
THE UNIVERSITY
*of* EDINBURGH

# Large Scale Simulations of Genome Organisation in Living Cells

J. Johnson


Doctor of Philosophy
The University of Edinburgh
March 2018

# Abstract

Within every human cell, approximately two meters of DNA must be compacted into a nucleus with a diameter of around ten micrometers. Alongside this daunting storage problem, the 3D organisation of the genome also helps determine which genes are up- or down-regulated, which in turn effects the functionality of the cell itself. While the organisational structure of the genome can be revealed using experimental techniques such as chromosome conformation capture and its high-throughput variant Hi-C, the mechanisms driving this organisation are still unclear.

The first two results chapters of this thesis use molecular dynamics simulations to investigate the effect of a potential organisational mechanisms for DNA known as the "bridging-induced attraction". This mechanism involves multivalent DNA-binding proteins bridging genomically distant regions of DNA, which in turn promotes further binding of proteins and compaction of the DNA.

In chapter 2 (the first results chapter) we look at a model where proteins can bind non-specifically to DNA, leading to cluster formation for suitable protein-DNA interaction strengths. We also show the effects of protein concentration on the DNA, with a collapse from a swollen to a globular phase observed for suitably high protein concentrations.

Chapter 3 develops this model further, using genomic data from the ENCODE project to simulate the "specific binding" of proteins to either active (euchromatin) or inactive (heterochromatin) regions. We were then able to compare contact maps for specific simulated chromosomes with the experimental Hi-C data, with our model reproducing well the topologically associated domains (TADs) seen in Hi-C contact maps.

In chapter 4 of the thesis we use numerical methods to study a model for the coupling between DNA topology (in particular, supercoiling in DNA and chromatin) and transcription in a genome. We present details of this model, where supercoiling flux is induced by gene transcription, and can diffuse along the DNA. The probability of transcription is also related to supercoiling, as regions of DNA which are negatively supercoiled have a greater likelihood of being transcribed. By changing the magnitude of supercoiling flux, we see a transition between a regime where transcription is random and a regime where transcription is highly correlated. We also find that divergent gene pairs show increased transcriptional activity, along with transcriptional waves and bursts in the highly correlated regime – all these features are associated with genomes of living organisms.

# Lay Summary

This thesis studies the properties of DNA in human and bacterial cells. In every cell nucleus there will be 46 chromosomes and therefore 46 different DNA molecules, as each chromosome consists of just 1 long DNA molecule. Along with DNA, a cell contains lots of different proteins that all behave in different ways - for example some might read DNA and others could act to connect different parts of a DNA molecule. Both the proteins and DNA are free to move around in the cell nucleus, allowing all these different types of interaction to take place.

The first two results chapters study what happens when DNA interacts with certain proteins which are also found in the cellular environment. By running computer simulations which model the behaviour of both the DNA and proteins, we were able to observe the 3D structures created by this DNA/protein interaction.

For our most simple model of DNA/protein interaction, proteins could bind to any part of a DNA molecule, which leads to the formation of protein clusters. This is due to something called "bridging-induced attraction", where a protein will bind to two distant parts of the same DNA molecule - forming a "bridge" between them. This in turn causes further protein bridges to bind nearby.

Our model was then made more realistic by the addition of genetic data, which allowed simulations where specific proteins bind to their target sites on the DNA. These simulation results compared well with experimental data for the 3D structures formed by DNA, suggesting this simple protein binding model could be enough to explain experimental results.

In the final results chapter we look at how supercoiling (a property of twisted DNA) can effect proteins which read genes and vice versa. This also reveals a relationship between the direction gene pairs are read and the frequency at which this reading process occurs.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in [8, 9, 11, 47].

(*J. Johnson, March 2018*)

# Acknowledgements

There are a lot of people I'd like to thank for their help, both while writing this as well as through the last 4 years I've been working on the PhD.

Firstly I'd like to thank my Dad, Robert and Penny for all their support, especially over the last few months when I've been stressed out about writing! It has meant so much to know I have people I can turn to no matter how big or small a problem might be.

On the academic side of things I owe a great deal to my supervisor Davide for putting up with me for almost 5 years now! I've learned a lot and it's always been a pleasure working with someone so consistently positive and helpful. All the other members of our group have helped me out in some way or another while I've been here, but especially Chris who has helped me through several projects over my time here - I think even starting way back when I did my masters project!

Finally, thanks to all my friends in Edinburgh for making my time here so much fun!

# Contents

# Introduction

DNA research will always interest people for it's own sake; when there exists one molecule which plays such a key role across all known life it is only natural to want to find out more about it! Advances in practical techniques such as gene editing with tools like CRISPR, genetic screening for diseases, and researchers using DNA as a method of digital storage show that there can be clear, direct benefits which arise from DNA research. However these exciting applications, as well as potential future ones, could not come about without an understanding of how DNA functions at a basic level. Similarly, an understanding of how DNA functions in the cellular environment and interacts with its surroundings is required before attempting to manipulate or alter the processes taking place in the cell. Although the work here does not necessarily have a direct application to practical methods, hopefully it can provide some understanding of the mechanics behind genome organisation and how this further relates to important cellular processes such as transcription.

As the fundamental molecule of living things DNA will always interest researchers across a wide range of different fields, each bringing their own different insights and methods to approach the many different unanswered questions relating to DNA. The main contribution to this effort will always be grounded in experimental work, as there is no substitute for performing experiments and seeing the behaviour in real life. However, methods from physics can provide a different perspective towards research, bringing a focus on universal or generic behaviours and minimal, first-principles models. While it is rare for a biophysical system to behave predictably enough that it can be fully captured in a simple mathematical model, often general underlying characteristics of the system can be uncovered. This type of approach can provide an extremely strong starting point when dealing with complex systems, with characteristics from a simplified model often still remaining applicable to the more complex "real" system.

A physics-based approach can also help to overcome some of the limitations which may accompany experimental methods. Sometimes, these limitations are a case of prioritising resources when experiments may take a long time or be expensive. In this case simulations can indicate whether a particular direction of research is worth investigating. At other times, experiments may simply be "difficult" in terms of complexity or experimental design. A simulation or numerical model may be simpler to set up, and it is often easier to make modifications to an ex-

perimental design in a mathematical model or simulation. For example, changing the binding energy of a protein in an experiment may require a significant experimental redesign, whereas in a simulation this could amount to just changing one number in an input file!

One of the other challenges when working in biophysics is making sure research is accessible to people who do not have much experience of either physics or biology. This thesis is a little more weighted towards the physics side, but hopefully should be comprehensible to non-physicists as well! Each chapter also has an introduction detailing the biological background material required.

This thesis focusses on the 3D structure of DNA and how this both affects and is affected by its interactions with various proteins. DNA is also an ideal subject for study using physics based methods, as it is well described by polymer physics models for the length scales we investigate. The first chapter contains detailed descriptions of these polymer models and how they are implemented in the molecular dynamics simulation program LAMMPS. There is also a brief introduction to the topological phenomenon known as supercoiling, which is re-visited in more detail in chapter 4.

Chapters 2 and 3 introduce a process we call "bridging-induced attraction", where protein bridges between different DNA regions promote further binding of proteins in the same region. This is simulated for protein/DNA models of increasing complexity, beginning in chapter 2 with a model where generic proteins bind non-specifically to DNA. In chapter 3 we then move on to a model where different types of protein bind to specific regions of the DNA, determined by genetic data from the ENCODE project. We found that by simulating this bridging-induced attraction mechanism we were able to reproduce results from Hi-C experiments, such as the location of topological domains in DNA. These chapters used molecular dynamics methods to study the protein-DNA system's evolution over time.

In chapter 4 we develop a numerical model which links supercoiling to gene transcription. These two factors are linked as negative supercoiling is known to promote transcription, while positive supercoiling reduces it. Meanwhile, transcription causes a flux of supercoiling by pushing positive supercoiling in the direction of transcription. Our numerical model found two distinct regimes, one at high flux where supercoiling regulates transcription and one at low flux - known as the relaxed regime. We also observed transcriptional bursts and waves, along with higher transcription rates at divergent gene pairs.

The code for all the simulations performed has been made available at `http://www2.ph.ed.ac.uk/~s0841882/downloads.html` and `http://www.jjthesis.co.uk/downloads.html`, in a form where (hopefully!) the simulation paramaters and behaviour can be easily modified. I've attempted to present the simulation programs in a way which makes them straightforward to use, so ideally people who don't normally do much computational work will be able to use them without feeling like they have to make a huge investment of their time! The programs provided give a way to independently verify the results in this thesis, as well as

allowing the user to perform a wider range of investigations than the ones detailed here. Ideally this will provide extra clarity when interpreting simulation results, as well providing an extra level of reproducibility for non-analytical results.

# Chapter 1

# Modelling DNA

## 1.1 DNA, Chromatin & Cells

DNA is the substance at the core of all known life on Earth. From single-celled bacteria to humans, every living thing relies on DNA to store the genomic data which, in a sense, makes them what they are. It is an extremely versatile material, which must allow genetic information to be accessible to the cellular machinery, whilst also being able to create almost error-free copies of itself during cell replication. This is not a simple task when a single DNA strand can contain up to 200 million individual pieces of genetic information, known as bases.

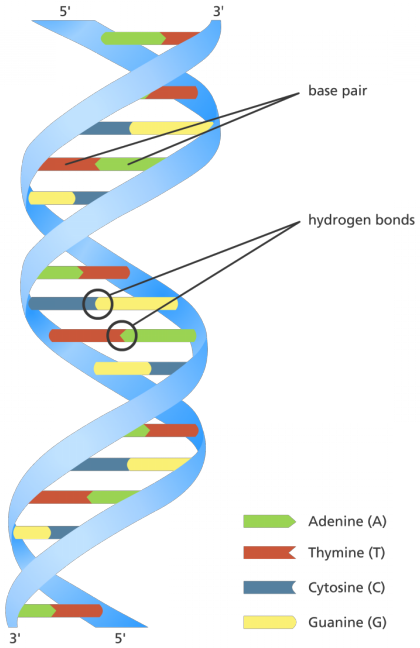



Figure 1.1: Image from [102]. A DNA double helix with labelled base pairs.

The DNA molecule itself has been a subject of continual study since its discovery in 1869, with its molecular structure identified as the iconic double-helix in 1953 by Crick, Watson

and Franklin. As shown in figure 1.1, DNA consists of base pairs which connect the twin helical backbones. There are four possible bases in DNA: these are adenine (A), cytosine (C), guanine (G) and thymine (T); A always pairs to T and C to G, so base pairs are denoted either AT or CG.

The main function of DNA is to contain the genes which code for specific proteins. The effect of these proteins is incredibly wide-ranging, from superficial things like hair and eye colour (the M1CR gene) to complex developmental processes (the SHH gene). However genes only make up 3% of the DNA in a cell, with the remaining 97% sometimes referred to as 'junk' DNA [15]. Although the remaining 97% does not have a direct effect on protein expression, it does play an important role in the way DNA is organised in the cell. In turn, DNA organisation influences which genes are expressed, and with what frequency.

Before going into futher detail on how DNA functions within the cell, it is worth taking a step back and explaining how a cell is structured and what its life-cycle looks like. An animal cell (figure 1.2) appears fairly disorganised at first glance, lacking the rigidity of their plant counterparts.

Within the cell itself the main component of interest for this thesis is the nucleus (see figure 1.3), which contains the cell's DNA and is where messenger RNA (mRNA) is transcribed. The nucleus is separated from the rest of the cell by a membrane which prevents unwanted molecules from interfering with the DNA, but also has pores so molecules which need to move in and out of the nucleus can pass across it. Outside of the nucleus there are various organelles such as mitochondria, which produces ATP for the cell and the endoplasmic reticulum, which acts as a transport network. Alongside these organelles, smaller molecules

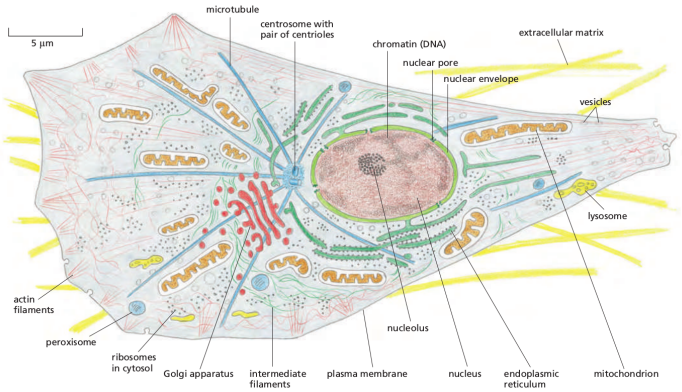



Figure 1.2: A diagram showing significant features in a typical animal cell. From [1].

such as RNA and proteins are also present in large numbers. However, for the simulations in this thesis we focus on DNA and so conditions outside the nucleus are not immediately relevant.

