significantly smaller than l_{knot}; typically by a factor of about 4. This indicates that only a fraction of the knotted region were effectively responsible for the topological friction.

Based on this, we believe it would be most interesting to probe l_{friction} experimentally (since both v_{drift} and D can be obtained from optical measurement) as it can provide a valuable information about key knot properties which would otherwise not be directly accessible experimentally (such as the knot length, l_{knot}).

2.4 AC field

As a complement of the DC field analysis, we further characterized the sliding motion of the knots in the presence of longitudinal AC field. The electric field was modulated as a square wave with amplitude equal to 80V/cm. In this way, the modulus of the applied field was, at all times, equal to the largest one used in the DC cases.

The effect of the AC field on the knot motion is illustrated in Fig. 2.6 which shows the knot position after the AC field was switched on at time $t = 0$. As it is apparent from the figure, the knot motion presented noticeable fluctuations during the trajectory, which indicates *a posteriori* that the applied field amplitude was not large enough to obliterate the stochastic motion.

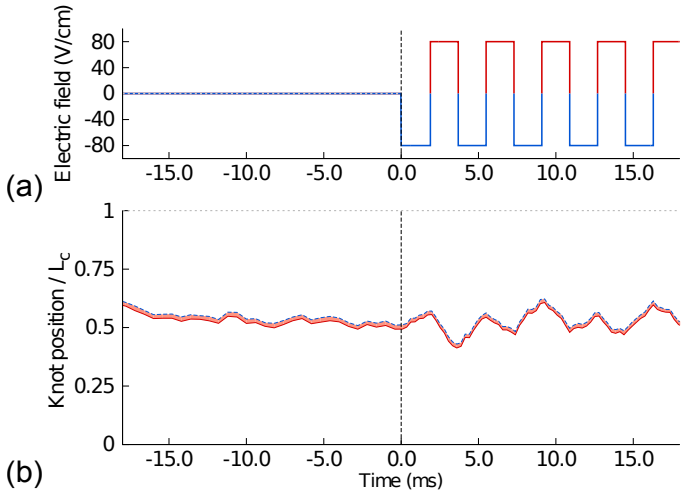


Figure 2.6: Effect of an AC electric field on a knotted tensioned DNA chain - Upon switching on an AC electric field (magnitude: 80V/cm, period: 3.6ms) at time $t = 0$, see panel (a), the knotted region moved stochastically along the chain contour responding to the oscillations of the dragging field, see panel (b).

Yet, even for this value of the field, which was chosen because it generates a force compatible with the one used in optical tweezers experiments, the AC field action affected the knot motion in a detectable way, to the point that one can observe a striking interplay of the AC driving period and the intrinsic timescale of the knot motion.

These effects were aptly characterized by a suitable analysis of the knot velocity (identified with the velocity of its central point), and particularly the deviations of its average from zero reference value in no field. Accordingly, we computed the average knot velocity at various stages of the AC driving period. The latter was quantified through the cycle index, $I(t)$,

$$I(t) = -\frac{1}{E_0}\frac{2}{\mathrm{T}}\int_0^t dt' E(t') \tag{2.15}$$

where E_0 and T are, respectively, the amplitude and the period of the electric field square wave whose instantaneous value at time t' is indicated with $E(t')$. At time $t = 0$ the field is at the beginning of the half-period with value -80V/cm and hence the $I(t)$ index varied periodically and without discontinuities in the [0:1] range during the AC cycle, following a saw-tooth profile.

The average knot drift velocity at various stages of the cycle is shown in Figure 2.7 for six different driving periods. At the longest period the v_{drift}-I curves show that, albeit with an initial delay, the knot velocity could adjust to the field reversal and attain its steady-state limiting value, see panel a. Reaching such steady state was more difficult as the period was progressively shortened, panels b–d), until it became impossible if the driving period was too short, see panel e and f.

It is readily noticed from Fig. 2.7 that the area enclosed by the closed curve, which is proportional to the work done by the field on the knot, is sizeable for the 3.6ms period and reduces to zero for the 0.225ms one. This difference is partly due to the fact that longer periods involved power dissipation over longer times.

To account for this effect we computed the dissipated power as the time-average of the product of the drift velocity and the applied field. Given the substantial stretching of the chain, we considered for simplicity the drift velocity along the contour in place of its projection on the stretching/field direction. Finally, for each knot we normalised the power by its value at the largest period ($\sim$ 3.6ms).

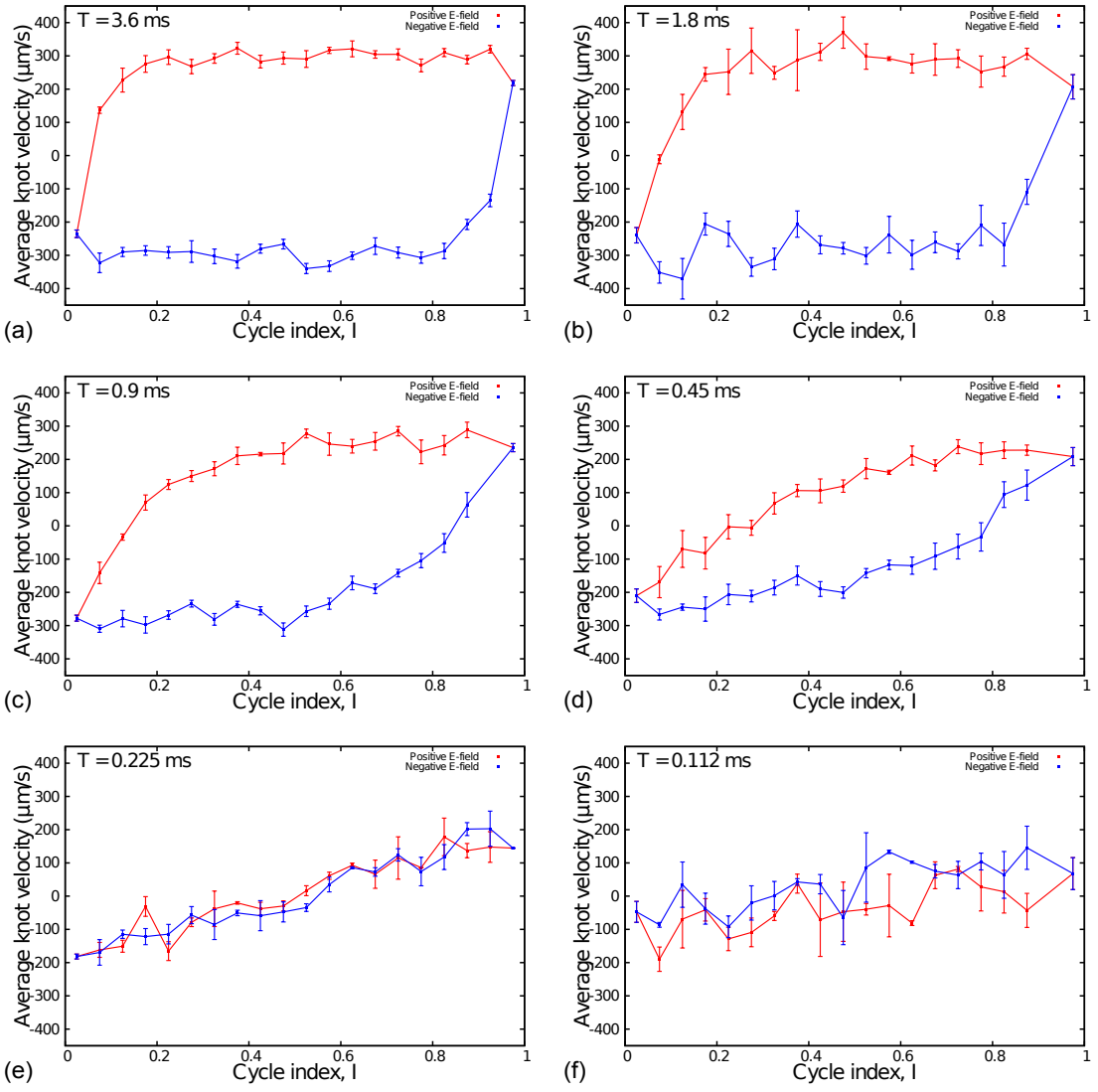


Figure 2.7: Average drift velocities of a trefoil knot in a tension DNA chain at various stages of the square-wave AC field cycle - The magnitude of the external field is 80V/cm, as in Fig 2.6a. Panels (a-f) refer to AC periods of 3.6ms, 1.8ms, 0.9ms, 0.45ms, 0.225ms and 0.112ms, respectively.

The normalised dissipated power is shown in Fig. 2.8 for 3_1 and 4_1 knots as a function of the AC period. One observes that both curves have a striking downward trend for decreasing AC period. The dissipated power is practically zero for periods smaller than about 0.25ms for both 3_1 and 4_1 knots.

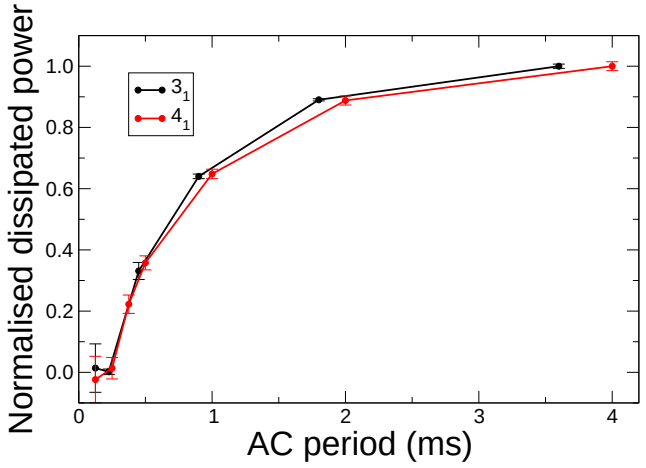


Figure 2.8: AC dissipation for 3_1 and 4_1 knots in AC electric fields - The dissipated power for 3_1 and 4_1 knots in tensioned DNA chains was computed as a function of the AC electric field period and, for ease of interpretation was normalised to the dissipated power at the slowest period, $T = 3.6$ms. Each data point was averaged over at least 100 AC cycles collected over different initial conditions.

The associated half period (during which the field acted consistently, before its directionality was reverse) of 0.5ms provides an intrinsic relaxation time for the knots. In fact, only for longer driving periods, knots can adjust and follow the repeated changes of directionality. By comparison with the data in Table 2.1 it is noticed that for both 3_1 and 4_1 knots the relaxation timescale was a fraction of the self-diffusion time (which bounds it from above since it provides a timescale over which the knot position along the chain is completely renewed).

2.5 Simulations with explicit counterions

The results presented so far were obtained for a model system where the action of counterions in solution was implicitly taken into account through the Debye-Hückel screened electrostatics.

To confirm the generality and robustness of the observed phenomenology we repeated the analysis for knotted polyelectrolyte chains in the presence of explicit counterions. To

keep the computational cost at a manageable level we considered chains parametrised as before except for the chain length and bead charge which were reduced to $N = 150$ beads and $q = -3e$, respectively. Each chain was placed within a periodic box with sides equal to 37.5, 37.5 and 375.0nm. To ensure the overall system neutrality, we added to the system 450 monovalent counter-ions with same size and mass as the chain beads, corresponding to a nominal salt concentration of 1.4mM and $\lambda_{DH} \sim 8$nm. All charges in the system, except the nearest neighbours along the chain, interacted via the standard unscreened Coulomb interaction.

Note that, similarly to the DNA case, the electrostatic persistence length [66] $4l_c = \frac{q^2 \lambda_{DH}^2}{4\pi\epsilon_0\epsilon_r \kappa_B T \sigma^2}$ (see Eq. 2.13) was equal to 64nm. The latter was again larger than the intrinsic persistence length $l_p = 50$nm and, according to the criteria of Dommersnes *et al.* [50], it was expected to have tight knots (at least as metastable states) also in the absence of a stretching force.

The model system was first studied by characterizing the free motion of 3_1 knots along the chains in zero electric field with their termini fixed at a separation of 325nm, corresponding to the same nominal tension of 10.0pN applied to the model DNA. The average knot length and self-diffusion time were found to be respectively equal to 35nm and 0.3ms.

The DC response was studied by introducing a longitudinal electric field producing a total dragging force of 20.0pN (again equal to that considered for DNA). Despite the reduced beads charge and the backflow of the counterions, the knot was found to acquire a net motion in the direction of the electric force, as for the DNA system without counterions.

The accord with earlier findings extended to the AC behaviour as well, as shown in Figure 2.9. In particular, it was seen that the dissipated power versus driving period had a trend analogous to the one of Figure 2.8. In absolute terms, the characteristic cutoff time for this system with reduced bead charges and explicit counterions was equal to 0.01ms and hence substantially smaller than for the model DNA case (0.25ms). In relative terms, the 0.01ms relaxation time was in this case also appreciably smaller than the upper bound provided by the self-diffusion time (0.3ms).

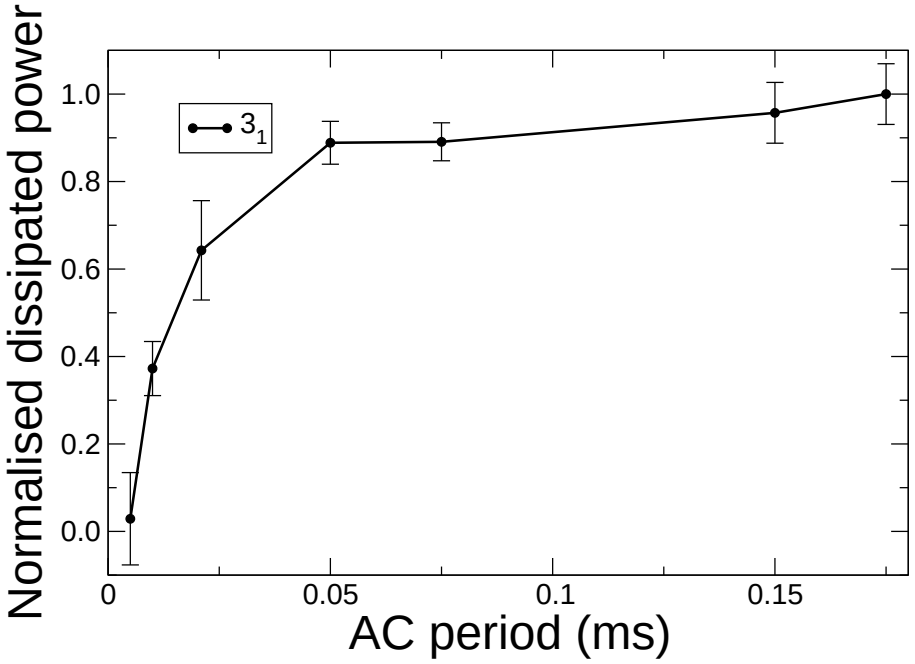


Figure 2.9: AC dissipation curves for a trefoil knot in a model tensioned polyelectrolyte chain with explicit counterions - The dissipated power for the 3_1 knot in tensioned DNA chains was computed for AC electric field with different period and, for ease of interpretation was normalised to the dissipated power at the slowest period. Each data point was averaged over at least 100 AC cycles collected over different initial conditions. The different time-scale with respect to Figure 2.8 is due to the change in the parameters of the model chain.

2.6 Summary

The effect of DC and AC electric fields on the dynamics of entangled polymers was investigated theoretically and numerically for mechanically-stretched polyelectrolytes. Specifically, we considered different physical knot types 3_1, 4_1, 5_1, 5_2 and 7_1 tied in model DNA chains subjected to mechanical tension applied at the two termini and to the action of an external electric field.

The system is interesting *per se* because of the interplay of several competing effects on the motion of the knotted region. These include the bending forces and electrostatic self-repulsion of the chain which may lead the knot to tighten or to swell depending to the ratio of the "mechanical" and electrostatic persistence lengths, as demonstrated in ref. [50], and the tension applied to the chain. Our study investigated in particular the tight knots regime.

We found that all considered knot types, including twist ones, maintained their ability to slide along the chain under an external electric field. In particular, knots were dragged in the direction of the electric force with a field-dependent drift velocity.

By measuring the friction associated to the drift velocity one could define a previously unexplored characteristic knot length scale, the frictional length. This length, which could also be computed from experimental measurements, was smaller than the nominal knot length since it captures the extent of self-interactions within the knotted region.

The rich phenomenology of the dynamics of self-entangled polyelectrolytes was aptly exposed by applying AC fields of various frequencies. In particular, for sufficiently low (and experimentally accessible) AC frequencies, the knot motion presented the typical signatures of dynamical systems with memory, such as a lag time in following the time-modulation of the external fields. For frequencies higher than a critical value, the knot became completely unable to follow the external field which did not do a net work on the knot. The transition from the dissipative regime to the "frozen" one can be rationalised in terms of a spontaneous, intrinsic relaxation time associated to the motion of the knots along the chain.

Chapter 3

Geometrical and topological entanglement in model polyelectrolyte chains

In the previous chapter, we studied several aspects related to the motion of physical knots along the contour of tensioned polyelectrolyte chains under external electric fields. We showed that knots acquire a systematic drift in the direction of the electric force and eventually escape from one end of the chain. Interestingly, the obtained results are largely independent on the description, implicit or explicit, of the ionic solution surrounding the chain.

Here, we shall discuss a further study we carried out on negatively-charged polyelectrolyte (PE) chains where we removed the constraints (tensile force or pinning) at the chain termini. The charged chains were studied in explicit ionic solution given by the dissociation of a trivalent salt. With this setup, we first studied the equilibrium properties of the polyelectrolytes and, next, their out-of-equilibrium dynamics under action of external AC/DC electric fields.

3.1 Equilibrium and out-of-equilibrium phenomenology of polyelectrolyte chains in explicit counterions

The conformation of a polyelectrolyte chain in equilibrium is strongly affected by the concentration and valency of counterions in solution. When the concentration of, say, monovalent counterions is too low to screen effectively the polyelectrolyte chard, the latter will self-repel and hence adopt a swollen conformation [68]. Upon increasing the

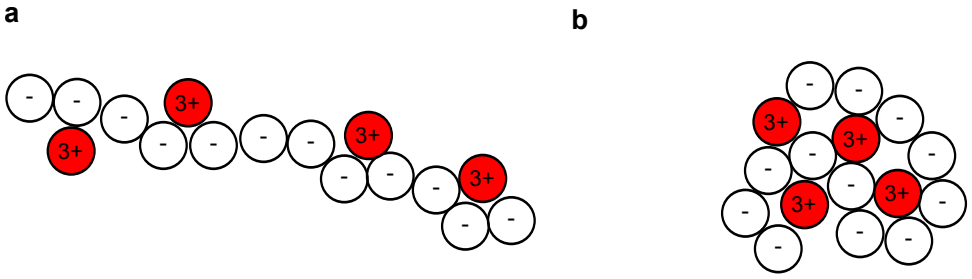


Figure 3.1: Illustration of the collapse of polyelectrolyte chain in trivalent salt solution. - (a) The trivalent counterions (*red beads*) are condensed on the monovalent charged beads of the chain. (b) The local overscreening of the chain promoted by the counterions favours the collapse of the chain.

concentration of the monovalent counterions, the latter will start to condense on the polyelectrolyte chain, thus reducing its effective charge density and, at the same time, they will screen the electrostatic self-interaction of the chain. The use of multivalent counterions instead of monovalent ones, leads to the interesting phenomenon of "over-screening" which amounts to inducing a self-attracts rather compact globular structures. This notable effect arises because the counterions act as "gluing" agent bridging various portions of the chain, see Figure 3.1.

This effect is strongly dependent on the short-range (local) attraction of the multivalent ions and the discrete charges on the polyelectrolyte chain and, hence, cannot be accounted for by mean-field theories and approximations, such as the Poisson-Boltzmann one [68, 69], which can otherwise be successfully used when dealing with e.g. monovalent counterions.

As a starting point for characterizing the properties of polyelectrolyte chains in ionic solutions, it is worth recalling that the strength of the electrostatic interaction with respect to thermal noise is captured by the Bjerrum length of the system:

$$l_B = e^2/(4\pi\epsilon_0\epsilon_r\kappa_B T) \tag{3.1}$$

This length scale, in fact, corresponds to the the spatial separation at which the interaction between two electronic charges is comparable in magnitude to the thermal energy $k_B T$.

In particular, in the context of polyelectrolyte chains modelled as chain of charged beads of radius σ (see Figure 3.1), it has been shown [6, 69] that if the ratio of the Bjerrum length l_B and the thickness of the polyelectrolyte chain, σ is larger than 1.8, the electrostatic effect dominates over the entropic one and provides the attractive energy for the formation of stable bundles. Accordingly, to study the phenomenology related to polyelectrolyte chains collapse one should consider a ratio l_B/σ larger than 1.8.

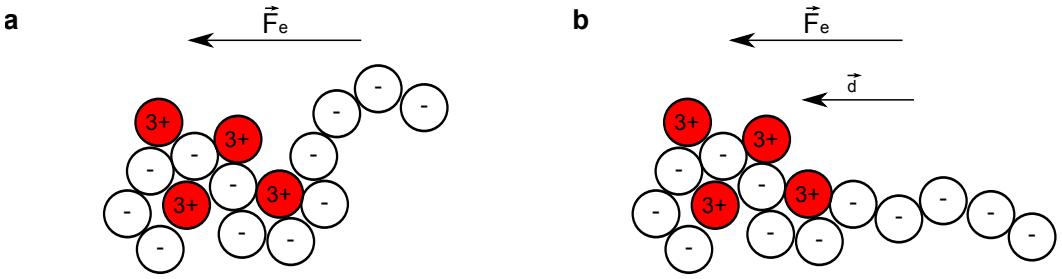


Figure 3.2: Illustration of polyelectrolyte elongation under external electric fields. - The external electric force $\vec{F}_E$ couples to the spontaneous dipole moment of the chain $\vec{d}$ leading to chain unfolding (a) and progressive elongation (b).

The out-of-equilibrium physical properties of polyelectrolytes interacting with counterions are, instead, strikingly represented by their capability to elongate along a uniform electric field. This phenomenon, which can be exploited in practical contexts such as DNA preconditioning for sequencing or molecular display experiments [70], was first predicted on a theoretical basis [5, 71–74]. Numerical simulations and theoretical arguments lead to formulate the hypothesis that a sufficiently strong electric field coupled to the spontaneous dipole moment of a collapsed polyelectrolyte (PE) can overcome the activation energy of longitudinal stretching modes leading to the progressive unfolding of the chain, see Figure 3.2.

This out-of-equilibrium effect has, indeed, been observed experimentally both in DC fields as well in AC ones of suitable frequency [51, 70].

A key feature of this effect is that the threshold strength of the critical field, E_c, decreases systematically with chain length, N. This is because longer chains can develop larger spontaneous dipoles and hence facilitate the activation of the soft longitudinal stretching modes. The expected monotonically-decreasing dependence of E_c with N has been recently investigated and characterized with numerical simulations for collapsed flexible PE chains of up to ~ 200 monomers both in DC and AC field [5, 71, 73, 74].

Here, we focus on one aspect that so far has not been previously considered, namely the impact of the spontaneous entanglement, and in particular knotting, of long PE chains on their capability to unfold in the external electric field both constant and time-dependent.

As a matter of fact, the emergence of knots has not been previously discussed for counterion-collapsed PE chains neither in this context nor in the more fundamental equilibrium situation. The possible scenarios opened by explicitly accounting for the spontaneous entanglement of PE chains are, *a priori*, most interesting both from the theoretical and the practical point of view. In fact, because the incidence of physical

knots grows rapidly both with chain length and degree of compactification, one may envisage that sufficiently long condensed PE chains can be highly entangled and knotted, and hence their unfolding can possibly be dynamically arrested at field strengths that are sufficiently strong to unfold unknotted ones.

To advance our understanding of this matter we first characterized the degree of self-entanglement of PE chains by studying the overall incidence of knots in PE chains and next considered how chains with different types of entanglement respond to DC fields.

3.2 Model details

We considered a single flexible polyelectrolyte chain of N beads. Each bead had a diameter σ, taken as the unit length, and carried a unitary electric charge $-e$. To account for the presence of trivalent salt at the equivalence point, we added to the system $N/3$ trivalent counterions and N monovalent co-ions. The latter, together with N additional monovalent counter-ions, guaranteed the system charge neutrality. The ions had the same diameter, σ, of the monomeric beads.

We studied, in particular, chains of length $N = 60$, 120, 240, 480 and 960 for a total of $M = 200$, 400, 800, 1600 and 3200 particles, respectively. The system was organized within a parallelepiped periodic simulation box. The latter had the z-side of length $N\sigma$, to allow the PE chain stretching, while each of the other two sides had length $100\,\sigma$ so to yield a PE monomer concentration of $10^{-4}\,\sigma^{-3}$ for all the values of N.

The interaction energies of the PE chain model were composed by the following terms:

$$\mathcal{H} = \mathcal{H}_{\mathrm{PE}} + U_{\mathrm{ext}} \tag{3.2}$$

The first term, $\mathcal{H}_{PE}$, whose definition was given in Section 1.2.2 of Chapter 1, accounts for the excluded volume interactions and chain connectivity of the PE chain. Because the chains are assumed to be fully flexible, we also have that the persistence length, l_p, is about equal to σ. The Lennard-Jones and the electrostatic energies described also the the ionic beads of the system. As anticipated we shall also set $l_B/\sigma = 1.8$ to promote the aggregation of the PE chain in the presence of the trivalent counterions [6].

The last term in Eq. 3.2, U_{ext}, described the interactions of the particles (both the PE beads and the ions) with a spatially-uniform, and possibly time-dependent, external field $E(t)$ directed along the $\hat{z}$ axes. The associated potential energy is:

$$U_{\mathrm{ext}} \quad = \quad \sum_{i=1}^{N} qE\,\hat{z}\cdot\vec{r}_i. \tag{3.3}$$

In our study we varied the magnitude of the electric field from 0.0 to a maximum of $0.22\epsilon/(e\sigma)$ in steps of $0.1\epsilon/(e\sigma)$, where ϵ is the unit energy of the system. The intensities of the field were chosen to be high enough to elongate the PE chains.

The system dynamics was described with a Langevin equation (see Section 1.3):

$$m\ddot{r}_{i\alpha} = -\gamma\dot{r}_{i\alpha} - \partial_{i\alpha}\mathcal{H} + \eta_{i\alpha}(t) \tag{3.4}$$

where m is the mass of the particles (the same for monomeric beads and ions), $i = 1, \ldots, M$ runs over the particles of the system and $\alpha = x, y, z$ is the index of the Cartesian component. The Gaussian white noise, $\eta_{i\alpha}(t)$, has the usual statistical properties $\langle\eta_{i\alpha}\rangle = 0$ and $\langle\eta(t)_{i\alpha}\,\eta_{j\beta}(t')\rangle = 2\kappa_B T\gamma\,\delta_{ij}\delta_{\alpha\beta}\,\delta(t-t')$, where δ_{ij} is the Kronecher delta, $\delta(t-t')$ is the Dirac delta, k_B the Boltzmann constant and T the temperature of the system (See Section 1.3). To time-integrate the equation of motion, we used the LAMMPS simulation package [29] with an integration time step $\Delta t = 0.012\,\tau_{LJ}$, where $\tau_{LJ} = \sigma\sqrt{m/\epsilon}$ is the Lennard-Jones time and the ratio m/γ is set equal to $10\,\tau_{LJ}$. Since in this chapter we present ongoing work, the results are presented in simulation units.

3.2.1 System preparation

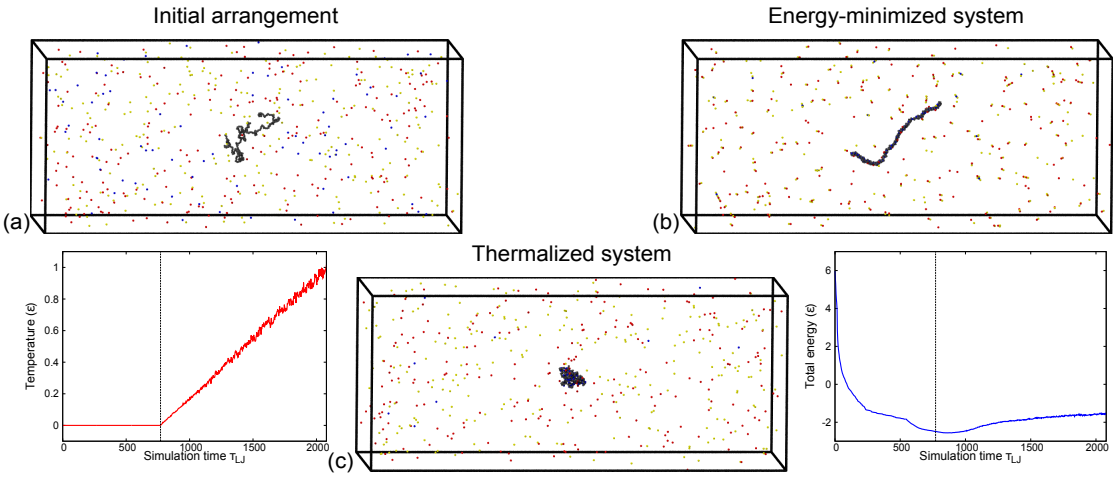


Figure 3.3: Illustration of the steps of the system preparation - A polyelectrolyte chain of $N = 240$ beads is shown in its simulation box along with $N/3$ trivalent counterions, N monovalent co-ions and N monovalent counterions (blue, yellow and red beads, respectively) at three different stages of the system preparation. The polyelectrolyte chain is colored in gray. (a) The PE was generated as an equilibrated Self-Avoiding Walk within the proper simulation box. The ionic particles were arranged in random non-overlapping positions. (b) The system was, next, driven to a minimum energy conformation with a minimization step to prevent unphysical fast condensation of the ions onto the polyelectrolyte chain. The minimization procedure was performed at zero temperature by using a standard Conjugated-Gradient algorithm until the interaction energy of the system reaches a minimum (indicated with a *dashed line*) in the inset. (c) Finally, the temperature of the system was linearly increased from 0 up to the final temperature $T = 1.0\,\epsilon$ (see inset) with a thermalization step.

The preparation of the model system was done via several steps to favour the relaxation of the polyelectrolyte chains to the equilibrium collapsed conformations. The strategy of the system preparation is described in the following and illustrated in Figure 3.3 for a chain of $N = 240$ beads.

The polyelectrolyte chain was initially generated as an equilibrated self avoiding walk (SAW) polymer fitting into the simulation box. Next, the surrounding ions were arranged within the box in random positions, avoiding overlaps with the other particles. A typical conformation obtained at this stage is shown in Figure 3.3a.

To control the condensation of the counter-ions on the polyelectrolyte chain, the system was, next, subjected to an energy minimization step at zero temperature. Specifically, the total energy of the system was minimized by a using a standard conjugated-gradient algorithm until the energy of the system reached a minimum value. The illustration of a typical configuration of the system at the minimum of the energy is shown in Figure 3.3b. The decreasing trend of the energy of the system is presented in the corresponding inset.

Finally, the temperature of the system was linearly increased up to the desired final temperature T, see inset of Figure 3.3c. This thermalization step was meant to favour the equilibration of the polyelectrolyte toward the [6] collapsed globular state. The latter is shown in Figure 3.3c.

3.3 Self-diffusion time

As a first step of our analysis, we estimated the self-diffusion time of the chains, τ_d, at equilibrium by using the standard diffusion equation in three-dimensions:

$$R_g^2 = 6D\tau_d \tag{3.5}$$

which relates τ_d to the mean square radius of gyration of the chain R_g^2, (see section 1.1). The latter is a standard and robust measure of the chain size and the diffusion coefficient of the center of mass of the chain, D. The self-diffusion time τ_d is, hence, the time required by the chain to diffuse in space by a distance equal to its size.

Specifically, R_g^2 was measured as an average over the trajectories at equilibrium and the diffusion coefficient D was obtained by fitting the curves of the mean square displacement as a function of time. In Figure 3.5, we show such curves with the respective linear fitting curves for chains of length 240 beads. The values of these quantities with their estimated errors are reported in second and third columns of Table 3.1.

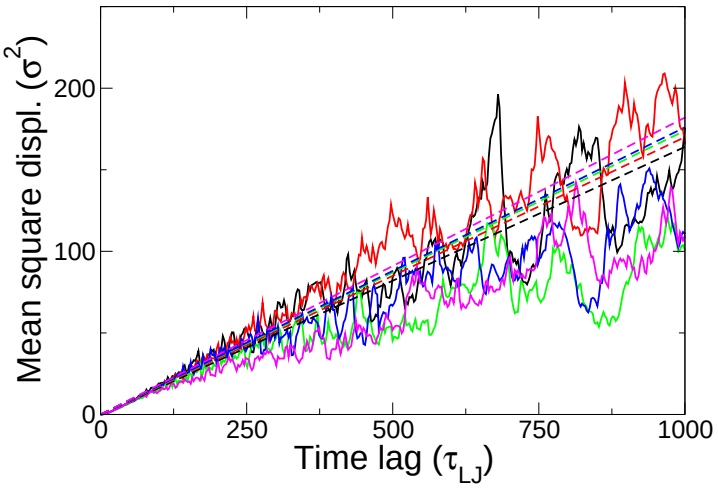


Figure 3.4: Diffusion at zero external electric field for chain of 240 **beads** - Time dependence of the mean-square displacement of the PE chains in space. The dashed lines indicate the linear fitting curves.