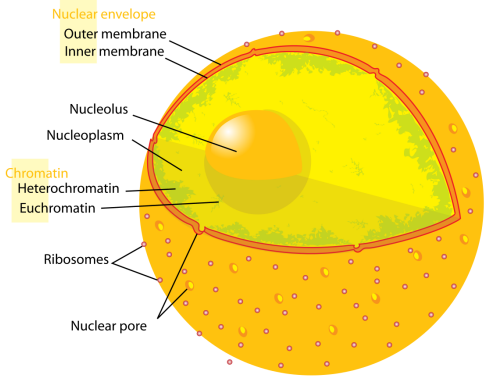



Figure 1.3: A diagram showing a cell nucleus. From [61].

The internal environment of a cell is also dependent on which stage of the cell cycle it currently occupies. The cell cycle (figure 1.4) describes how the cell grows, as well as how DNA configurations in the nucleus change between cell division events, which occur approximately once every 24 hours. The conformation of the cell's DNA changes quite significantly during this process, going from densely packed, X-shaped chromosomes in the mitotic phase to looser, less structured forms during interphase. The simulations in the thesis all take place while the cell is in interphase.

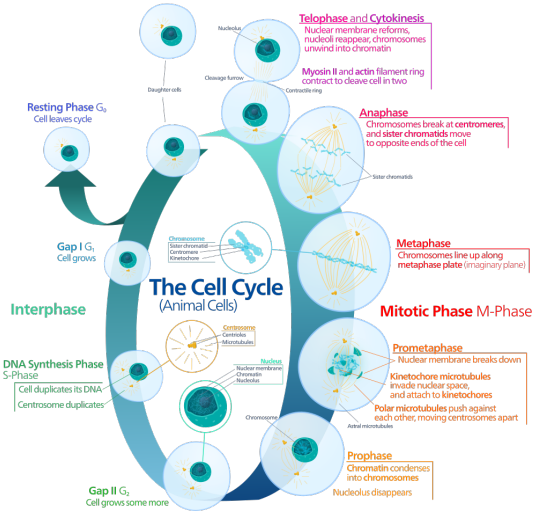



Figure 1.4: A diagram of the cell cycle, from [61].

Every cell in the human body contains a complete copy of our entire genetic code, which contains the genes they need to function correctly. This requirement alone creates a daunting storage problem, as up to 2m of DNA must be compacted into cells with a diameter of order $10^{-5}$m [14].

Figure 1.5: Image from [1]. Cartoon showing a the "beads on a string" model for the 10nm chromatin fibre.

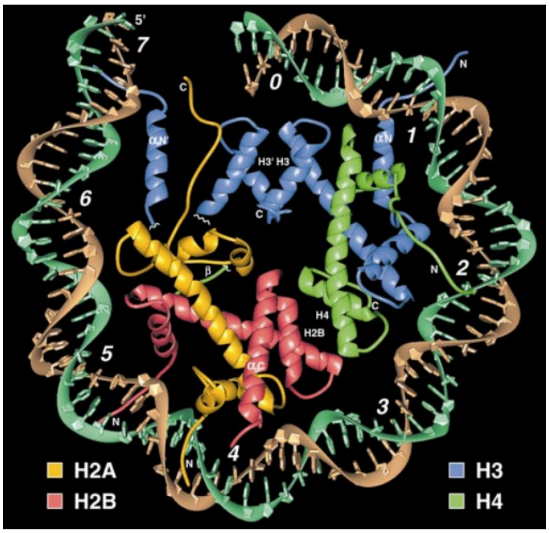



Figure 1.6: Image from [60]. A nucleosome with different histone proteins labelled. The numerals alongside the DNA indicate the number of double-helix turns, with a 72 bp length of DNA shown here.

While 2m of DNA could fit into a cell by rolling it up into a solid ball, this would render any genes contained near the centre inaccessible to the proteins which transcribe DNA. Instead, DNA molecules are compacted by forming a complex with multiple histone proteins which is known as chromatin. At the smallest length scale chromatin consists of repeating units of DNA wrapped around a nucleosome, forming a 10 nm fibre (Figure 1.5).

Each repeating unit contains ∼200bp, of which 147bp is wrapped around the nucleosome proteins while the rest links neighbouring nucleosomes. The nucleosomes themselves are made up of eight histone proteins, two molecules each of the four core histone proteins (H2A,H2B,H3 and H4). The structure, determined by X-ray crystallography [60] is shown in figure 1.6.

At physiological salt concentrations (∼ 100 mM of a buffer containing a monovalent salt such as NaCl), the 10 nm fibre is compacted further into a 30 nm fibre. The exact structure of this 30 nm fibre is still not completely clear, with a few competing models describing potential structures [91, 93].

While *in vitro* experimental data from x-ray crystallography of nucleosomes suggests a conformation like the zig-zag model in figure 1.7, cryo-electron microscopy (cryo-EM) of longer strings of nucleosomes supports the solenoidal model. Also,

while the 30 nm fibre has been observed *in vitro*, as well as *in vivo* for certain cell types, it is not necessarily true that it exists for all eukaryotes.

In fact, cryo-EM experiments which sought to characterise the structure of the chromatin fibre *in vivo* found no evidence of regular 30nm structures [78]. In the absence of 30nm structures, an alternative hypothesis is that the 10 nm fibre could be compacted into a irregularly structured fibre [62]. Recent work has also provided support for human chromatin being less regularly structured than assumed in the "textbook" hierarchical model, where 10 nm chromatin fibres fold into 30 nm, 120 nm and 300-700 nm fibres (Figure 1.8). Using a technique known as ChromEMT (Chromatin EM Tomography) fibres between 5 and 24 nm diameter were found, but regularly structured chains of greater diameter were not present [79].

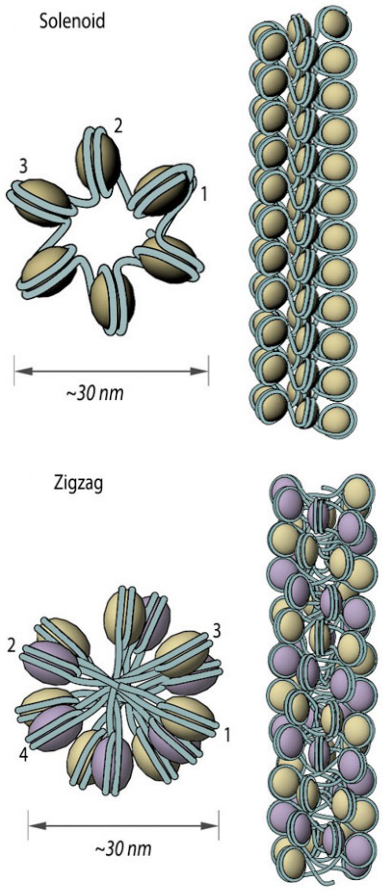



Figure 1.7: Images from [1]. Nucleosomes are represented by yellow/blue ellipsoids and DNA by the light blue string. Top: The solenoidal model for the 30 nm chromatin fibre. Bottom: The zig-zag model for the 30 nm chromatin fibre.

Although the exact nature of the chromatin fibre is an important area of study, the results here should not have too significant an impact on the design of our computational model for DNA. The two parameters in our model most likely to be affected by the structure of the fibre are persistence length and DNA packing density (the amount of DNA contained in a given length of chromatin). The persistence length ($l_p$) is a measure of how flexible a polymer is, with a short persistence length meaning a more flexible polymer. It can be calculated using $exp(-\frac{s}{l_p}) = \langle \mathbf{u}(s')\mathbf{u}(s+s') \rangle$, where $\mathbf{u}(s')$ is the tangent vector to the polymer at a point $s'$.

The value used for chromatin's persistence length in simulations (90 nm) is based on experimental measurements which put it between 30 - 200 nm. Experiments

performing scanning force microscopy analysis of the end-to-end distances of chromatin fibres on mica give values for the persistence length of 30-50 nm. As noted in *Langowski et al* [2], the details of the fibre/mica interaction can influence the measured persistence length. Similar results were derived from studies of recombination frequencies in human cells [87] and formaldehyde cross-linking probabilities in yeast [26].

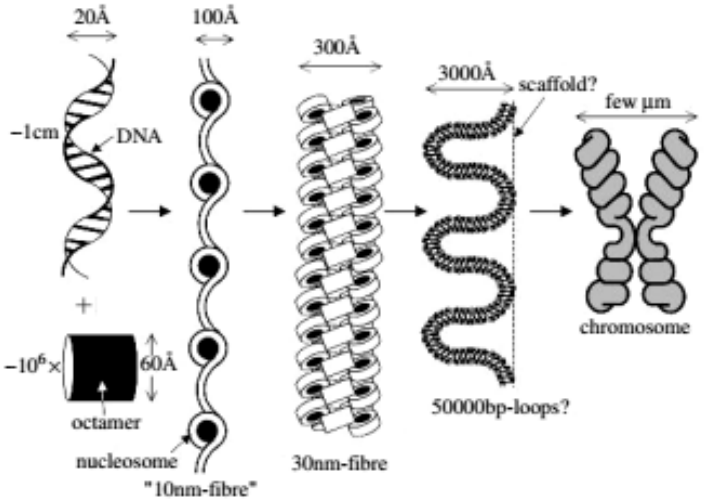



Figure 1.8: Image from [90]. A hierarchical model for chromatin structure, where the higher-order ($\geq$ 3000 Å) chromatin structures are thought to rely on an external protein scaffold rather than self-organisation alone. While putative component proteins for this scaffold have been identified, the scaffold itself has only been observed *in vitro* [81].

Further experiments based on measuring the distances between genetic markers in human fibroblast nuclei using Fluorescence in Situ Hybridisation (FiSH) gives persistence lengths between 100-200 nm [83]. In our simulations the chromatin fibre beads have a 30 nm diameter, meaning the persistence length is set at 3 beads.

An irregular chromatin fibre would mean that the persistence length varies between different parts of the fibre, although on average the persistence length should still be within the range of experimentally measured values. It is unlikely that replacing the constant value used in simulations with a distribution of possible values would make much difference to the simulations as the range of possible values is not likely to be particularly large. Similarly, while the packing density could vary for different sections of an irregular fibre, the average packing density would be similar to that of the regular fibre. As there is currently no data indicating the exact form of the packing density for an irregular fibre, we continue

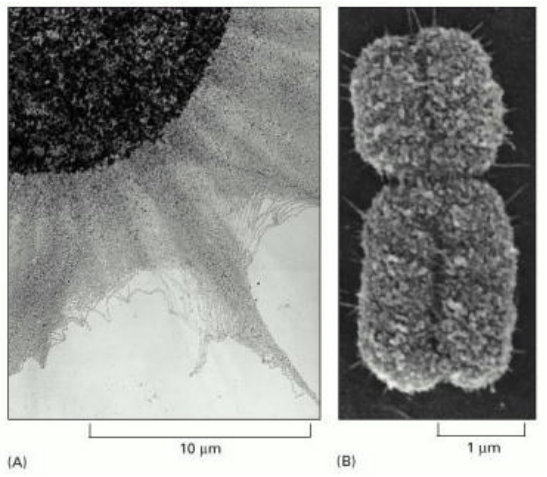



Figure 1.9: Image from [1]. A: An electron micrograph showing chromatin in the interphase state, escaping from a lysed nucleus (a nucleus where the membrane has broken down). B: A scanning electron micrograph of a mitotic chromosome.

under this assumption for the simulations in this thesis.

Chromatin displays further levels of structure beyond the nm scale (Figure 1.10), eventually leading to the familiar X-shaped chromosomes seen in figure 1.9. It is important to note that this X-shape structure is only present during phases of the cell cycle where the cell is close to replication (i.e. during mitosis). For the majority of the time chromatin remains in the non-dividing "interphase" state, which is also pictured in figure 1.9.

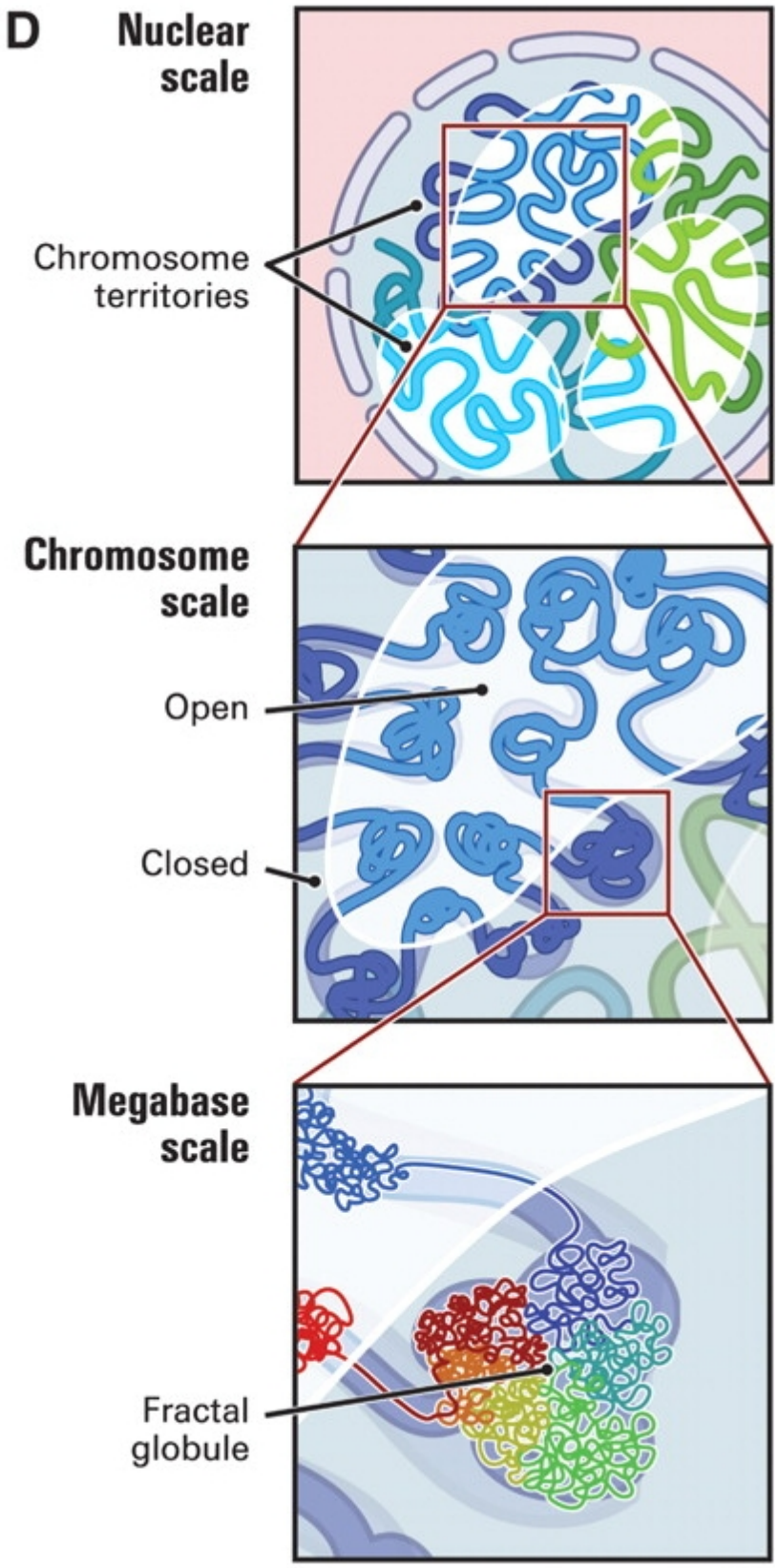



Figure 1.10: Image from [57]. In the top panel, different colours represent separate chromosomes which tend to remain in specific territories inside the nucleus [71]. The scales we work at in this thesis are most similar to the middle panel.

## 1.2 Polymer Models for DNA

While DNA is a polymer which has been studied in incredible detail, this does not mean that our mathematical models are required to incorporate every single feature of this molecule. Instead, the choice of model will depend on the characteristics we are seeking to study. Models do exist which move toward providing a base-pair [33] or atomistic [54] level of detail and would represent an "ideal" implementation of DNA in a computer simulation. However, the computational cost of this level of detail means that simulations necessarily represent a smaller time period (picoseconds/timestep) and generally model smaller sections of DNA.

This type of model can be used to study finer details of the mechanical properties of DNA, with one example being the response under a stretching force. Through computer simulation of short DNA strands, two studies [50, 56] were able to replicate experimentally derived force-extension curves. These curves showed that DNA would initially stretch elastically before transitioning to a regime where it stretched plastically (at constant force). At even higher extensions it would transition back to the elastic regime before breaking. The simulations were able to characterise the transition from elastic to plastic as a conformational change affecting the base stacking interactions in DNA, which were intra-strand in the elastic regime, but inter-strand in the plastic regime.

As we aim to study large sections of both chromatin and DNA, we require a model which is simpler but still representative of DNA. This necessitates some coarse-graining, and in the model a DNA molecule is represented as a polymer made up of monomer units with a semi-flexible connector between them. The extent to which the connector is flexible defines the persistence length $L_p$ for the DNA molecule, with less flexible connectors giving longer persistence lengths. For histone-free, or "naked" DNA, this persistence length has been measured experimentally at 50 nm [44]. As noted in [44], this can be done a few different ways - including electron microscopy. More recently, DNA persistence lengths have been calculated by tethering a nanoparticle to a substrate with DNA and tracking the motion of the nanoparticle [13]. The sequencing of the DNA being measured can also have an effect on persistence lengths, with regions containing more GC bases being less flexible.

In our model, each monomer can be thought of as being a set length of DNA and having the average properties of that length of DNA. As an example, in the simu-

lations in section 2.4 a DNA monomer has a size of 2.5 nm (∼7 bp). This averaging means that interactions between base pairs are not considered, so simulations where DNA is denatured (unzipped) would not be possible. The characteristic double helix motif can also still be represented by a string of monomers, as we can model its effects by setting appropriate interactions between the monomers.

Within this basic polymer model, there are two possible versions depending on the characteristics of the individual monomers. In the random walk (RW) model all individual monomers are assumed to have no volume and are able to occupy the same spatial position. Instead, in the self-avoiding walk (SAW) model, as the name suggests, this is not possible. This leads to an extra "excluded volume" interaction which increases the mean square end-to-end distance $< R^2 >$ of the chain due to the reduced number of possible chain conformations. The relationship between the number of monomers $N$ and $< R^2 >$ for both model types is given in (1.1) and (1.2), where the scaling exponent $v$ is 0.5 (RW) and 0.588 (SAW). Experimental measurements of $v$ for DNA give values in good agreement with the SAW model, with a measurement of $v = 0.571 \pm 0.014$ for a linear dna strand [95].

$$\langle R^2 \rangle = L_p^2 N^{2v} \quad \texttt{(RW)} \tag{1.1}$$

$$\langle R^2 \rangle = f(L_p) N^{2v} \quad \texttt{(SAW)} \tag{1.2}$$

The simplicity of the random walk model makes for more straightforward analytical calculations, but clearly is not so representative of a "real" DNA molecule. The computational cost of calculating steric interactions between monomers as in the SAW is also not particularly large, so it makes sense to design simulations which more closely resemble a SAW.

This base model can also be applied to simulations of chromatin or other polymer-like materials, with the same caveat that details below the chosen monomer size will not be present.

This model does not consider electrostatic effects between monomers. As DNA is negatively charged, there will be a repulsive force in effect between DNA sections. However at physiological salt concentrations ($\approx$ 100 mM of monovalent salt buffer) DNA has a Debye length of 1 nm [52], so as long as the monomer size is not smaller than this electrostatic effects will not have a significant effect on the simulation behaviour. For lower salt concentrations, or more detailed models at

physiological concentration, this electrostatic effect can be significant. For example, a DNA molecule has an effective thickness of 12 nm at 10 mM concentration and $\approx$ 50 nm at 0.1 mM [66].

## 1.3 Langevin Dynamics

As the intention is to simulate a DNA molecule in a cell, the DNA model used must also have some way of representing the internal cellular environment, which is composed primarily of water (70%), proteins (15%) and RNA (6%) [102]. As larger molecules such as proteins and DNA will be modelled explicitly, the cellular environment can be modelled by implementing the effect of an aqueous solvent on the molecules being simulated. This does neglect molecular crowding effects from the molecules not explicitly in the simulation. While this is done for computational efficiency reasons, the effect can be approximated by adding an attractive potential between DNA monomers. A trial run can then be performed to test the effects of neglecting this interaction.

The Langevin equation (equation 1.3) provides a mathematical description of this phenomenon. In addition to the non-solvent forces on a molecule, there is a noise term $\eta$ which represents the effect of random collisions with solvent molecules and a drag term which is dependent on the viscosity of the solvent $\gamma$ and the velocity of the molecule.

$$F_i = -\gamma v_i + \sqrt{2k_B T\gamma}\eta_i(t) + F_{others} \tag{1.3}$$

In the above equation, $\eta_i(t)$ is random uncorrelated noise, meaning it has the properties $\langle\eta_i(t)\rangle = 0$ and $\langle\eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t-t')$.

The motion of a larger particle due to collisions with the constituent particles of a fluid is known as Brownian motion. For particles with low mass we can neglect inertia and set $F_i$ to zero, recovering an equation for Brownian dynamics (equation 1.4).

$$\gamma v_i = \sqrt{2k_B T\gamma}\eta_i(t) + F_{others} \tag{1.4}$$

The dependence on $\gamma$ for both the drag and noise terms is obtained by writing

a Fokker-Planck equation for our system and imposing the condition that the equilibrium probability distribution coincides with the Maxwell distribution. A fluctuation-dissipation relation can then be used to link the drag and noise terms. A derivation for this in the Brownian motion case can be found in [51].

### 1.3.1 Further Hydrodynamics

Our model does not take into account hydrodynamic interactions between DNA segments, mainly for reasons of computational efficiency. The consequences of this can be seen by comparing two similar models for polymer dynamics, the Rouse and Zimm models. The Rouse model neglects hydrodynamic interaction and is used for systems with a high polymer concentration, where any hydrodynamic effects are screened out. In contrast, the Zimm model does include hydrodynamic effects and is appropriate for dilute polymers.

The different scaling behaviour for relaxation time and diffusion coefficient is shown below.

$$\texttt{(Rouse)} \quad \tau_R \propto N^{1+2\nu}, \quad D \propto \frac{1}{N} \quad \texttt{(Zimm)} \quad \tau_R \propto N^{3\nu}, \quad D \propto \frac{1}{N^{\nu}} \tag{1.5}$$

As $\nu \approx \frac{3}{5}$ for a good solvent, this means the Rouse model will have a longer relaxation time and smaller diffusion coefficient [30]. As our computational model is similar to the more simplified Rouse model, simulations will unfold a little more slowly than is realistic. However, this is a trade off worth making for efficiency reasons as implementing hydrodynamics can be very computationally costly. While hydrodynamic forces do have an effect in the cellular enivronment, they are less significant at the length scales of the simulation.

## 1.4 Implementation Using LAMMPS

LAMMPS is a classical molecular dynamics software package and was used for all of the molecular dynamics (MD) simulations in this thesis. While there are other MD packages available such as openMM, NAMD and AMBER, we chose LAMMPS because of its good parallel performance, open-source nature and already having some familiarity with the existing LAMMPS codebase.

All of the aforementioned software packages, including LAMMPS, function in more or less the same manner. All particles in the simulation have predefined force fields and the force on each particle is calculated at each timestep, which then allows the spatial position for each particle to be updated.

Since LAMMPS was not specifically designed for one particular style of simulation there are few restrictions on the form a force field can take, similarly particles can be representative of a wide variety of objects. In our coarse-grained simulations a particle will generally represent an amount of DNA, but in smaller scale simulations a particle could represent an actual atom.