Chain length	$\langle R_g^2 \rangle$ σ^2	D σ^2/τ_{LJ}	Self diffusion time τ_{LJ}
60	19.7 ± 1.0	0.115 ± 0.0083	28.6 ± 6.5
120	29.1 ± 1.7	0.1733 ± 0.0041	83 ± 25
240	48.2 ± 3.4	0.0578 ± 0.00045	277 ± 30
480	57.0 ± 0.8	0.02911 ± 0.00028	651 ± 100
960	84.0 ± 0.7	0.007473 ± 0.000073	1873 ± 210

Table 3.1: Properties of the polyelectrolyte chains at zero external electric field - Equilibrium mean square radius of gyration R_g^2, diffusion coefficient D and self diffusion time τ_d are shown for various chain length. To estimate these quantities, we collect data over 5 independent simulations at equilibrium for each chain length N. Next, we compute 5 average values (one per each run) of the mean square radius of gyration R_g^2. The error on this measure is standard deviation of the 5 values of the replicas. The diffusion coefficient was obtained by fitting the curves of the mean square displacement as a function of time, see Figure 3.5. Finally, the self-diffusion time was obtained as $\tau_s = R_g^2/6D$ and its error was computed by using the standard expression for error propagation.

The values of the self-diffusion times, calculated from Equation 3.5 are reported in Table 3.1 plotted in Figure 3.3 as a function of the chain contour length N. The monotonic increase of R^2 with N is shown by the fitting curve.

We point out that, *a posteriori*, both the energy minimization and the thermalization steps have been carried on for time spans much longer than the chain self-diffusion time for all the values of the chain length, see for instance the insets in Figure 3.3 for the chain length $N = 240$ beads.

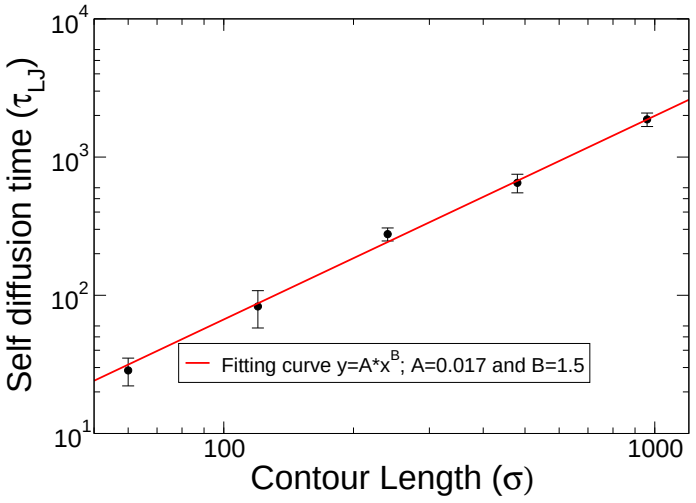


Figure 3.5: Self-diffusion time as a function of the chain length - The Self-diffusion times are shown here for various knot lengths together with the fitting curve. The self-diffusion time increases with the length, N, of the chain as $\sim N^{1.5}$.

3.4 Spontaneous chain knotting in polyelectrolyte chains

3.4.1 Knotting probability

As a first characterization of the topological entanglement that can spontaneously arise in the estimated the equilibrium knotting probability of the chains.

Specifically, we collect simulations over time spans much longer than the chain self-diffusion time and next searched for the presence and location of physical knots.

To look for knots in the simulated PE chains we adopted the method described in ref. [64]. The strategy involves a bottom-up search of the knotted region. Specifically, we starts by considering all possible chain segments involving as few as 5 beads and check whether whether any of them accommodates a physical knot. If no knot is found, the search is expanded to segments of six beads and so on, until one identifies the shortest knotted region. We note that, in order to detect whether a linear portion of the chain accommodates a physical knot, it is necessary to close it into a circle, so to trap its entanglement in proper topological state. For this purpose we used the minimally-interfering closure scheme of ref. [64].

The results of the knot identification analysis are shown in Figure 3.6 as the percentage of knotted conformation for all the chain lengths.

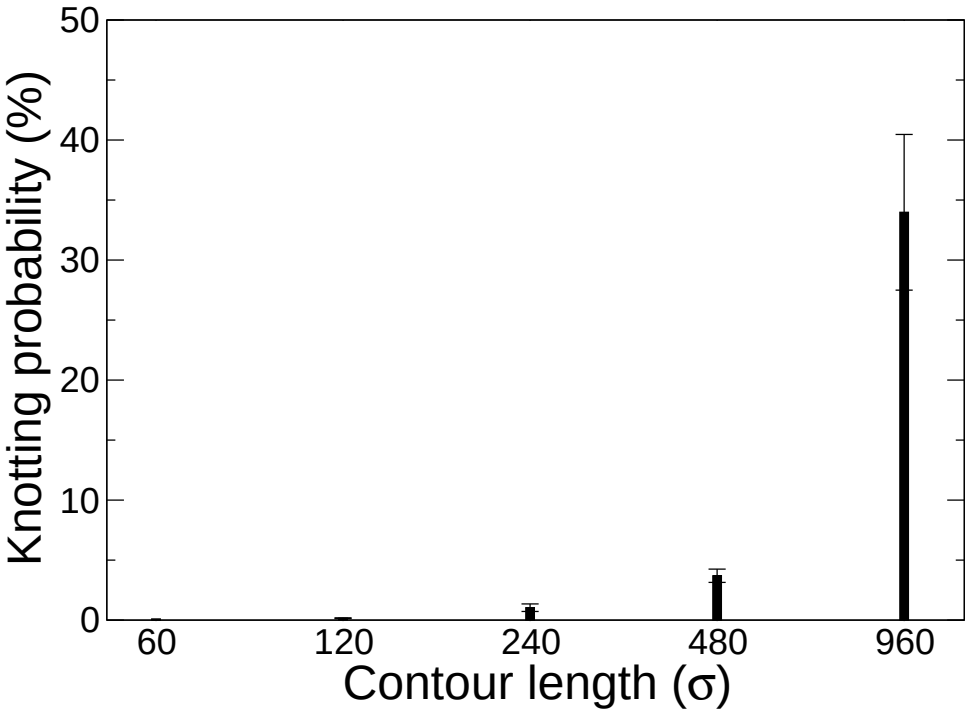


Figure 3.6: Percentage of the spontaneous occurrence of various types of physical knots as a function of the PE chain length - The fraction of spontaneously knotted conformations for model PE chains of increasing length. Each point on the graph reports on the average and the standard error over 5 measures from independent runs.

It is important to notice that the knotting probability is essentially negligible up to the chain length $N_0 = 240$, which corresponds to about the chain length considered in previous numerical studies of PE unfolding in electric fields [5, 71, 73, 74]. Therefore, we can argue that in the previous investigations the incidence of knots was negligible and poorly affecting the equilibrium and out-of-equilibrium behaviour of the polyelectrolyte chain.

Beyond N_0, the knotting probability increases very rapidly, up to reach 30% for the longest chains considered. This reflects in the fact that, for lengths $N = 960$ beads the incidence of topological knots is so high that one out of 3 chain conformations is knotted. We can therefore expect that, at least for the longest chains considered here, the geometrical and topological entanglement of the chains can affect their dynamical behaviour.

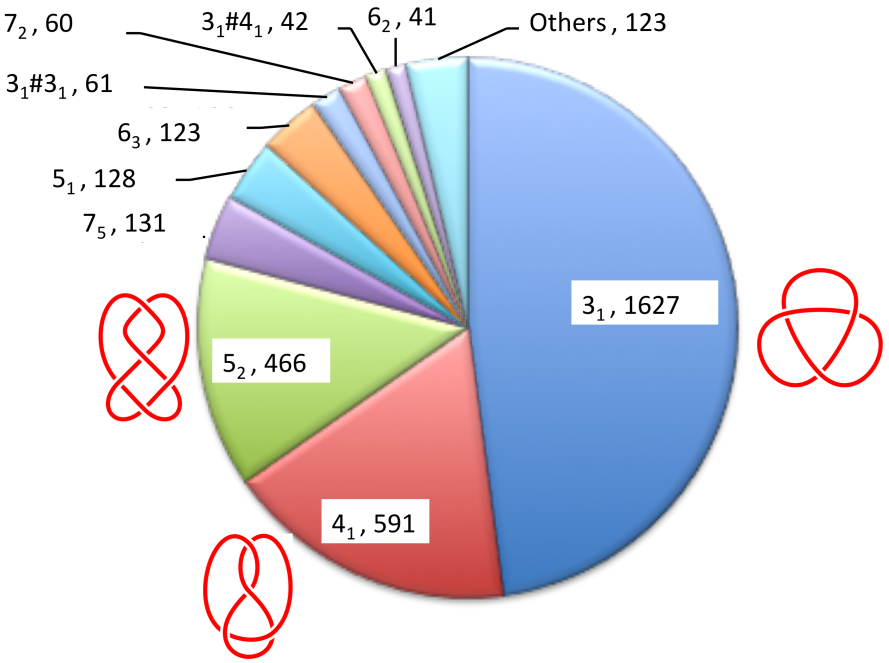


Figure 3.7: Knot types distribution for chains of length 960 beads - Probability of occurrence of the first types of knots in polyelectrolyte chain of $N = 960$ beads. Prime knots are labeled with the standard nomenclature as in Figure 2.1. The composite knots ($3_1\#3_1$ and $3_1\#4_1$) are knotted chains which host more than one prime knot. Accordingly, they are described in terms of the constitutive prime knots. Thus, the notation $3_1\#3_1$ denotes a knot built from two trefoil knots (3_1). Twist knots prevail, as evident from the predominance of 3_1, 4_1 and 5_2 knots.

3.4.2 Characterizing knot complexity

To study the complexity of the spontaneously knotted conformations, we next looked at the distribution of the populated non-trivial topologies for the longest chains considered, $N = 960$ beads.

The distribution of the 10 knot types with the highest occurrence is shown in the pie chart of Figure 3.7 together with the knot diagrams for the first three knot types. The latter are, in particular, the knots with minor complexity: the trefoil knot (3_1), the figure-of-eight knot (4_1) and the 5_2 knot. It is interesting to notice that they are three instances of twist knots. In particular, we note that the 5_2 twist knot has a 4-fold higher incidence than its torus counterpart with the same nominal complexity, i.e. the 5_1 knot.

Twist knots are a knot family that can be simply obtained by repeatedly twisting a closed ring and then linking the ends together with a single strand passage. The steps to tying the simplest twist knot 3_1 are shown in Figure 3.8. By reversing the procedure it is clear that twist knots can be untied by a single, suitable chosen strand passage.

Torus knots are, instead, kind of knots that can be drawn as continuous closed curves that run on the surface of a torus without self intersecting, see Figure 3.9. Notice that

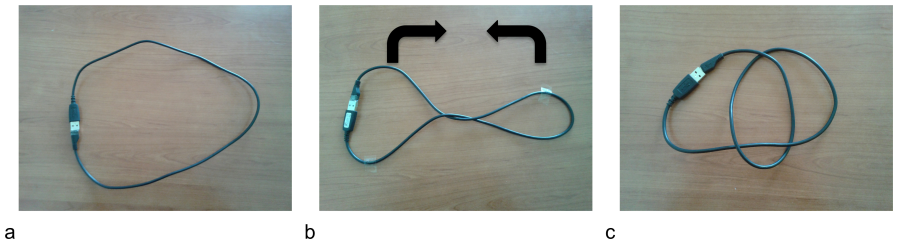


Figure 3.8: Illustration of the general mechanism to tie twist knots - Twist knots can be tied by taking a ring (a), twisting it keeping two loops (b) and clamping together the two loops (c). In this figure we show the case for a trefoil 3_1 knot.

Figure 3.9: Illustration of the 5_1 torus knot - The torus knots can be drawn as a non-intersecting continuous closed curve on the surface of a torus.

the simplest knot, the 3_1 or trefoil knot, is both a torus and a twist knot, a property that does not hold for any other torus knot. In particular, it should be noted that, except for 3_1, more than one strand passage is always required to untie torus knots.

We observe a systematically higher incidence of twist knots over torus knots. This fact, is consistent with what is typically found for both unconstrained and spatially-confined flexible (uncharged) polymer chains [1, 31, 38].

The knot distribution also features several *composite* knots, formed by the concatenation of several prime knotted components. Specifically, in our case composite topologies are represented by the $3_1\#3_1$ and the $3_1\#4_1$ knots. Notice that both of the prime components involved in composite knots are twist knots. Again this confirms the abundance of these knot types.

3.5 DC electric field

3.5.1 Validation in DC electric field

After characterizing the impact of entanglement, in particular knots, on free polyelectrolyte chains, we subjected the charged polymers to the action of an external DC electric field. With this setup, we aim to reproduce the known phenomenon of elongation of PE

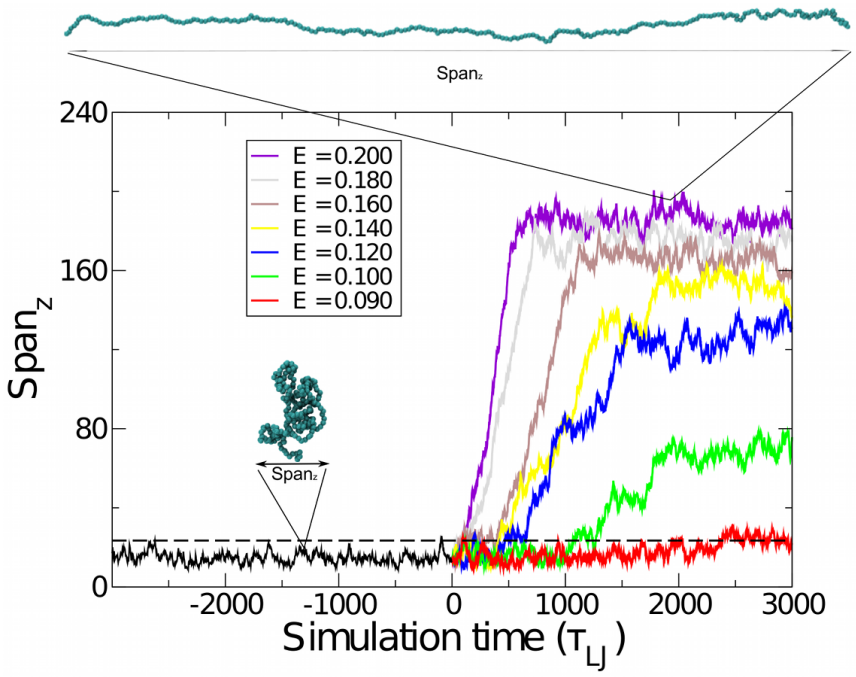


Figure 3.10: Elongation curves of PE chains of $N = 240$ beads induced with DC external electric fields - After equilibration, we attempted the elongation of the PE chains by applying external electric DC electric fields along the $\hat{z}$ direction. The intensity of the electric field is decreased gradually, as indicated in the legend. The critical value of electric field is defined as the magnitude at which, during the simulation run (spanning ~ 10 self-diffusion times), the external field is not able to induce an elongation significantly larger than the typical one at equilibrium. In this plot the dashed line is the average elongation at zero field and the red curve corresponds to the estimated critical electric field of $E = 0.090\epsilon/(e\sigma)$. Typical snapshots of the equilibrated globular conformation and the totally elongated one are shown as insets.

chains under electric fields summarised in section 3.1. This validation step is also instrumental to estimating the critical amplitude of the electric field for which the onset of elongation is observed in out model chains.

As a practical criterion to identify the critical electric field, we adopt the strategy illustrated in Figure 3.10, for a chain of length $N = 240$ beads.

Specifically, we collect simulations of the free PE chains over time spans much longer than the chain self-diffusion time to estimate the average value of the z projection, $\langle span_z \rangle$ of the polyelectrolyte chain.

The value $\langle span_z \rangle$ is, next, used as an elongation threshold for the study in external electric field, see dashed line in figure 3.10. In fact, we started by applying electric fields of magnitude $\sim 0.2\epsilon/(e\sigma)$, which is certainly sufficiently strong to elongate the chain. Next we progressively reduce the field magnitude in steps of $0.01\epsilon/(e\sigma)$ until we reach from above the critical electric field, E_c, for which no appreciable elongation of the chain, i.e. larger than $\langle span_z \rangle$, is observed over time spans as long as $10\tau_d$.

For the largest values of the field, the PE chain is completely elongated reaching about $\sim 80\%$ of its contour length. By decreasing the electric field the elongation of the chain at the end of the DC runs is reduced to smaller values: for instance it is equal to 50% for $E = 0.12\epsilon/(e\sigma)$ and to 30% for $E_{DC} = 0.09\epsilon/(e\sigma)$. The critical electric field in this case is $E = 0.09\epsilon/(e\sigma)$, because such a filed is not able to elongate the polyelectrolyte chain more than 10%, that is the average value of $\langle Span_z \rangle$ at equilibrium.

By applying the same procedure to all system replicas, we obtained the average critical electric fields as a function of the chain length. In Figure 3.11, we show that the values of critical electric field as a function of the length of the chain are fitted by the power-low decay $N^{-0.574}$, which is fairly compatible with the $N^{-0.5}$ decay expected from theoretical arguments and numerical simulations in refs. [5, 72].

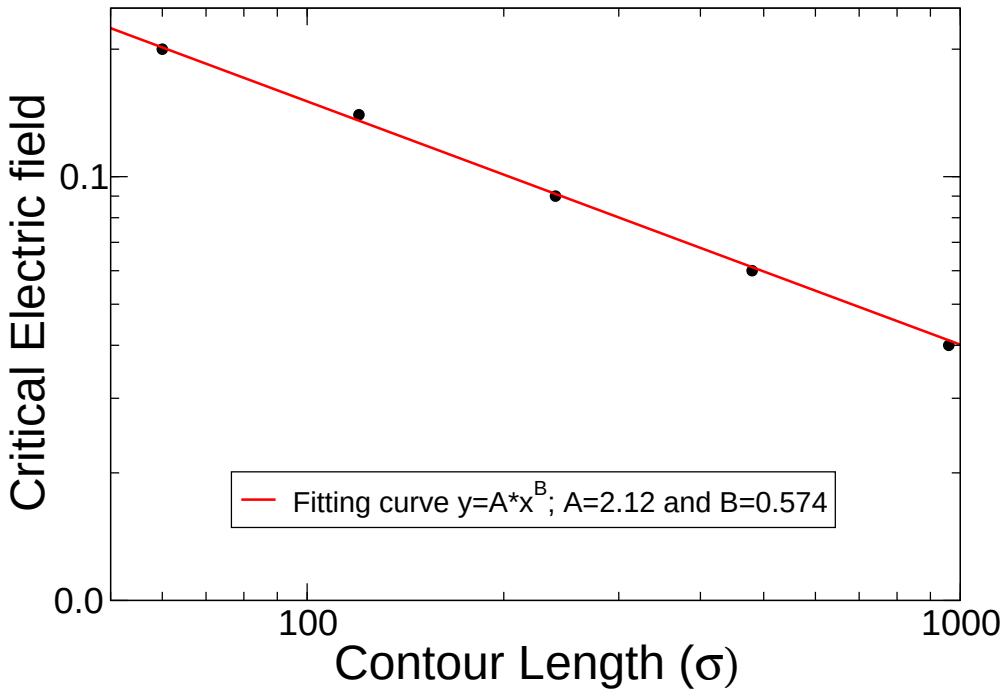


Figure 3.11: Critical values of the elongating electric field as a function of the chain length and the respective fitting curve.

It should be noted that the latter theoretical result was obtained by extrapolating for increasing chain lengths the properties of short chains up to about 200 monomers, which are clearly free of knots. Here we confirm that a similar trend is retained also for longer chains, which spontaneously develop an higher degree of entanglement, as shown in Section 3.4.

This still leaves open the question of whether the entanglement, which inevitably originates for longer chains, affects the kinetics of chain elongation.

3.5.2 Kinetics of unravelling of polyelectrolyte chains in DC electric fields

To analyze this point, we considered two instances, A and B, of chains $N = 960$ beads and studied their dynamical response to a DC electric field as a function of the simulation time. Specifically, we used an electric field of strength $0.080\epsilon/(e\sigma)$ to elongate the equilibrated PE chains. This strength of the field was chosen because it is twice larger than the critical electric field for $N = 960$ ($E_c = 0.040\epsilon/(e\sigma)$).

In the upper part of Figure 3.12 we show the elongation profile ($span_z$) for the two instances (panels a and b). The vertical lines, which indicate the time intervals where the chain is physically knotted, reveal that both chains remains unknotted almost for the entire time span of the simulations.

However, the difference in the kinetics of unravelling of the two chains is, strikingly illustrated by the behaviour of the $span_z$ as a function of time. In particular, in case A the polyelectrolytes elongates about 70% of its total extension in about $10,000\tau_{LJ}$, which corresponds to about $6\tau_d$. In case B the charged chain is, instead, unraveled after the much longer time-span of $\sim 36\tau_d$ which is $6 - times$ larger than the one in case A.

The much larger unravelling time in case B is mainly due to a long-lived meta-stable state where the chain is only partially elongated at about 25% of its total contour. To further illustrate where this slowing down of the chain elongation originates, four snapshots of the elongation trajectories are shown in panels c and d of Figure 3.12 for each of the two elongation trajectories. The snapshots are taken at regular time intervals of $20,000\tau_{LJ}$.

We point out that in both cases A and B, the unravelling mechanism of the charged chain is promoted by the migration of the chain (as a whole) in the direction of the electric force.

In case A, the unfolding dynamics originates from with a single end of the chain, which protrudes from the globular structure of the chain and starts to unravel the polyelectrolyte. The elongation process carries on with the progressive realising of larger and larger portions of the chain from the globule. The chain is, finally, elongated to about $\sim 70\%$ of its contour.

In case B, instead, in the initial stage of the elongation process a chain loop protrudes out of the chain globule and starts to unfold (see panel d). Arguably, the loop formation

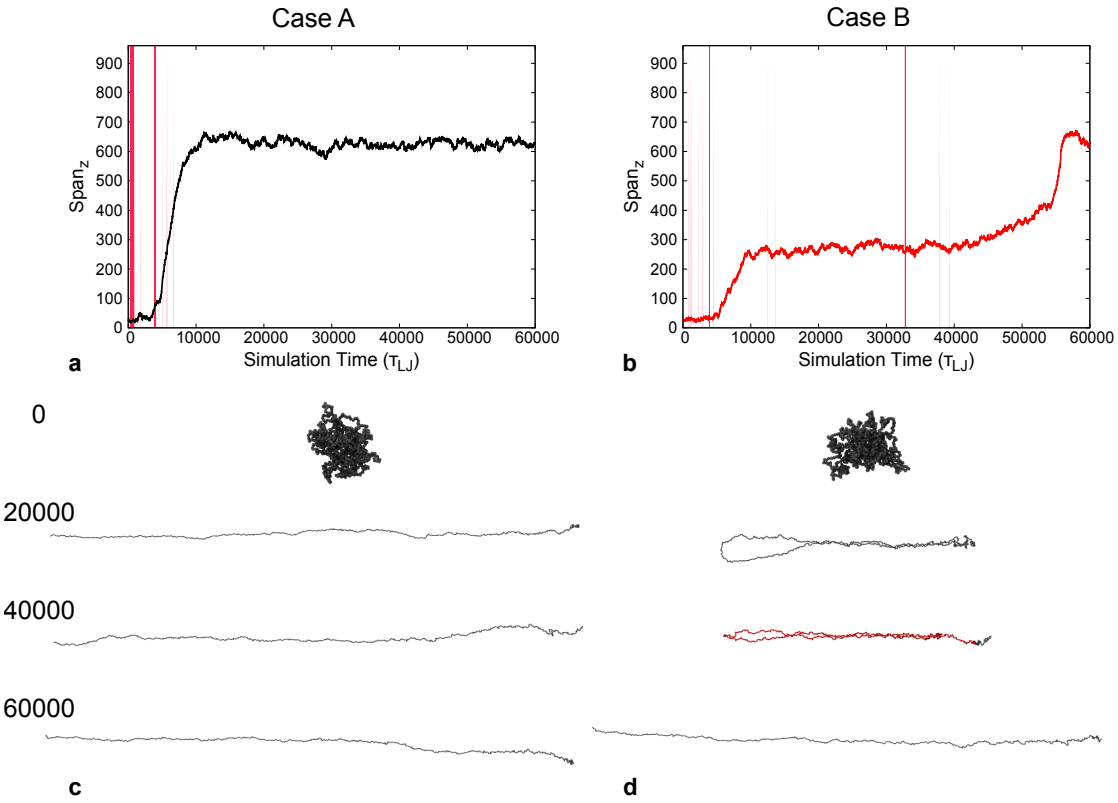


Figure 3.12: DC elongation of two instances of polyelectrolyte chains of length $N = 960$ beads

is triggered by the fact that the two ends of the PE chain are deeply trapped within the collapsed globule and, hence, cannot easily spill out. It is rather intuitive to connect this event to the geometrical complexity of the collapsed state of the PE chain.

The elongation process carries on with the progressive enlargement of the chain loop, which encompasses a larger and larger fraction of the chain. The effect of the loop enlargement is to slow down significantly the kinetics of chain elongation. The long-lived plateau at about 25% elongation in Figure 3.12b corresponds to the time in which the loop is enlarged by the progressive sliding of the two juxtaposed regions of the PE chain. The latter are connected by the trivalent counterions, which bridge together the two negatively charged strands.

At various stages of the elongation process, the loop structures is apparently maintained by physical knots. In fact, it is found that at $40000\tau_{LJ}$ of the elongation trajectory a trefoil (3_1) knot is holding together the two ends of the looped region. This phenomenology is illustrated in panel d of Figure 3.12 and in Figure 3.13, where the knotted region is highlighted in red.

However, because the knotted region encompasses a large fraction of the chain contour

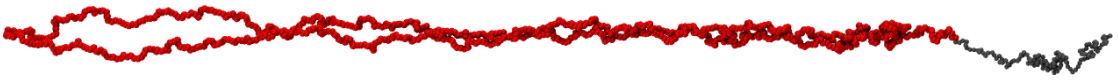


Figure 3.13: Close-up view of the knotted conformation in Figure 3.12d - The beads knotted region which account for 91% of the chain are enlarged for visual clarity and highlighted in red

length ($\sim 91\%$) the application of the electric force does not lead to a progressive tightening of the knot, which could stall indefinitely the kinetic evolution of the chain (as it happens in several contexts), but progressively unties the knot and favours the chain elongation.

Finally, the external field is still capable of fully unravel the PE chain. In fact, the PE chain is eventually elongated to 70% of its contour.

3.6 Unravelling of polyelectrolyte chains in AC electric field

To further characterize these effects, we studied also the action of AC external electric field on the polyelectrolyte chain.

Thus, we applied to the initial conformation of the previously-discussed case B, a square-wave AC electric field of magnitude equal to $0.080\epsilon/(e\sigma)$ and of varying period, p. Specifically, as a reference period we took $p = 12,000\tau_{LJ}$, which corresponds to a time span which is on average sufficient to lead to the complete elongation of the chain in DC field. We next considered two other periods of duration $24,000\tau_{LJ}$ and $6,000\tau_{LJ}$, which are respectively 2-times and one half the reference one. Analogously to the DC case the field is directed along the $\hat{z}$ axes.

The elongation process of the polyelectrolyte chain in AC electric field is shown in Figure 3.14. The profiles of the projection along $\hat{z}$, $span_z$ as a function of time for the three runs at different AC periods show a fast elongation of the chain, which is not hindered and slowed down as in the DC case.

This striking results shows that the AC electric field seems to be more suitable, with respect to the DC one to unravel collapse polyelectrolyte conformations.

This qualitative difference of behaviour, which has not been reported before, ought to be verifiable experimentally and could be exploited in practical contexts to optimize the manipulations conditions of PE chains so to avoid (or harness) the presence of knots.

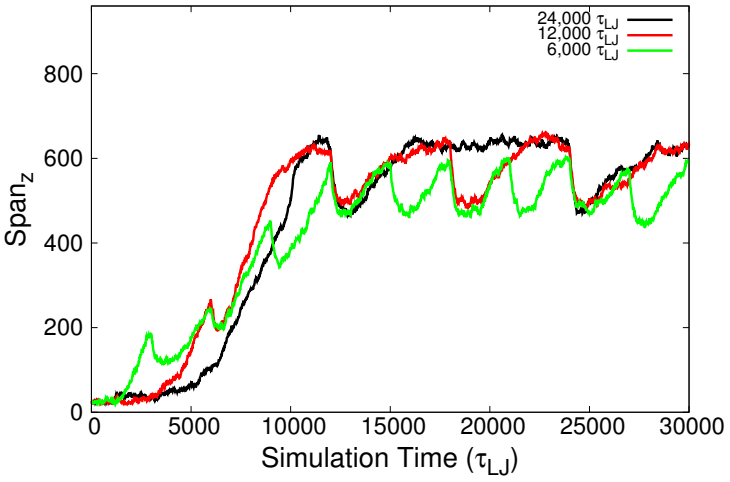


Figure 3.14: AC elongation of two instances of polyelectrolyte chains of length $N = 960$ beads.

3.7 Summary

We reported on a systematic study of the equilibrium and out-of-equilibrium behaviour of polyelectrolyte chains interacting with counterions.

After reproducing well-established equilibrium properties, such as the propensity of the PE chains to collapse in solution with trivalent salt, we studied the spontaneous occurrence of entanglement, and in particular knotting, in collapsed PE chains. Specifically, we measured the spontaneous knotting probability for PE chains of length from 60 up to 960 beads. At equilibrium, the presence of knots is negligible for chains up to ~ 200 beads and then starts to increase for longer chains. At the largest considered contour length, $N = 960$, the knotting probability is as high as 33%, meaning that one out of 3 chain conformations is knotted. To the best of our knowledge this is the first time that the occurrence of nontrivial entanglement trapped in the form of physical knots is ascertained and quantified for a general model of PE chains.

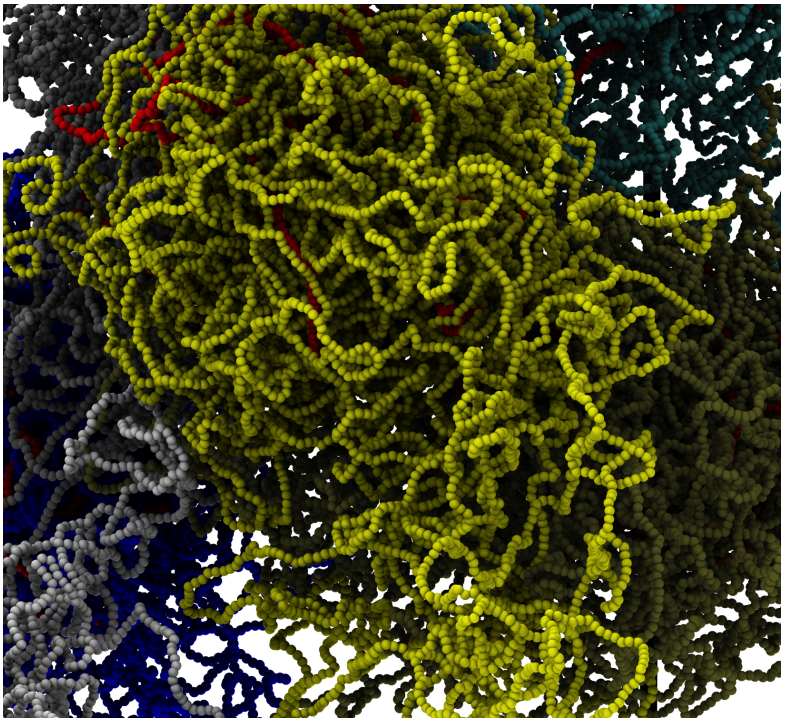
This study highlighted also the impact of geometrical entanglement on the dynamics of elongation of PE chains under the action of an external DC electric field. In particular, we showed that the entanglement does not appreciably alter the "steady-state" elongation of the charged chains. We found, in fact, that the critical electric field, E_c, needed to elongate the chain is only minimally affected by the geometrical entanglement of the polyelectrolyte.

Interestingly, we found that the compaction of the chain affects the transient leading to the elongated steady-state. Specifically, the kinetics of chain unfolding and elongation is significantly slowed down (albeit not halted) due to long living meta-stable states promoted by the formation of loops protruding from the chain. These results pose the interesting challenge of extending the investigation to longer chains, where the effect of geometrical and topological entanglement is expected to be even larger and where the transient states could arguably last larger time spans.

Finally, and at variance with the DC case, we found that AC electric fields of suitable frequency can unravel the chains without any significant slowing down of the elongation dynamics. The AC manipulation might, therefore, provide a novel strategy to control the self-entanglement polyelectrolytes (such as DNA). This might be valuable in contexts where are needs to precondition the geometry and topology of the chains [75].

Chapter 4

Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19



Close-up view of the spatial organization of human chromosome 19 as obtained from coarse-grained molecular dynamics simulations - The chromatin filament was described as a string of beads (30nm in diameter). The chromosomal arrangement resulted from the application of knowledge-based spatial proximity constraints between coregulated genes on the chromatin fiber. It was shown that most coregulated gene pairs can be brought into contacts.

The advent of innovative fluorescence-based techniques has provided an unprecedented insight into the organization of eukaryotic chromosomes during various phases of the cell cycle [7, 8]. A notable example is given by the demonstration - based on imaging techniques - that when the tightly packed mitotic chromosomes enter interphase they swell and occupy specific nuclear regions, aptly termed "territories" [7]. More recently, the salient local and global spatial properties of chromatin fibers inside these territories have been addressed by the so-called "chromosome conformation capture" techniques [9, 76–79], which allow for probing the cis/trans contact propensity of various chromosomal *loci*.

The recent systematic application of these experimental techniques is providing increasing evidence that chromosomes are organized in functionally-heterogeneous macrodomains with different molecular and genetic composition [9, 16, 17].

Several efforts are being spent to clarify the functionally-oriented implications of such chromosomal organization. Towards this goal, in ref. [80] V. Belcastro and D. di Bernardo, with whom I collaborated for the study described in this chapter, carried out a comprehensive bioinformatic survey of data gathered in more than $20,000$ gene expression profiles measured for several cell lines in different human tissues. It was thus established that genes can be grouped into large clusters based on significant pairwise correlations (mutual information) of their expression patterns. In addition, the matrix of pairwise gene expression correlations displayed features qualitatively similar to the matrix of pairwise gene contacts inferred from the HiC [9].

Furthermore, for various model organisms, specific sets of genes that are systematically coexpressed were shown to be in spatial contact too [14, 15, 81]. This is the so-called "gene-kissing" mechanism [14, 15]. A chief example is provided by the human IFN$-\beta$ gene, an ≈ 800 base pairs-long region on human chromosome 9. This gene, during virus infection, induces colocalization and coexpression of 3 distant NF$-\kappa$B bound genomic *loci* [82].

While not all sets of coexpressed or coregulated genes are expected to be nearby in space [83], several arguments and model calculations have consistently indicated that the simultaneous colocalization of multiple genes can occur with appreciable probability even when the genes are far apart along a chromosome and in the presence of a crowded nuclear environment [84, 85]. Indeed, it has been argued that the cooperative colocalization of various genes can provide a very efficient means for achieving their functional coregulation [86, 87].

These considerations, motivated our numerical study in ref. [88] where a knowledge-based coarse-grained model of eukaryotic chromosome 19 was used to ascertain whether

the large number of coregulated gene pairs on a given chromosome can be actually colocalized in space. The analysis therefore complemented recent efforts through which the organization of model chromatin fibers was investigated by bringing distant regions into contact by using attractive interactions, which either mimicked the effect of transcription factories [85] or 5C-based distance restraints [89].

Our investigation, was carried out for human chromosome 19 (Chr19). This chromosome was chosen because it has the highest gene density and extensive gene expression data are available for it. By analysing the mutual information content of thousands of such expression profiles we identified hundreds of coregulated gene pairs for Chr19. These coregulated gene pairs were next mapped onto a previously-validated model for interphase chromosomes (where the chromatin filament was coarse-grained at a resolution of $\approx 30\,\mathrm{nm}$) and their pairwise colocalization was enforced using a steered molecular dynamics scheme. The protocol was applied to various initial chromosome configurations where the degree of entanglement was comparable to that expected for chromosomes *in vivo* (based on the crumpled-globule interpretation of HiC data [9, 11]) or much higher (as in equilibrated polymer chains). Further terms of comparisons were obtained by randomizing the positions or pairings of the *loci* to be colocalized.

4.1 Coregulation networks for Chr19 genes

4.1.1 Collection of gene expression data

To identify the set of significantly coregulated gene pairs on Chr19 we relied on a gene expression survey carried out by V. Belcastro and D. di Bernardo that were collaborating with us in this project and provided us with the extensive data they had published in ref. [80].

Specifically, such data had been compiled from an extensive dataset of gene expression data for human cells starting from the public ArrayExpress database [90]. The data set consisted of $20,255$ gene expression profiles of $22,283$ human probe sets[1] probed on HG-U133A Affymetrix chip [91] measured in 591 distinct experiments. The expression profiles pertained to heterogeneous experiments, in fact they referred to different human cell types or mutants, diverse tissues, and were possibly taken in various experimental conditions.

For this investigation, we considered in particular chromosome 19 (Chr19) which is $\approx$ 60Mbp long, because it has the highest gene density compared to other chromosomes [92].

[1] As customary we shall hereafter refer to the probe sets simply as genes or *loci*.

The data collection was restricted to the set of $1,278$ genes which exclusively target a single sequence (i.e. an uninterrupted stretch) of chromosome 19. Next, to perform a robust comparison between the differently normalized gene expression profiles, all expression levels were coarse-grained all to one of three discrete states only: low, medium and high, which include 33% of the entries each, see in ref. [80].

4.1.2 Measuring correlation with mutual information

To look for correlations between the expression data of gene pairs along the experiments, next, the mutual information content (MI) was computed for all the possible pairs of gene expression profiles, I and J:

$$MI_{\mathrm{IJ}} = \sum_i \sum_j \pi_{ij} \ln\left(\frac{\pi_{ij}}{\pi_{i+}\,\pi_{+j}}\right) \tag{4.1}$$

where i [j] runs over the three coarse-grained expression levels for gene I [J]. In Eq. 4.1, π_{ij} is the joint probability that, in a given experiment, the expression levels i and j are respectively observed for genes I and J, while the quantities $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are the probabilities to observe expression level i [j] for gene I [J] (marginal probabilities).

To understand which kind of correlations can be captured by mutual information with respect to other methods that measure, for instance, linear correlations, let us to consider some examples of two variables generated with different degrees of correlation.

In all the cases presented below, X is a continuous random variable uniformly distributed between 0 and 1 and the second variable Y, is chosen such to highlight some of the possible outcomes of correlation analysis done with mutual information.

A In the first case we consider the variable Y to be independent on X, hence $\pi_{ij} = \pi_{i+}\pi_{+j}$. From Equation 4.1, one readily sees that the value of mutual information is zero. To verify it, we generate the random variable Y independently on the X. Next, we compile the discretized joint probability matrix π_{ij} of X and Y and compute their mutual information content. As shown in Figure 4.1A, the π_{ij} is uniform, which results in a zero mutual information value.

B Now, we move to analyzing a case in which the variable Y is highly correlated with X. We generate, in fact, the values of second random variable Y as $y = 1 - x$ that is, by construction, linearly anti-correlated with the values x of the variable X, see Figure 4.1B. Discretizing the continuous random patterns and computing the joint probability matrix we obtain an anti-diagonal matrix with all the entries equal to

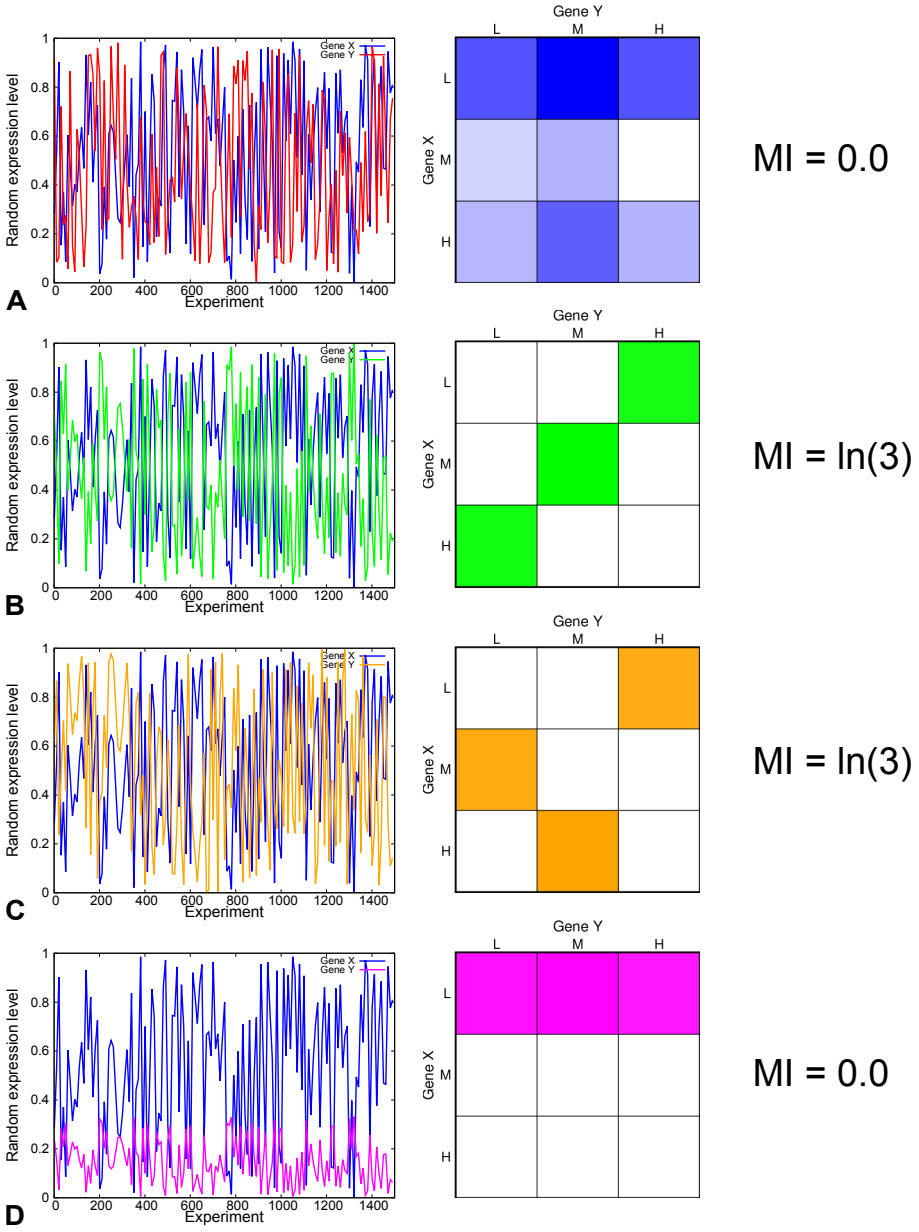


Figure 4.1: Illustration of some examples of mutual information content between random variables - Realizations of pairs of random variables are shown together with their tripartite joint probability matrices and the mutual information contents. In (A) the two variables are independent, in (B) they are anti-correlated, in (C) there is a one-to-one correspondence between the discretized values of the two variables, and in (D) the second variable Y varies in a smaller range with respect to X.

1/3. The resulting value of mutual information is: $MI = \ln(3)$. As a matter of fact, such value is the largest that the mutual information of tripartite data sets can possibly attain. So, an high value of mutual information can, in principle, capture the linear anti-correlation (and analogously correlation) between any pair of gene expression patterns.

C To highlight the fact that mutual information can capture non-linear correlations,

we consider the case of a variable Y, whose values y are a non-linear function of the values x of X. Specifically, they are equal to $(x + 0.3)$ $mod\, 1$, where mod returns the residual of the integer division by 1. By discretizing the values of variables X and Y and computing the joint probability matrix we obtain a matrix with only 3 non-zero entries all equal to 1/3, see Figure 4.1C. The entries are positioned one per row and per column, suggesting that for each value of X there is one (and only one) correspondent value for Y. This situation describes one of the possible situations in which the amount of information one knows about the variable Y, knowing the value of the variable X, (that is the definition of mutual information) is maximum. The measure of MI is, in fact, $\ln(3)$. It is important to notice that in this case, there is not a linear dependence between the gene expression patterns but, the MI can capture their correlation content.

D Finally, we discuss a limiting case where the mutual information analysis can fail to pick up significant correlations between two data sets. Such situations typically occur when the values of one, or both variables, fluctuate only within one of the discretized levels. Consider, for instance, the case in which $y = (1 - x)/3$. Since the values of x vary in $[0 : 1.0]$, y fluctuates, but limitedly to the range $[0 : 1/3]$. In this situation, the linear estimator would capture an obvious anti-correlation between the two variables, but the mutual information, results in a zero value. Proceeding as in the previous cases, we end up with a joint probability matrix with the following three elements equal to 1/3: π_{LH}, π_{LM}, and π_{LH}. The value of mutual information is:

$$\begin{aligned} MI_{12} &= \pi_{LL} \ln(\pi_{LL}) + \pi_{LM} \ln(\pi_{LM}) + \pi_{LH} \ln(\pi_{LH}) \\ &= 3\frac{1}{3} \ln(1) = 0 \end{aligned} \tag{4.2}$$

In summary, the mutual information can capture correlation between random variables independently from their specific functional dependence, but we are also aware that sometime MI measurements cannot capture correlations when expression patterns vary with small fluctuations with respect to the whole range of variability.

4.1.3 Assessing the statistical significance of MI values

Another reason to use mutual information for the correlation analysis is that there exists a way to establish the statistical significance of the obtained MI contents, which is not always available for other estimators of correlations. To single out the pairs of genes with statistically-significant coexpression we proceeded, in fact, according to the procedure described below and summarized graphically in Figure 4.2.

As shown in ref. [93], the distribution of pairwise of MI values expected for two random variables (expression of the two genes) assuming 3 possible distinct values (low, medium and high) can be well approximated with the analytical expression $f(x) = a\,x\,e^{-b\,x}$, where x are the mutual information values 4.2A and a and b are the free fitting parameters. The reference, null distribution was used to define the mutual information threshold above which at most one false-positive entry is expected to occur.

Since the values of MI had to be homogeneous to guarantee the applicability of the statistical reference model, we subdivided the gene pairs in 15 groups to account for the expected dependence of gene coregulation on genomic distance. The first, second, etc. group gathered pairs of genes whose central bases had a genomic distance falling in the intervals $0-4$Mbp, $4-8$Mbp, etc. Next, for each group we fitted the histogram of MI values with the null reference $f(x)$, as shown in Figure 4.2B. The values of the fitting parameters (a and b) and the thresholds are reported in Table 4.1 for each of the 15 groups.

In each group, all gene pairs, whose MI content exceeded the stringent threshold, were retained (see panels C and D of Figure 4.2).

Group index	**Range of genomic distances (Mbp)**	**a**	**b**	**MI Threshold**
1	0-4	2887800	160.0	0.0884464
2	4-8	2702200	171.2	0.0806886
3	8-12	2269700	178.0	0.0773173
4	12-16	1613300	177.9	0.0738272
5	16-20	1460000	204.1	0.0683442
6	20-24	1221900	214.6	0.0644916
7	24-28	1418200	212.3	0.0673494
8	28-32	1623700	191.9	0.0681704
9	32-36	2033900	195.3	0.0713540
10	36-40	1953200	188.7	0.0712516
11	40-44	1860300	195.7	0.0672588
12	44-48	1393600	187.6	0.0660945
13	48-52	1025100	202.4	0.0602505
14	52-56	645280	207.0	0.0481679
15	56-60	302880	265.9	0.0331916

Table 4.1: Parameters of the fitting curves $f(x) = ae^{-b/x}$ and thresholds of mutual information for the plots in Figure 4.2B

The number of selected pairs for each bin ranged from 59 to 334, for a total of $1,991$ probe pairs. It should be noted that several of these pairs involve chromosome regions that were highly overlapping and were hence degenerate (or nearly degenerate). To eliminate this redundancy, we grouped together the pairs of coregulated genes that assured the coregulation of regions whose central beads are separated by less than 300nm (which

corresponds to the chromatin fiber statistical (Kuhn) length [11]). For each of these groups, we retained only the pair with the largest MI value. This filtering procedure returned $1,487$ non-degenerate probe set pairs, that involved 412 genes (native case). As customary, the significant degree of coexpression of such pairs was deemed indicative of their coregulation [94].

The obtained *native case* coregulation network is shown in Figure 4.2E, where the 412 genes involved in coregulation are represented as red dots positioned onto the schematic circular chromosome. The gray region indicates the centromere. The coregulatory relationships are described as links and distributed over three diagrams depending on the genomic separation of the involved *loci*. It is important to notice that there are links along all the possible range of genomic distances even spanning the entire length of Chr19 (rightmost diagram).

4.2 Modelling chromosome structure and dynamics

The feasibility to colocalize in space the $1,487$ pairs of genes was explored using the coarse-grained model chromosome and the steering molecular dynamics protocol described in the following sections.

4.2.1 The chromosome polymer model

The chromatin fibers of human chromosome was described as a semi-flexible chain of N beads with thickness $\sigma = 30$nm and persistence length, l_p, equal to 150nm [2, 11]. Accordingly with the mapping in ref. [11], each bead spans $\approx 3,000$ base pairs and, hence, to account for the total contour length $L_c = 59.13$Mbp of human chromosome 19, the number of beads N was set equal to $19,710$.

Each chain was described with the Kremer and Grest polymer model [24] presented in section 1.2.1:

$$\mathcal{H}_{\mathrm{KG}} = U_{\mathrm{FENE}} + U_{\mathrm{KP}} + U_{\mathrm{LJ}}. \tag{4.3}$$

We recall that the three terms acted, which within each copy of the chromosome chain, correspond to the FENE chain-connectivity interaction, the Kratky-Porod bending energy, and the pairwise Lennard-Jones excluded volume term.
The latter term controlled also the inter-chain excluded volume interaction and ensured that any two regions cannot pass through each other. In this way, intra- and inter-chain

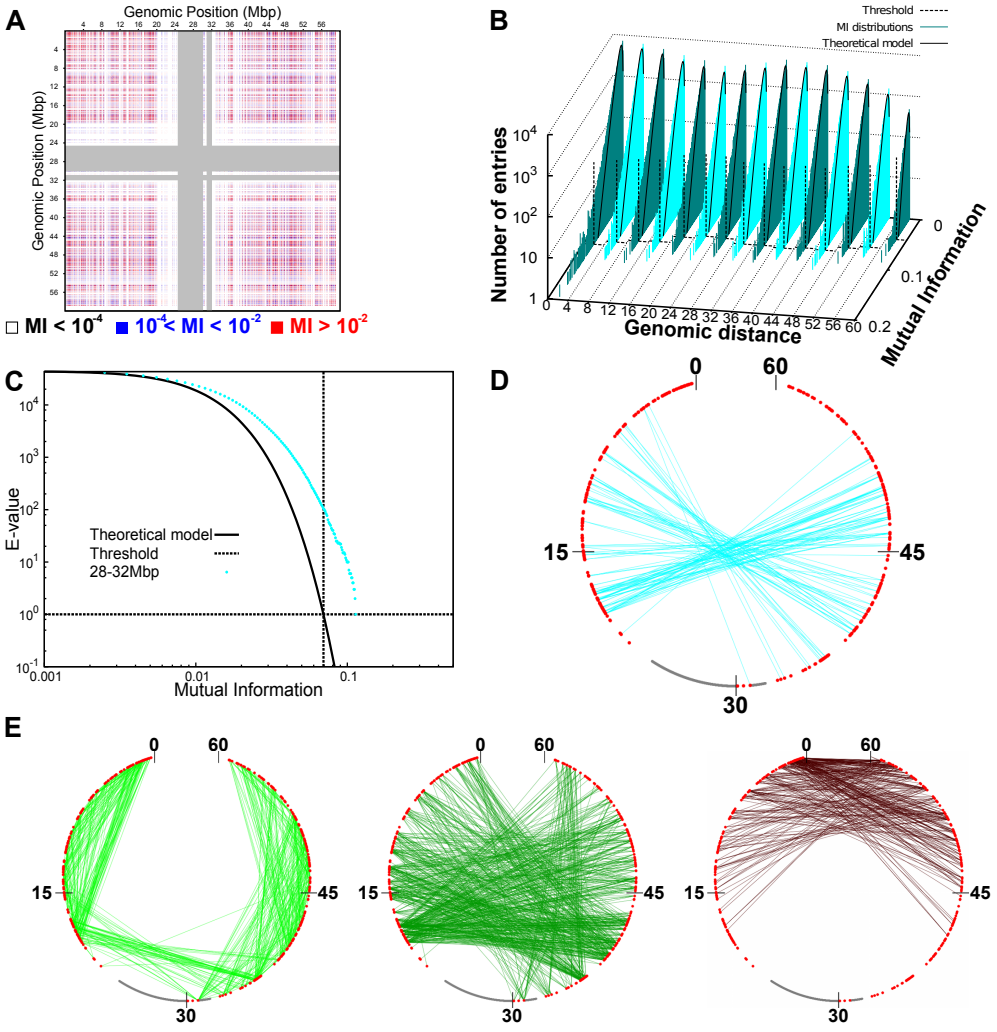


Figure 4.2: Statistical analysis of mutual information - (A) mutual information values for any pairs of genes on Chr19. The middle point of each gene identifies its position along the chromosome. The gray stripes correspond to the centromere. (B) Histograms of values of mutual information for pairs of genes located at various intervals of their genomic separation. The black lines correspond to fitting the histograms with the theoretical (null case) MI distribution [93]. The vertical black dashed lines correspond to the estimated threshold values (see next and main text). (C) Example of E-value (expected number of false positives) distribution for gene pairs located at genomic separation in the range $28 - 32$Mbp. The threshold is the value of mutual information at which the E-value is equal to 1.0. For different genomic separations, analogous curves were obtained. (D) Network of coregulated pairs of genes at $28 - 32$Mbp separation. The analysis illustrated in (C) singles out significantly-high values of mutual information. These contributions corresponds to connections (*cyan links*) between coregulated gene pairs (*red dots*). The scale is in μm. (E) Networks of coregulated pairs of loci used to fix the spatial constraints between corresponding regions of the model chromosomes. For the sake of clarity, the whole network has been represented as three sub-networks for pairs of loci at genomic separations of 0-20Mbp (*left*), 20-40Mbp (*middle*) and 40-60Mbp (*right*), respectively.

topology was preserved during the dynamical evolution of the system. The three energy terms were parametrized as in section 1.2.1.

This model for chromatin filaments has been previously shown to be capable of accounting for the fractal-like organization observed for eukaryotic chromosomes [9, 11, 95–99].

4.2.2 Steered molecular dynamics protocol

The colocalization of the 1,487 coregulated genes was attempted by using a steered molecular dynamics protocol which progressively favoured the spatial proximity of the pairs of genes in each of the six model chromosomes.

Specifically, we mapped each pair of selected genes, A and B, onto the discrete beads using the Affymetrix annotation table ([91]) and added to the system energy an harmonic constraint:

$$U_{\mathrm{H}} = \frac{1}{2} k\left(t\right) d_{A,B}^{2} \tag{4.4}$$

where $d_{A,B}$ is the distance of the centers of mass of the chromosome stretches covered by the two genes. The stiffness of the harmonic constraint was controlled by the time-dependent parameter $k(t)$.

The chain dynamics was described with an underdamped Langevin equation, see section 1.3:

$$m\ddot{r}_{i\alpha} = -\partial_{i\alpha}\mathcal{H} - \gamma r_{i\alpha} + \eta_{i\alpha}\left(t\right) \tag{4.5}$$

where $\mathcal{H}$ is the Hamiltonian of the system, i runs over the particles in the system and α over the Cartesian coordinates. The Gaussian white noise $\eta_{i\alpha}\left(t\right)$, introduced in section 1.3, had the statistical properties $\langle \eta_{i\alpha} \rangle = 0$ and $\langle \eta_{i\alpha}(t)\, \eta_{j\beta}(t') \rangle = 2\kappa_B T \gamma\, \delta_{ij}\delta_{\alpha\beta}\delta(t - t')$, where δ_{ij} is the Kronecher delta, $\delta(t-t')$ is the Dirac delta, k_B is the Boltzmann constant and T the temperature.

The Langevin equation was integrated numerically with the LAMMPS molecular dynamics software package [29] with an integration time step equal to $t_{int} = 0.012\tau_{MD}$, where $\tau_{MD} = \sigma(m/\epsilon)^{1/2}$ is the Lennard-Jones time and m is the bead mass which was set equal to the LAMMPS default value.

The simultaneous application of the 1,487 constraints of the coregulation networks (see section 4.1) to each of the six chromosome copies was implemented using the PLUMED plugin for LAMMPS [100]. Specifically, the steering protocol consisted in the progressive cranking up of the the time-dependent stiffness of the harmonic constraints $k\left(t\right)$. The latter was, in fact, ramped exponentially in time from the initial value $k\left(t=0\right) = 0.001\epsilon/\sigma^2$ up to the value $k\left(T_{end}\right) = 16.384\epsilon/\sigma^2$. The total duration of the steered dynamics was $T_{end} = 10^7 t_{int}$. This protocol favoured the progressive reduction of the width of the distribution of probe set distances from the initially generous value of $\approx 50\sigma$ (see 4.8) down to $\approx 0.4\sigma$.

The protocol was sufficiently mild that no crossings of the chains should occur. This was checked by running the steering protocol on a circularized variants of the mitotic conformation shown in Figure 4.3A, and checking that the initially unknotted topological state was maintained [64].

4.2.3 Preparation of the initial chromosome conformation

To mimic inter-chromosome interactions in the dense nuclear environment, we considered a system where six copies of Chr19, accounting for a total of $\approx$ 360Mbp (59.1Mbp per copy), are placed in a cubic simulation box (with periodic boundary conditions) of side equal to 3μm. The overall system density is therefore 0.012bp/nm^3, which matches the typical genomic one in human cells, where $\approx 6 \cdot 10^9$bp are packed in a spherical nucleus $\sim 5\mu$m in radius [11].

To mimic the mitotic state, each model chromosome was initially prepared in an elongated solenoidal-like configuration [11], and the six copies were placed in a random, but non-overlapping arrangement inside the cubic simulation box as shown in Figure 4.3A. To remove any excessive intra-chain strain of the orderly designed mitotic arrangement, the model chromosomes of Figure 4.3A were briefly evolved with a standard push-off protocol for $10^5 t_{int}$. The resulting relaxed mitotic configuration is shown in Figure 4.3B.

This mitotic arrangement was further evolved for a much longer simulation time, roughly corresponding to 7 hours in "real-time" [11], to obtain the fully decondensed arrangement shown in Figure 4.3C. Such configuration exhibits the same power-law decay of contact probabilities versus genomic separation as observed in HiC experiments [9, 101], see inset of Figure 4.3C. The model system therefore aptly reproduces the salient experimentally-observed features of interphase chromosomes.

4.3 Colocalization of coregulated genes in human chromosome 19

After setting up the mitotic and interphase systems, we next applied the steered molecular dynamics protocol presented in section 4.2.2 to each of them to promote the spatial proximity of regions corresponding to coregulated gene pairs.

To characterize the establishment of the coregulation contact we computed the *percentage of established target contacts*, Q as a function of the time evolution of the system.

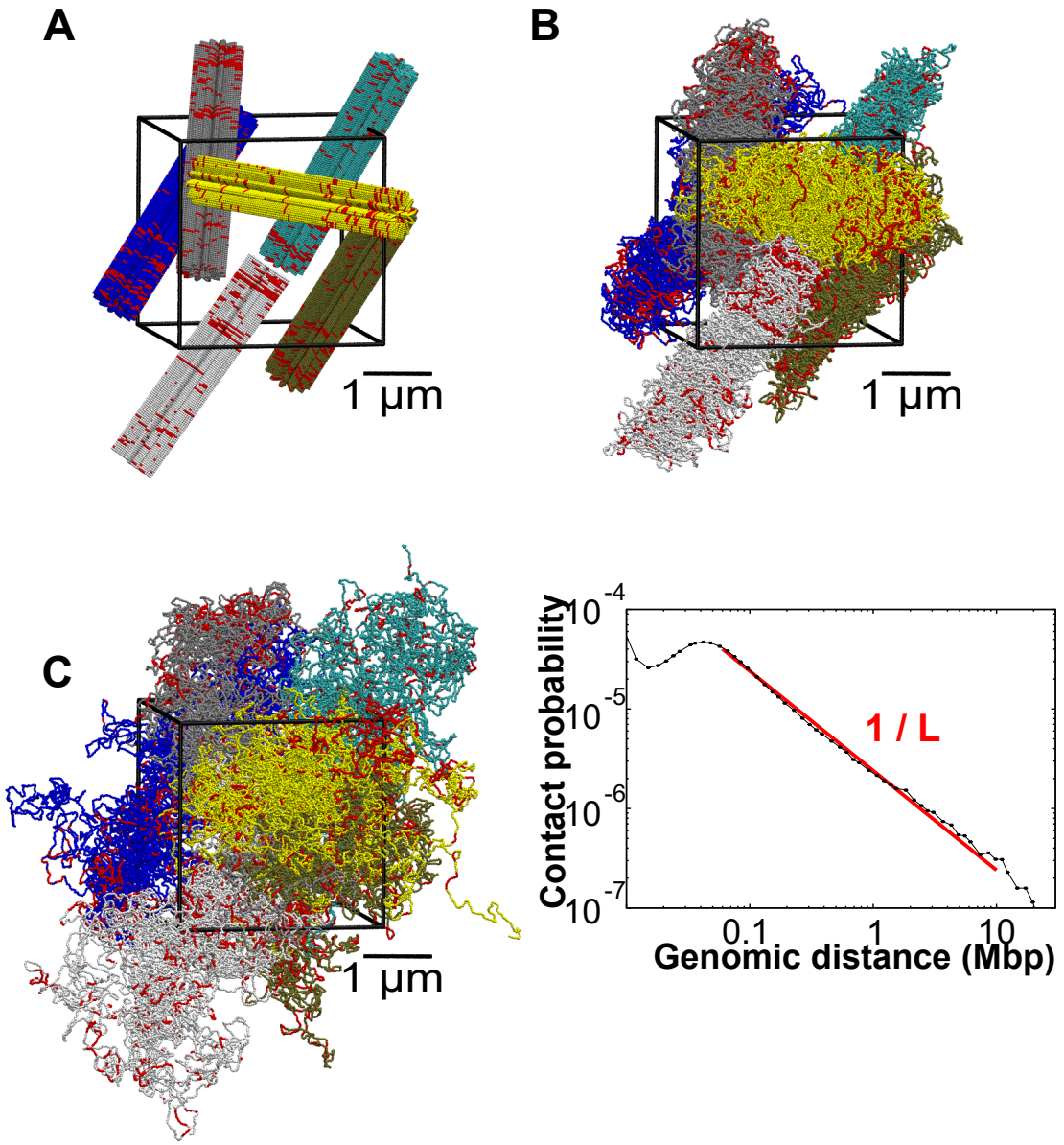


Figure 4.3: Mitotic and interphase configurations of the model system chromosomes - (A) Initial mitotic-like arrangement, constituted by 6 copies of model human chromosome 19. Following ref. [11], the chromatin fiber was helicoidally arranged into loops of ≈ 50kbp each, and departing radially from a central axis. The six solenoidal arrangements were next placed in a random, but non-overlapping manner inside a cubic simulation box of side equal to 3.0μm and with periodic boundary conditions. (B) Chromosome spatial arrangement after short relaxation with a standard push-off protocol of $10^5 t_{int}$. (C) Interphase-like configuration obtained by evolving the initial mitotic configuration for $10^8 t_{int}$ MD time steps (approximately corresponding to 7 hours in "real-time" [11]). (*Inset*) The corresponding contact probabilities between *loci* of model interphase chromosomes decays as a power law of the genomic distance, $\approx L^{-1}$, consistently with recent experimental observations [9, 101]. In all panels, chromosome regions involved in the coregulatory network are highlighted in red.

This parameter was calculated as:

$$Q = \frac{1}{G} \sum_{(A,B)} \Theta\left(r_c - d_{A,B}\right) \times 100 \ . \tag{4.6}$$

In the above expression, the sum runs over the coregulated pairs of genes, A and B which are in total $G = 8,922$ (i.e. $1,487$ for each of the six chromosome copies), $d_{A,B}$ is the distance of their centers of mass. $\Theta(x)$ is the Heaviside step which takes a values of 1 if $x > 0$ and 0 otherwise. Θ was used to restrict the sum to those gene pairs that are at distance within the contact range, $r_c = 120$nm. This cutoff distance was chosen because it is about equal to the typical size of a *transcription factory* [87].

The compliance of the two systems to the steering protocol is illustrated, in fact, in Figure 4.4 which shows the increase of Q during the simulation time.

It is striking to observe that for both system it was possible to simultaneously colocalize a very high fraction of the target pairs, namely 80% of them (averaged over the six chromosome copies). The conformations reached at the end of the steering protocol are shown in the right panels of Figure 4.4.

Considering the relatively-high density of the simulated system of chromosomes and that most of the coregulated pairs lie at large genomic distances, the results point to an unexpectedly high degree of plasticity of the mitotic and interphase conformations, which is presumably ascribable to their fractal-like metric properties which keeps at a minimum the entanglement of the chromatin fiber [9, 11, 95–99, 102].