In general, a molecular dynamics simulation consists of a number of particles which move around a simulation area and interact with each other. Usually there will be different 'types' of particle, with each type representing a different object - an example could be for particles of one type to be positively charged, and another type representing negatively charged particles. This representation then informs the interaction between the particles. In this case it would be a coulomb force, along with a repulsive force preventing particles from occupying the same region of space.

The implementation of the SAW model for DNA/chromatin consists of four force fields, which act on particles representing an amount of DNA/chromatin.

These forces can be described as follows:

- A steric interaction between particles using a Lennard-Jones or Weeks-Chandler-Anderson potential. The two are equivalent for the cutoff distance used below, which is $2^{\frac{1}{6}}\sigma$. This potential is defined below,

$$\begin{cases} E = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) & \text{if} \quad r < 2^{\frac{1}{6}}\sigma \\ 0 \quad \text{otherwise.} \end{cases} \tag{1.6}$$

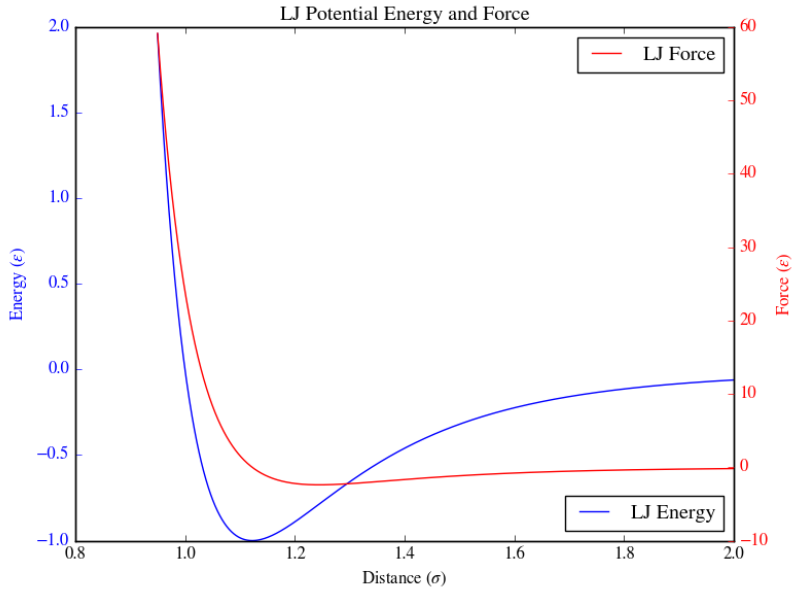



Figure 1.11: The force and energy contribution from a Lennard-Jones potential

The resulting inter-particle force between particles $i$ and $j$, $\mathbf{F}_{ij}$, is

$$\mathbf{F}_{ij} = -\frac{24\epsilon}{\sigma^2}\left(2\left(\frac{\sigma}{r}\right)^{14} - \left(\frac{\sigma}{r}\right)^{8}\right)\mathbf{r_{ij}} \tag{1.7}$$

In these formulas, $\sigma$ is the size of the particle and $\epsilon$ the interaction strength. The interaction cuts off at $2^{\frac{1}{6}}\sigma$ ($\sim 1.12\,\sigma$), the minimum of the Lennard-Jones potential. $\mathbf{F}_{ij}$ is the force on the $i$th particle from the $j$th particle and $\mathbf{r_{ij}}$ a vector directed from the $i$th particle to the $j$th.

- A permanent bond between adjacent particles, known as a Finite Extensible Non-linear Elastic (FENE) bond. The energy of a FENE bond is,

$$E = -\frac{1}{2}KR_0^2 \log\left[1 - \left(\frac{r}{R_0}\right)^2\right] + \epsilon, \tag{1.8}$$

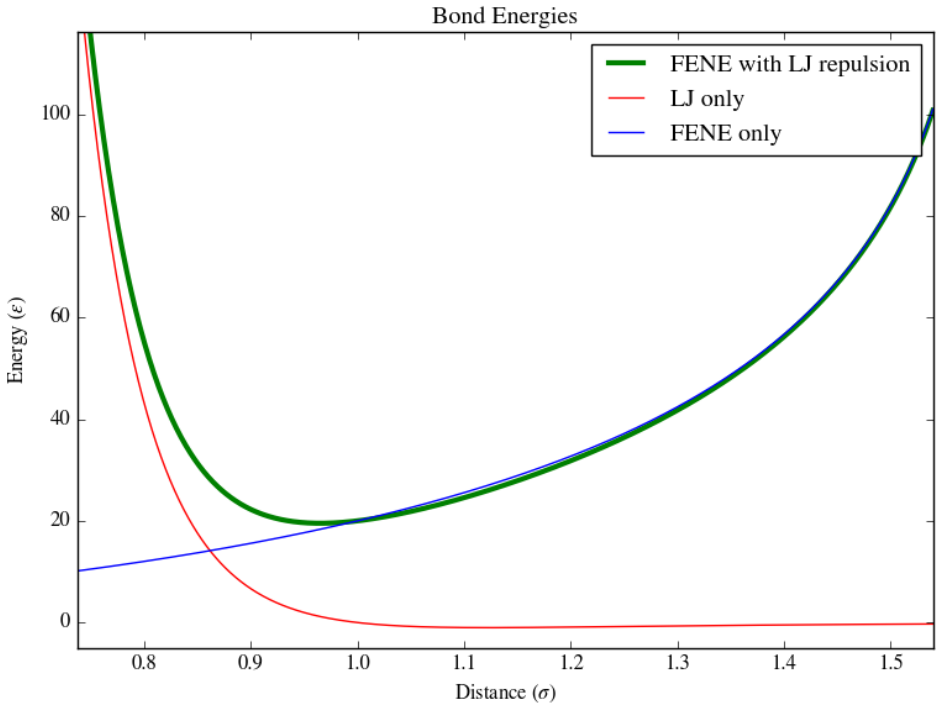



Figure 1.12: The energetic contribution of the FENE and Lennard-Jones Potentials

While the force is given by,

$$\mathbf{F}_{ij} = -K\frac{r}{1-(\frac{r}{R_0})^2}\mathbf{r_{ij}}, \tag{1.9}$$

where $R_0$ is the maximum bond length and $K$ is a factor determining bond strength with units $energy/distance^2$. For values of $r > R_0$, the bond length is large enough to consider the bond "broken" and the simulation is halted. $\mathbf{F}_{ij}$ and $\mathbf{r_{ij}}$ are defined as above. In LAMMPS, the FENE bond consists of this force along with the steric Lennard-Jones interaction above. For bonded particles, this interaction replaces the existing pair potential between the particles.

- A bending energy for the monomer chain, which is given by a cosine angle potential. This potential is defined by the formula,

$$E = K[1 + cos(\theta)] \tag{1.10}$$

where $\theta$ is the angle between three consecutive beads (Figure 1.14). The coefficient $K$ sets the persistence length for a DNA chain, with $L_p = K\sigma$. For example, in the case of DNA we have $\sigma = 2.5$nm and $K = 20$ in order

to get a persistence length matching the experimental value for DNA of $L_p = 50$ nm. The forces generated by this potential act on three particles (i, j, k) rather than pairwise, so deriving them is a little more convoluted. We require $\mathbf{F_i}+\mathbf{F_j}+\mathbf{F_k} = 0$ for there to be no net force on the system, along with there being no net torque on particle j. If this were not the case the angular potential would induce a drift and rotation to the system. Finding the force can be a little more tricky than previously, due to the reliance of both the potential energy and relative positions of $i$ and $k$ on $\theta$.

The forces on the particles end up being:

$$\mathbf{F_i} = \frac{Kcos(\theta)}{|\mathbf{r_{ji}}|^2}\mathbf{r_{ji}} - \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{jk}}$$

$$\mathbf{F_j} = \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{ji}} - \frac{Kcos(\theta)}{|\mathbf{r_{jk}}|^2}\mathbf{r_{jk}} + \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{jk}} - \frac{Kcos(\theta)}{|\mathbf{r_{ji}}|^2}\mathbf{r_{ji}}$$

$$\mathbf{F_k} = \frac{Kcos(\theta)}{|\mathbf{r_{jk}}|^2}\mathbf{r_{jk}} - \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{ji}}$$

Which is the result also found in the LAMMPS files for the potential. A derivation of this result is shown in A.1.

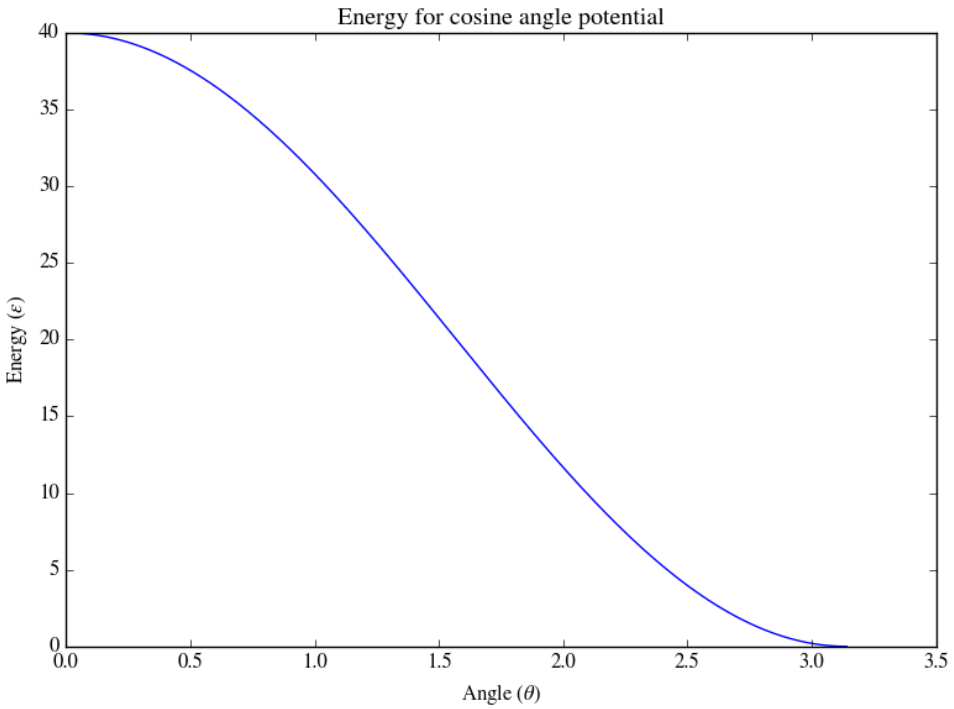



Figure 1.13: The energy contribution of a cosine angle potential, with $\theta$ given in radians.

- A force implementing Brownian dynamics for the $i_{th}$ bead in the simulation, as discussed in section 1.3.

$$F_i = -\gamma v_i + \sqrt{2k_B T\gamma}\eta_i(t) \tag{1.3}$$

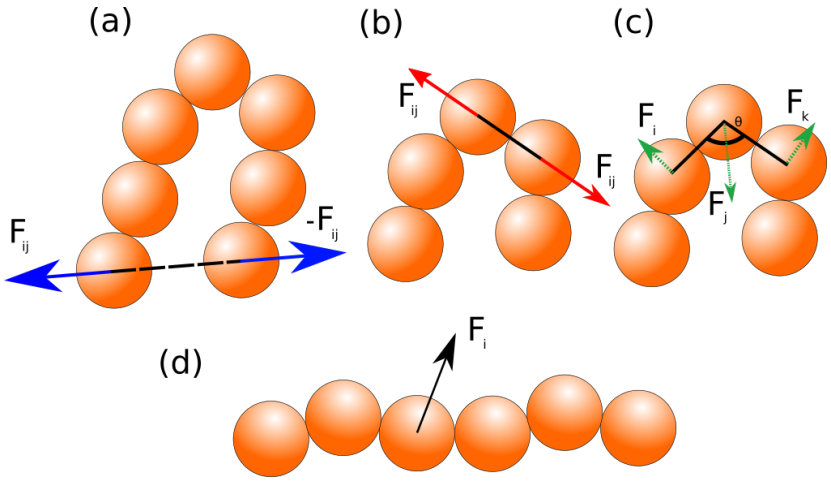



Figure 1.14: A cartoon showing the way the LAMMPS force fields act on particles in a chain of monomers representing DNA. The forcefields in this model are (a) A Lennard-Jones pair potential, (b) A FENE bond, (c) A cosine angle potential, (d) A Langevin thermostat.