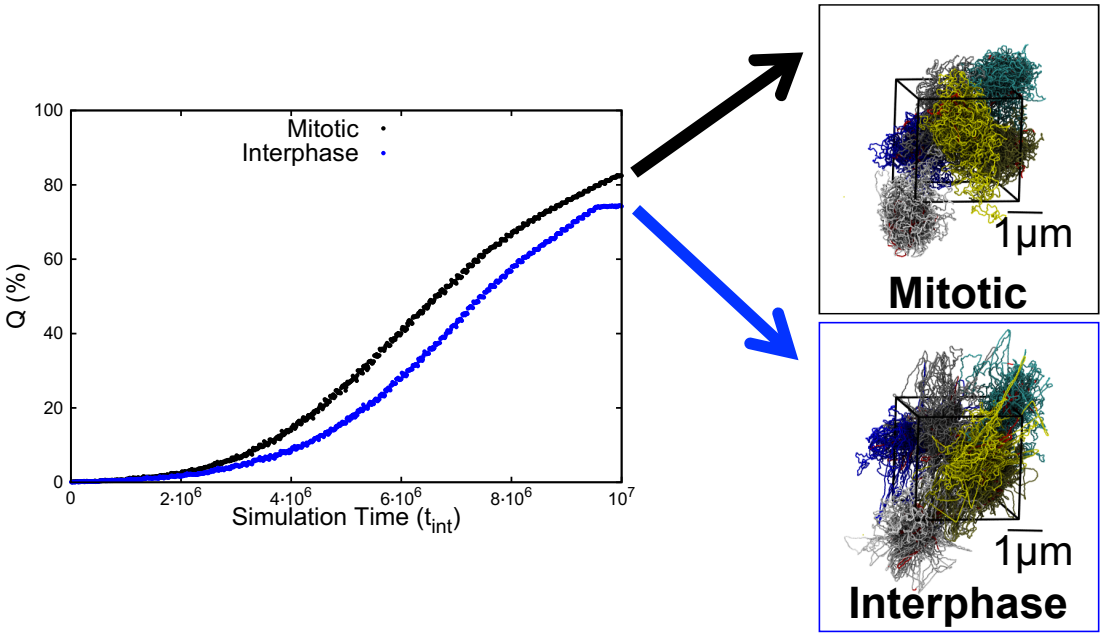


Figure 4.4: Increase of the percentage, Q, of Chr19 coregulated pairs which colocalized during the MD steering protocol - The two curves reflect different initial conditions corresponding to the mitotic and the interphase conformations of panels (B) and (C) of Figure 4.3. The final configurations, corresponding to $Q \approx 80\%$ are shown on the right. Chromosome regions involved in the coregulatory network are highlighted in red. These and other graphical representations of model chromosomes were rendered with the VMD graphical package [63].

A second noteworthy feature of the results of Figure 4.4 emerges considering the diversity of the sources used to derive the knowledge-based coregulation data. In fact, granted the validity of the coregulation–colocalization hypothesis, one might have envisaged *a priori* that the chromosomal configurations corresponding to different tissues or experimental conditions would be so heterogeneous that it would be impossible to satisfy the cumulated set of colocalization constraints. By contrast, the results of Figure 4.4 demonstrate *a posteriori*, that the set of pairwise colocalization constraints are largely mutually compatible because most of them can be simultaneously satisfied.

The findings are therefore not only consistent with the coregulation–colocalization hypothesis but, based on such hypothesis, also suggest that the conformations adopted by a chromosome in various conditions can share a common underlying pattern of colocalized genes.

4.4 Spatial macrodomains: comparison with data based on HiC maps

To further characterize the overall organization of the steered conformations shown in Figure 4.4 we identified their spatial macrodomains and compared them with those inferred from the analysis of HiC data collected by Dixon *et al.* [17].

4.4.1 Constructing the contact maps for comparison

In both the experimental and the numerical conformations, the starting point of the analysis was the construction of a binary chromosome contact map, C_{ij}, with a 60kbp resolution, which is commensurate with both the experimental resolution (20kbp) and the bead equivalent contour length (3kbp).

The generic matrix entry $C_{i,j}$ takes on the value 1 or 0 according to whether the ith and jth 60kbp-long segments (equivalent to 20 beads) were in spatial proximity or not. The recent high-resolution HiC measurements of Dixon et al. [17] were used to derive the experimental, reference contact map. Specifically, for every significant HiC entry (i.e. normalized contact enrichment ≥ 1) the corresponding contact-matrix elements were set equal to 1. The resulting HiC-based contact map was sparse in that only 5% of its entries were non-zero. For an equal footing comparison, we next populated the theoretical contact maps by considering in spatial contacts (entries equal to 1) only the top 5% 60kbp-strands ranked for increasing average distance. The distance average was

taken over the six Chr19 copies at the end of the steering protocol. Both binary matrices are shown in Figure 4.5.

4.4.2 Clustering analysis of the contact maps

Next to find the optimal partition of Chr19 in macro-domains we used a clustering analysis. Specifically, we following the K-medoids clustering strategy in ref. [103], in which each domain spans an uninterrupted stretch of the chromosome and one domain always matches the centromere region.

Accordingly the optimal domain partitioning was identified by minimizing the total intra-domain dissimilarity. The latter, for one domain r, covering the chain interval from i to j, was measured as:

$$\Delta_r = \sum_{l=i}^{j} (1 - C_{l,r}) \tag{4.7}$$

where C is the contact matrix and r is the domain representative, i.e. is the element belonging to the i–j interval for which Δ_r is minimum. Consistently, the dissimilarity score, Δ, takes on small (or large) values if respectively many domain members are in contact ($C_{l,r} = 1$) or not ($C_{l,r} = 0$) with the representative.

For a given number of domains K, the optimal domain partitioning is the one that minimizes the sum of the Δ_r scores for the domains, S:

$$S = \sum_{r=1}^{K} \Delta_r \tag{4.8}$$

The clustering analysis of the contact maps was used to subdivide Chr19 into up to $K = 10$ spatial macro-domains. For both maps the consensus domain boundaries were well-captured by the subdivision into $K = 8$ spatial domains, see Figure 4.6. The corresponding macro-domain partitions are overlaid on the contact maps of Figure 4.5.

4.4.3 Computing the overlap between the experimental and the model partitions

For a given number of domains, the consistency of the steered-MD and HiC-based subdivisions was measured by establishing a one-to-one correspondence of each domain in the two cases and next measuring the percentage of elements, q, having identical domain assignment. The one-to-one domain correspondence was identified by exploring

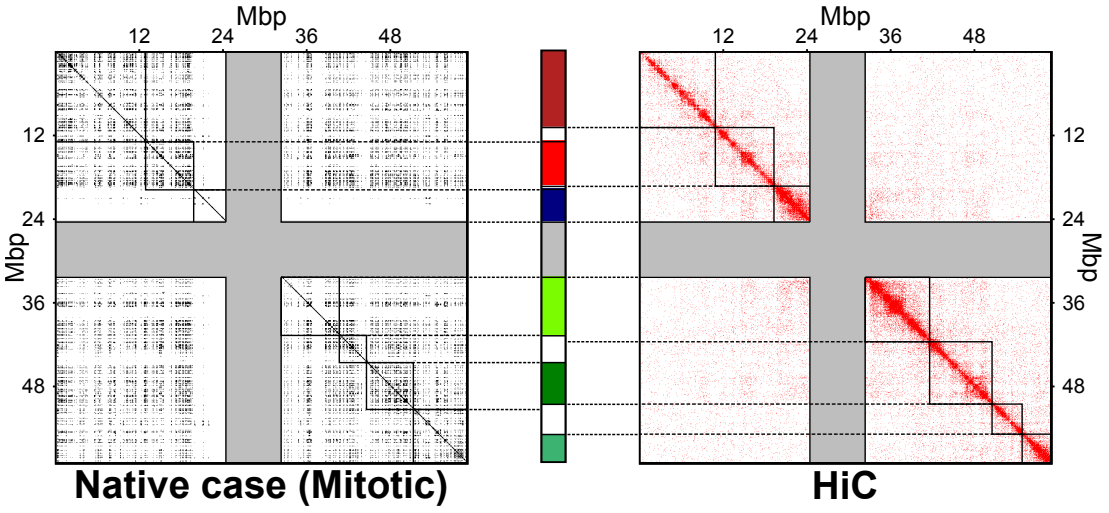


Figure 4.5: Spatial macrodomains - The contact maps for Chr19 obtained at the end of the steered-MD simulations and inferred from HiC data are shown on the left and right, respectively. The grey bands mark entries involving the centromere region. The boundaries of the 8 principal spatial domains, identified with a clustering analysis of the contact maps, are overlaid on the matrices. The consistency of the two macro-domain subdivisions is visually conveyed in the chromosome sketch at the center. The overlapping portions of the domain subdivisions are colored (different colors are used for different domains). Non-overlapping regions are shown in white, while the centromere region is shown in grey. The overlapping regions accounts for 79% of the chromosome (centromere excluded).

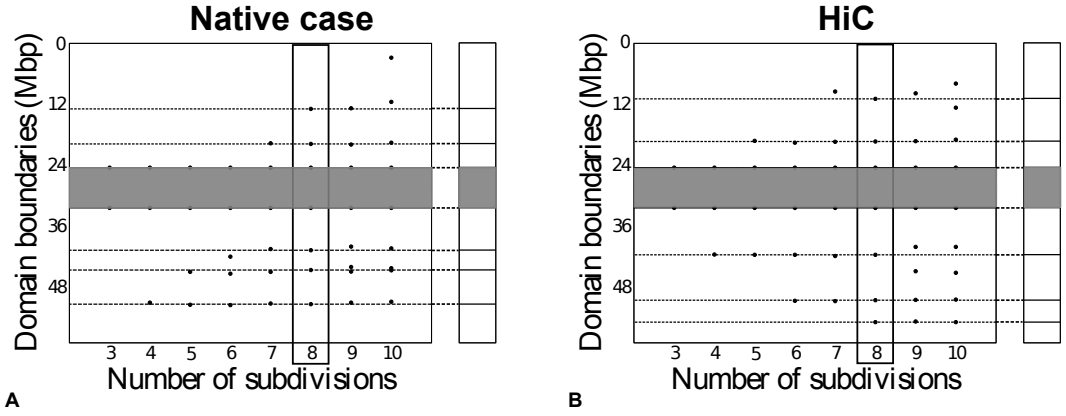


Figure 4.6: Chr19 spatial macrodomains - The filled circles mark the boundaries of the Chr19 spatial macrodomains obtained from the clustering analysis of the steered-MD contact maps (A) and inferred from HiC data (B). The number of imposed macrodomains is shown on the x axis. In all cases, one domain was fixed to correspond to the centromere (for which no HiC data are available) which is shown in grey. The dashed guidelines mark the subdivision into eight macrodomains which, by visual inspection provides robust, consensual boundaries in both cases. For clarity, the eight-domain subdivision is also reported on the chromosome sketch on the right.

the combinatorial space of correspondences and picking the one yielding the largest value of q.

The good consistency of the domains found using HiC-based and steered-MD contacts maps is visually conveyed by the matching colored regions in the schematic chromosome

partitioning of Figure 4.5. It is interesting to notice that the two domain subdivisions consistently indicate larger domains for the upper arm. Quantitatively, the overlap of the two subdivisions is 0.79, which has a p-value smaller than 0.03. This means that random partitions of the chromosome into eight domains (one always being the centromere) yields overlaps ≥ 0.79 in less than 3% of the cases, see Figure 4.12. The quantitative comparison therefore indicates a statistically-significant consistency of the spatial macrodomains arising in the steered chromosome conformations and those inferred from experimental data.

4.5 Chromosome entanglement, regulatory network properties and gene colocalizability

Besides the previous considerations, the results of Figure 4.4 prompted the question of whether, and to what extent the feasibility to colocalize a significant fraction of the coregulated gene pairs depended on distinctive chromosomal features, such as the spatial arrangement of the mitotic and decondensed states or the network of coregulated genes.

To address these issues we re-applied the steering protocol starting from 3 different initial conditions, which corresponded to "ad hoc" designed variants of the model chromosomes. Specifically, the three systems were:

1. A *random-walk-like* chromosome arrangement as shown and described in Figure 4.7A.

2. A mitotic-like spatial arrangement but with *randomized gene pairings*. The chromosome spatial configuration was the same as in Figure 4.3B, but the $1,487$ pairings of the native network between the considered set of 412 genes were randomly reshuffled while preserving the native number of pairings for each gene. This alternative set of gene pairs was obtained by applying the iterative randomization method described in ref. [104]. The asymptotic fraction of randomized gene pairs matching the native ones was $\approx 10\%$.

3. A mitotic-like spatial arrangement but with *randomized gene positions*, see Figure 4.7C. As in case 2 above, the chromosome spatial configuration was again the same as in Figure 4.3B, but the positions of the 412 genes involved in the native coregulatory network were randomly assigned along the chromosome (except for the centromeric region). The repositioned genes inherited the native coregulatory pairings.

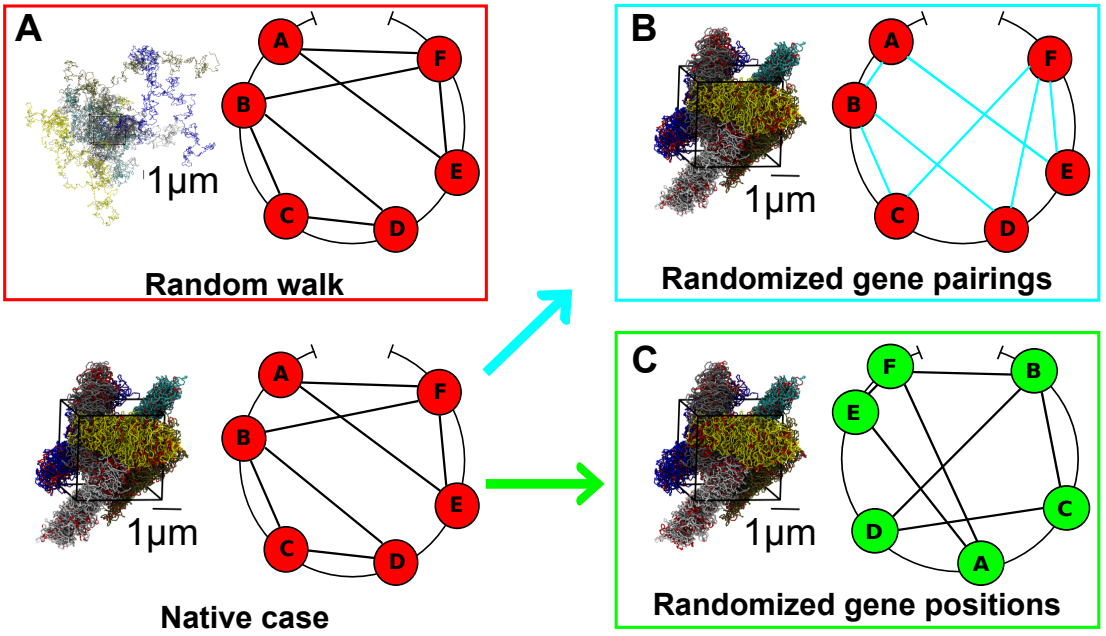


Figure 4.7: Variant systems subjected to the MD steering protocol - (A) Initial configuration of 6 random-walk like chains the linear size the model chromosome 19. (B) Model chromosomes were initially arranged as in the mitotic-like configuration of Figure 4.3B, but the pairings between genes were randomized. The randomization preserved the number of pairs that each gene takes part to. (C) Model chromosomes were initially arranged as in the mitotic-like configuration of Figure 4.3B, but the gene positions along the chromosome were randomized. The randomization preserved the native pairings of the genes. In all panels chromosome regions involved in the native or randomized coregulatory network are highlighted in red. For all the three systems considered the same physical conditions of fiber density, stiffness and excluded volume interactions of the original system apply.

We stress that the three variants were prepared so to preserve the native overall density, number of coregulated genes and also the number of coregulated pairs to which a selected gene takes part to. They nevertheless presented major differences which allow for probing the impact of different system properties on gene "colocalizability".

In particular, the random-walk-like arrangement had a much higher degree of intra- and inter-chain entanglement than all other arrangements, as illustrated by the much wider distribution of gene pairwise distances in the initial configuration, see Figure 4.8. For randomly-paired and randomly-repositioned genes, instead, the distributions of genomic distances of the target genes to be paired was similar to the native one. This is clearly shown by the distributions in Figure 4.8.

However, the same figure clarifies that the two randomized cases differed markedly from the native one for the clustering coefficient. The latter, CC, was used to characterize connectivity properties of graphs. In the present case the graph of coregulation of pairs of genes. Each gene was represented by a node in the graph. Pairs of coregulated genes were represented by a link connecting the two corresponding nodes, see Figure 4.2E.

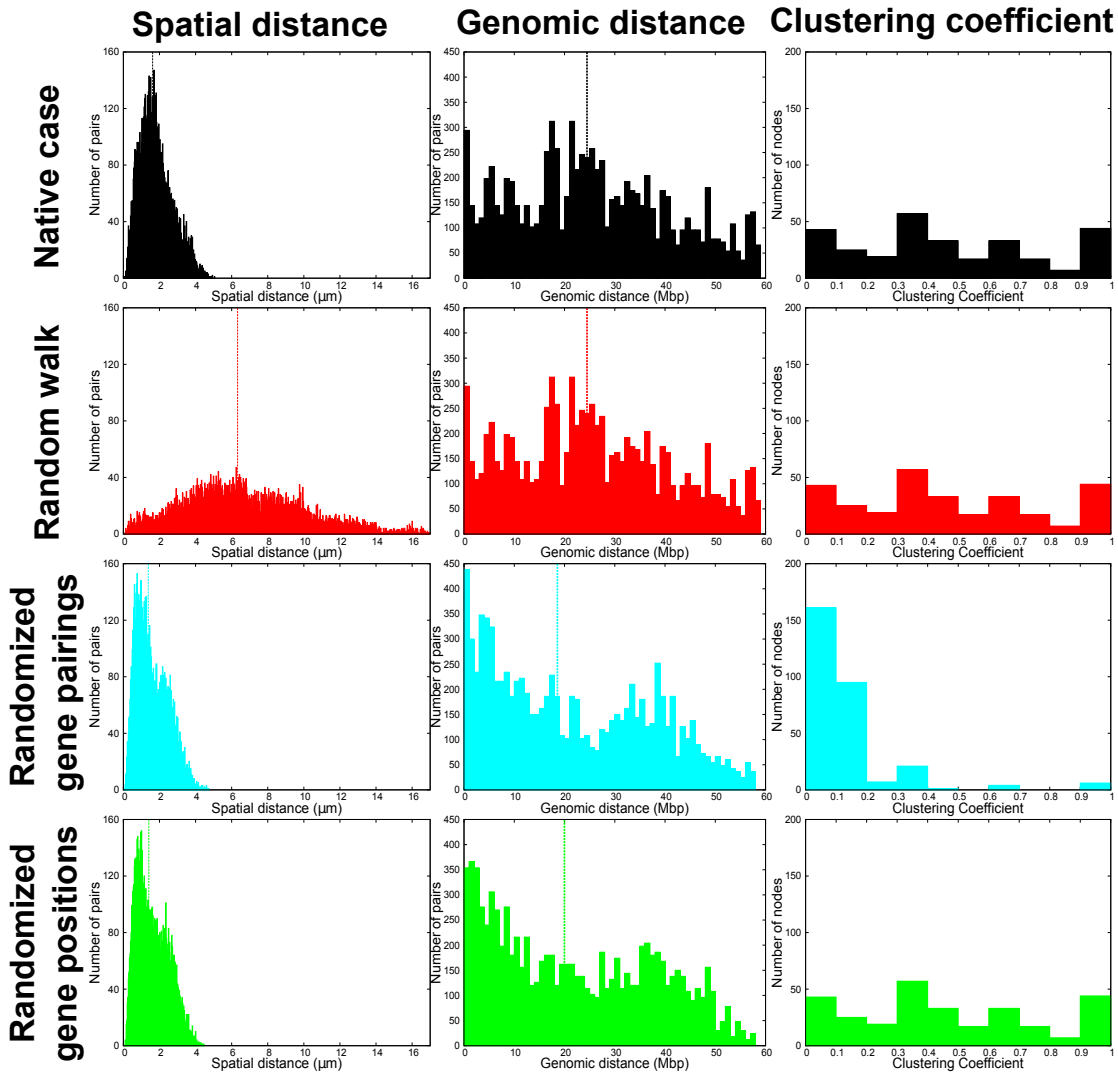


Figure 4.8: Summary of the structural properties of the native system (Figure 4.3B) and its three variants (Figure 4.7) - (First column) Distribution of the *spatial* distances between steered *loci*. The distribution of the random-walk-like is broader than the native case one. The randomized position and randomized pairs cases have instead a similar distribution with respect to the native case. (Second column) Distribution of the *genomic* distances between steered *loci*. (Third column) Clustering coefficients of the corresponding networks of pairings between steered *loci*. Dashed lines correspond to the median values. The results are cumulated over all 6 chromosome copies in the simulation box.

The clustering coefficient of the individual ith node in the graph was defined as [105, 106]:

$$c_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{i,l} a_{l,m} a_{mi}}{(a_{i,l})^2 - \sum_{l \neq i} a_{i,l}^2} \tag{4.9}$$

where $A = [a_{i,j}]$ is the adjacency matrix. The latter is a symmetric matrix with entries 1 or 0 depending on whether or not two node are connected. The clustering coefficient can assume values between 0 and 1 and, as it is illustrated in Figure 4.9, the case $c_i = 1$ corresponds to the case in which all the neighbors of the node i are all connected between each other.

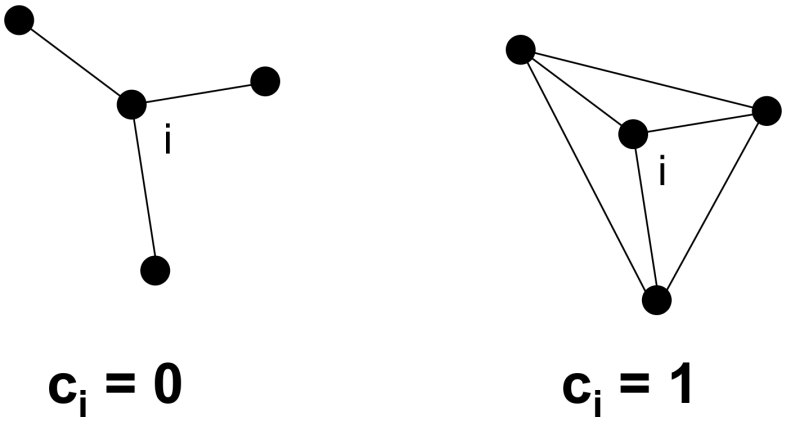


Figure 4.9: Illustration of two nodes with minimum and maximum clustering coefficient.

Hence, the clustering coefficient captures the degree of cooperativity of the coregulatory network in that it measures how frequently two genes that are both coregulated with a third one, are themselves coregulated too.

The inspection of the rightmost graphs in Figure 4.8 therefore indicates that the clustering coefficient distribution of the randomly-paired system was shifted towards much smaller values than the others, which all inherited the native pairings network. This fact indicates that the clustering coefficient of the native network was significantly larger than random. This implies that genes could frequently interact concertedly in groups of three or more.

4.5.1 Gene colocalizability in randomized systems

As for the native network of target gene pairs, we report on the properties measured at the end of the steering protocol after averaging them over the six chromosome copies in the simulation cell.

The results of the steering protocol applied to the three system variants are shown in Figure 4.10. The data indicate that:

(i) for random-walk-like chromosomes only a minute fraction ($< 1\%$) of the target contacts could be satisfied;

(ii) for randomly-paired genes about 47% of the gene pairs could be colocalized;

(iii) for randomly-repositioned genes about 75% of the gene pairs could be colocalized, similarly to the native case (Figure 4.4).

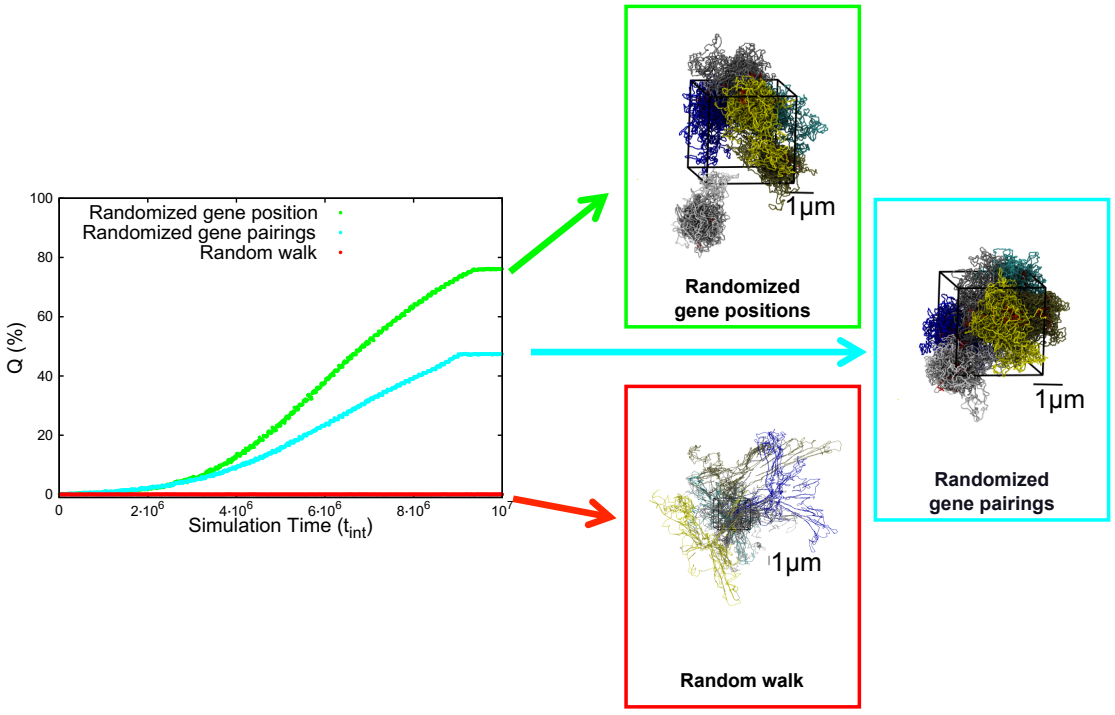


Figure 4.10: Increase of the percentage, Q, of Chr19 coregulated pairs which colocalize during the MD steering protocol, for the three variants of the native systems - The configurations reached at the end of the steering protocol are shown on the right. Chromosome regions that took part to the pairs of loci to be colocalized are highlighted in red.

These findings provided valuable clues for interpreting the high degree of "colocalizability" of coregulated genes observed in Figure 4.4 for the mitotic and interphase arrangements.

In particular, the very low asymptotic value of the percentage of successfully colocalized gene pairs for the random-walk-like system clarifies that the low intra- and inter-chromosome entanglement of both the mitotic and decondensed configurations is crucial for bringing into contact the coregulated gene pairs.

Furthermore, the comparison of the randomly-paired and randomly-repositioned gene cases shows that the connectivity properties of the native coregulatory network appear even more important than the detailed positioning of the coregulated genes along the chromosomes. In fact, the randomly-repositioned genes – which retain the same native coregulatory graph – have the same high degree of colocalizability of the native system. By converse, the randomly-paired gene case – corresponding to a significant disruption of the original network – reflects in an appreciably lower value of percentage of successfully colocalized gene pairs. It is also worth noticing that, in all cases, a significant fraction of gene pairs brought in contact are at large genomic distances ($>$ 20Mbp), see Figure 4.11.

Finally, to understand how the network randomization effected on the spatial organization of the steered conformations, we measured the overlap of their spatial macrodomains

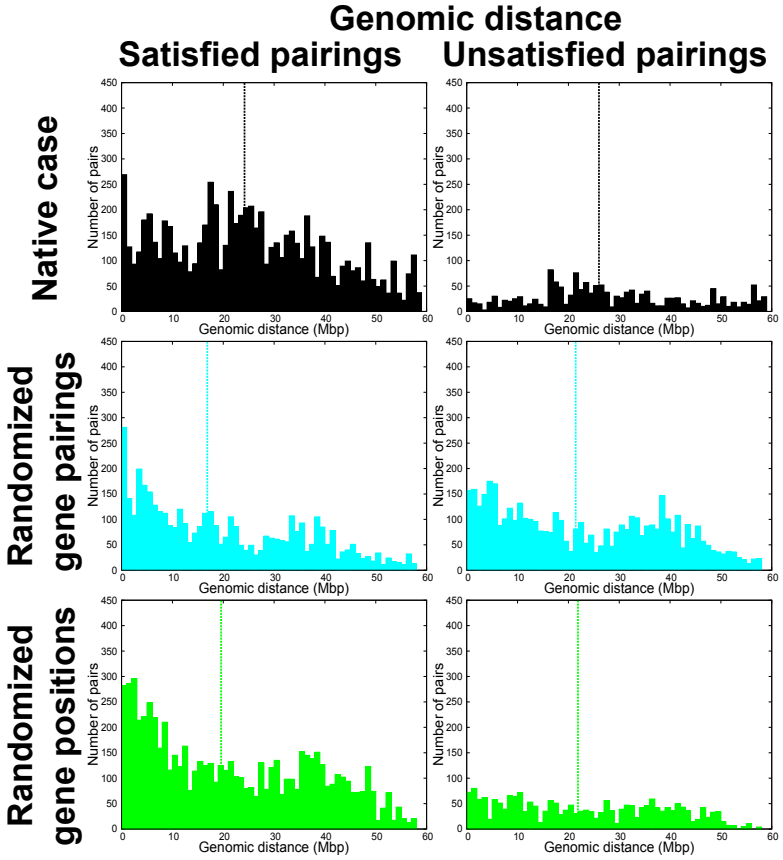


Figure 4.11: Genomic distance distribution for the target gene pairings established at the end of the steering protocol - The plots on the left provide the genomic distance distributions of target gene pairings that are actually satisfied at the end of the steering protocols for the native and randomized cases. The analogous distribution for non-satisfied pairings is shown on the right. Dashed lines correspond to the median values. The results are cumulated over all 6 chromosomes copies in the simulation box.

with those established from HiC data. We recall that for chromosome subdivisions into eight macrodomains, the native case overlap was 0.79. For the randomized gene positions and randomized gene pairings we instead observed the lower values 0.73 and 0.63, respectively. These values clearly had a much lower statistical significance than the native case; their p-values being respectively 0.113 and 0.490, see Figure 4.12. Their non-significant similarity with the reference, HiC-data based macro-domain subdivisions underscored that randomized, non-native constraints resulted in appreciably-different, and less realistic, chromosomal features.

To further validate the effect of cliques on the gene contact satisfiability, we considered an additional target network for the steered-MD simulations. This network was obtained by a partial randomization of the native gene pairings and its average clustering coefficient was 30%, which is intermediate to the native one (47%) and the fully-randomized case (12%) discussed previously. As shown in Figure 4.13, 64% of the target colocalization constraints were satisfied. This value was intermediate between the native and

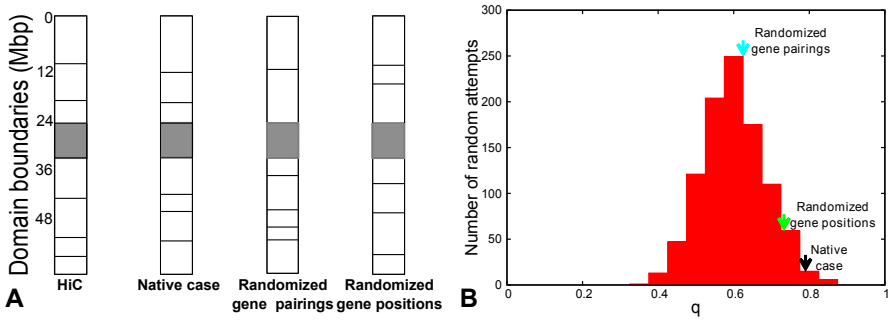


Figure 4.12: Comparison of macro-domain subdivisions - (A). Schematic representation of the Chr19 partitioning in 8 macrodomains (one being the centromere) based on the clustering analysis of contact maps inferred from HiC data and from steered-MD simulations on the native and randomized versions of the gene pairing network. In all cases, one domain was constrained to match the centromere (shown in grey). The overlap, q and associated p-value of the steered-MD subdivisions against the reference HiC-data based one are as follows, (i) native case: $q = 0.79$, p-value= 0.027; (ii) randomized gene positions: $q = 0.73$, p-value=0.113; (iii) randomized gene pairings: $q = 0.63$, p-value=0.49. The p-values were computed by comparing the observed overlap against a reference distribution of overlaps of 1000 random chromosome partitions into 8 domains (one always corresponding to the centromere). The reference distribution is shown in panel B. The arrows indicate the overlaps of the native and randomized cases.

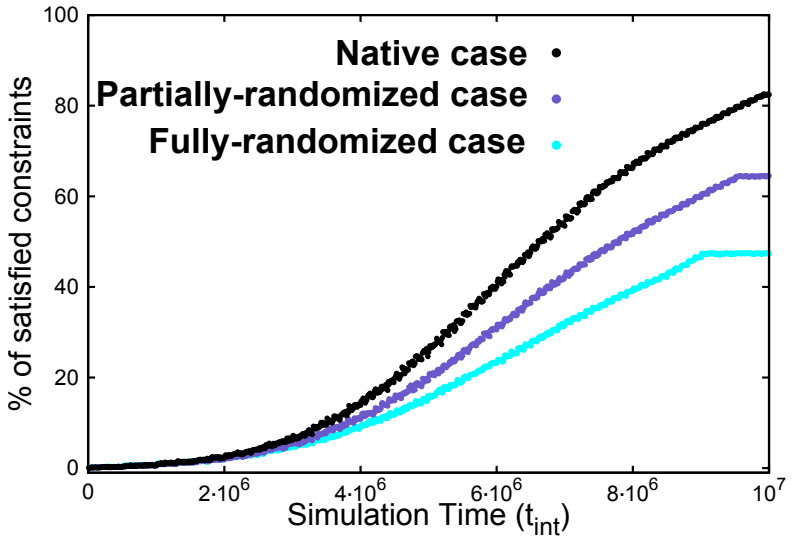


Figure 4.13: Gene colocalizability and gene network cliquishness. The time evolution of the fraction of satisfied gene pairings for three different steered-MD simulations - The target gene pairing networks for the simulations are: the native network and two variants of it obtained by partial and full randomizations of gene pairings. The curves for the native and fully-randomized cases are the same as in Figure 4.10. The different cliquishness of the three target networks is captured by their clustering coefficient: 0.47 for the native case, 0.30 for the partially-randomized case and 0.12 for the fully-randomized case. The fraction of established pairings shows a clear monotonic (increasing) dependence with the clustering coefficient.

fully-randomized case (82% and 47%, respectively) and hence supported the existence of a meaningful correlation between gene colocalizability and the regulatory network cliquishness.

4.6 Summary

Recent experimental advancements have provided unprecedented insight into the occurrence of concerted transcription of multiple genes. In particular, it has been reported that the chromatin fiber can rearrange so that genes, concertedly transcribed upon activation, are found nearby in space too.

Because of its important ramifications, the possible existence of a general relationship between gene coregulation and gene colocalization, the so called "gene-kissing" mechanism [14, 15], has been a subject of very active research.

This standing question was addressed here numerically by carrying out molecular dynamics simulations of a knowledge-based coarse-grained model of human chromosome 19. The model consisted of a coarse-grained representation (30nm resolution) of the chromatin fiber complemented by the knowledge-based information of the loci corresponding to (≈ 1500) coregulated gene pairs. These pairs were identified from the analysis of extensive sets of publicly-available gene expression profiles. To mimic the crowded nuclear environment, we considered a system where several copies of the model chromosome 19 were packed at typical nuclear densities. The colocalization of the coregulated gene pairs was finally imposed by applying a steered molecular dynamics protocol.

It was found that most ($\approx 80\%$) of the coregulated pairs could be colocalized in space when the steering protocol was applied to chromosomes initially prepared in mitotic-like and interphase-like arrangements, see Figure 4.4. Notably, the pattern of intra-chromosome contacts established for the steered conformations exhibited significant similarities with that of experimental contact propensities [9, 79] of chromosome 19. Furthermore, the overall chromosomal organization into spatial macrodomains showed significant similarities with that inferred from experimental HiC data.

By converse, the percentage of colocalized target pairs decreased substantially (or vanished altogether) when the system was initially prepared in a random-walk like arrangement, or if the genes to be colocalized were randomly paired or displaced along the chromosome. Likewise, the macro-domain organization of these alternative systems was found to be much less similar to the HiC-data based one.

The findings allowed to draw several conclusions. First, the data in Figure 4.4 demonstrated that, even in a densely packed system of mitotic or interphase chromosomes it is physically feasible to achieve the simultaneous colocalization of a large number of pairs of loci that can be very far apart along a chromosome. This result is therefore well compatible with the gene coregulation–colocalization hypothesis. In fact, the findings can be read as adding support to the hypothesis in consideration of the fact that if no

meaningful relationship existed between coregulation and colocalization one might have expected the unfeasibility of bringing into simultaneous contact so many coregulated pairs.

The much poorer compliance of alternative systems (random-walk-like chromosome conformations, randomized gene pairings and positions) to the steering protocol provides valuable insight into the native chromosomal properties that allow for gene colocalization.

The first and most important property is the low degree of entanglement that mitotic or interphase chromosomes are known to have compared to equilibrated polymer solutions of equivalent density [1, 9, 11, 95–99, 102]. The second property is that the number of gene cliques that is present in the native gene regulatory network of chromosome 19 is much higher than for the equivalent random network. In this respect it is worth pointing out that the atypically large number of cliques found in biological regulatory networks has also been observed and pointed out in different contexts and for a different set of chromosomes [107].

Chapter 5

Gene coregulation/colocalization in *Drosophila melanogaster*: direct comparison of mutual information contents and HiC contacts

In the previous chapter, we discussed a novel computational approach to test the *gene-kissing* hypothesis [14, 15] and exploited it for pinning down plausible three-dimensional models for human chromosome 19 [88]. We recall that our strategy relied on the extensive collection and analysis of gene expression data, the measure of correlations (mutual information content) between gene expression patterns and statistical analysis to infer significant gene pairs coregulation. Next, we employed steered molecular dynamics on a three-dimensional polymer model of chromatin fiber in order to enforce the colocalization of the chromosome regions hosting coexpressed gene pairs.

Our investigations showed a large compliance of gene pairs coregulation and colocalization for human chromosome 19 ($\sim$ 80% of the target contacts were established). Furthermore, chromosome organizations obtained from the steering procedure were shown to be arranged in spatial macro-domains which largely match the internal subdivisions inferred from HiC data [17].

The results obtained on human chromosome 19 prompted us to carry out the study presented in this chapter, where we extended the exploration of the relationship between coregulation and colocalization to an entire eukaryotic chromosome system, in

particular the genome of the common fruit fly, *Drosophila melanogaster*. Specifically, we characterized the gene pairs coregulation by measuring mutual information content and the colocalization by using recent HiC data [16]. Next, we applied bio-informatics and statistical inference techniques to study the possible correlations between the extensive gene-expression and HiC data sets for this model organism.

In this chapter, I first review the aim and the main steps of DNA microarray technology to introduce our analysis of gene expression data.

5.1 Major steps in microarray experiments

Microarray experiments measure the amounts of expressed genes in a population of cells. The set of the measured expression values for an ensemble of genes is called gene expression profile (GEP). The technique helps to answer one of the fundamental question in biology such as in which quantities genes of interest are transcribed in a cell population. The latter information is, in fact, indicative of the cell activity.

Microarray technology is based on seminal experimental studies published in the middle of the 70's, in which investigators showed that by anchoring nucleic acid probes to a solid surface and hybridizing to them labeled transcripts, it is possible to monitor quantitatively the expression of genes within the studied cell population. However, this technology has become practically available only in the mid 90's when it started to be used to profile the full gene expression of simple organisms, such as *Saccharomyces cerevisiae* [108, 109].

Today, the microarray technology has been further developed up to the point that it is possible to measure in a single experiment the amounts of expressed molecules for thousands of genes, covering for instance the entire genomes of humans and *Drosophila melanogaster*.

Part of the results that I shall present in this chapter is based on the extensive analysis of gene expression data obtained with a particular technology, namely Affymetrix chip. For this reason, I will hereafter recall the salient features and properties of Affymetrix-based gene expression profiling.

5.1.1 Production of the gene chip

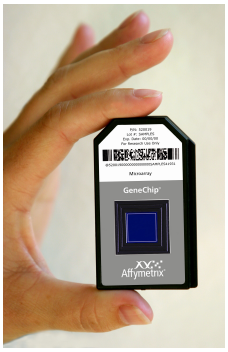


Figure 5.1: Example of an Affymetrix chip - Reproduction of a high-density oligonucleotide microarray often referred to as a chip. The small dimensions of the chip ($1.28\,cm^2$) are compared to a human hand. The image is courtesy of Affymetrix, Inc., Santa Clara, CA, USA.

The basic device for microarray experiments is a gene chip. As shown in Figure 5.1, the latter is a quartz wafer of about $1.28\,cm^2$ which contains 25-nucleotide long single-stranded DNA molecules, the so-called *probes* in preselected positions (spots). The basic technique of manufacturing chip is the use of photo-lithography and combinatorial chemistry to assemble the molecules directly onto the wafer (*in situ* technique). A chip contains typically around one million different DNA probes and in each spot of the chip there are millions of identical copies of the same probe.

Each probe is represented in the chip with a *probe pair*, consisting of the so-called *perfect match* (PM) and *mismatch* (MM) sequences, see Figure 5.2. Specifically, the PM sequence is an exact copy of the reference target one, while the MM sequence one has a changed nucleotide in the middle (position 13). The presence of a probe pair is meant to capture the occurrence of non specific binding with the target sequence.

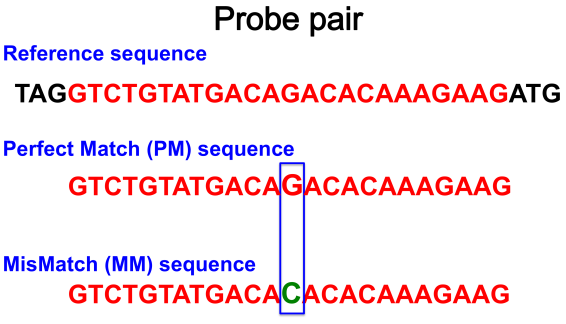


Figure 5.2: Illustration a probe pair - Each probe sequence is long 25 nucleotides and it is represented in the chip by two versions the perfect match (PM) and the mismatch (MM). The PM sequence is a copy of the target and the mismatch (MM) one differs for the $13th$ nucleotide, which is changed in to its complement base (Watson-Crick base-pairing). The MM probe measurements are aimed to capture most of the nonspecific binding of transcripts to the probes.

Groups of $11 - 20$ probes are organized in *probe sets*, which are designed to hybridize to various parts of the messenger RNA (mRNA) generated from the transcription of a single gene.

Once the chip has been manufactured, the microarray experiment can be performed. In the following, we briefly describe the major steps to carry out a DNA microarray experiment by using the Affymetrix technology.

5.1.2 Sample preparation and labelling

The aim of this step is to produce a sample of RNA that is representative of the transcripts contained in the cell population of interest. The sample needs also to be labelled for future chemical staining.

The sample preparation is, accordingly, done by isolating and amplifying the RNA from the cells, in particular, messenger RNA molecules which represent the amounts of expressed genes at the time of sample collection.

Next, the extracted mRNA is subjected to rounds of reverse transcription, whose final product is a sample of chemically-labeled (biotin) complementary RNA (cRNA) strands.

Since the DNA probe in the chip are long 25-nucleotide, the cRNA is fragmented to obtain molecules of homogeneous lengths (in the $35 - 400$bp range), before being transferred onto the array for hybridization.

5.1.3 Hybridization reaction between sample and probe cDNA molecules

Hybridization is the process of joining two complementary strands of nucleic acids (DNA or RNA molecules) to form a double-stranded filament. Here, the labelled cRNA is hybridized against the DNA probes on the gene chip. Each molecule in the sample of cRNA binds to its appropriate complementary target sequence on the immobilised array, as illustrated in Figure 5.3.

5.1.4 Image acquisition

After hybridization, the chip is stained with a fluorescent molecule that binds to biotin. When the chip is next scanned with a confocal laser, fluorescent molecules emit light and the distribution pattern of the signal in the array is recorded as an image, see Figure 5.4.

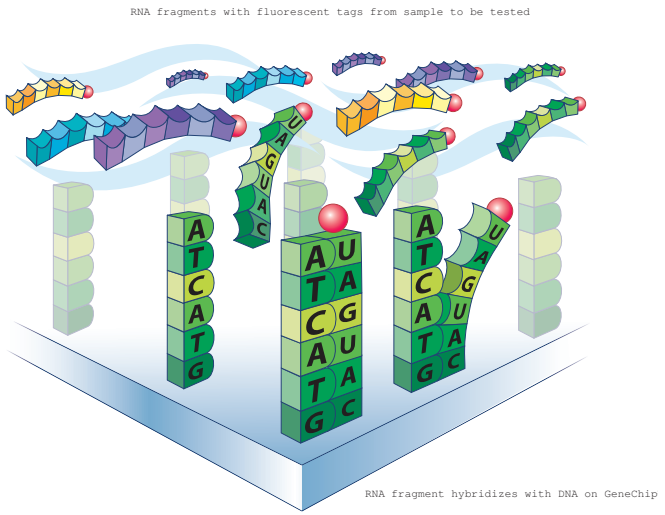


Figure 5.3: Chip Hybridization - Cartoon depicting hybridization of biotin-labeled probes to Affymetrix chip microarray. The cDNA molecules spotted on the chip reacts with the labeled complementary RNA strands of the sample and form and hybrid double stranded molecules. Image courtesy of Affymetrix.

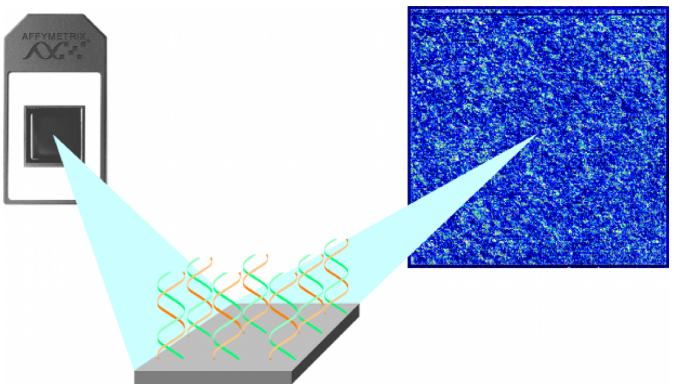


Figure 5.4: Gene chip scanning - The chip is light irradiated and the fluorescence intensities emitted by the labeled cRNA molecules bound to the probes is measured. The light intensity emitted by a single spot on the chip is proportional to the quantity of hybridized cRNA molecules in it and, hence, is a quantitative measure of the expression value for that specific probe. Image courtesy of Affymetrix, Inc., Santa Clara, CA, USA.

The light intensity emitted by a single spot on the chip is proportional to the quantity of hybridized cRNA strands in it and, hence, is a quantitative measure of the expression for that specific probe.

The values of the fluorescence intensities for each spot are stored in a *.CEL* computer file which contains the level of intensity and the location for each probe in the chip.

5.2 Data analysis: from probe fluorescence intensities to the gene expression values

The scope of the microarray experiment is to estimate the expression levels for the probe sets. To do this, it is necessary to covert the information at the probe-level to the probe set one. This process can be done by following different strategies.

In the following, I discuss one of them, the *robust multi-array* (RMA) protocol [110, 111], that has become *de facto* a standard procedure for processing expression data from arrays.

The RMA strategy proceeds through three main steps, which will be explained in the following: background correction, normalization and summarization.

5.2.1 Background correction

The first step is the correction of the measured probe intensities signal for experimental background noise. The procedure I present here is motivated by looking at the distribution of probe intensities. Figure 5.5 shows the probe intensity distribution for the array *GSM397756* in the experiment tagged as *E-GEOD-15825*.

The procedure is based on the main hypothesis that the observed intensity value of each *perfect match* probe I is the sum of the true signal S and a random noise N: $I = S + N$.

It is further assumed, that S is non-negative and exponentially distributed:

$$f_S(s) = \alpha e^{-\alpha s}, \text{ s} \geq 0 \tag{5.1}$$

where $f_S(s)$ is the density function of S and the rate α of the exponential.

The random noise N, instead, is assumed to be Gaussian:

$$f_N(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(n-\mu)^2/2\sigma^2} \tag{5.2}$$

where μ are the mean value and σ^2 the variance.

It is somewhat troublesome to estimate the parameters α, μ and σ. An *ad hoc* approach is used in many cases to estimate the parameters. First, μ is taken as the mode of the measured PM intensity values. We recall that the mode is the value that appears most often in the distribution (see Figure 5.5). Next, σ and α are obtained by fitting procedures. Specifically, σ is obtained by fitting the lower tail of the distribution with

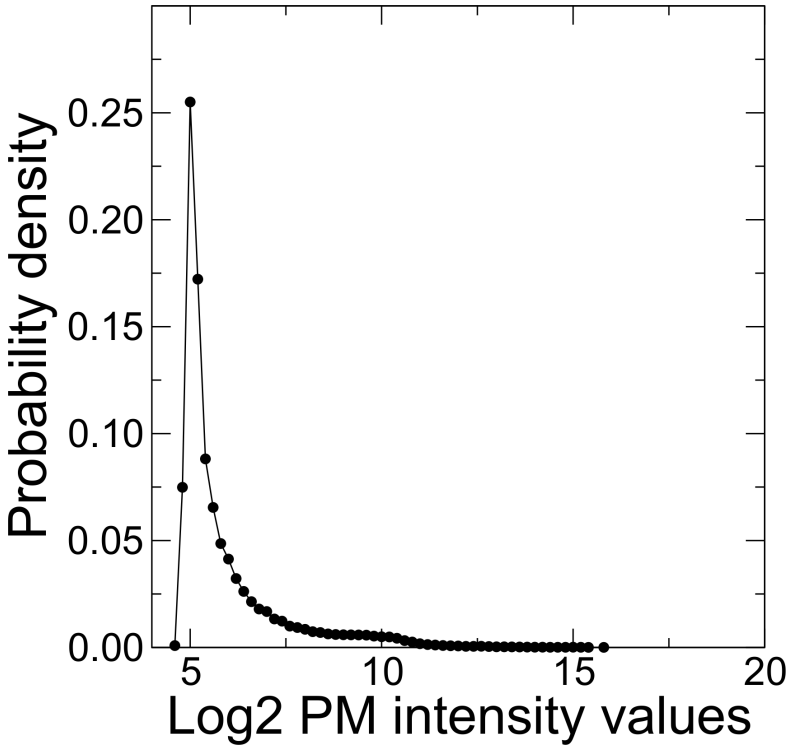


Figure 5.5: Distribution of the PM probe intensities for the array *GSM397756* **form experiment *E-GEOD-15825*** - Since the probe intensity values vary over several order of magnitudes, their values are customarily *Log*2 transformed.

the Gaussian form in Equation 5.2 and α by fitting the right tail of the exponential decay in expression 5.1.

Given the hypothesis above, the scope of the background correction procedure is to find the value of the true intensity S by averaging over all possible values of the noise N given the measured intensity $I = i$, $E(N|I = i)$.

From the definition of $N = I - S$, the joint density function of S and I is:

$$\begin{aligned} f(s,i) &= f_S(i) f_N(i-s) \\ &= \alpha e^{-\alpha s} \frac{1}{\sqrt{2\pi}\sigma} e^{-((i-s)-\mu)^2/2\sigma^2} \\ &= \alpha e^{-(i-\mu)\alpha+\alpha^2\sigma^2/2} \frac{1}{\sqrt{2\pi}\sigma} e^{(s-(i-\mu-\alpha\sigma^2))^2/2\sigma^2}, 0 < s < i. \end{aligned} \tag{5.3}$$

Setting $a = i - \mu - \alpha\sigma^2$ and $c = \alpha e^{-(i-\mu)\alpha+\alpha^2\sigma^2/2}$, which are independent on s, we can compute the marginal distribution of i as:

$$\begin{aligned} f(i) &= \int_0^i ds f_S(s) f_N(i-s) \\ &= c \frac{1}{\sqrt{2\pi}\sigma} \int_0^i ds e^{(s-a)^2/2\sigma^2} \end{aligned} \tag{5.4}$$

By substituting $t = (i-a)/\sigma$ and $dt = dx/\sigma$ and recalling the definition of the cumulative distribution function of the standard Gaussian, $\Phi(x)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} dt e^{-t^2/2} \tag{5.5}$$

for which $\Phi(-x) = 1 - \Phi(x)$, the marginal distribution $f(s)$ is:

$$\begin{aligned} f(i) &= c\frac{1}{\sqrt{2\pi}} \int_{-a/\sigma}^{(i-a)/\sigma} \frac{ds}{\sigma} e^{(s-a)^2/2\sigma^2} = c\frac{1}{\sqrt{2\pi}} \int_{-a/\sigma}^{(i-a)/\sigma} e^{t^2/2} dt \\ &= c\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(i-a)/\sigma} e^{t^2/2} dt - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-a/\sigma} e^{t^2/2} dt \right) \\ &= c\left[\Phi\left(\frac{i-a}{\sigma}\right) - \Phi\left(\frac{-a}{\sigma}\right) \right] = c\left[\Phi\left(\frac{i-a}{\sigma}\right) + \Phi\left(\frac{a}{\sigma}\right) - 1 \right] \end{aligned} \tag{5.6}$$

Hence, the desired quantity, the average value of variable S given a value i for I, is:

$$\begin{aligned} E\left(S|I=i\right) &= \int_0^i s f\left(s|i\right) ds = \int_0^s s \frac{f\left(s,i\right)}{f\left(i\right)} ds \\ &= \frac{1}{f(i)} \frac{c}{\sqrt{2\pi}\sigma} \int_0^i s e^{-(s-a)^2/2\sigma^2} ds \end{aligned} \tag{5.7}$$

By substituting $z(s) = (s-a)^2/2\sigma^2$ and $dz = (s-a)ds/\sigma^2$:

$$\begin{aligned} E\left(S|I=i\right) &= \frac{1}{f(i)} \frac{c}{\sqrt{2\pi}\sigma} \sigma^2 \left(\int_0^i \frac{s-a+a}{\sigma^2} e^{-(s-a)^2/2\sigma^2} ds \right) \\ &= \frac{1}{f(i)} \frac{c}{\sqrt{2\pi}\sigma} \sigma^2 \left[-\left(e^{-z(i)} - e^{-z(0)} \right) + \frac{a}{\sigma^2} \int_0^i \sigma^2 e^{-(s-a)^2/2\sigma^2} ds \right] \end{aligned} \tag{5.8}$$

Finally, by using Eq. 5.4 in the latter expression, we obtain:

$$\begin{aligned} E\left(S|I=i\right) &= \frac{1}{f(i)} \frac{c}{\sqrt{2\pi}\sigma} \sigma^2 \left[-\left(e^{-z(i)} - e^{-z(0)} \right) + \frac{a}{\sigma^2} \frac{f(s)\sqrt{2\pi}\sigma}{c} \right] \\ &= a + \frac{c}{f(i)} \sigma \left(\frac{1}{\sqrt{2\pi}} e^{-a^2/2\sigma^2} - \frac{1}{\sqrt{2\pi}} e^{-(i-a)^2/2\sigma^2} \right) \\ &= a + \sigma \frac{\phi(a/\sigma) - \phi((i-a)/\sigma)}{\Phi(a/\sigma) + \Phi((i-a)/\sigma) - 1} \end{aligned} \tag{5.9}$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ is the density function of the standard Gaussian and, as above, $a = i - \mu - \alpha\sigma^2$.

In most microarray applications $\phi((i-a)/\sigma)$ is negligible and $\Phi((i-a)/\sigma)$ is close to one. So, in practice, it is only necessary to compute the first term in the numerator and the first term in the denominator to make the correction [110, 111].

To summarize, Equation 5.9 gives the value of the true intensity S of the probe, given the measured intensity level i and the parameters α, μ and σ, which can be all estimated from the measured values of PM intensities.
Since the probe intensity values vary over several order of magnitudes, their values are customarily *Log*2 transformed.

As an illustration of the procedure, we show in Figure 5.6 the distribution of the *Log2* transformed probe intensity values before (A) and after (B) the background correction procedure for the 3 arrays of experiment tagged as *E-GEOD-15825*.

5.2.2 Quantile normalization method

Since the aim of a microarray experiment is to compare the expression levels of the genes in different arrays, it is necessary to use a procedure to normalize together different distributions. Since the normalizations based on the maximum (or minimum) values of the distributions could be subjected to statistical fluctuations, it is customary to use the more robust quantile-normalization.

A

Probe	Assay		
	α	β	γ
A	10	3	8
B	3	8	7
C	5	9	6
D	7	2	5
E	3	1	4
F	6	7	3

B

Rank	Assay			Average
	α	β	γ	
I	3	1	3	2.33
II	3	2	4	3.00
III	5	3	5	4.33
IV	6	7	6	6.33
V	7	8	7	7.33
VI	10	9	8	9.00

D

Probe	Assay		
	α	β	γ
A	9.00	4.33	9.00
B	2.33	7.33	7.33
C	4.33	9.00	6.33
D	7.33	3.00	4.33
E	3.00	2.33	3.00
F	6.33	6.33	2.33

C

Rank	Assay			Average
	α	β	γ	
I	2.33	2.33	2.33	2.33
II	3.00	3.00	3.00	3.00
III	4.33	4.33	4.33	4.33
IV	6.33	6.33	6.33	6.33
V	7.33	7.33	7.33	7.33
VI	9.00	9.00	9.00	9.00

Figure 5.6: Example of quantile normalization analysis applied to a small data set - To understand how the quantile normalization procedure works let us to consider (A) a sample of 3 different sets α, β, and γ (*columns*) with 6 entries (*rows*) each, denoted as A, B, C, D, E, and F. (B) The values are, first, sorted within each set and the average expression value for each rank position is computed. (C) Each ranked entries is, next, substituted by the average value at the corresponding rank. (D) The normalized data set is then constituted by 3 sets whose entries follow the same statistical distribution.

In particular, to quantile-normalize two or more data sets, one proceeds following the strategy illustrated in Figure 5.6 for the small data set in panel A and discussed below. First, each set is sorted separately and the arithmetical average is computed for each ranked position (see Figure 5.6)B. Each element is next set to the arithmetical average

of the corresponding ranked position (see Figure 5.6)C). In this way, the highest value in each set becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on. Finally, the original order of the values in each data set is restored, see Figure 5.6D.

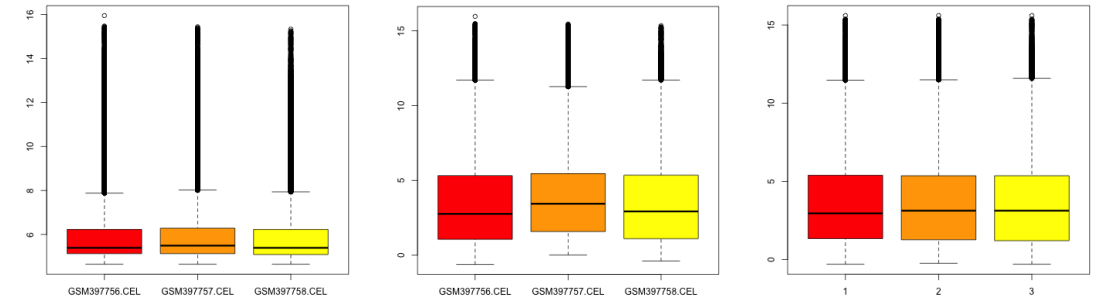


Figure 5.7: Box-whisker plot of the raw probe intensity values (A), after background correction (B), and after the quantile normalization (C) - In each panel, the bottom and the top of the box indicate the values corresponding to the first and third quartile. Those are the values below which there are 25% and 75% of the data. The center line is the median value. The whisker (vertical bar) span the range from the minimum value to the maximum value excluding outliers (see Appendix A). The plotted data refer to the 3 sets of the experiment tagged as *E-GEOD-15825*.

For example, by applying the quantile normalization, to the data of experiment tagged as *E-GEOD-15825*, the distributions of background corrected values for the PM probes in Figure 5.7B, are set in register Figure 5.7C.

5.2.3 Summarization

The background-corrected and normalized probe PM fluorescence intensity values need to be combined so that the intensities of a specific probe set in a given array is conveyed by a single numerical value.

The summarization is based on the assumption that for each probe set, the perfect-match PM probes intensities, Y are decomposed as the linear summation of three terms, each describing the contribution to the intensity value given by different players:

$$Y_{ij} = \mu_i + \alpha_j + \epsilon_{ij} \tag{5.10}$$

where the index i runs over the arrays from 1 to the total number I and j is the probe index ranging from 1 to the number of probes J included in the probe set.

Specifically, in Equation 5.10, μ_i describes the contribution to the total intensity of the probe set (gene) in the array i. Actually, this is the expression value of the probe set, i.e. what we would like to know at the end of the procedure.

The term α_j describes, instead, the contribution that is specific for the probe j. This term is aimed to remove from the intensity values of the probe set μ_i the effect that can possibly come from the different affinity of each probe sequence for the hybridization reaction. It is also assumed that the values of α_j sum to zero, $\sum_j \alpha_j = 0$, for all probe sets, because are expected to be representative (on average) of the associated genes expression without introducing any probe set-specific bias.

Next, ϵ_{ij} represents a uniformly distributed error term with mean 0.

Given the hypothesis above, it is necessary to introduce a robust procedure to find the values of the terms in Equations 5.10.

Customarily, it is applied at this stage the median polish (MP) iterative strategy, which is preferred because it is based on median calculations and, hence, it is robust against outlier values. The MP strategy is an exploratory data analysis procedure proposed by the statistician J. Tukey [112] to find an additively-fit model for data in a two-way table, where each entry of the table Y_{ij} is affected by a global effect m, a row effect r_j, a column effect c_j and a noise term ϵ_{ij}. The final outcome of the procedure are all the values of the *effects*.

The illustration of the various steps of the procedure are discussed in the following:

1. For each probe set, one, first, constructs a matrix of the background-corrected, normalized, and log-transformed probe intensity values Y such that the probes j for $j = 1, \ldots, J$ are in rows and the arrays i for $i = 1, \ldots, I$ are in columns.

$$\begin{pmatrix} Y_{11} & \ldots & Y_{I1} & r_1 \\ \vdots & & \vdots & \vdots \\ Y_{1J} & \ldots & Y_{IJ} & r_J \\ c_1 & \ldots & c_I & m \end{pmatrix}$$

   Initially all rs, cs and m values are set to zero.

2. Next, each row is swept by taking the median across columns (ignoring the last column) subtracting it from each element in that row and adding it to the final column $(r_1, \ldots, r_J, m)$.

3. Then the columns are swept in a similar manner by taking medians across rows, subtracting them from each element in those rows and then adding them to the bottom row of the column effects $(c_1, \ldots, c_I, m)$.

4. The procedure is carried on, by iterating row sweeps followed by column sweeps, until convergence.

At the end of this iterative scheme, one obtains the values of m, c_i and r_j. Besides, the median polish provides the residual ϵ_{ij} in each cell of the table, which tells how far apart that particular cell is from the value predicted by the model.

The estimate of the expression level of the probe set in array i, μ_i, is finally the sum of the whole matrix effect m and the column (array) effect, c_i:

$$\mu_i = m + c_i. \tag{5.11}$$

5.3 Characterization of the gene expression dataset for *D. melanogaster*

Now, I move to present the specific analysis of microarray experimental results I have done for *D. melanogaster* data.

As a first step, I carried an extensive search of gene profiling experiments performed on the model organism, *Drosophila melanogaster* and whose raw data had been deposited on the public database of microarray experiments *ArrayExpress* [90]. Specifically, I considered all the experiments performed on the Affymetrix GeneChip Genome 2.0 (Array Design REF *A-AFFY-35*), which included the larger number of experiments (182) at the date of collection (November 2012). By using the GeneChip Genome 2.0, one can measure the level of expression of 18,952 *probe sets* accounting for all the known genes of *D. melanogaster*.

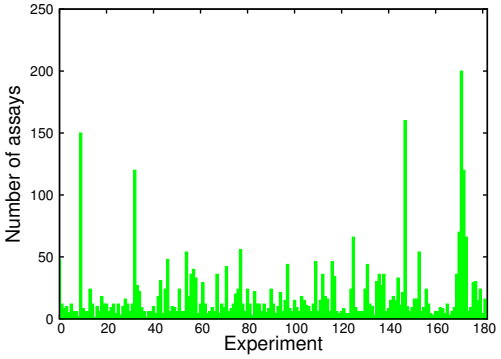


Figure 5.8: Distribution of the number of assays per experiment

Each experiment usually includes several measures of gene expression levels the so-called *assays*. Each assay is, in fact, the account of all the probe expression levels that are measured for a given cell population. It is, in fact, common practice to carry out at least two measures of gene expression levels within the same experiment. One is performed

on the reference (or control) sample and the other on a specific variant of it. The latter could differ for the state of the cell population (e.g. in case of diseases) or for the different treatment of the sample (e.g. different feeding of the fruit fly or incubation conditions of the cells).

The distribution of the number of arrays over the experiments is shown in Figure 5.8 and account for a total of 3,495 assays.

The common feature of the gene expression data set is the fact that they probe the expression level of the same set of *D. melanogaster* genes, which are contained in the chip Genome 2.0. Otherwise the dataset is heterogeneous, because, for instance, the assays pertain to different cell lines and developmental stage of *D. melanogaster*.

In Figure 5.9, I show the weights with which different cell lines are represented in the data set. The database contains experiments which have been carried out on, at least, 3 different cell lines. Specifically, I present: Schneider 2 S2 cells which are derived from a primary culture of late stage (20–24 hours old) embryonic cells; kc cells derived from embryonic cells in the dorsal closure stage and the larval ML-DmDG3-C2 cells.

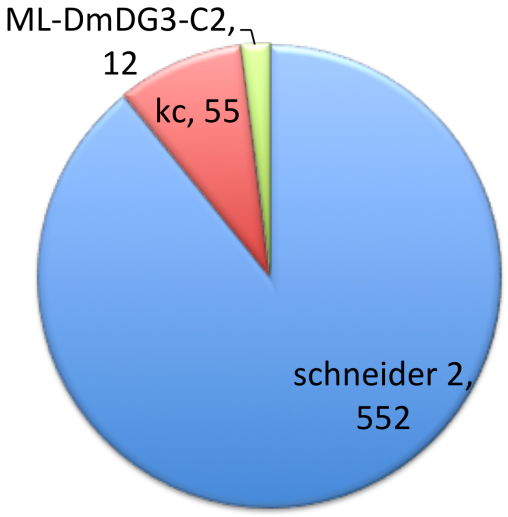


Figure 5.9: Pie-chart describing the occurrence of different cell lines in the considered experiments - The pie-chart shows the number of assays which probe a sample obtained from the indicated cell line. The pie-chart has been constructed by looking at the SDRF file, which is a tab-delimited file describing the relationships between samples, assays, data, and other objects used or produced in a microarray investigation. Specifically, I considered all the experiments in which one or more of the following fields has been filled *characteristics[cell line]*, *characteristics[line]*, *characteristics[strain or line]*, *factorvalue[cell line]*, *factorvalue [line]*, *factorvalue[strain or line]*. Within this categories, I look for the keyword to single out the 3 cell lines: *Schneider 2*, *kc* and *ML-DmDG3-C2*. The large majority of the experiments (2,876) do not provide the explicit tag indicating the used cell line.

A second feature which largely varies in the data set is the developmental stage of the studied *D. melanogaster* sample.