## 1.5 Timescales and Units

When running computer simulations it is often useful to use reduced (dimensionless) units for measured quantities, rather than SI units. This is usually done so that simulation results are not given by numbers with very large or very small orders of magnitude. Since simulations of DNA involve distances of nm length, it makes sense to use reduced units here. In LAMMPS, this is done by initially setting the fundamental (and also dimensionless!) quantities mass, $\sigma$, $\epsilon$, and the Boltzmann constant equal to 1.

Table 1.1 shows how to convert reduced units to real units; LAMMPS uses the quantities $\sigma$ for distance and $\epsilon$ for energy.

It is also useful to compare natural timescales of the system to simulation timesteps. The simulation time units are defined in terms of the friction $\gamma$ from equation 1.3, with $\gamma$ having units of $s^{-1}$. In turn $\gamma$ is related to the diffusion constant of a monomer of size $\sigma$ by the Einstein-Smoluchowski relation $D = \frac{k_B T}{m\gamma}$. For simulations of DNA in the cellular environment, we generally have $k_B T = 1$ and $\gamma = 1$

| Quantity | Reduced (*) and Real Units |
|---|---|
| Mass | $m^* = \frac{m}{m}$ |
| Distance | $r^* = \frac{r}{\sigma}$ |
| Energy | $E^* = \frac{E}{\epsilon}$ |
| Temperature | $T^* = \frac{k_B T}{\epsilon}$ |
| Time | $t^* = t\left(\frac{\epsilon\sigma^2}{m}\right)^{\frac{1}{2}}$ |
| Force | $F^* = \frac{f\sigma}{\epsilon}$ |

Table 1.1: Reduced units used in LAMMPS

with the value of $\gamma$ being similar for both water and cytosol, the fluid within a cell.

For both DNA and chromatin simulations we can look at the brownian time $\tau_B = \frac{\sigma^2}{D}$, which is the order of magnitude of the time taken for a monomer to diffuse across a distance equal to its own size. As we use different values of sigma for simulations of chromatin ($\sigma = 30$ nm) and DNA ($\sigma = 2.5$ nm), one brownian time in simulation units corresponds to different timescales, with $\tau_B \approx 0.6$ ms and $\approx 36$ ns respectively. This allows us to run simulations of total length $\approx 200$ seconds for chromatin.

We also make the assumption that the DNA or chromatin is in an equilibrium configuration at the beginning of the simulation run. In real cells, chromosomes would take a prohibitively long time ($\approx$ 500 years!) to disentangle [88] if they were to behave as an equilibrated polymer solution. However, clearly cells do not live for hundreds of years and equally, chromosomes are not found to be tangled within the cell. The resolution to this given in [88] is that chromosomes never equilibrate and behave like unentangled ring polymers in a semi-dilute regime, which are known to be topologically segregated. This means our molecular dynamics simulations, which take place in a dilute, equilibrated polymer solution may take longer than is realistic to run. Since the chromosomes are segregated we do not have any issues arising from simulating a chromosome in isolation, but the chromsome will be more spread out than it should be within the cell. This will mean that the mechanisms we want to study may appear to take longer.

## 1.6 Supercoiling

Since we have discussed the effects of persistence length and bending energy on DNA, it makes sense that we also consider the effects of torsional stresses on DNA. These arise due to the helical nature of DNA, which can vary between DNA forms - some may be more tightly wound than others and while most helices are right-handed, left-handed helices do exist. Three of the most common forms (A,B and Z-DNA) are shown in figure 1.17 along with an explanation of their differences.


Figure 1.15: Supercoiling through the ages. Like DNA phone cord has an in-built curvature which leads to the crossed-over supercoils in the diagram. As for the games controller, I must have somehow managed to twist it round over the years - it probably shouldn't look like this!

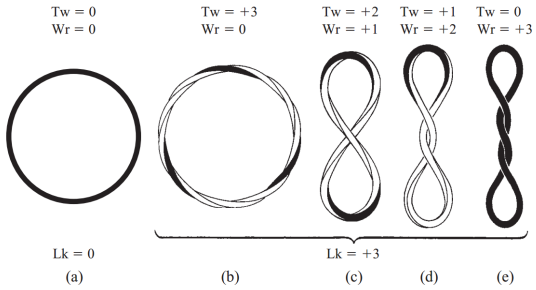



Figure 1.16: Image from [14]. A DNA molecule with a Linking Number (Lk) of 0 and four topologically equivalent molecules with $Lk = 3$.

The reason DNA has a helical structure at all is due to the molecular configuration of the A,T,C and G bases. While an untwisted 'ladder' of base pairs would seem the simplest possible structure for DNA, the length of the bonds between adjacent base pairs would leave significant gaps between them. By twisting the bases into a helical structure, this bond also has a horizontal component - meaning the base pairs are moved closer together. In cells, DNA is always found to be underwound, with one missing turn of twist for every 17 turns of stable, right-handed double helix [14]. This happens as DNA will always coil around proteins in the cell nucleus in a left-handed toroidal spiral, which gives a negative Lk. These torsional stresses, combined with the interchangablity of twist and writhe lead to the phenomenon known as supercoiling.

The effects of torsional stress are similar even at very different length scales,

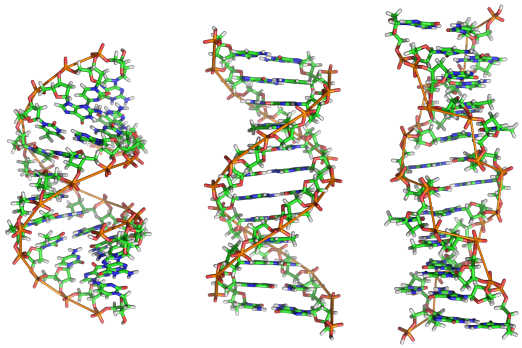


Figure 1.17: From left to right: A, B and Z-DNA. A and B-DNA are both right handed but B is slightly more twisted, with 10 phosphates per helical turn compared to 11 for A. Z-DNA is left handed with 12 phosphates per helical turn.

so we can look at more familiar objects to get an idea of what supercoiling is. Figure 1.15 shows the effects of supercoiling in two cables, with crossovers in the cable forming due to the system being torsionally stressed. As an aside, in DNA a supercoil with a shared loop base and interwound coils extending from it is known as a plectoneme.

Increasing the number of twists in a chain or loop leads to the formation of supercoils, since there is a point where conformational changes are more favourable than the torsional energy cost. Whether the system creates a supercoil in order to reduce twisting is determined by three factors - the entropic cost of stabilising a loop at the base of the plectoneme, the energetic cost of bending and the energetic cost of twisting the DNA. The fact that this is an option at all is due to the topological equivalence of the quantities twist ($Tw$) and writhe ($Wr$). To a first approximation, twist represents how many $360^o$ turns a cable would make when forced to lie in a planar, circular configuration, while writhe represents the number of self-crossings a cable makes over itself. This is most easily seen in figure 1.16. We also refer to the sum of twist and writhe as the Linking Number ($Lk = Tw + Wr$) as this quantity is conserved in loops or linear chains with fixed ends. Linking number, twist and writhe can all be positive or negative depending on the direction (right or left handed) of the twist. Right handed (clockwise) twists are taken to be positive, though this is a completely arbitrary choice. Twist and writhe do not have to be integer values either and while it is obvious that a chain can have a half twist, it is less clear what a non-integer writhe looks like. While we know a loop with $Wr = 2.5$ would look like a hybrid of (d) and (e) in figure 1.16, unfortunately it is also a difficult concept to represent diagrammatically!

In nature DNA supercoils are generally underwound, meaning they have a negative linking number (figure 1.19 shows increasing levels of underwinding). This can be fairly large even for short DNA loops, a 7000 bp loop ($\approx 2.4\mu m$) has a linking number of $-40$. While figure 1.19 shows negatively supercoiled DNA, the equivalent amount of positive supercoiling would look exactly the same. As mentioned previously, overwound DNA supercoils tend not to be found in nature.

Supercoiling can also be confined to regions within a DNA strand, which makes sense as linear DNA strands exhibit supercoiling, despite being free to rotate at their ends. Regions of supercoiling have been observed experimentally, however the exact mechanism behind their formation is not completely understood. The boundaries of these supercoiled regions also often coincide with topologically associated domain (TAD) boundaries, where a TAD represents sections of DNA which are spatially close (see figure 1.18). These supercoiling boundaries often have an increased amount of binding sites for the protein CTCF (CCCTC-binding factor), suggesting CTCF could act as a barrier to supercoiling [40].

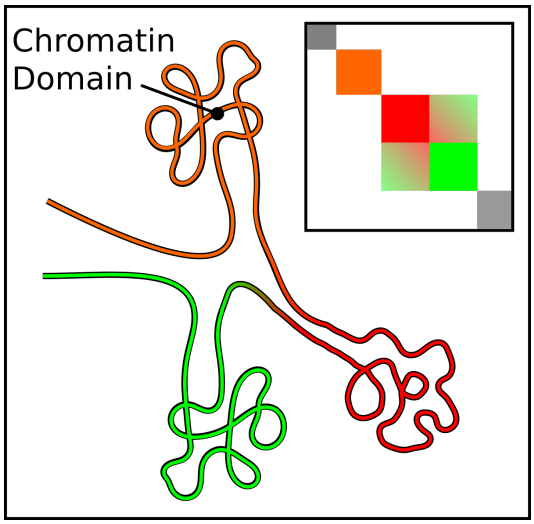



Figure 1.18: A cartoon showing a chromatin fibre folded into TADs. Inset: A possible contact map for this system.

Implementing supercoiling in LAMMPS can also be a difficult task, as the base particles in a molecular dynamics simulation are usually isotropic spheres which do not have an orientation, in order to make calulations as simplified as they can possibly be. Adding orientation to the particles usually requires modifying or building the program with specific extra packages (in LAMMPS this is the ASPHERE package). A suitable interaction representing a twisting potential between neighbouring particles must also be implemented, as well as a separate potential for particles which are barriers to supercoiling. While in this thesis we consider numerical simulations of a 1-D model for supercoiling rather than MD, the 3-D implementation may be of interest to some readers. A potential which can be used in LAMMPS is availible from the paper by *Brackley et al* in the Journal of Chemical Physics [10].

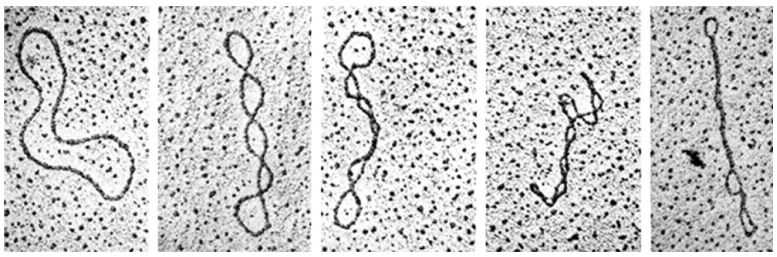


Figure 1.19: An electron micrograph of DNA loops with successively greater levels of supercoiling. Despite this image being widely used I could not find the source paper, but it is very similar to the images from *Vinograd et al (1965)*[97]

## 1.7 Running a DNA only simulation

If you are interested in running a simulation, example code and an installer for all required software is available at `http://www2.ph.ed.ac.uk/~s0841882/downloads.html` or `http://www.jjthesis.co.uk/downloads.html`. There are also videos and graphs from an example run of the DNA only model available at `http://www2.ph.ed.ac.uk/~s0841882/simulations.html` or `http://www.jjthesis.co.uk/simulations.html`.

You can run simulations of supercoiled and non-supercoiled DNA (or a mixture!), along with other simulations detailed in the other chapters.

# Chapter 2

# Bridging Induced Attraction

The following chapter is based around the J. Phys Condensed Matter paper "A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction" [47].

## 2.1 DNA-Binding Proteins

Within both bacterial or eukaryotic cells, DNA does not exist in isolation. Proteins make up approximately half the dry weight of a cell in the *E-Coli* bacterium, corresponding to around $10^6$ proteins per cell [68, 76].