In Figure 5.10, I show an illustration of the main stages of fruit fly development (panel A) and a pie-chart with the distribution of the occurrences of each of them in the data set (panel B). The distribution includes exclusively the subset of $2,317$ arrays for which the developmental stage is specifically indicated in the annotation file.

In particular, the majority of the experiments $(1,363)$ are carried out on adult cells and only a small number (42) on *Drosophila* eggs. Each of the 5 stages, including also embryos, larvae and pupae, is associated to large transformations in the cells of the organism and, in particular, to a complex variation of the gene activity. The expected different gene activity arguably makes the database of gene expression levels heterogeneous.

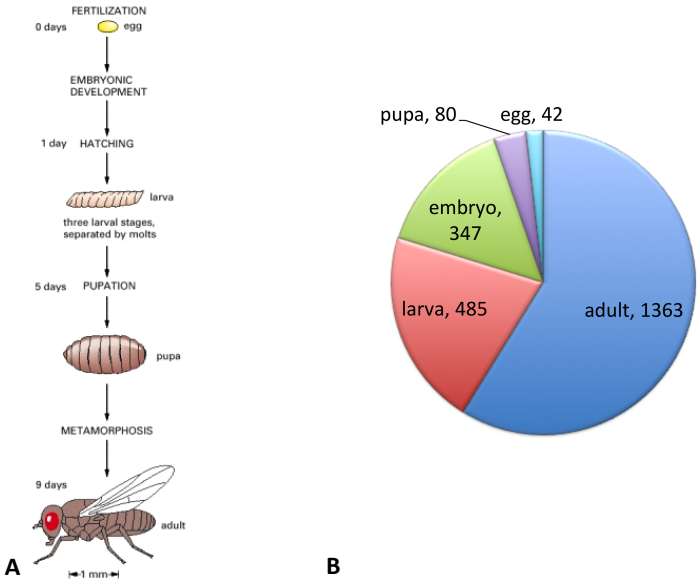


Figure 5.10: Pie-chart of the occurrence of different developmental stages - A sketch of the main developmental stages on *D. melanogaster* is shown in panel (A). The pie-chart in panel (B) shows the number of assays which probe a sample obtained from organisms in the specified developmental stage. The pie-chart has been constructed by looking at the SDRF file. Specifically, we considered all the experiments in which the investigators filled one or more of the fields *characteristics[developmental stage]*, *characteristics[stage]*, *factorvalue[developmental stage]*, *factorvalue [embryonic stage]*, *factorvalue[stage]*. Within this categories, we look for some keyword to capture each developmental stage: *adult* for adult fruit flies; *embryo* or *stage*, or the number of the stage for embryos, *larva* or *instar* for organism in the larval stage, *pupa* and *egg*.

The heterogeneity of the gene expression data set should be considered in connection with the analysis that I will carry out in the present chapter. As I said in section 4.1.2, I use mutual information content to measure the correlation between pairs of gene expression patterns, which we showed to be mainly suitable to capture correlations between largely varying gene expression pattern. Given the heterogeneous set of assays considered, the gene expression pattern variability is expect to increase and produce a final data set of patterns that is suitable for mutual information analysis.

The data set heterogeneity should also be considered in connection with the aim of this chapter, that is, as stated above, to compare the information about gene coregulation (mutual information) with the HiC-based information about contacting chromosome loci in ref. [16]. In this regard, I recall that HiC data on *D. melanogaster* have been collected only for a specific developmental stage (embryos) of *D. melanogaster*. This difference in data source should be considered when discussing the obtained results at the end of the chapter.

5.4 Mutual information analysis

After having retrieved the gene expression data, I normalized them by using the RMA strategy presented in section 5.2. The implementation of the analysis has been carried out by using the standard Bioconductor routines in R [113]. Next, I proceed to perform the correlation analysis on the gene expression patterns, similarly to what I have presented in section 4.1.2 of this Thesis.

Specifically, I started by filtering the $18,952$ *probe sets* of GeneChip Genome 2.0 of *D. melanogaster* to single out only those $11,882$ probe sets which exclusively target a single sequence (i.e. an interrupted stretch of the genome), spanning a maximum of 3 kilobasepairs (kbp) in length. This choice prevents from having very long probe sets, whose expression signal can be non-specific.

The set of considered probe sets contains stretches of length between 120 and $2,990$ nucleotides (nt). The distribution of the number of probe sets as a function of the length of the probed genes is shown in Figure 5.11.

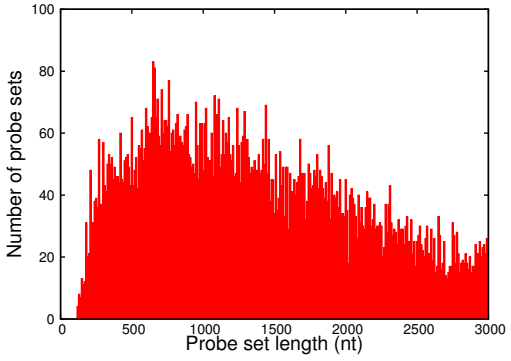


Figure 5.11: Distribution of probe set lengths - Each bin spans 10 nucleotides (nt).

I checked that, *a posteriori*, the reduced set of $11,882$ probe sets offers an almost complete probing (98%) of the regions of *D. melanogaster* genome, see Figure 5.12.

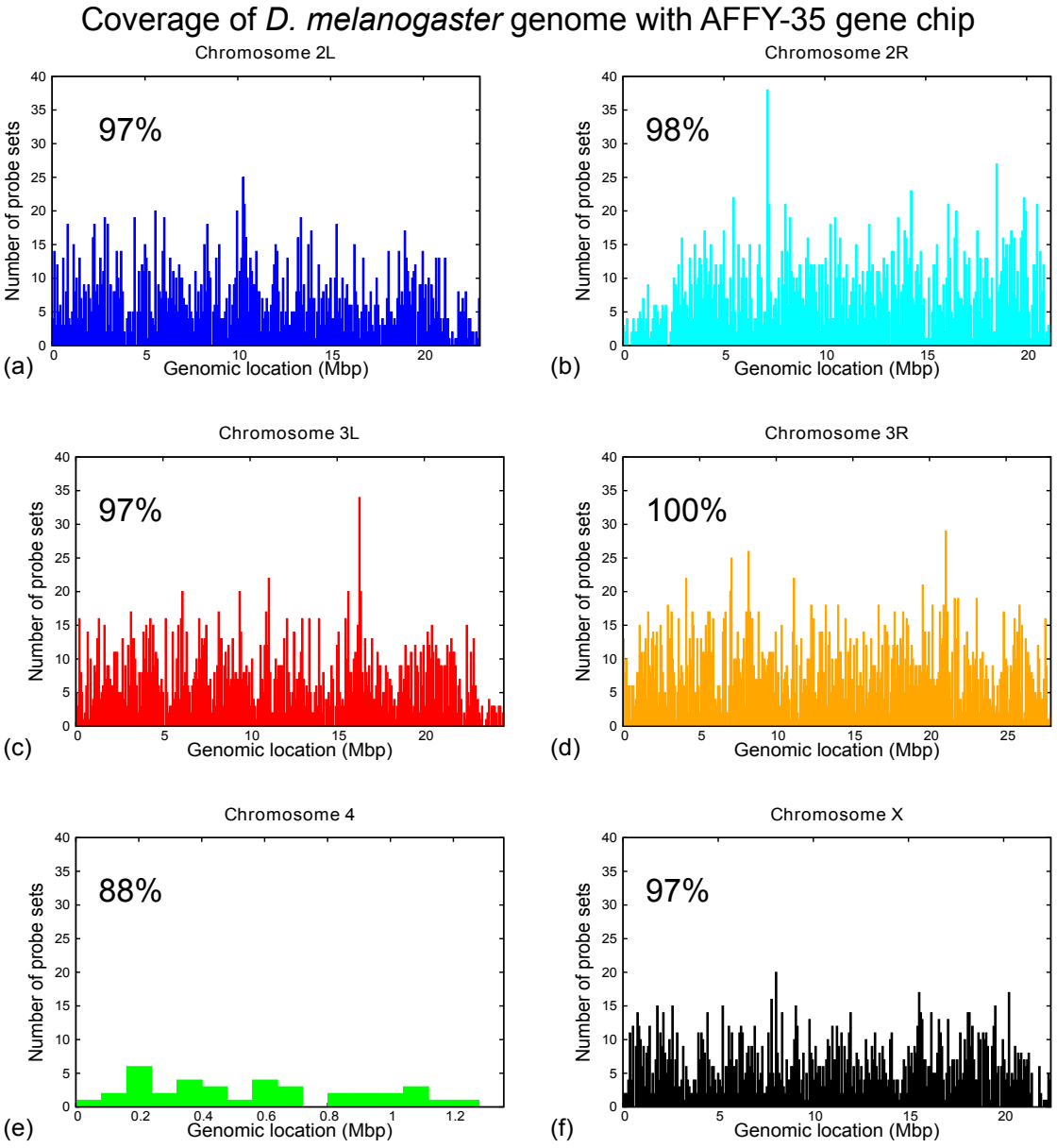


Figure 5.12: Coverage of the genome by the restricted probe set ensemble - The number of probe sets insisting on each stretch of 80kbp is presented for *D. melanogaster* chromosome. The probes span a large fraction of each chromosome arm from the 88% of chromosome 4 to the complete (100%) chromosome arm $3R$. The probe sets account for a coverage of 98% of the whole *D. melanogaster* genome. Since the centromeric regions are not probed by the microarray chip (neither by HiC experiments) they are excluded from our analysis.

Next, to perform a robust comparison of the heterogeneous gene expression profiles we coarse-grain all expression levels within each of the 182 experiments to one of three discrete states only: low, medium and high, as done in ref. [80]. The three levels contains 33% of the expression levels each. For each possible probe set pair, I and J, I next computed the mutual information content (MI) of the expression profiles:

$$\mathrm{MI}_{IJ} = \sum_i \sum_j \pi_{ij} \ln\left(\frac{\pi_{ij}}{\pi_{i+}\,\pi_{+j}}\right) \tag{5.12}$$

where i [j] runs over the three coarse-grained expression levels for probe set I [J]. I recall that in Eq. 5.12, π_{ij} is the joint probability that expression levels i and j are respectively observed for probe sets I and J over the experiments, while the quantities $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are the marginal probabilities.

5.5 Pairwise gene coregulation and colocalization: analysis of HiC and gene expression data

In this section, we present the results of the comparison between the gene coregulation analysis above and gene colocalization characterized by recent HiC experiments [16].

We recall that HiC experiments involve the extensive, non-specific fixation of genome-wide chromosome contacts by formaldehyde, followed by the pinpointing of cross-linked pairs of loci by using high-throughput sequencing [9, 76]. Since HiC data are collected over a large ensemble of cells [9], they provide quantitative insight into the most typical (average) *intra-* and *inter*-chromosome contacts patterns. The HiC data used in this study are collected by Sexton *et al.* on *D. melanogaster* embryo cells [16] at 80 *kbp*-resolution. The typical intra-chromosome pairwise contact propensity of fruit fly chromosomes is illustrated in the left column of Figure 5.13.

For comparison with HiC contact matrices at 80 kbp of resolution [16], we grouped together the measures of MI involving pairs of probe sets within the same regions of 80 *kbp* ($\sim$ 800nm of chromatin contour length) and to pick up the average values of those. This procedure results in $1,475$ distinct segments spanning 98% of the total extension of the *D. melanogaster* chromosome arms (excluding centromeric regions) and $1,088,550$ distinct mutual information measures between all the possible pairs of $80kbp-$strands. The MI content for each *D. melanogaster* chromosome is shown in Fig. 5.13 (*right column*).

The HiC matrix has a noticeable block character reflecting the much higher incidence of intra-arm contacts with respect to inter-arm ones. Furthermore, within each arm the contact probability decreases fairly uniformly as one moves away from the diagonal. This is equivalent to stating that the HiC-averaged contact propensity of two loci decays monotonically with the loci sequence separation.

It is striking to see that the noticeable sequence-separation modulation of the HiC matrix has no manifest counterpart in the mutual information one. The latter, in fact, has a more diffuse and scattered character, with its top entries distributed over a wide range of sequence separations.

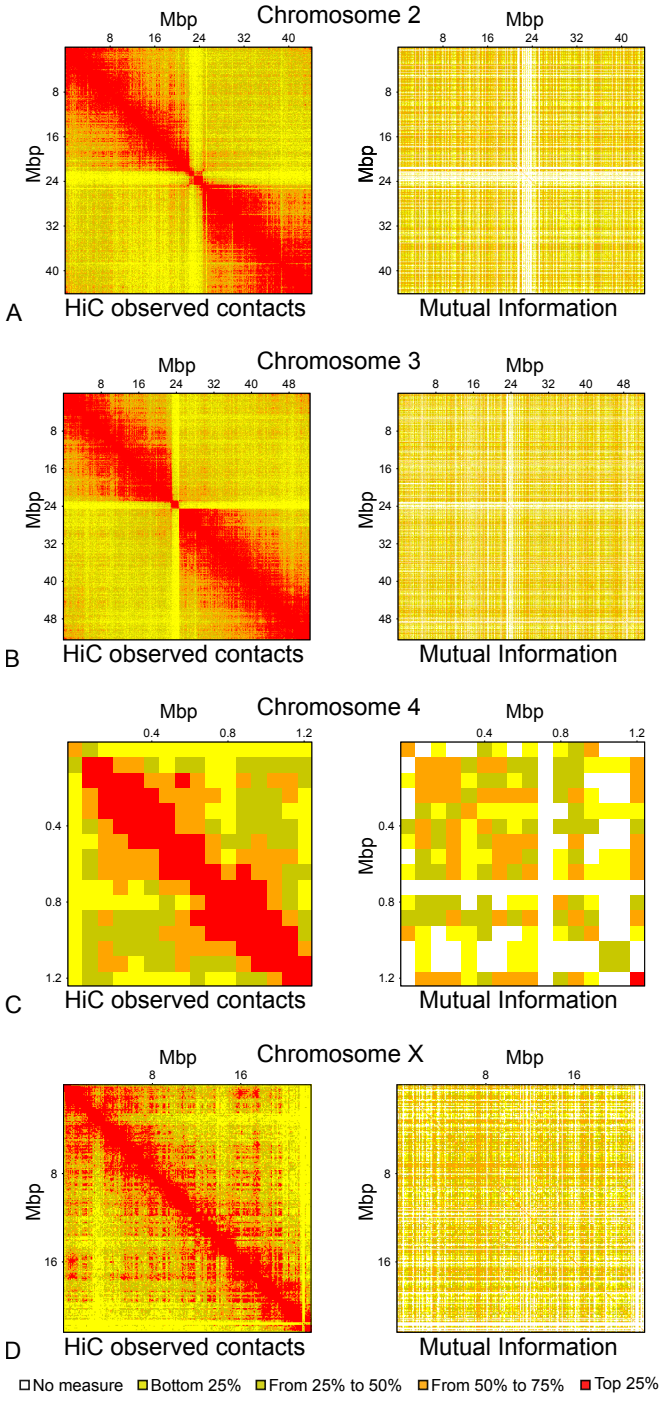


Figure 5.13: HiC contact matrix (*left*) and MI matrix (*right*) for *D. melanogaster* **chromosomes.** - The color code reflects various percentile levels of the ranked entries in the two matrices, see legend.

The striking visual difference of the two matrices in Fig. 5.13 contrasts with what previously observed for specific human chromosomes, particularly chromosomes 19. For such chromosomes, in fact, the connection of HiC and MI matrices was sufficiently strong that noticeable qualitative similarities emerged from the "block-correlation" analysis (also termed plaid pattern analysis) of the two matrices [80].

To go beyond the overall qualitative assessment that would be obtained from a block correlation analysis, we characterized the correlation between colocalization (HiC) and coregulation (MI) data by adopting the Spearman rank test (see Figure 5.14). This is a non-parametric test and can be aptly used to compare inhomogeneous quantities, such as MI and HiC data, because it is insensitive to their shifts or arbitrary monotonic transformations (log, power, additive constant etc.) of either or both variables [30].

Specifically, we rank the whole set of $80kbp$-long pairs of chromatin fragments for increasing values of HiC reads and MI, and then measure the linear correlation coefficients between the two ranks for various fragment pairs. To minimize the incidence of correlated entries due to the chain connectivity constraint, the rank-correlation coefficient is not computed across all fragment pairs, but only on subsets of them. More precisely, we picked randomly $10,000$ sets of $1,000$ pairs each and compared the resulting distribution of the correlation coefficients to the null one, obtained by picking an equal number of non-corresponding entries from the two matrices.

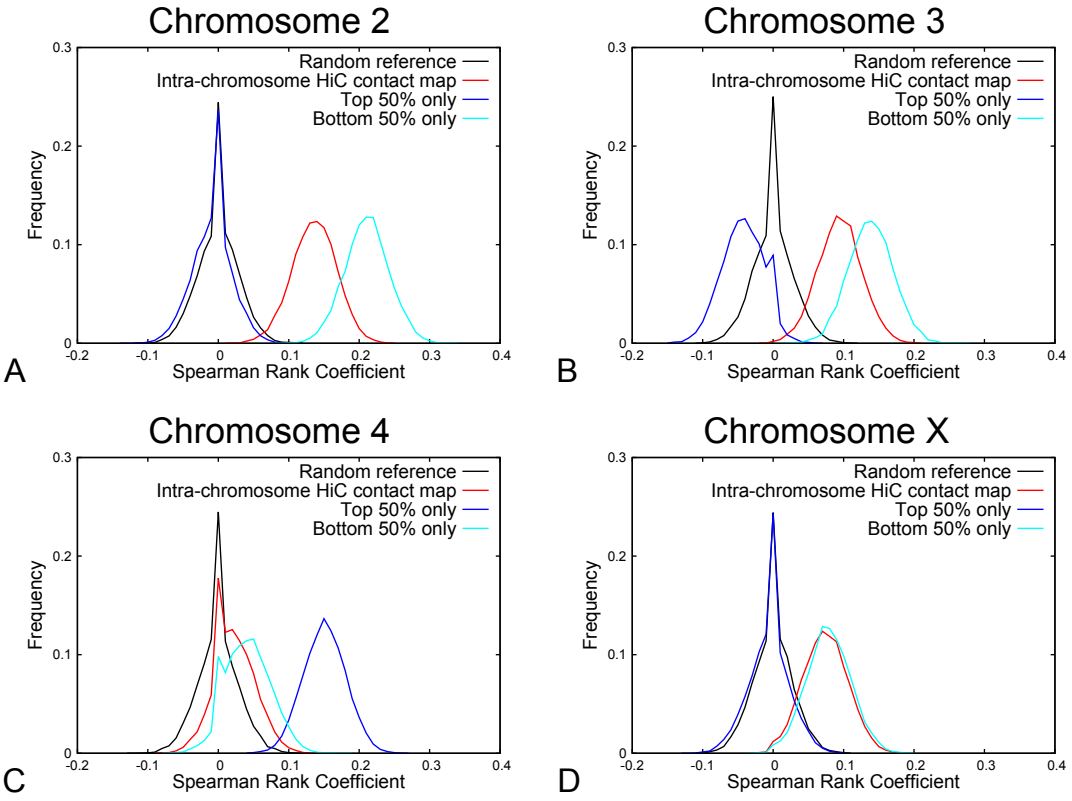


Figure 5.14: Distributions of Spearman rank correlation coefficient calculated over various sets of corresponding HiC and MI entries. Shown data pertains to all chromosomes. Notice that the highest correlation is observed for the low-ranking entries in the two types of matrices.

The results are shown in the panels of Figure 5.14 all chromosomes (*red and black (random reference) curves*). They show that for all chromosomes the random reference distributions are centered around zero and have a large peak around the central value.

The distributions computed on corresponding entries of the intra-chromosome maps are, instead, shifted towards positive values, indicating a positive correlation between MI content and HiC contact propensity. The only exception is chromosome 4 (in panel C) which span a small portion ($\sim$ 1.2Mbp) of the genome.

Furthermore, the distributions of chromosomes 2, 3 and X are clearly distinguishable from the null distributions with only small overlaps. This evidence indicate that the correlation between HiC and MI entries, although not immediately apparent from the visual inspection of the matrices in Fig. 5.13, is still significant from the statistical one.

We, further, investigate this point by picking the corresponding entries exclusively within the top or bottom 50% of the HiC map entries. This analysis is meant to clarify if the compliance between coregulation and colocalization is mainly due to the fact that significantly coregulated gene pairs (high MI) are prone also to be in contact (high number of HiC observed contacts), or vice versa.

The data shown in the panels of Figure 5.14 with *blue (top 50%)* and *cyan (bottom 50%) curves* indicate that the correlation between high MI and high HiC entries is marginal, because the Spearman rank correlation distributions have large overlaps with the null distributions for all the chromosomes. In fact, the significant correlation mostly originates from the fact that entries with low MI tend to be far apart (low HiC).

The fact that high mutual information values do not correspond to high HiC entry can arguably reflect the different sources of the data for the gene expression and the colocalization analysis. I recall, in fact, that the data set of the gene expression values, on which the coregulation is based, is heterogeneous, see section 5.3, while the HiC experiment has been done, instead, on a sample of embryonic cells.

5.6 Summary

We recall that the investigation of the gene coregulation/colocalization hypothesis presented in chapter 4 was articulated over several steps. In particular, the first part of our approach included the extensive collection of gene expression data from public experiment repositories and statistical analysis to single out significant coregulated gene pairs. Next, the pairs of coregulated genes were used as spatial restraints in knowledge-based numerical simulations in a three-dimensional model of chromosomes.

In this chapter, we discussed in more detail the first part of our gene pair coregulation/-colocalization approach. Specifically, we reported on the data mining and processing strategies followed to obtain an extensive and consistent data set of gene expression for

D. melanogaster. The data were next used to identify significantly coregulated pairs by computing the mutual information content between all the possible pairs of genes in the data set.

Next, we addressed the relationship of gene pairs coregulation and colocalization for *D. melanogaster* with a complementary approach with respect to chapter 4. In particular, we used the robust Spearman rank statistical correlation analysis to compare directly the mutual information content of the gene pairs, (which is a measure of coregulation) with the established gene contacts observed in recent HiC experiments on *D. melanogaster* [16] (which is a measure of colocalization).

Interestingly, this test showed that the consistency of gene pairs coregulation/colocalization, while it is not readily noticeable at the visual inspection of the MI and HiC matrices, is nonetheless statistically significant. In particular, we found that the pairs with low contact propensity are typically not coregulated. The vice versa, however, is not necessarily true.

The results point to a less stringent relationship of gene coregulation and colocalization in *D. melanogaster* with respect to what was found in human chromosome 19. This difference may be due, at least in part, to the fact that the high resolution colocalization (HiC) data were collected for embryo cells, which may have a more diffuse and hence atypical organization than adult cells for which most of the coregulation data were obtained.

This interesting coregulation/colocalization relationship can be followed along several directions for future investigations. For instance it will be natural to repeat the analysis for human chromosome 19 (see chapter 4) by using three-dimensional computer models and knowledge-based simulation.

Chapter 6

Charting chromosome territories with constraints based on HiC data

In chapter 4, we investigated the spatial organization of human chromosome 19 by building on the gene-kissing hypothesis. Specifically, we obtained putative three-dimensional conformations for this chromosome by enforcing the colocalization of significantly coregulated gene pairs. Finally, we compared the structural features of the obtained models against available HiC data showing significant similarity of the organization in spatial macrodomains.

However, as we pointed out in the previous chapter, there are potential limitations to the general use of gene coregulation as a proxy of gene colocalization for all chromosomes and all cell lines of a given organism. By studying gene coregulation/colocalization in *Drosophila melanogaster*, we showed, for instance, that the most significantly coregulated gene pairs (high mutual information content) do not necessarily correspond to colocalization contacts measured with HiC.

Here, we present a study on human chromosomes where we tried to overcome such limitations by using knowledge-based constraints that makes direct use of HiC data. The study was carried out in collaboration with group of E. Hovig in Oslo who complemented our expertise on chromosome modelling with state-of-the-art statistical analysis of significant HiC contacts as well as the preference of specific chromosome regions to be located in the central or peripheral regions of the nucleus.

6.1 Significant HiC contacts in the hESC cell-line

As mentioned above, the significant HiC contacts for the human chromosomes, were identified by J. Paulsen, T.G. Lien and E. Hovig by using their powerful statistical analysis technique introduced in ref. [114].

Specifically, the starting point of such analysis was the retrieving of the recent HiC contact matrix obtained on human embryonic stem cells (hESC) at resolution 100kbp [17]. As customarily, the contact map was adjusted for technical bias using the method of Imakaev *et al* in ref. [115]. Finally, the HiC contact data set was restricted to the intra-chromosome contacts only because of the expected high incidence of false positives in inter-chromosome contacts.

Reflecting the fractal-like organization observed for eukaryotic chromosomes [9, 11, 95–99], the intra-chromosome HiC contact maps were characterized by a large number of entries equal to zero. The statistical analysis on them was, accordingly, carried out by modeling the data set with the *zero inflated negative binomial (ZiNB)* [116] distribution. The latter is, in fact, a statistical model based on the usual binomial distribution extended for negative values, which assigns particular weights to the entries equal to zero.

The analysis allowed to single out intra-chromosome contacts in the set where the expected *false discovery rate* (FDR) was less or equal than 1% across the entire set of chromosomes. This means that in the set of significant HiC contacts at most one entry every 100 could be on average a false positive.

The set of significant contacts accounted for a total of $14,928$ pairs of 100kbp chromosome regions. The contacts per each chromosome range from 39 for chromosome 13 to $1,475$ for chromosome 1. The summary of the detailed number of significant pairs of loci, N_{tp}, is shown in the rightmost column of Table 6.1.

6.2 Modelling chromosome structure and dynamics

The feasibility to colocalize simultaneously the significant HiC contacts was explored by using coarse-grained model chromosomes and steered molecular dynamics simulations. The latter are described in the following sections.

Chr	Length (Mbp)	N_b	Centromeric beads	N_{tp} for hESC FDR=1%
Chr1	249.3	82,269	40,095–42,537	1,475
Chr2	243.2	80,256	29,865–32,944	1,390
Chr3	198.0	65,340	29,007–30,987	214
Chr4	191.2	63,096	15,906–17,391	161
Chr5	180.9	59,697	15,213–16.731	296
Chr6	171.1	56,463	19,371–20,889	304
Chr7	159.1	52,503	19,140–20,361	1,464
Chr8	146.4	48,312	14,223–15,873	453
Chr9	141.2	46,596	15,609–16,731	1,187
Chr10	135.5	44,715	12,540–13,959	631
Chr11	135.0	44,550	17,028–18,381	597
Chr12	133.9	44,187	10,989–12,606	249
Chr13	115.2	38,016	5,379–6,435	39
Chr14	107.3	35,409	5,313–6,303	124
Chr15	102.5	33,825	5,214–6,831	436
Chr16	90.4	29,832	11,418–12,738	1,152
Chr17	81.2	26,796	7,326–8,514	642
Chr18	78.1	25,773	5,082–6,270	48
Chr19	59.1	19,503	8,052–9,438	1,311
Chr20	63.0	20,790	8,448–9,702	46
Chr21	48.1	15,873	3,597–4,719	76
Chr22	51.3	16,929	4,026–5,907	143
ChrX	155.3	51,249	19,173–20,790	2,490
Total	3,036.3	1,001,979	/	14,928

Table 6.1: Summary of relevant data for the chromosome models - The first, second, third and fourth columns indicate respectively: the chromosome index, its length in Mbp, the corresponding number of beads in the coarse-grained model, N_b, and the range of beads pertaining to the centromeric regions. The latter as been determined according to the hg19 human genome assembly [117]. The remaining column indicates the number of target pairs of loci to be colocalized, N_{tp}, for the cell line hESC for $FDR = 0.01$. The cases highlighted in yellow are discussed in details in this chapter.

6.2.1 The chromosome polymer model

As for the case discussed in chapter 4, the various human chromosomes were modelled as semi-flexible chains of beads with thickness $\sigma = 30$nm. The chains were endowed with a bending rigidity appropriate to reproduce the chromatin nominal persistence length, $l_p = 150\,nm$ [11].

Each model chain accounted for the total contour length of the corresponding human chromosome, as summarized in the second column of Table 6.1. The contour length length of each chromosome was reported on model chains by mapping each region of 100kbp of chromatin fiber onto 33beads. Hence, each bead spans $\approx 3,000$ base

pairs [118]. The resulting number of beads, N_b, associated to each of the 23 model human chromosomes is given in the third column of Table 6.1.

Each chain was described with the Kremer and Grest polymer model [24] presented in Section 1.2.1:

$$\mathcal{H}_{\mathrm{KG}} = U_{\mathrm{FENE}} + U_{\mathrm{KP}} + U_{\mathrm{LJ}}. \tag{6.1}$$

We recall that the three terms, which act within each chromosome chain, correspond to the FENE chain-connectivity interaction, the Kratky-Porod bending energy, and the pairwise Lennard-Jones excluded volume term. The latter term controlled also the inter-chain excluded volume effect.

Here, the parametrization of the three terms of the Hamiltonian 6.1 differs from the standard Kremer and Grest one of Section 1.2.1. Specifically, we set the strengths of the FENE and of the Lennard-Jones potentials respectively to $300\epsilon/\sigma^2$ and 10ϵ for the nearest neighbor beads interactions. As will be discussed later, this choice was found to avoid the bond overstretching, which might otherwise result in unphysical crossings of the chain strands. We also checked *a posteriori* that these changes do not affect the average bond length of the polymer chain, which remained $\sim 1\sigma$.

Furthermore, to account for the compact aspect of the centromeric chromatin shown in imaging experiments [7], the centromeric regions, whose locations in the model chains are shown in Table 6.1, were compactified (before steering) by introducing a nonspecific Lennard-Jones attractive interaction between all pairs of beads. In particular, we accounted for the attracting part of the LJ potential by increasing the cutoff radius to $r_c = 2.5\sigma$, where the interaction strength is about 1/60th of its minimum value.

6.2.2 Description of the free chain dynamics

The steered molecular dynamics of the chromosomes was modeled with a two steps strategy. Namely, the free dynamics of the chains was described with an underdamped Langevin equation, while the steering process was guided by using pairwise harmonic constraints, analogously to what described in Chapter 4.

Specifically, the underdamped Langevin equation, see Section 1.3, was:

$$m\ddot{r}_{i\alpha} = -\partial_{i\alpha}\mathcal{H} - \gamma\dot{r}_{i\alpha} + \eta_{i\alpha}(t) \tag{6.2}$$

where m is the bead mass which was set equal to the LAMMPS default value, $\mathcal{H}$ is the system energy in Equation 6.1, i is runs over all the particles in the system, and $\alpha =$

(x, y, z) indicates the Cartesian coordinate. The random term $\eta_\alpha(t)$, whose definition was given in Section 1.3, has the statistical properties $\langle \eta_{i\alpha} \rangle = 0$ and $\langle \eta_{i\alpha}(t)\, \eta_{j\beta}(t') \rangle = 2\kappa_B T \gamma\, \delta_{ij}\delta_{\alpha\beta}\delta(t-t')$, where k_B is the Boltzmann constant, T the temperature, δ_{ij} the Kronecher delta and $\delta(t-t')$ the Dirac delta .

The Langevin equation was integrated numerically with the LAMMPS molecular dynamics software package [29] with an integration time step equal to $\Delta t = 0.006\tau_{MD}$, where $\tau_{MD} = \sigma(m/\epsilon)^{1/2}$ is the Lennard-Jones time.

6.2.3 Steered molecular dynamics protocol

The colocalization of the significant HiC pairs of 100 *kbp*-long chromosome stretches was promoted by using a steered molecular dynamics protocol which progressively favoured the spatial proximity of the target pairs in each model chromosome.

Specifically, we mapped each pair of selected regions, A and B, onto the corresponding 33beads long stretches of the chromosome chain and added to the system energy an harmonic constrain:

$$U_{\mathrm{H}} = \frac{1}{2} k\,(L,t)\, d_{A,B}^2 \tag{6.3}$$

where $d_{A,B}$ is the distance of the centers of mass of the chromosome stretches. The stiffness of the harmonic constraint was controlled by the spring constant $k\,(L,t)$, which is dependent on the sequence separation between the two regions, L, and the time t.

In fact, to avoid having the steering process being dominated by the target pairs at the largest sequence separation we did not use a single spring constant for all target pairs. Rather we made the spring constant dependent on the sequence separation L of the target pairs so that, in the initial mitotic-like state, all pairs are pulled together with the same average force despite of their sequence separation.

To accomplish this balancing of the spring constant, we used a statistical reweighting approach. In particular, we computed the distributions of the spatial distances between stretches of 100kbp in the initial conformation before steering (which will be discussed in detail in section 6.3) for different values of L and found the values of the spring constants k to provide the harmonic energy necessary to bring 90% of the pairs below the contact radius.