Of these, many proteins will directly interact with DNA and are involved in processes such as gene-regulation, transcription and genome organisation. There are $\approx 3 \times 10^4$ DNA-binding proteins in an *E-Coli* cell. Since the *E-Coli* genome is $\approx 4.6$ Mbp this means we have one binding protein every $\approx 150$ bp.

Binding between proteins and DNA takes place when amino acids in the protein come into contact with base pairs in a DNA molecule. Individual amino acids within the protein will bind to a particular base (Figure 2.1) and with a large enough binding energy the protein will remain in place. A typical binding protein in both eukaryotic and prokaryotic cells can have 10-20 contacts, meaning that they will only be able to bind to certain matching DNA sequences. DNA-binding proteins tend to be structured so that they have positively charged amino acids facing the negatively charged phosphate backbone of the DNA [1].

An example of such a protein in bacteria is the "histone-like" H-NS protein, which forms dimers that bind to AT-rich DNA. This type of interaction is known as non-specific binding, since the protein does not target any particular gene, promoter or other identifiable marker. There are many other proteins which display this type of behaviour, a further example being the polycomb repressive complex (PRC1) protein in Drosophila. This complex binds to chromatin at many locations, while also having a sub unit which self-polymerises and thus allows PRC1 to bridge different regions of the chromatin fibre.

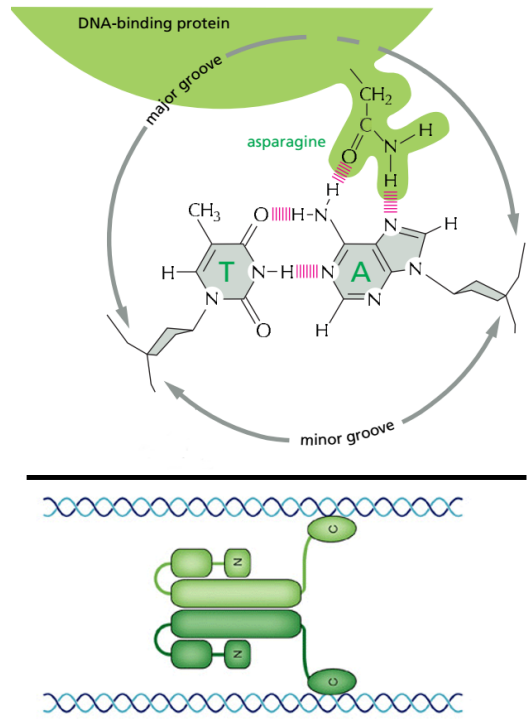



Figure 2.1: Both images from [1]. Upper: The amino acid asparagine within a hypothetical protein binding to an adenosine (A) base within DNA. Lower: Cartoon showing a H-NS dimer bridging two regions of a DNA molecule. Image from [31].

Similarly, since both parts of the H-NS dimer can bind to DNA simultaneously (Figure 2.1), H-NS binding can bridge genomically distant regions of the bacterial chromosome, though there will also be bridging between regions which are already reasonably close.

Proteins which bind non-specifically tend to make fewer contacts with the DNA, greatly increasing the likelihood of finding matching bases in a given region. For example, the "Zif finger 2" protein will bind to two consecutive Guanine (G) bases in DNA, a pattern which is likely to occur extremely regularly.

In a eukaryotic cell, DNA binds with histones to form chromatin as discussed in section 1.1. As with H-NS in bacteria, this binding is non-specific. Alongside histones there are also many other DNA-binding proteins, each with differing effects on gene regulation. For example, the protein HP1 compacts the chromatin fibre into denser, transcriptionally-inactive heterochromatin, while CTCF proteins bind to specific genetic sequences known as promoters. Binding to a specific genetic sequence is possible as proteins are effectively able to "read" which bases are present from the outside of the DNA double helix.

CTCF will bind to specific locations on the genome while also being capable of binding to two DNA sections simultaneously. This has the effect of bridging two promoter-containing regions and providing easier access for polymerases, which results in increased transcription of the gene associated to that promoter. CTCF is part of a group of proteins called transcription factors; a eukaryotic cell will contain $O(10^4)$ of this type of protein.

## 2.2 The Effects of Non Specific Binding

While specific binding is clearly important to cell functionality, we initially studied the effects of non-specific binding proteins on both DNA and chromatin. Due to their abundance in both bacterial and eukaryotic cells, it is important to understand the collective behaviour of these proteins and how it influences their interactions with DNA.

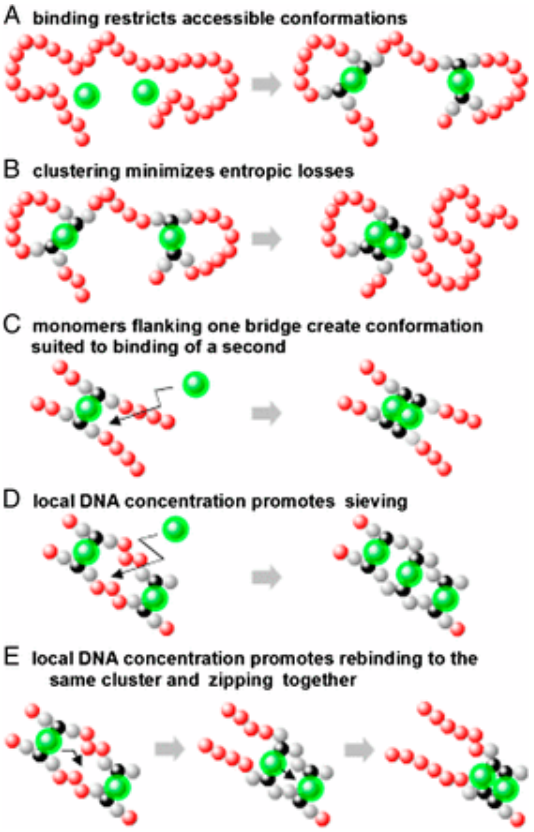



Figure 2.2: Image from [12]. Potential mechanisms by which bridging proteins can facilitate clustering.

Previous work from our group [12] showed via simulation the existence of an effect known as "Bridging Induced Attraction". This refers to the way bridging proteins will tend to promote further protein binding at locations where there is already a higher concentration of proteins. This happens despite there being no interaction (other than steric effects) between the proteins themselves. A protein binding does not directly change the way additional proteins interact with the DNA either, the interaction strength remains the same.

Instead it is the effect the bridging protein has on the 3D organisation of the DNA which causes these clusters of proteins. Bridging distorts the DNA locally

in a number of ways, either by bringing distant DNA sections together, creating loops, or by bending or straightening DNA. As shown in figure 2.2 there a few different effects arising from the bridging proteins, which can be discussed in a little more detail.

Parts A and B in the figure can be taken as two parts of the same process. While having a protein bind to the DNA is clearly energetically favourable, it restricts the number of available conformations of the DNA as the binding requires two sections of DNA which are distant genomically to be close spatially. As we add more binding proteins to one DNA molecule the number of available conformations is lowered even further, however this can be addressed via the mechanism shown in part B. By moving the proteins close together the system retains the energetic favourability from binding, but the number of possible DNA conformations is not reduced by as much as before.

Part C shows how the binding of one protein means local DNA beads are left at a suitable distance apart for further protein binding, while parts D and E address the effect of higher local DNA concentration. These can either increase the likelihood of a new protein binding to a region with high DNA concentration as in part D, or provide a pathway for two proteins to move closer together as in part E.

## 2.3 Experimental Studies of Protein Clustering

While the work in [12] looked at systems *in silico*, protein clustering has been observed in several different biological contexts. In the fly genome, transcription factor proteins form clusters at specific points along the genome, while molecules of RNA polymerase are known to cluster around transcription factories [6, 80]. Experiments *in vitro* using synthetic gold nanoparticles designed to be representative of histones also show clustering behaviour [105, 106], as do experiments *in vivo* on the H-NS protein in bacteria which found the formation of row-like clusters [24, 100]. As mentioned in section 2.1, the PRC1 protein has also been observed to cause clustering behaviour in *Drosophila* cells [101].

## 2.4 Model and LAMMPS Implementation

We extend the DNA/Chromatin polymer model from section 1.2 by adding particles representing the proteins discussed above. The protein particles are spheres which can simultaneously interact with two or more DNA monomers. These particles should be thought of as representing a generic, non-specifically binding protein (Figure 2.3).

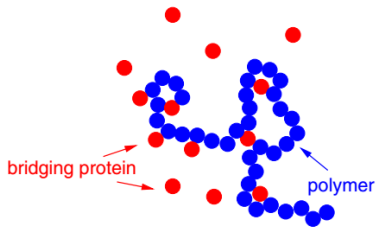



Figure 2.3: Image from [47]. A schematic of the protein-DNA model which shows how protein binding can affect the DNA conformation.

The LAMMPS implementation of the attractive interaction between proteins and DNA uses the Lennard-Jones pair potential again, but with the cutoff distance $r_{cut} = 1.8\sigma > 2^{\frac{1}{6}}\sigma$ meaning there is an attractive potential whenever the protein-DNA separation $r$ is $2^{\frac{1}{6}}\sigma < r < r_{cut}$.

The interaction strength $\epsilon_l$ can also be set, and was varied in order to view the effect on protein clustering and polymer compaction. The results in this section have $\epsilon_l$ in the range $0.5\,k_bT < \epsilon_l < 5.0\,k_bT$ as setting $\epsilon_l$ to values much less than $0.5\,k_bT$ or much greater than $5.0\,k_bT$ makes no qualitative change to the simulation outcomes. To give some context to these numbers, the adsorption energy for DNA-histone binding is estimated to be approximately $6\,k_bT$, so we are considering values in line with reasonable binding energies [90].

One technical note is that the L-J potential in LAMMPS is shifted by adding a positive constant $\epsilon_{shift}$, which is chosen so the potential is zero at the cutoff of the potential. This means the value supplied in lammps scripts $\epsilon_l$ is slightly greater than the true value $\epsilon$.

Another important parameter in the simulations is the concentration of proteins $c_p$ and DNA $c_d$. This is varied between simulations, although rather than changing $c_p$ and $c_d$ independently, $c_d$ is fixed and $x = c_p/c_d$ is varied.

A fixed value for $c_d$ also means that the simulation box size $L$ is fixed. In simulations $L = 200\sigma$ meaning the DNA is in the dilute ($R_g \ll L$) or semi-dilute ($R_g \approx L$) regimes, depending how strongly the polymer is compacted. For simplicity, the size of the proteins was set at $\sigma$, the same as DNA monomers. Consid-

ering different sized proteins would be more representative of real DNA-binding proteins, but would again make no qualitative change to the simulation outcomes. In the DNA simulations our protein size (considering H-NS) is 15.6kDa, which correspons to a radius of gyration of about 2nm [73] making them approximately the same size as the DNA monomers (2.5nm). For comparison, an "average" protein size is approximately 27kDa for *E-Coli* and 43kDa for humans [69] meaning the size for both is of the order of nanometers (this conversion assumes 1 amino acid = 100Da). However when we consider the chromatin simulations and their larger monomer size of 30nm, the protein size does not match up quite so well - in simulation our proteins will be larger than is realistic. Despite this, it is stil reasonable to model the proteins as being of size $\sigma$ since all the protein-DNA interactions specified in figure 2.2 will still apply, just with a larger separation between DNA monomers then would be expected with realistic proteins. The important feature of the model is that the protein causes the two bound DNA monomers to be spatially close and this still occurs with over-sized proteins.

If the reverse situation was true and the proteins were in fact much larger than how they were represented in simulation this would cause a problem, as the larger proteins would cause changes to the DNA conformation (e.g. wrapping the protein or bridging a large distance between DNA) which would not be seen in simulation without modelling the proteins as their true size.

## 2.5 Results - Chromatin

Figure 2.4 shows a typical simulation run for a 15 Mbp (5000 monomers) chromatin fibre and 1000 proteins. The fibre has a persistence length of 3 monomers ($\approx$ 90 nm). For the value of $\epsilon_l$ used here bridges stick almost irreversibly to the chromatin and form many small clusters. These clusters then combine until only a single cluster remains. This occurs by fusion of clusters which meet stochastically (coalescence), rather than one cluster growing at the expense of another (Ostwald ripening). This process also leads to the compaction of the chromatin fibre, which can be seen by measuring the radius of gyration ($R_g$) of the DNA. $R_g$ is useful as a measure of polymer size, as well as being a quantity which is accesible experimentally. Chromatin compaction is also strongly dependent on the protein concentration of the system, as shown in figure 2.5.