Specifically, we computed the distribution of the square spatial distances between all the pairs of 100kbp-long chromosome stretches. We subdivided them in 5 groups with the first, second, etc. group gathering pairs at genomic distances in the $0-15$Mbp,

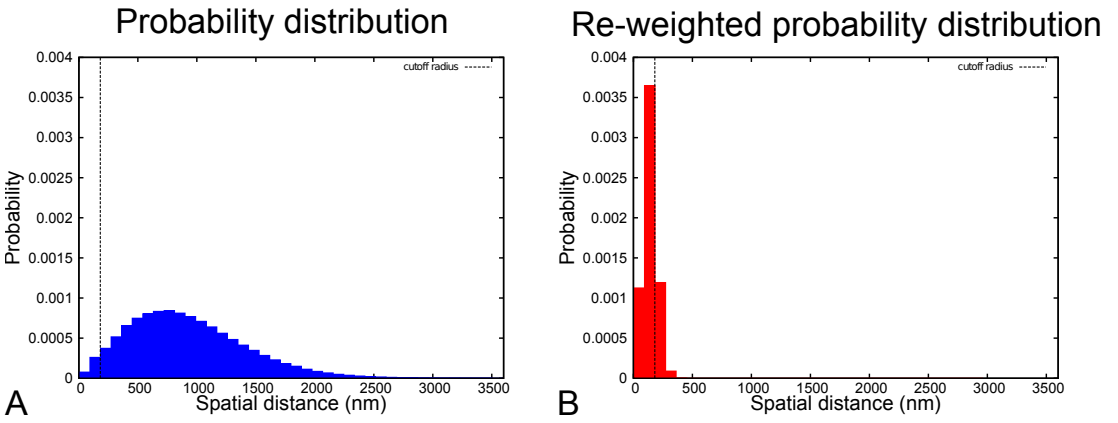


Figure 6.1: Initial (A) and re-weighted (B) distributions of the spatial distances for chromosome strand pairs in group 1.

$15-30$Mbp ranges , etc. Given the typical length of human chromosomes of ~ 43Mbp, all the entries above 60Mbp are included in the last 5th group (see Table 6.2). Within each group, we computed the normalized distribution of the spatial distances, d, between al the pairs of 100kbp chromosome strands in the conformation of the human haploid system before steering. The latter distribution for bin 1 is shown in Figure 6.1A.

The gathered distance statistics was next reweighted with a Gaussian function:

$$w\left(k\right) = e^{-k/2(d-d_0/2)^2}$$

where $d_0 = 180$nm is the cutoff radius for defining a contact between target pairs. By doing so, we find the value of the spring constant k to yielding at least 90.0% of the pairwise distances below the contact radius. The re-weighted probability distribution for bin 1 is shown in Figure 6.1B.

The obtained values of the spring constants are shown in Table 6.2 for each of the 5 bins. To start with an even milder steering, we set the initial spring constants to 10% of those values.

Bin	**Range of sequence separation (Mbp)**	**Spring constant for** 90% ϵ/σ^2
1	0 -15	0.16059
2	15-30	0.45363
3	30-45	0.32675
4	45-60	0.28126
5	60 or larger	0.14944

Table 6.2: Spring constant values to yielding at least 90.0% **of the pairwise distances below the contact radius.**

The simultaneous application of the N_{tp} constraints to each chromosome was implemented by using the PLUMED plugin for LAMMPS [100]. The spring constants were gradually ramped linearly every $10^3\Delta t$ of steered simulation so to avoid driving the system significantly out of equilibrium: $k\,(L,t) = k(L,0)t/(1000\Delta t)$ for each value of L. Moreover, the resulting pulling force between each constrained pair was controlled every $1000\Delta t$ and, if it exceeded a maximum pulling force of $300\,\epsilon/\sigma$, we set it to this maximum value.

This maximum force was low enough and the simulation time-step Δt short enough to avoid appreciable overstretching of the bond connecting the beads, as this may result in unphysical passages of the strands through each other during a numerical integration time step.

The typical sequence separation of the target pairs was $\sim 43\,Mbp$ which was a sizable fraction of the chromosomes' length and was also much larger than the contour length usually associated to chromosome domains [17]. One thus could expect that bringing the target pairs into spatial proximity requires overcoming of very significant entropic barriers. It is therefore not obvious *a priori* that one can simultaneously satisfy most of the target colocalization constraints.

6.2.4 Calculation of the cutoff for chromosome contacts.

Before accounting for the results of the steering protocol, which will be discussed in section 6.3, I shall introduce the criterion we used to determine the contact radius for two constrained chromosome stretches.

Specifically, we considered the expression of the mean square gyration radius R_g^2, which has been established by Benoit and Doty in ref. [119] for a worm-like chain (WLC):

$$\left\langle R_g^2\,(M)\right\rangle = \frac{M\,l_k^2}{6} - \frac{l_k^2}{4} + \frac{l_k^2}{4\,M} - \frac{l_k^2}{8\,M^2}\left(1 - e^{-2M}\right) \tag{6.4}$$

where L_c is the contour length of the chain, l_K is the Kuhn-length and $M = L_c/l_K$ is the number of Kuhn-lengths.

A heuristic use of expression 6.4 is to estimate the effective size of the region occupied in equilibrium by portions of contour length L_c from a long polymer chain. We therefore considered the occupied region to be spherical, centered on the center of mass of the segment and with radius equal to $\sqrt{\left\langle R_g^2\,(M)\right\rangle}$.

The criterion to define an established spatial contact between a pair of segments of contour length L_c should be based on the volume accounted by the overlap of the two

spheres spanned separately by the two stretches. In particular we considered, in this chapter, any positive overlap of the volume of each individual sphere. Accordingly, we set the contact radius r_c being equal to the square root of the gyration radius.

For the chromosome chains studied here, the contour length L_c of the stretches to colocalize accounts for 100kbp which map onto $\sim 1\mu m$ [11]. Given the Kuhn-length of the chromosome fiber $l_K = 300\,nm$ [11], M results to be equal to ~ 3.3. This corresponds to a mean square gyration radius from Equation 6.4:

$$\left\langle R_g^2\,(3.3)\right\rangle \simeq 32786.5nm^2 \tag{6.5}$$

which gives a contact radius:

$$r_c = \sqrt{\left\langle R_g^2\,(3.3)\right\rangle} \sim 180nm$$

As a term of comparison, by measuring the same quantity on conformations of the simulated chromatin fiber (study on human Chr19 presented in chapter 4), we obtained:

- $159.9\,nm$ for the partially-decondensed (mitotic-like) state,
- $163.5\,nm$ for the more decondensed state,
- $173.4\,nm$ for the conformation at the end of the steering protocol.

which are in fair agreement with the approximate theoretical estimate.

6.3 Human haploid chromosome system with periodic boundary conditions

As a first step, we modelled 23 human chromosomes, accounting for the human haploid chromosome system. Specifically, each chromosome was initially prepared in a rod-like arrangement to mimic the mitotic shape, as done in [11, 88] and in chapter 4 of this thesis. Each rod corresponded to a cylindrical solenoidal arrangement. The 23 rod-like chromosomes are next placed in a cubic simulation box with periodic boundary conditions. The relative positions of the cylinders are initially picked in a random, though non-overlapping, manner inside a fairly large cubic simulation box, see Figure 6.2A. The initial conformations are allowed to relax with a free dynamical evolution except for the beads of the centromere region which are subjected to the nonspecific attractive interaction presented above to promote their compactness. During this evolution, the

side of the simulation box is also progressively shortened down to $\sim 6.3\,\mu m$, so to attain the typical nuclear density, $\sim 0.012 bp/nm^3$, see Section 4.2.3.

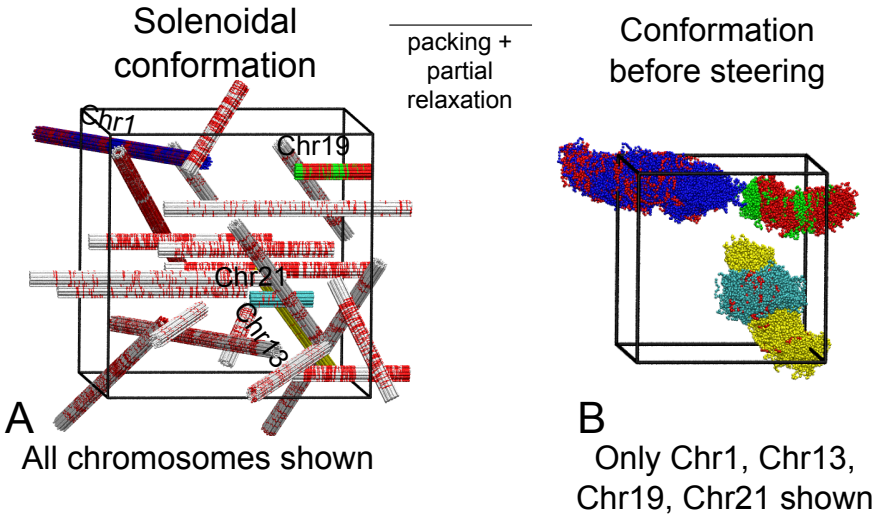


Figure 6.2: Initial conformation of the haploid system - **A** The 23 chromosomes of the human haploid system are arranged in a random, though non-overlapping manner in a cubic box. Periodic boundary conditions apply. Each molecule is prepared in a rod–like hierarchical solenoidal arrangement. Specifically, each major turn of the solenoid is constituted by a finer rosette-like arrangement of the chain (12 loops each 50kbp long [11]). **B** The chromosomes are next relaxed from the initial arrangement and reach an elongated mitotic–like structure. The confining box is simultaneously shrunk until the side is $\sim 6.3\,\mu m$ long for which the typical nuclear density is achieved. The red beads in panels A-B mark the regions involved in the target pairings. In panels B, we show only four chromosome Chr1, Chr13, Chr19 and Chr21 for visual clarity.

The resulting arrangement of the 23 chromosome system is shown in Fig. 6.2B and is next used as the initial conformation for the steering procedure.

In the following, I show the results of the steering molecular dynamics protocol applied to the haploid system to promote the spatial proximity of target pairs corresponding to significant HiC contacts.

In particular, I discuss the results of the steering process for four chromosome that are of particular interest for their size or richness in target pairs:

- Chr1, which is the longest chromosome;
- Chr21, which is the shortest one;
- Chr19, which has the highest density of target pairs per unit length;
- Chr13, which has the lowest density of target pairs per unit length.

6.3.1 Colocalization of the significant HiC pairs

The compliance of the system to the steering protocol is illustrated, in Figure 6.3 which shows the increase of the *fraction of established contacts*, during the simulation time for the haploid system.

It is striking to observe that it is possible to simultaneously colocalize about 30% of the all target contacts within a contact cutoff distance of 180nm (Fig. 6.3A). The conformations reached at the end of the steering protocol are shown in Figure 6.3B for the previously-mentioned selected set of chromosomes.

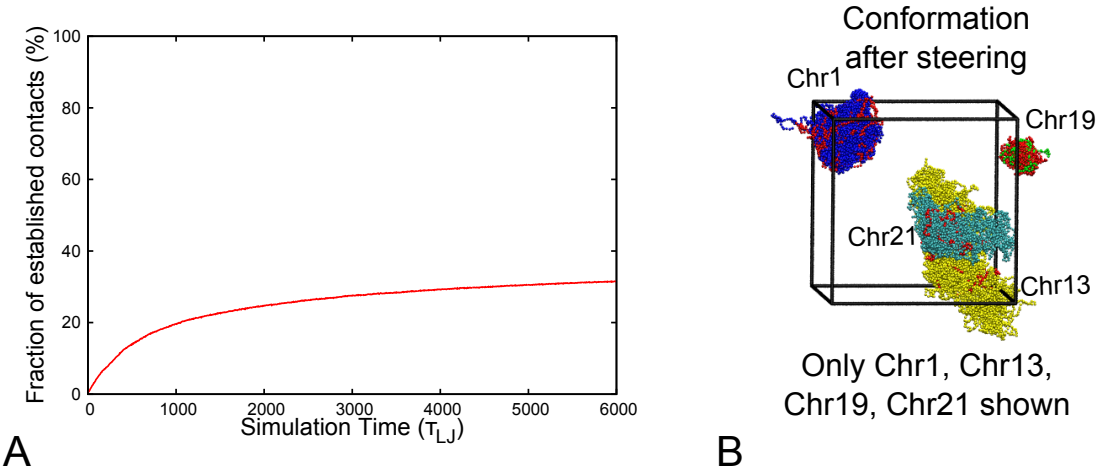


Figure 6.3: Illustration of the fraction of established target contacts for all the chromosome ensemble (A) and the resulting chromosome arrangement (B) after steering for a selected subset of chromosomes.

6.3.2 Local density of the model chromosomes

The application of the steering protocol introduces major changes in the structural organization of the model chromosomes. In Figure 6.2, we illustrate this point by highlighting with a red colour all the beads involved in the target pairings. Notice that this colouring can only provide point-wise information while the target constraints are pairwise and cannot be conveniently visualized in an uncluttered manner.

At the end of the steering protocol (panel B of Figure 6.3) almost none of the red beads are visible because they have coalesced in a core region that is relatively dense compared to the "fluffier" outer halo. The latter mostly consists of unconstrained loops protruding out of the core. To illustrate this point we present a cut-through view of the conformation of Chr13 and Chr19 in Figure 6.4 panels A and C. These two chromosomes were chosen because Chr13 has the lowest density of target pairs per unit length (~ 0.3 target contacts per Mbp) while Chr19 the highest (~ 5 target contacts per Mbp).

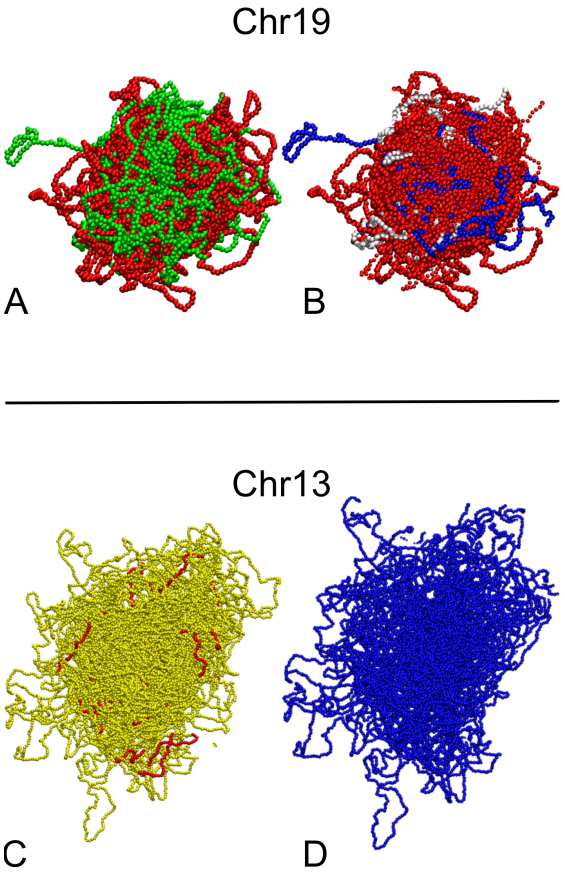


Figure 6.4: Cut-through of the final conformation of chromosomes 13 and 19 - A–C In the cut-through, the beads taking part to the target pairings are colored in red, while the remainder beads are shown in yellow for Chr13 and green for Chr19. **B–D** The beads colouring scheme indicates the different local density of the chromosome. The local density is measured by computing the coordination number (number of contacts) for each $100kb$-long stretch. Specifically, low density regions (*blue*) corresponds to stretches which perform a maximum of 30 contacts, medium (*white*) between 30 and 50 and high (*red*) for larger coordination numbers. The difference number of target contacts of Chr19 and Chr13 reflects in a very different compactness of their core regions.

The resulting inhomogeneous density of the chromosomes is conveyed in Figure 6.4B and D where the beads of chromosome 19 and 13 have been coloured according to the density of their local environment (blue, white and red correspond respectively to low, medium and high local density). As a simple and intuitive local measure of density we took the coordination number, that is the number of established contacts (target and non-target), for each stretch of $100\,kbp$ within the interaction cutoff distance of $180nm$. The fact that Chr19 and Chr13 have very different densities of beads to be colocalized reverberates in a much more compact core region for Chr19 (panel B) with respect to Chr13 (panel D).

6.3.3 Analysis of the contact maps.

To characterize the pattern of established target and non-target contacts in the model chromosomes, it is interesting to inspect the chromosome contact matrices at the end of the steering dynamics. Figure 6.5 shows such matrices for Chr13 (panel A) and Chr19 (panel B) for a sequence resolution of 100 *kbp* and interaction cutoff distance of 180nm. In these matrices the thick red dots indicate the established target contacts, while the

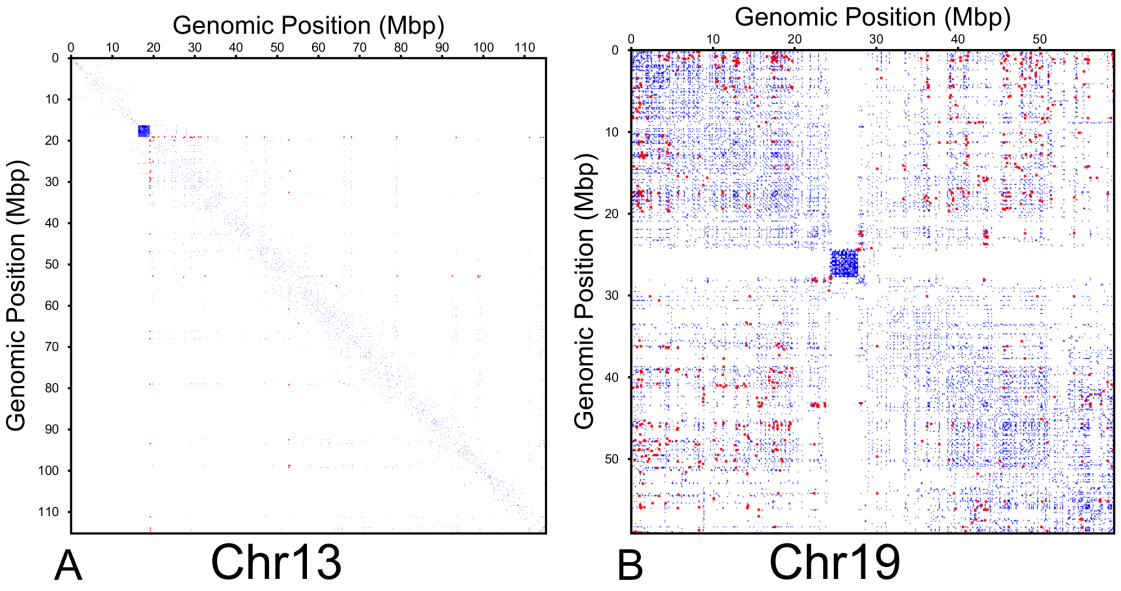


Figure 6.5: Contact matrix for Chr13 and Chr19 - Contact matrix of (A) Chr13 and (B) Chr19 at 100-kb resolution. Chr13 and Chr19 are the chromosomes with, respectively, the lowest and highest linear density of loci involved in target contacts. At the end of the steering protocol both target and non-target contacts are established. The former are shown with thick red dots while the later are shown with thin blue ones.

thin blue points pertain to all other (non-target) contacting pairs. In both matrices the centromere appears as a noticeable blue square.

The matrix of Chr13 appears to be very sparse. This reflects the very few target contacts, only 39, which are all established at the end of the steering protocol. By contrast, the matrix of Chr19 is much more densely populated. Of the $1,311$ target pairs of this chromosome, 370 are established at the end of the steering protocol, corresponding to a 28% yield.

The information provided by these matrices is well complemented by the data presented in Figure 6.6 which shows the gradual increase of the target and non-target contacts during the steering protocol. From the difference of the black and red curves it is seen that the absolute incidence of non-target contacts is particularly evident for the more constrained chromosomes, such as Chr19. Overall, at the end of the steering protocol,

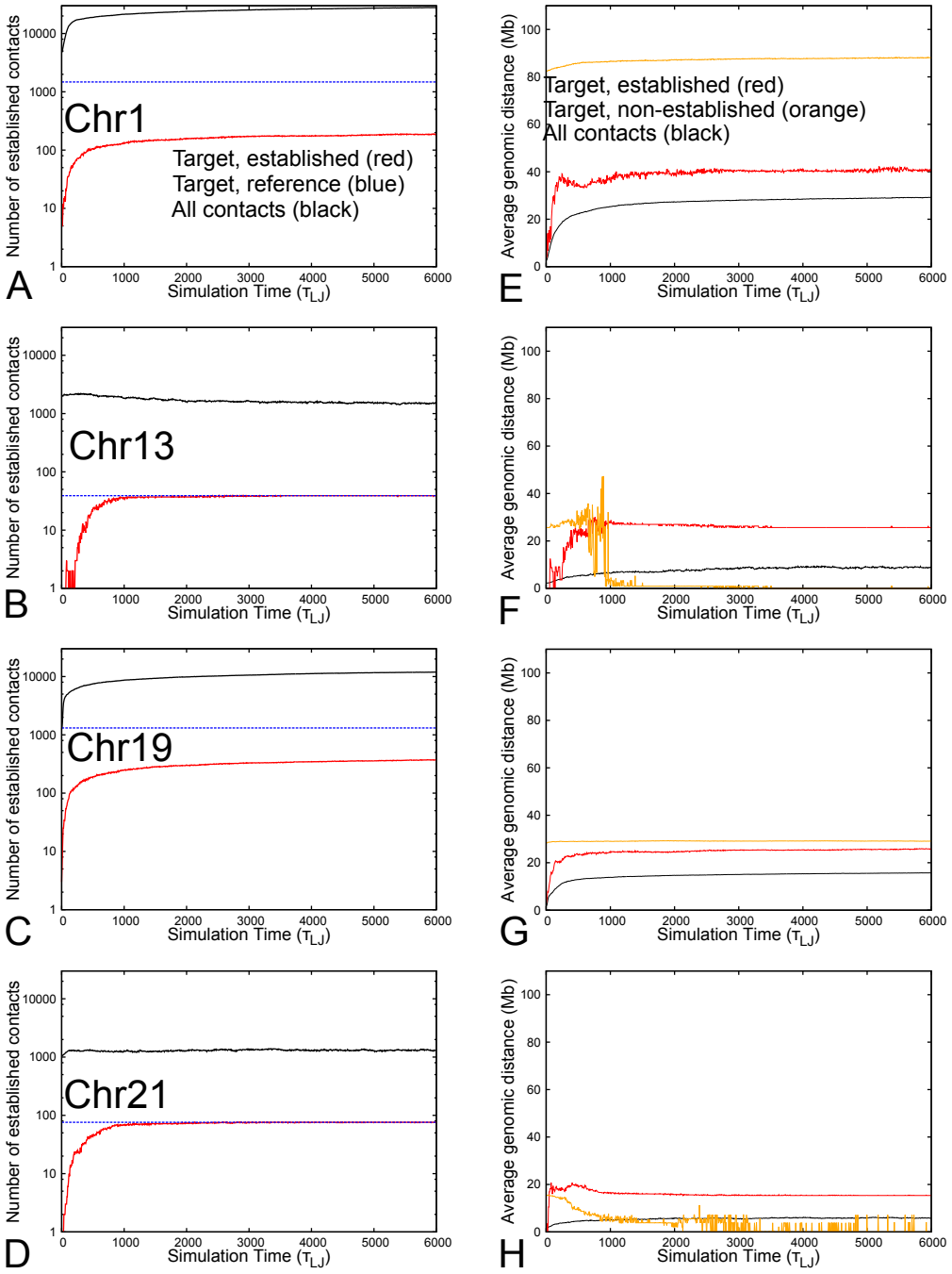


Figure 6.6: Progressive formation of target contacts - **A-D** Time evolution of the numbers of established contacts for various chromosomes. The black curve indicates the total number of contacts, while the red and the blue ones pertain to the satisfied and total target contacts. **E-H** Time evolution of the average genomic distance of established contacts for various chromosomes. The black curve pertains to the overall contacts, the red one relates to the established target contacts and the orange one to the non-established target contacts.

the established target contacts are only $\sim 3\%$ of the total number of formed contacts in both Chr13 and Chr19.

The rightmost panels E–H show the time evolution of the average genomic distance of:

- target pairs that are in contact, red curve
- target pairs that are not in contact, orange curve,
- all contacting pairs (target and non-target), black curve.

The data in Fig. 6.6 confirms the intuitive expectation that the overall number of established contacts increases during the steering protocol and is accompanied by an increase of their average genomic separation. Target contacts start forming at a certain stage of the steering protocol and, possibly due to their limited number, their degree of locality fluctuates significantly. We point out that for most chromosomes the final number of established contacts and target contacts has about a 30:1 ratio.

6.3.4 Distribution of lamina associated domains (LADs) within the chromosome territories

An interesting strategy to validate the viability of the obtained chromosome conformations after steering is to look at the positioning of the *lamina associated domains* (LADs) within the conformation of each chromosome obtained after steering.

LADs are expected, in fact, to be found at the periphery of the chromosome structures, because in most cases they interact with the lamina which is close to the nuclear membrane [120]. To check if the LAD domains occupy the periphery of the chromosome models is, indeed, very informative, because this property is not included in the set of constraints used for the simulations. In fact it is important to point out that the knowledge-based HiC target contacts, which we enforce as spatial proximity constraints in our simulations are oblivious of any LAD information.

In Figure 6.7, we show the results of the analysis of LAD positioning in model chromosomes. The latter analysis was specifically carried out by the partner group of E. Hovig based on the steering simulation data shown in Figure 6.7.

Panels A–D show the positions of the central beads of 33-beads chromosome strands (corresponding to 100kbp) with respect to the center of the chromosome conformation which is located in the origin of the Cartesian axes. In all these plots, red dots pertain to strands which are associated to LAD, while black dots indicate the other chromosome portions following the recent experimental characterization in ref. [121] for human fibroblasts. In the insets are shown the result of the statistical analysis to test whether the most central beads are preferentially associated to LADs with respect to the most peripheral ones. In particular, a $p-value$ close to zero indicate that the most peripheral beads are significantly identified as LADs.

Panels E–H show the distribution of the spatial distance of the LADs and non-LADs strands from the chromosome centers.

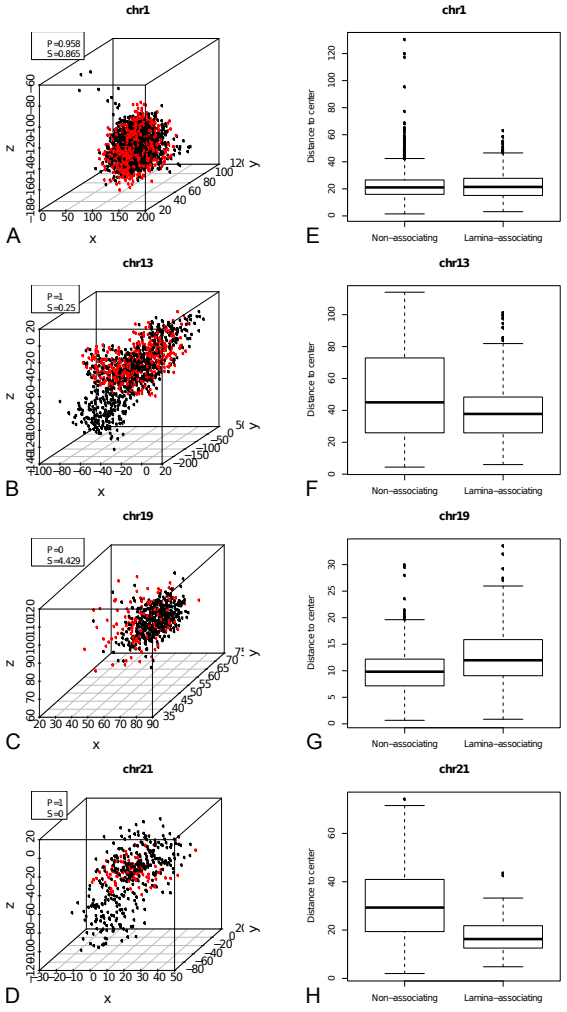


Figure 6.7: Illustration of the positioning of LADs on the steered chromosome structures - The analysis of LADs was done by considering the strands of 100kbp ($\sim 33\sigma$) of the chromosome chain at the end of steering procedure and associating them to be a LAD following recent experimental investigations on human fibroblasts [121]. In panels A–D the central bead positions of strands associated to LADs are shown as red dots, while the other as black ones. Next, the distance of each central beads of the 33σ strands to the chromosome center of mass is measured. Panels E–H show the distributions of the distances for LADs and non-LADs chromosome strands as box-whisker plots. The meaning of the thresholds indicated in the latter plots are illustrated in Appendix A. In the insets of panels A–D we report the results of the statistical analysis that has been done to ascertain whether the location of the LADs domains is significantly peripheral within the chromosome model conformation. In particular, the most 10% peripheral stretches and the most 10% central stretches are considered. Then, the Fisher's exact test [30] was used to test if the central beads are more associated to LADs than the peripheral ones. The results are provided as the score S and the associated *p-value*, p. Specifically, a *p-value* close to 0 indicates that central domains stretches are associated to non-LADs stretches, while a *p-value* close to 1 indicates the opposite.

It is particularly interesting to consider the cases of Chr19 and Chr13. In the former

chromosome, there exists a significant propensity for most peripheral beads to be associated with LADs than the central one. We shall notice that, as a matter of fact, this is in agreement with the location expected for the lamina associated domains. It is important to notice also that Chr19 is the chromosome with the higher density of significant HiC contacts and, as a consequences, its structure is expected to be largely perturbed by the steering process. To find a good compliance for Chr19 is therefore an important result.

The case of Chr13 is also interesting, because, while the most peripheral beads are not significantly associated with LADs, the visual inspection of the spatial positions of the lamina-associated strands in Figure 6.7B reveals a propensity for one side of the chromosome to be associated to the lamina.

This indication, which is present also in other chromosomes, could be arguably indicative of a relevant and biologically meaningful feature, even though the positioning of the LADs is not significantly peripheral.

We should conclude this analysis, commenting on the results obtained for Chr1 and Chr21, in which central regions of the chromosomes tend to be LADs associated, differently from what is expected, see panels A and C in Figure 6.7. We should say in this respect that the location of LADs used for this analysis refer to human fibroblasts. Since the HiC data used for the chromosome modelling pertain to hESC cells, this fact can reflect in false-positives entries for the locations of the LADs.

6.4 Human diploid chromosome system in spherical confinement

As a further investigation on human chromosomes, I will now discuss the more realistic case of a diploid model system, where two copies for each chromosomes are confined to the same packing density as the discussed haploid case. As a further realistic element, we replaced the periodic boundary conditions with a spherical confinement. The latter choice is, in fact, motivated by the fact that embryonic stem cells have an approximately spherical nucleus.

Accordingly, I organized the 46 chromosome chains in a solenoidal rod-like arrangement, as above, and, next, placed them inside a fairly large sphere. The confining surface is modeled as a fixed rigid spherical wall, whose interactions with the chain beads are described with a purely repulsive Lennard-Jones potential.

In this regard, we noted that several studies, including those of refs. [122, 123] have shown that there is a propensity for centers of the gene-rich/short chromosomes to be located

in the nuclear center, while gene-poor/long chromosomes have a propensity to stay at the nuclear periphery. This effect was accordingly modelled by positioning the centers of model chains to match the experimental average radial distribution of chromosome centers in ref. [122], see the black curve in Figure 6.8.

To accomplish this chromosome location, we applied the following procedure by positioning each rod-like solenoidal chromosome starting from the longest one. For each chromosome we generated hundreds of tentative positions in the nucleus by randomly picking the location of its center and the direction of its longitudinal axis. Of these tentative arrangements, we discarded those where the considered chromosome was protruding outside the confining spherical nucleus or that had steric clashes with chromosomes that were previously placed in the sphere. After collecting about 100 of these sterically-viable tentative configurations, we kept the one whose distance from the nucleus center gave the best match with the experimental average radial location of the chromosome centers.

It is seen *a posteriori* that such stochastic, yet data guided, placement process process works pretty well for most chromosomes, except the longest ones (Chr1 and Chr2) that cannot be positioned too close to the nuclear periphery, see blue curve in Figure 6.8.

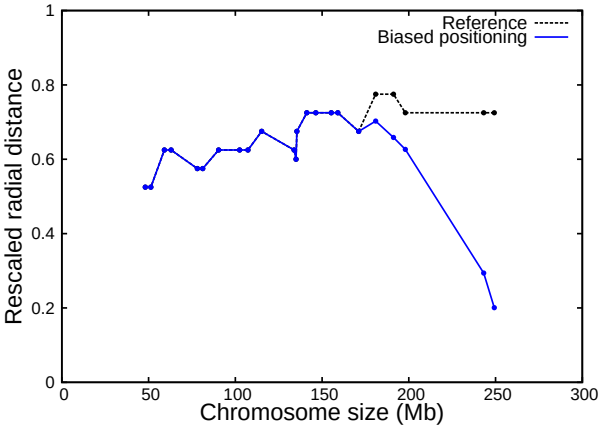


Figure 6.8: Radial positions of the chromosome centers in the confining sphere - The rescaled average distance of the chromosome centers to the center of the sphere is shown as a function of the chromosome size. The imaging experiments in ref. [122] showed a non random radial distribution of the chromosome centers in hESC spherical nuclei. In particular, gene-rich/short chromosomes are shown to be located in the nuclear center, while gene-poor/long chromosomes have a propensity to stay at the nuclear periphery (*dashed black curve*). I therefore tried to place the center of the chromosomes at the known experimental target distance (discretized in 6 bins for convenience). I sequentially generated randomly the chromosome center position for 100 times always starting from the longest one and pick the trial that is closest to the target. All the trials respect the self-avoidance with the chromosomes already placed. The placement matches well the target position except for the two copies of the largest chromosomes (Chr1 and Chr2) which can only be accommodated close to the nuclear center (*blue curve*).