The fraction of proteins located in a cluster is $\simeq 1$ for all concentrations, meaning

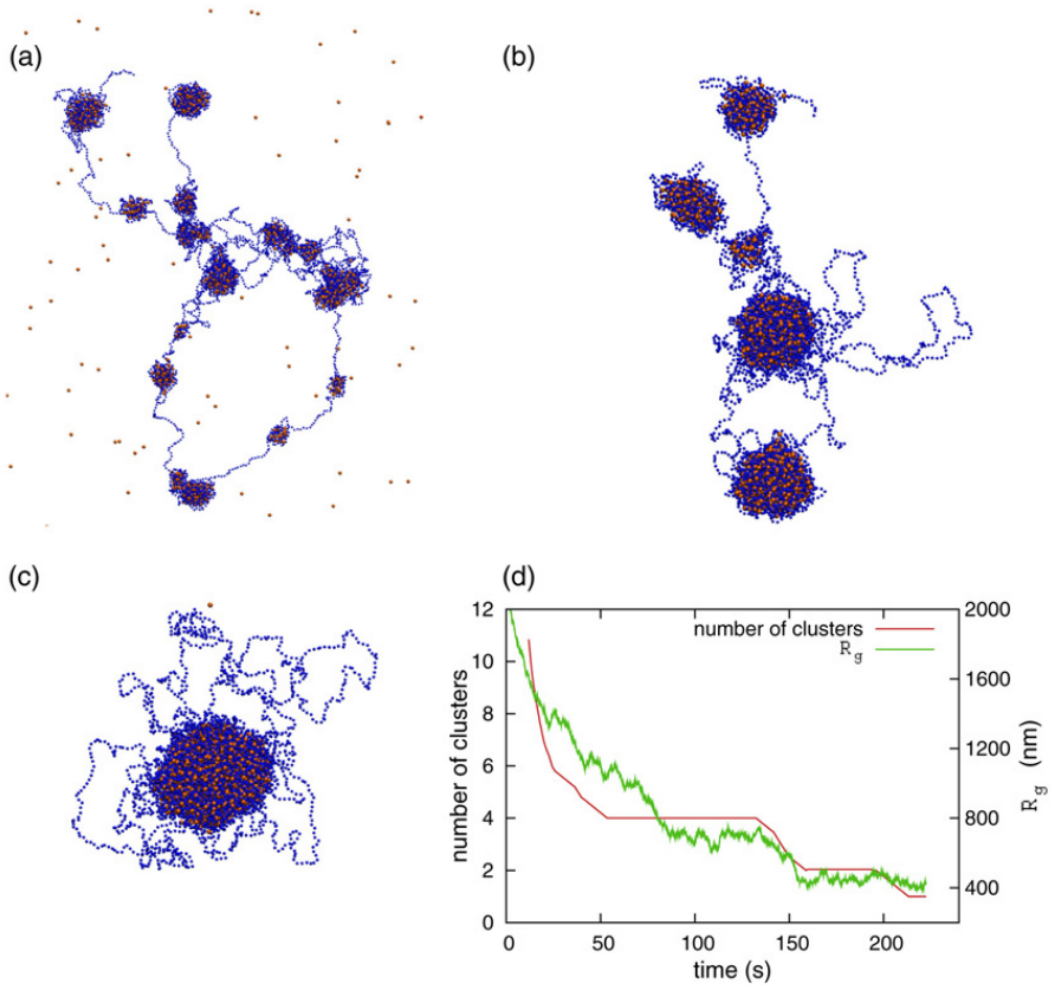



Figure 2.4: Image from [47]. (a)-(c): Snapshots at increasing times from a simulation with a 5000 monomer chromatin fibre and 1000 proteins. (d): The relationship between $R_g$, number of clusters and time. The simulation has parameters $\epsilon_l = 3\,k_bT$ ($\epsilon = 2.83\,k_bT$) and $r_{cut} = 60.6\,\text{nm}\,(2.02\sigma)$. The number of clusters in (d) is a local time-average.

there are very few isolated bridging proteins. This is most likely due to the mechanisms illustrated in figure 2.2. For all values of $x$ in the range $0.1 \leq x \leq 0.5$ we expect to see coarsening until a single cluster remains, as we expect any pair of clusters will at some point be spatially close for large enough values of t. However for intermediate values of $x$ the simulation time ($10^6$ Brownian Times) is not sufficient to complete the process.

This is because the kinetics of coarsening become much slower as the initial clusters increase in size. While clusters fusing is still a stochastic process, the probability of two large, distant clusters moving close and fusing in a given time period becomes small. For small values of $x$ the initial clusters are small enough

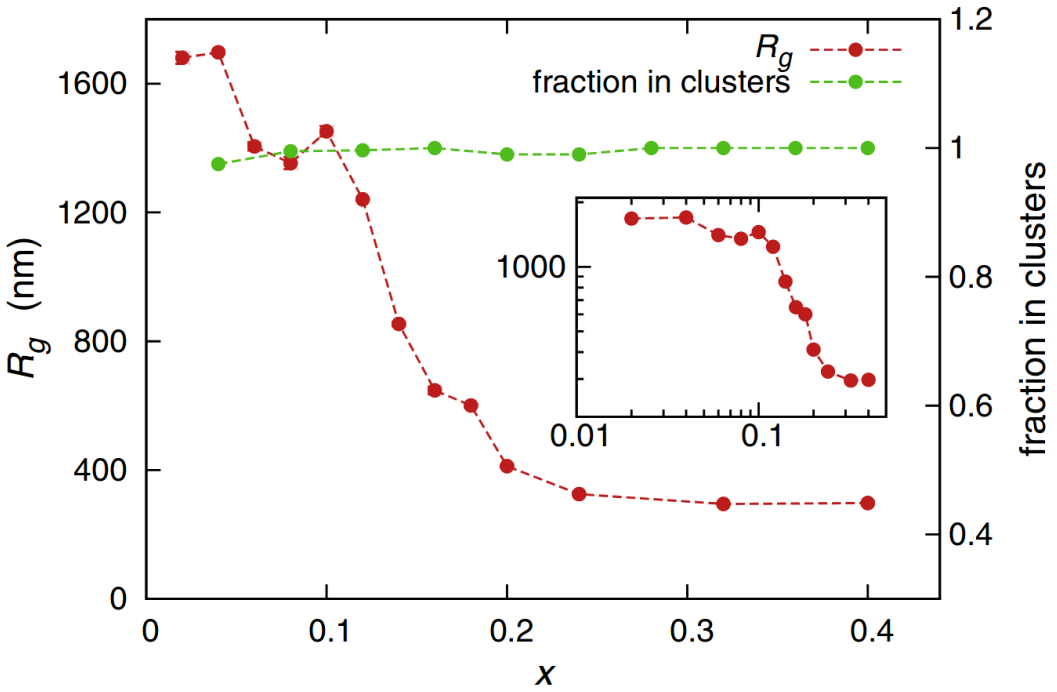



Figure 2.5: Image from [47]. The radius of gyration ($R_g$) and fraction of proteins in clusters for the chromatin fibre at simulation run end ($t > 200$s) for different values of $x = c_p/c_d$. The value plotted is an average taken over the final 60 ms of the run. The parameters $\epsilon$ and $r_{cut}$ are as in figure 2.4. Inset: log-log scale plot of the same graph.

that they can still diffuse quickly and form a single aggregate in the given time. Whereas for large values of $x$ binding will occur all over the chromatin fibre and the distance between initial clusters will be relatively small, speeding up the coarsening process. While it seems more likely that the clusters would all eventually combine for intermediate values of $x$, it is also possible that clusters at intermediate values could be dynamically stabilised - following trajectories where they do not interact, even across long timescales.

Increasing the value of $x$ leads to a transition from an open phase with large $R_g$ to a more compact structure with small $R_g$. For the parameters from figure 2.5, the equilibrium state consists of a co-existing open region and a denser globular region formed by the bridging induced attraction of the proteins. The volume of this globule scales linearly with $x$, at least until $x$ is so large that the globule encompasses the entire chromatin fibre. This type of bridging induced attraction leading to chromatin compaction has been observed in simulation-based studies by *Nicodemi et al* [77] and *Barbieri et al* [3].

Both the aforementioned papers and our simulations show a similar 'switch-like'

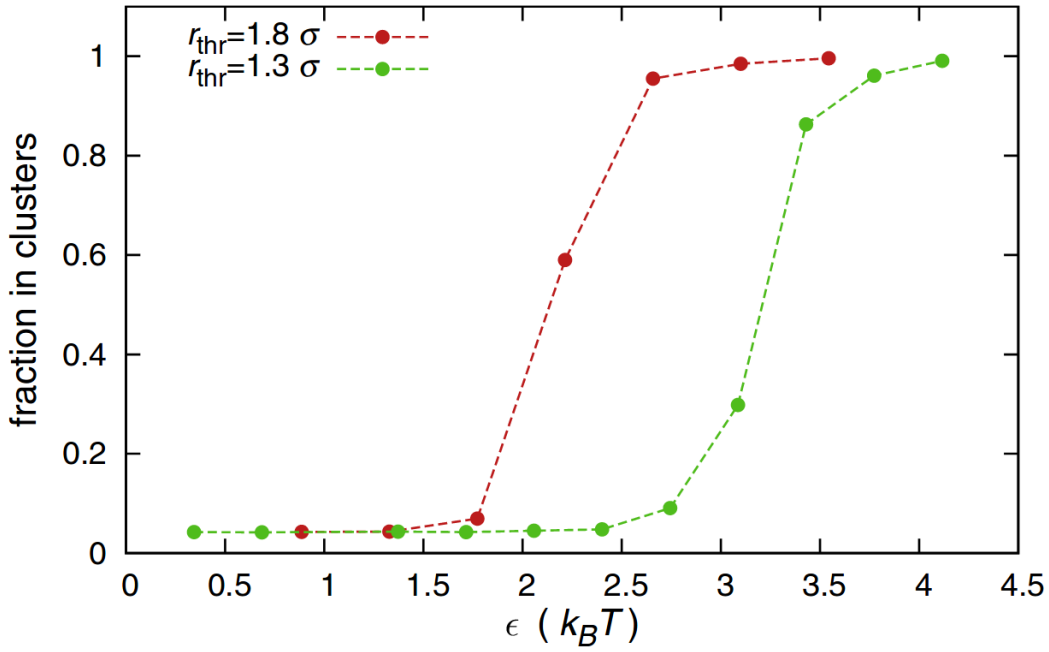



Figure 2.6: Image from [47]. Fraction of proteins in a cluster at simulation run end. For all of the simulations runs used, $x = 0.08$ (5000 DNA monomers and 400 proteins). The graph shows a fairly sharp transition between a regime where few proteins are bound and a regime where almost all proteins are bound.

transition from swollen to globular when increasing protein concentration or binding energy. The main difference between these works and the simulations carried out in this thesis is the simulation scale, with our simulations using polymers of $\approx 10$ times greater length. There are also some technical differences, as the model in *Nicodemi et al* and *Barbieri et al* is lattice-based and performs Monte Carlo simulations using the Metropolis algorithm, where our simulations use a Molecular Dynamics model.

The reorganisation of the polymer which occurs as a result of the bridging-induced attraction can be compared with experimental observations of chromatin. It is known that chromatin fibres are disordered, with compact heterochromatic regions interspersed amongst open euchromatic ones [78]. The coexistence of a cluster or globule state with a more open region reported here provides a generic pathway to drive segregation of different chromatin states.

Protein binding, and consequently chromatin compaction also has a dependence on $\epsilon$ and $r_{thr}$ which is shown in figure 2.6.