This is because for simplicity our chromosomes were initially modelled as rigid bodies and therefore the largest ones, which barely fit inside the nucleus, cannot be accommodated at the periphery of the spherical nucleus. We plan to overcome this limitation in future studies by introducing some flexibility in the initial chromosome configurations.

Once the initial conformation was obtained, see Figure 6.9A, the chromosome chains were relaxed with a free dynamical evolution except for the beads of the centromeric region whose compactness was promoted by a nonspecific attractive interaction (see section 6.2.1). During this evolution, the radius of the confining sphere is also progressively shrunk down to $\sim 5\,\mu$ yielding the typical nuclear density of $\sim 0.012\,bp/nm^3$ [11].

After relaxation, the obtained conformation, shown in Figure 6.9B is used as initial states for the steering procedure.

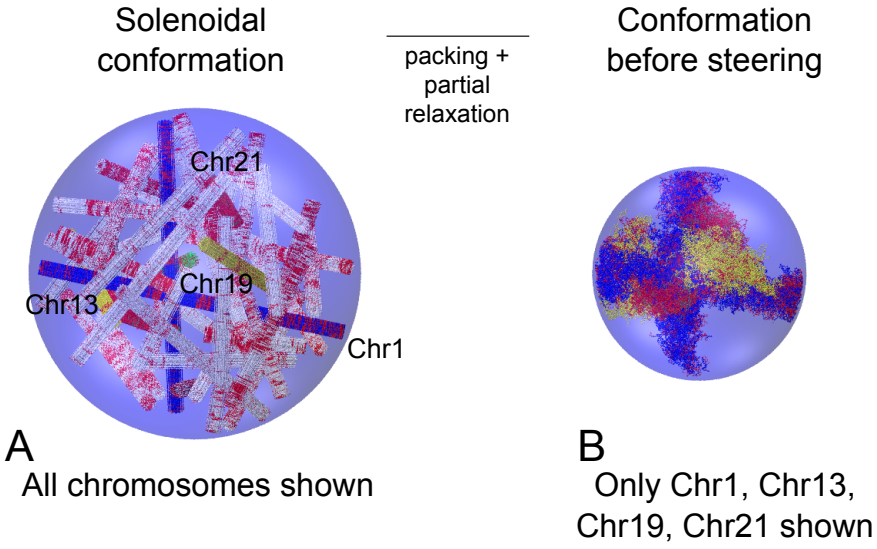


Figure 6.9: Initial conformation of the diploid chromosome system inside the confining sphere - A The 46 chromosomes accounting for the human diploid chromosome system are arranged in a non-overlapping manner in a sphere mimicking the nuclear confinement. Each molecule is prepared in a rod–like hierarchical solenoidal arrangement, similarly to Figure 6.2. **B** The chromosomes are next relaxed from the initial arrangement and reach an elongated mitotic–like structure. The confining sphere is simultaneously shrunk until its radius is $\sim 5\,\mu m$ long (for which the typical nuclear density is achieved). The red beads in panels A–B mark the regions involved in the target pairings. In panels B, we show only the copies of four chromosome Chr1, Chr13, Chr19 and Chr21 for visual clarity.

6.4.1 Preliminary results of the steered dynamics

In this section I present the result of the steering for the diploid system. Since the analysis of the trajectories for this system is still ongoing, I shall discuss here only preliminary results.

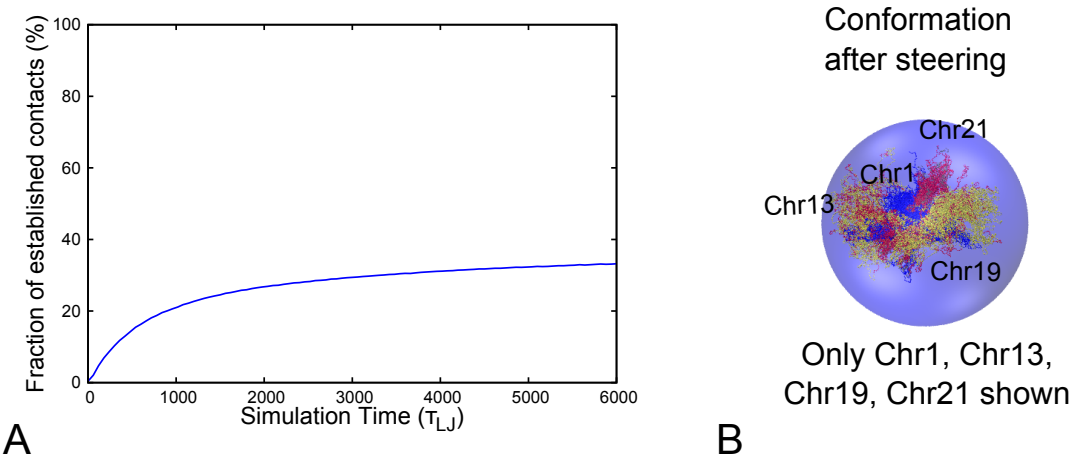


Figure 6.10: Illustration of the fraction of established target contacts for all the chromosome ensemble (A) and the resulting chromosome arrangement (B) after steering for a subset of chromosomes.

The compliance of the diploid system to the steering protocol is shown in Figure 6.10A. It is found that at the end of the steering process about 33% of the all target contacts are colocalized within a contact radius of 180nm. The conformations obtained at the end of the steering process are shown in Figure 6.10B for a selection of chromosomes.

It is important to notice the similarity of this results with the haploid system for which we can establish an almost equal fraction of contacts ($\sim 30\%$). In this respect, the striking similarity of the haploid and diploid systems is an important result in connection with the fact that the present diploid system complements the previous haploid one in three respects first being the number of chromosomes. Second the boundary conditions, which are a spherical confinement for the diploid and periodic for the haploid, and finally the choice of the rationale behind the positioning of the chromosomes. The latter is a random choice for haploid system and a biased towards experimental data for the diploid one.

All these three differences should be considered in connection with the fact that in this chapter we are are considering only intra-chromosome contacts. It can be expected that taking into account also the inter-chromosome ones the differences in the two systems could play a major role. In particular, the chromosome positioning in the initial conformation can be expected to have a major impact in the pattern of inter-chromosome neighbors and, hence, to reverberate on the compliance of the systems to the steering.

The similarity in the pattern of established contacts in the diploid system with respect to the haploid case is evident by looking at the contact maps of chromosomes 13 and 19. Here, the matrices, shown in Figure 6.11, account for the contacts established in at least one of the two copies of the chromosomes.

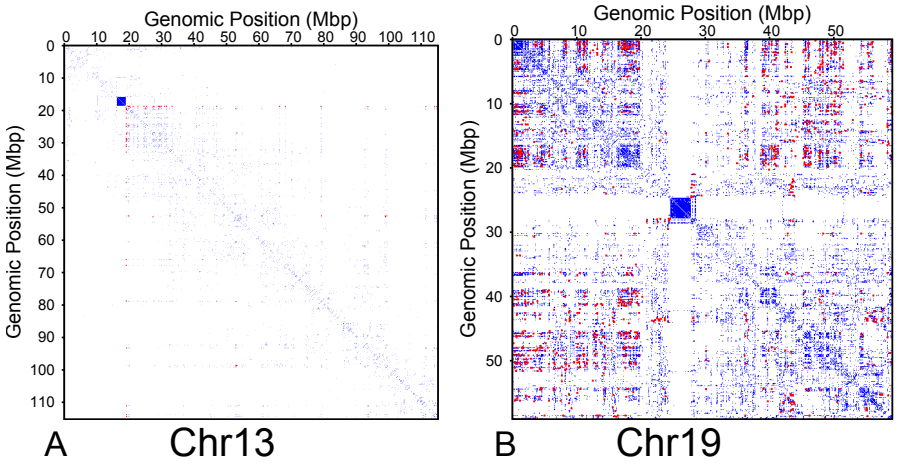


Figure 6.11: Contact matrix for Chr13 and Chr19 of the diploid system - Contact matrix of (A) Chr13 and (B) Chr19 at 100-kb resolution. At the end of the steering protocol both target and non-target contacts are established. The former are shown with thick red dots while the later are shown with thin blue ones. The matrices account for the total of the contacts established in the two copies of the chromosomes in the diploid system.

In particular, the matrix of Chr13 is, again, very sparse reflecting the small number of target contacts 39, of which 37 are established at the end of the steering protocol. On the other hand, the Chr19 map is much more rich in contacts. In particular, we can establish 364 (28%) of the target contacts, that is again similar to the haploid case.

To conclude, we wish to mention that, given the promising results obtained on the haploid system with the LAD analysis and the similar compliance of the diploid system to the steering protocol, it would be very informative to carry out the analysis of lamina-associated domains on the latest chromosome conformations to validate our modeling approach. This step will be carried out by the partner group in Oslo based on the numerical data generated by the steering protocol.

6.5 Summary

In this chapter, we presented an ongoing work on the modelling of the spatial organization of human chromosomes obtained by using a coarse-grained model of chromatin and knowledge-based molecular dynamics simulations. Specifically, we singled out pairs of 100kbp chromosome regions which showed significant HiC contact propensities for hESC cells and used them as spatial restraints to steer their colocalization in the model chromosome.

We found that about 30% of the HiC-based target pairs could actually be colocalized simultaneously.

Interestingly, the resulting chromosome structures at the end of the steering protocol show a bipartite arrangement: a dense core and an outer, “fluffy”, region.

Interesting features were found also for the positioning of the regions that are known to correspond to the LADs. The latter are, in general, expected to interact with the nuclear envelop and, accordingly, ought to be placed at the periphery of the chromosome conformation. We found that Chr19, which has the largest (linear) density of constraints, has a pronounced tendency of LADs to be associated to the chromosome periphery.

The robustness of these results is underscored by the fact that analogous features emerge when extending the analysis for the haploid system to the diploid one, where we considered twice the number of chromosomes, spherical spatial nuclear confinement and more realistic radial positioning of chromosome centers.

Concluding remarks

Genomic DNA filaments are typically confined in compartments whose linear size is much shorter than their contour lengths. Clearly, the resulting geometrical and topological entanglement has a severe impact on the conformational and dynamical properties of the DNA chains.

In this thesis I considered various contexts where this relationship between entanglement and the dynamical and conformational plasticity of DNA could be analyzed and characterized by theoretical and computational means.

In particular, I first considered two case of knotted semiflexible polyelectrolytes, such as DNA molecules, which are subjected external AC/DC electric fields. The first system was constituted by DNA molecules pulled at the chain termini as it is done in typical optical tweezers experiments. In this conditions (see chapter 2) the knotted region can be driven along the DNA contour in the direction of the external electric force both in DC and AC conditions.
External electric fields can be also used to manipulate untensioned polyelectrolyte chains in solution with trivalent counterions. We discussed (see chapter 3) the spontaneous occurrence of entanglement, and in particular knotting, in collapsed PE chains. To the best of our knowledge, this is the first time that these aspects have been characterized for these systems. Interestingly, we showed that the geometrical entanglement slows down the elongation dynamics of PE chains under the action of an external DC electric field, but the same structures are, instead, elongated under external AC electric fields in a much shorter time span. This study might, hence, provide a novel strategy to precondition DNA molecules to avoid or minimise their self-entanglement.

Finally, we moved to study the more complex and challenging case of eukaryotic chromosomes where we analysed chromosome conformational plasticity in connection with the gene coregulation–colocalization hypothesis [14, 15] for human chromosome 19. By using steered molecular dynamics simulations we showed that the the gene pairs coregulation is largely compatible with their spatial colocalization in the model chromosomes. Interestingly, the obtained chromosome arrangements showed a partition in macrodomains

similar to the one inferred from recent HiC data. The investigation of the coregulation-colocalization relationship was next carried out also for *D. melanogaster* at the genome wide level, albeit by solely using HiC and gene expression data, that is with no reference to an explicit modelling of chromosome conformations. For such system, we found a statistically significant correlation between gene coregulation/colocalization, which is due to the fact that gene pairs with low mutual information tend to have a low propensity to be in spatial contact.

Finally, I presented (see chapter 6) a set of preliminary results on modeling of human chromosomes structural organization by using HiC significant contact propensities and knowledge-based computer simulations of model chromosomes. The modelled chromosome structures are robust to significant variations of the system properties, such as dealing with haploid or diploids nuclei, and a good correlation is observed between the predicted and measured regions that occupy central of peripheral regions of the chromosomes.

These results, like those presented in other chapters of the thesis, underscore the fact that suitable coarse-grained models and simulation techniques can be aptly used to elucidate the complex interplay of structure and function of the densely packed, and hence highly entangled, DNA filaments *in vivo* and hence offer a valuable starting point for future extensions towards theoretical and computational characterizations of eukaryotic nuclei with even more detail and realistic properties.

Appendix A

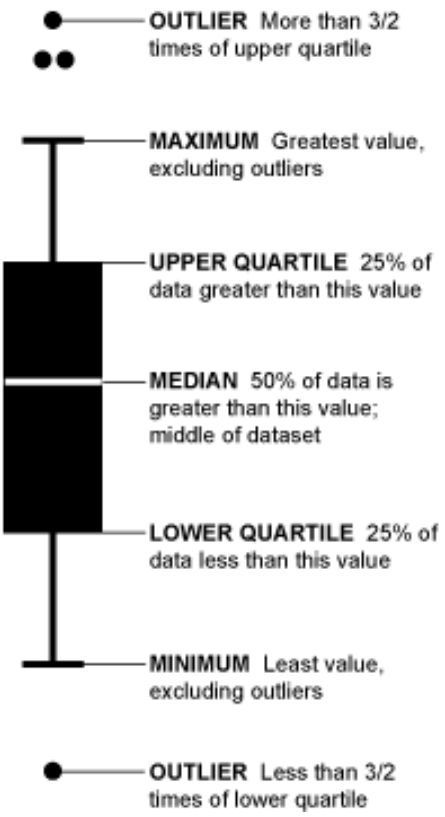


Figure 6.12: Standard box-whisker plot - Illustration of the statistical threshold indicated in the box-whisker plots presented in this Thesis in Figures 5.7 and 6.7. Specifically, the bottom and the top of the boxes indicate the values corresponding to the first and third quartile. Those are the values below which there are 25% and 75% of the data. The centerline is the median value. The whisker (vertical bar) span the range from the minimum value to the maximum value excluding outliers. The latter are represented with points.

References

[1] D Marenduzzo, C Micheletti, and E Orlandini. Biopolymer organization upon confinement. *Journal of Physics: Condensed Matter*, 22(28):283102, 2010.

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

[3] X.R. Bao, H.J. Lee, and S.R. Quake. Behavior of complex knots in single DNA molecules. *Physical review letters*, 91(26): 265506, 2003.

[4] Kleanthes Koniaris and M. Muthukumar. Knottedness in ring polymers. *Phys. Rev. Lett.*, 66:2211–2214, Apr 1991.

[5] RR Netz. Nonequilibrium unfolding of polyelectrolyte condensates in electric fields. *Physical review letters*, 90(12):128104, 2003.

[6] Mehmet Sayar and Christian Holm. Equilibrium polyelectrolyte bundles with different multivalent counterion concentrations. *Physical Review E*, 82(3):031901, 2010.

[7] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.*, 2:292, 2001.

[8] M. R. Branco and A. Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.*, 4:e138, 2006.

[9] E. Lieberman-Aiden et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326:289, 2009.

[10] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[11] A. Rosa and R. Everaers. Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.*, 4:e1000153, 2008.

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

[13] Mariano Barbieri, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Josée Dostie, Ana Pombo, and Mario Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, 109(40):16173–16178, 2012.

[14] C. G. Spilianakis, M. D. Lalioti, T. Town, G. R. Lee, and R. A. Flavell. Interchromosomal associations between alternatively expressed loci. *Nature*, 435:637–645, 2005.

[15] G. Cavalli. Chromosome kissing. *Curr. Opin. Genet. Dev.*, 17:443, 2007.

[16] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148:458, 2012.

[17] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes

identified by analysis of chromatin interactions. *Nature*, 485:376–380, 2012.

[18] Samuel Frederick Edwards and M Doi. *The Theory of Polymer Dynamics*. Clarendon, Oxford, 1986.

[19] Michael Rubinstein and RH Colby. *Polymers Physics*. Oxford, 2003.

[20] B Alberts et al. Molecular biology of the cell in cell 4th, 2002.

[21] Helmut Schiessel. *Biophysics for beginners: a journey through the cell nucleus*. CRC Press, 2013.

[22] John F Marko and Eric D Siggia. Stretching dna. *Macromolecules*, 28(26):8759–8770, 1995.

[23] Marian Walhout, Marc Vidal, and Job Dekker. *Handbook of systems biology: concepts and insights*. Academic Press, 2012.

[24] K. Kremer and G. S. Grest. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.*, 92: 5057, 1990.

[25] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.

[26] Kurt Binder. *Monte Carlo and Molecular Dynamics Simulations Polymer*. Oxford University Press, Inc., 1995.

[27] P. Nelson. *Biological physics*. WH Freeman New York, 2004.

[28] Christopher Maffeo, Robert Schöpflin, Hergen Brutzer, René Stehr, Aleksei Aksimentiev, Gero Wedemann, and Ralf Seidel. Dna˘DNA interactions in tight supercoils are described by a small effective charge density. *Phys. Rev. Lett.*, 105:158101, Oct 2010. doi: 10.1103/PhysRevLett.105.158101.

[29] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.*, 117:1, 1995.

[30] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007. ISBN 0521880688, 9780521880688.

[31] C. Micheletti, D. Marenduzzo, and E. Orlandini. Polymers with spatial or topological constraints: theoretical and computational results. *Physics Reports*, pages 1–73, 2011.

[32] A. Rosa, E. Orlandini, L. Tubiana, C. Micheletti. Structure and dynamics of ring polymers: entanglement effects due to solution density and ring topology. *Macromol.*, 44:8668–8680, 2011.

[33] M Delbrück. Dnotting problems in biology. In R E Bellman, editor, *Mathematical problems in biological sciences*, volume 14 of *Proc. Symp. Appl. Math*, page 55, 1962.

[34] E. Orlandini and S. G. Whittington. Statistical topology of closed curves: some applications in polymer physics. *Rev. Mod. Phys.*, 79:611, 2007.

[35] M. Kardar. The elusiveness of polymer knots. *The European Physical Journal B-Condensed Matter and Complex Systems*, 64(3):519–523, 2008.

[36] AY Grosberg. A few notes about polymer knots. *Polymer Science Series A*, 51(1):70–79, 2009.

[37] A.Y. Grosberg. Critical exponents for random knots. *Physical review letters*, 85(18): 3858–3861, 2000.

[38] Vsevolod Katritch, Wilma K. Olson, Alexander Vologodskii, Jacques Dubochet, and Andrzej Stasiak. Tightness of random knotting. *Phys. Rev. E*, 61:5545–5549, May 2000. doi: 10.1103/PhysRevE.61.5545.

[39] O. Farago, Y. Kantor, and M. Kardar. Pulling knotted polymers. *Europhys. Lett.*, 60:53–59, 2002.

[40] Peter Virnau, Yacov Kantor, and Mehran Kardar. Knots in globule and coil phases of a model polyethylene. *Journal of the American Chemical Society*, 127(43):15102–15106, 2005. doi: 10.1021/ja052438a.

[41] B. Marcone, E. Orlandini, A.L. Stella, and F. Zonta. Size of knots in ring polymers. *Physical Review E*, 75(4):041105, 2007.

[42] M. Baiesi, E. Orlandini, A. L. Stella, and F. Zonta. Topological signatures of globular polymers. *Phys. Rev. Lett.*, 106:258301, Jun 2011. doi: 10.1103/PhysRevLett.106.258301.

[43] D. Meluzzi, D.E. Smith, and G. Arya. Biophysics of knotting. *Annual review of biophysics*, 39:349–366, 2010.

[44] A. Rosa, M. Di Ventra, and C. Micheletti. Topological jamming of spontaneously knotted polyelectrolyte chains driven through a nanopore. *Physical review letters*, 109(11): 118301, Sep 2012. ISSN 1079-7114.

[45] W.R. Taylor. A deeply knotted protein structure and how it might fold. *Nature*, 406 (6798):916–919, 2000.

[46] Daniel Bölinger, Joanna I Sułkowska, Hsiao-Ping Hsu, Leonid A Mirny, Mehran Kardar, José N Onuchic, and Peter Virnau. A stevedore's protein knot. *PLoS computational biology*, 6(4):e1000731, 2010.

[47] Peter Virnau, Anna Mallam, and Sophie Jackson. Structures and folding pathways of topologically knotted proteins. *Journal of Physics: Condensed Matter*, 23(3):033101, 2011.

[48] Tatjana Škrbić, Cristian Micheletti, and Pietro Faccioli. The role of non-native interactions in the folding of knotted proteins. *PLoS computational biology*, 8(6):e1002504, 2012.

[49] L. Tubiana, A. Rosa, F. Fragiacomo, and C. Micheletti. Spontaneous knotting and unknotting of flexible linear polymers: Equilibrium and kinetic aspects. *Macromolecules*, 46(9):3669–3678, 2013. doi: 10.1021/ma4002963.

[50] P. G. Dommersnes, Y. Kantor, and M. Kardar. Knots in charged polymers. *Physical Review E*, 66(3):031802, 2002.

[51] Jing Tang, Ning Du, and Patrick S. Doyle. Compression and self-entanglement of single DNA molecules under uniform electric field. *Proceedings of the National Academy of Sciences*, 108(39):16153–16158, 2011. doi: 10.1073/pnas.1105547108.

[52] E. Orlandini, A. L. Stella, C. Vanderzande, and F. Zonta. Slow topological time scale of knotted polymers. *J. Phys. A: Math. Theor.*, 41:122002, 2008.

[53] Piotr Szymczak. Tight knots in proteins: can they block the mitochondrial pores? *Biochemical Society transactions*, 41(2):620–624, Apr 2013. ISSN 1470-8752. doi: 10.1042/BST20120261.

[54] Lei Huang and Dmitrii E. Makarov. Translocation of a knotted polypeptide through a pore. *The Journal of chemical physics*, 129 (12):121107, Sep 2008. ISSN 1089-7690. doi: 10.1063/1.2968554.

[55] F Tessier, J Labrie, and G Slater. Electrophoretic separation of long polyelectrolytes in submolecular-size constrictions: A monte carlo study. *Macromolecules*, 35:4791–4800, 2002.

[56] Davide Marenduzzo, Enzo Orlandini, Andrzej Stasiak, De Witt Sumners, Luca Tubiana, and Cristian Micheletti. Dna-dna interactions in bacteriophage capsids are responsible for the observed dna knotting. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52): 22269–22274, Dec 2009. ISSN 1091-6490. doi: 10.1073/pnas.0907524106.

[57] Davide Marenduzzo, Cristian Micheletti, Enzo Orlandini, and De Witt Sumners. Topological friction strongly affects viral

dna ejection. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):20081–20086, Dec 2013. ISSN 1091-6490. doi: 10.1073/pnas.1306601110.

[58] Y. Arai, R. Yasuda, K. Akashi, Y. Harada, H. Miyata, T. Kinosita, and H. Itoh. Tying a molecular knot with optical tweezers. *Nature*, 399:446–448, Jun 1999.

[59] A. Vologodskii. Brownian Dynamics Simulation of Knot Diffusion along a Stretched DNA molecule. *Biophys. J.*, 90:1594, 2006.

[60] R. Matthews, AA Louis, and JM Yeomans. Effect of topology on dynamics of knots in polymers under tension. *EPL (Europhysics Letters)*, 89:20001, 2010.

[61] Lei Huang and Dmitrii E. Makarov. Langevin dynamics simulations of the diffusion of molecular knots in tensioned polymer chains†. *The Journal of Physical Chemistry A*, 111(41):10338–10344, 2007. PMID: 17637045.

[62] Marco Di Stefano, Luca Tubiana, Massimiliano Di Ventra, and Cristian Micheletti. Driving knots on dna with ac/dc electric fields: topological friction and memory effects. *Soft matter*, 10(34):6491–6498, 2014.

[63] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[64] Luca Tubiana, Enzo Orlandini, and Cristian Micheletti. Probing the entanglement and locating knots in ring polymers: A comparative study of different arc closure schemes. *Progress of Theoretical Physics Supplement*, 191:192–204, 2011. doi: 10.1143/PTPS.191.192.

[65] A. M. Saitta, P. D. Soper, E. Wasserman, and M. L. Klein. Influence of a knot on the strength of a polymer strand. *Nature*, 399:46–48, 1999.

[66] T. Odijk. Polyelectrolytes near the rod limit. *Journal of Polymer Science: Polymer Physics Edition*, 15(3):477–483, 1977. ISSN 1542-9385. doi: 10.1002/pol.1977.180150307.

[67] Jeffrey Skolnick and Marshall Fixman. Electrostatic persistence length of a wormlike polyelectrolyte. *Macromolecules*, 10(5):944–948, 1977. doi: 10.1021/ma60059a011.

[68] Arindam Kundagrami and M Muthukumar. Theory of competitive counterion adsorption on flexible polyelectrolytes: Divalent salts. *The Journal of chemical physics*, 128(24):244901, 2008.

[69] Mehmet Sayar and Christian Holm. Finite-size polyelectrolyte bundles at thermodynamic equilibrium. *EPL (Europhysics Letters)*, 77(1):16001, 2007.

[70] A Balducci and PS Doyle. Conformational preconditioning by electrophoresis of dna through a finite obstacle array. *Macromolecules*, 41(14):5485–5492, 2008.

[71] RR Netz. Polyelectrolytes in electric fields. *The Journal of Physical Chemistry B*, 107 (32):8208–8217, 2003.

[72] Adam E Cohen. Force-extension curve of a polymer in a high-frequency electric field. *Physical review letters*, 91(23):235506, 2003.

[73] Pai-Yi Hsiao and Erik Luijten. Salt-induced collapse and reexpansion of highly charged flexible polyelectrolytes. *Physical review letters*, 97(14):148301, 2006.

[74] Pai-Yi Hsiao, Yu-Fu Wei, and Hsueh-Chia Chang. Unfolding collapsed polyelectrolytes in alternating-current electric fields. *Soft Matter*, 7(3):1207–1213, 2011.

[75] Yanwei Wang, Wes F Reinhart, Douglas R Tree, and Kevin D Dorfman. Resolution limit for dna barcodes in the odijk regime. *Biomicrofluidics*, 6(1):014101, 2012.

[76] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295:1306, 2002.

[77] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38:1348, 2006.

[78] J. Dostie et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16: 1299, 2006.

[79] E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43:1059, 2011.

[80] V. Belcastro, V. Siciliano, F. Gregoretti, P. Mithbaokar, G. Dharmalingam, S. Berlingieri, F. Iorio, G. Oliva, R. Polishchuck, N. Brunetti-Pierri, and D. di Bernardo. Transcriptional gene network inference from a massive dataset elucidates trascriptome organization and gene function. *Nucleic Acids Research*, 39: 8677, 2011.

[81] Melissa J. Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G. Y. Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N. Ariyaratne, Vinsensius B. Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K. D. Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V. Desai, Jane S. Thomsen, Yew Kok Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G. Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T. Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462: 58–64, 2009.

[82] E. Apostolou and D. Thanos. Virus Infection Induces NF-kB-Dependent Interchromosomal Associations Mediating Monoallelic *IFN-β Gene Expression*. *Cell*, 134:85–96, 2008.

[83] S. Kocanova, E. A. Kerr, S. Rafique, S. Boyle, E. Katz, S. Caze-Subra, W. A. Bickmore, and K. Bystricky. Activation of estrogen-responsive genes does not require their nuclear co-localization. *PLoS Genet*, 6: e1000922, 2010.

[84] D. Marenduzzo, C. Micheletti, and P. R. Cook. Entropy-driven genome organization. *Biophys. J.*, 90:3712–3721, 2006.

[85] I. Junier, O. Martin, and F. Képès. Spatial and Topological Organization of DNA Chains Induced by Gene Co-localization. *Plos Comput. Biol.*, 6:e1000678, 2010.

[86] P. R. Cook. The organization of replication and transcription. *Science*, 284:1790–1795, 1999.

[87] P. R. Cook. A model for all genomes: The role of transcription factories. *J. Mol. Biol.*, 395:1, 2010.

[88] M. Di Stefano, A. Rosa, V. Belcastro, D. di Bernardo, and C. Micheletti. Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. *PLoS Comput. Biol.*, 9: e1003019, 2013.

[89] D. Bau and M. A. Marti-Renom. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome Res.*, 19:25–35, 2011.

[90] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. Arrayexpress-a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, 35, 2007.

[91] http://www.affymetrix.com.

[92] J. Grimwood et al. The DNA sequence and biology of human chromosome 19. *Nature*, 428:529, 2004.

[93] B. Goebel, Z. Dawy, Z. Hagenauer, and J. Mueller. An approximation to the distribution of finite sample size mutual information estimates. *IEEE International Conference on Communications*, 2:1102, 2005.

[94] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 5:415, 2000.

[95] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer. Crumpled globule model of the three-dimensional structure of dna. *Europhys. Lett.*, 23:373, 1993.

[96] H. Albiez, M. Cremer, C. Tiberi, L. Vecchio, L. Schermelleh, S. Dittrich, K. K/"upper, B. Joffe, T. Thormeyer, J. von Hase, S. Yang, K. Rohr, H. Leonhardt, I. Solovei, C. Cremer, S. Fakan, and T. Cremer. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome Res.*, 14:707–733, 2006.

[97] T. Vettorel, A. Y. Grosberg, and K. Kremer. Statistics of polymer rings in the melt: a numerical simulation study. *Phys. Biol.*, 6: 025013, 2009.

[98] L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, 19:37–51, 2011.

[99] M. A. Marti-Renom and L. A. Mirny. Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. *PLoS Comput Biol*, 7:e1002125, 2011.

[100] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello. Plumed: a portable plugin for free-energy calculations with molecular dynamics. *Comp. Phys. Comm.*, 180:1961, 2009.

[101] A. Rosa, N. B. Becker, and R. Everaers. Looping probabilities in model interphase chromosomes. *Biophys. J.*, 98:2410, 2010.

[102] G. Fudenberg and L. A. Mirny. Higher-order chromatin structure: bridging physics and biology. *Genet. & Develop.*, 22:115–124, 2012.

[103] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2001.

[104] C. Micheletti, G. Lattanzi, and A. Maritan. Elastic properties of proteins: Insight on the folding process and evolutionary selection of native structures. *J. Mol. Biol.*, 321:909, 2002.

[105] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440, 1996.

[106] M. Kaiser. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New J. Phys.*, 10:083042, 2008.

[107] S. Djebali et al. Evidence for transcript networks composed of chimeric rnas in human cells. *PLoS ONE*, 7:e28213, 2012.

[108] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270 (5235):467–470, 1995.

[109] Deval A Lashkari, Joseph L DeRisi, John H McCusker, Allen F Namath, Cristl Gentile, Seung Y Hwang, Patrick O Brown, and Ronald W Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062, 1997.

[110] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of

affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15, 2003.

[111] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2): 185–193, 2003.

[112] J. W Tukey. *Exploratory Data Analysis*. Reading: Addison-Wesley, 1970.

[113] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. ISSN 1367-4803. doi: http://dx.doi.org/10.1093/bioinformatics/btg405.

[114] Jonas Paulsen, Tonje G Lien, Geir Kjetil Sandve, Lars Holden, Ørnulf Borgan, Ingrid K Glad, and Eivind Hovig. Handling realistic assumptions in hypothesis testing of 3d co-localization of genomic elements. *Nucleic acids research*, page gkt227, 2013.

[115] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[116] Kelvin KW Yau, Kui Wang, and Andy H Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452, 2003.

[117] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[118] J. T. Finch and A. Klug. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. USA*, 73:1897, 1976.

[119] H. Benoit and P. Doty. Light scattering from non-gaussian chains. *The Journal of Physical Chemistry*, 57(9):958–963, 1953. doi: 10.1021/j150510a025.

[120] Tom Misteli. Beyond the sequence: cellular organization of genome function. *Cell*, 128 (4):787–800, 2007.

[121] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, 2008.

[122] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157, 2005.

[123] Shelagh Boyle, Susan Gilchrist, Joanna M Bridger, Nicola L Mahy, Juliet A Ellis, and Wendy A Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, 10(3):211–219, 2001.

Acknowledgements

This thesis is the compendium of the work that I have carried out in my four years of studying in SISSA. Here, instead, I have to thank all the people that supported me in this rich and beautiful experience.

The two people that I wish to thank, first, are Angelo and Cristian. It is not simple to say how I am grateful to them for their teaching, patience, mentorship and friendship.

My deepest gratitude goes to all those people with whom I had the opportunity to collaborate. A special mention goes to Luca Tubiana for being my office mate during the first two years of real work. His being supportive and encouraging has been precious also later on.

Since my first days in Trieste I've been blessed by the friendship of three marvellous people: Ina, Francesco, and Francesca. Your friendship has been really important for me and I take you with me wherever I shall go.

I sincerely thank all the other members of mSBP Sector and, in particular, my office mates Guido, Jessica, and Manon. I wish to thank also the Organizers of the mSBP sharing group and Sandro who shared with me this experience last year.

I thank also the other friends from Sissa and friends of the tennis course of *mister* Matteo at Polisportiva in Opicina.

Many people who do not live in Trieste have been also a fundamental presence in the last four years. I wish to mention my friends in Pantan Monastero (Rome) and Turin.

Finally, I thank my family, to whom this thesis is dedicated: Vincenzo, Lucia, Angelo&Ilaria, Chiara&Fabio, and my grandma Giovanna. You have been gifts for me since I was born and every single day I feel these gifts enlarged and renewed: Thank you!