Both values of $r_{thr}$ show almost no binding for low values of $\epsilon$, up until a point $\epsilon_c$ where there is a sharp increase in binding probability. For $r_{thr} = 1.8\sigma$ (54 nm),

$\epsilon_c \approx 2.2\, k_B T$; while for $r_{thr} = 1.3\,\sigma$ (39 nm), $\epsilon_c \approx 3.1\, k_B T$. Since the values for $\epsilon$ are low enough that individual beads can dissociate after binding to the polymer, increasing $\epsilon$ leads to an increase in the average time a protein stays bound to a region of DNA. This increases the rate at which the stochastic bridging-induced attraction process occurs. If this is too low this process may never get started and leave just a minimal number of proteins bound. As this is not a phase transition, there are values of $\epsilon$ where bridging induced attraction occurs but is partially balanced out by proteins detaching from the DNA. The value of $\epsilon_c$ will also have some dependence on $c_p$ and $c_d$, again with an increased $c_p$ giving a higher binding probability.

## 2.6 Results - DNA

The simulations shown above were also run for the case of a semi-flexible fibre, representing naked DNA rather than chromatin. As noted in section 1.2, the model for DNA used has $\sigma = 2.5$ nm and a persistence length of 20 monomers (50 nm). These altered parameters mean that while the simulations in both sections have 5000 monomers and run for the same number of timesteps, we are actually considering different length- and time-scales. A chain of 5000 monomers corresponds to 36.8kbp in the semi-flexible (DNA) case as opposed to 15Mbp for chromatin. Similarly, a simulation running for $10^6$ Brownian times corresponds to around 10 ms for DNA but around 200 s for chromatin.

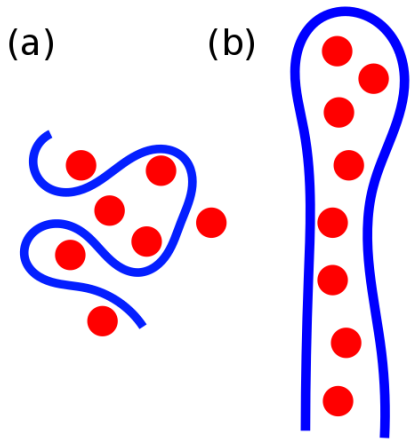



Figure 2.7: A cartoon showing a potential configuration of the system when bending energy is (a): less than protein binding energy, or (b): greater than protein binding energy. DNA is represented by the blue line and proteins by the red circles.

The clusters formed by bridging-induced attraction of proteins are now cylindrical, due to an apparent increased stiffness of the fibre. However, this is really due to the fact that we are considering a very different length scale than before - nearly 1000 times smaller. The entirety of this 5000 monomer DNA simulation contains the same amount of DNA as $\approx 12$ monomers in the chromatin simulations. Bridging between genomically local DNA monomers now carries a greater energy cost as the DNA is likely to have to bend substantially to accommodate this, whereas if bridging occurs between more distant DNA monomers the DNA may not have to bend as sharply. In addition to this, if proteins end up arranged in rows many bridges between two DNA segments may be formed for the "cost" of only one bend in the DNA (Figure 2.7). The clusters seen here are also qualitatively similar to those seen experimentally in [24, 100].

While the system quickly settles into a configuration where there are only a small number of large clusters, it coarsens much more slowly than in the chromatin case (Figure 2.8). This could be because the increased bending energy makes it more difficult for distant clusters to move close together and combine.

As with the chromatin model, increasing $x$ leads to increased compaction of the chain. This happens much more gradually (Figure 2.9) when compared to the sharp fall at $x \approx 0.15$ seen in figure 2.5. This is most likely since proteins now induce longer range contacts between DNA monomers, which can more efficiently compact the fibre than the local contacts seen in the more flexible chromatin fibre.

Changing the protein-DNA interaction energy (Figure 2.10) gives an effect similar to the one seen in figure 2.6, with the critical threshold beyond which clustering sets in slightly higher at $\epsilon_c = 3.4\,k_BT$ for $r_{th} = 3.25\,nm$ $(1.3\,\sigma)$. The 'switch' type behaviour is for the same reasons as in chromatin, while the higher energy requirement is due to the lower flexibility of the polymer.

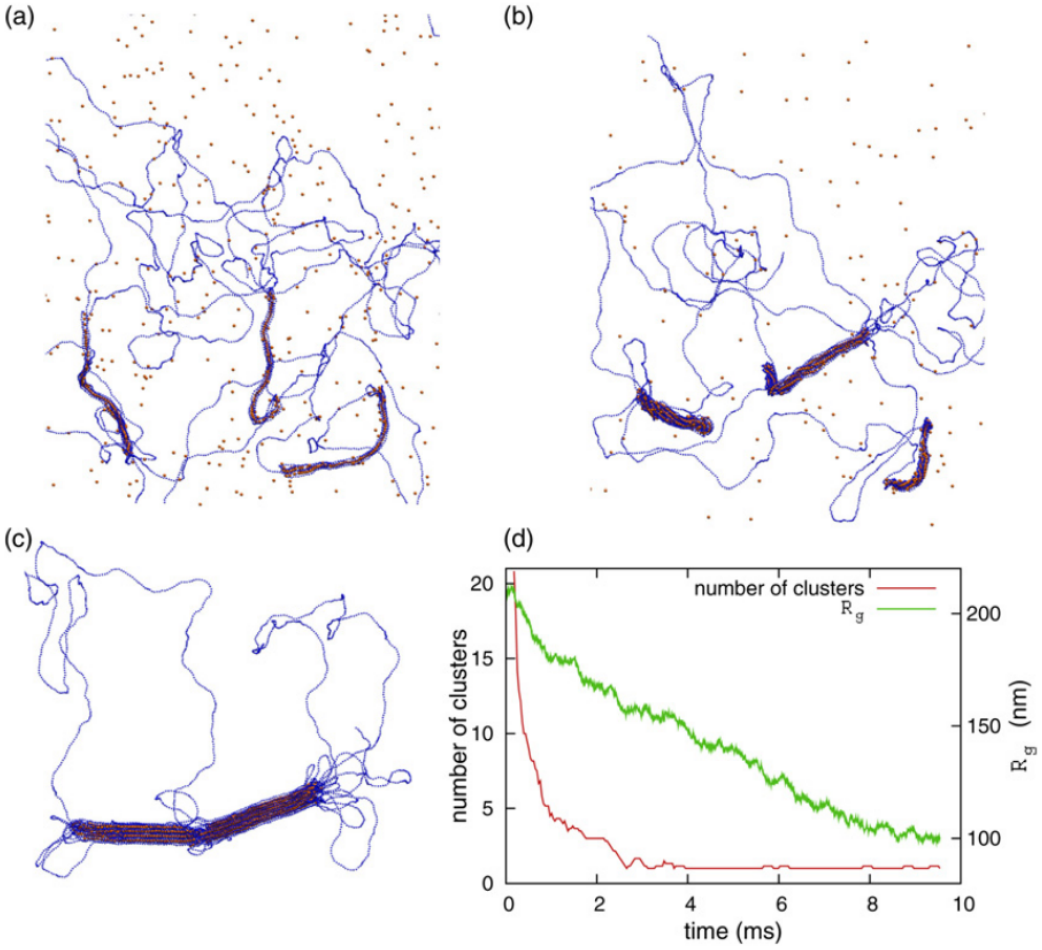



Figure 2.8: As Figure 2.4, but for naked DNA. (a)-(c): Snapshots at increasing times from a simulation with a 5000 atom DNA fibre and 1000 proteins. (d): Relationship between $R_g$, number of clusters and time. The simulation has parameters $\epsilon_l = 3\,k_bT$ $(\epsilon = 2.83\,k_bT)$ and $r_{cut} = 5.05nm\,(2.02\,\sigma)$. The number of clusters in (d) is a local time-average.

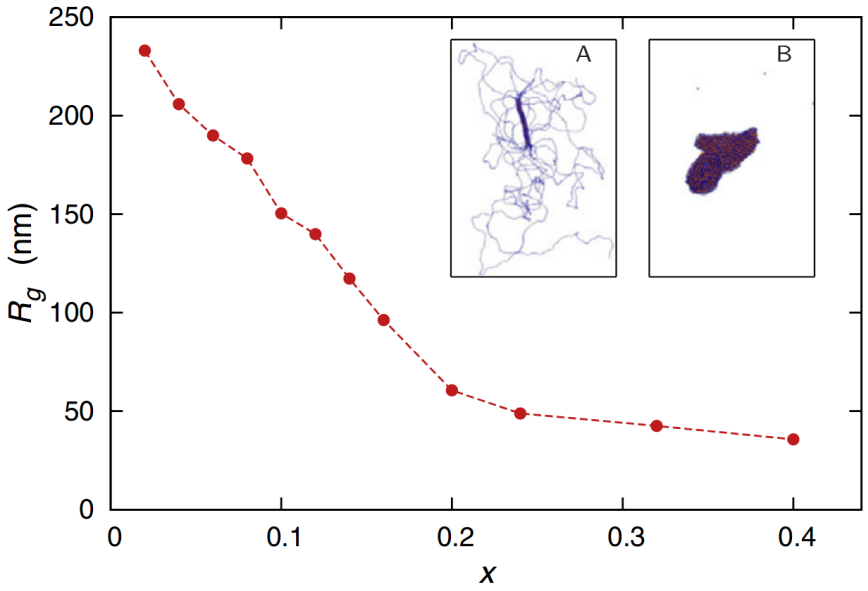



Figure 2.9: Radius of Gyration ($R_g$) at simulation run end ($t > 18$ms) for different values of $x = c_p/c_d$. The value plotted is an average taken over the final 3 ms of the run. The parameters $\epsilon$ and $r_{cut}$ are as in figure 2.8. Insets: A - At end of run with $x = 0.04$, B - At end of run with $x = 0.4$

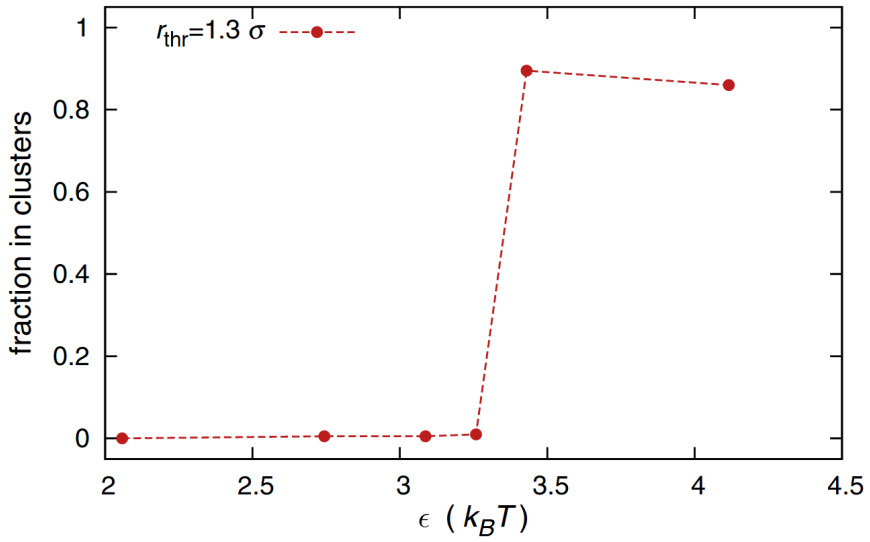



Figure 2.10: As figure 2.6, but with naked DNA. The interaction range $r_{cut}$ is set at 3.25 nm ($1.3\,\sigma$). Similar to the results in figure 2.6, there is a sharp transition between the regime where proteins bind to the DNA and a regime where they fail to do so.

## 2.7 Running the Simulations and Videos

A copy of all required software and the python script to run or modify the simulations can be found at my university webpage `http://www2.ph.ed.ac.uk/~s0841882/` or at `http://www.jjthesis.co.uk/`, along with a standalone lammps input script if you already have LAMMPS available on your computer. Videos showing both a flexible and semi-flexible simulation run are online at `http://www2.ph.ed.ac.uk/~s0841882/chapter1.html` or `http://www.jjthesis.co.uk/chapter1.html`.

## 2.8 Summary

The simple protein-DNA bridging model studied here provides a generic mechanism for cluster formation among bacterial DNA and chromatin. Even when the interaction is completely non-specific, there is a qualitative similarity between the results of the simulations here and the observations of experimental studies. For example, we see the same clustering behaviour as seen in experiments with DNA and nanoparticles but using different levels of particle concentrations and particle sizes. We can see a clear link between protein and polymer concentration $x$ and the degree the polymer is compacted, with low protein concentrations causing only local DNA compaction and leaving co-existing globular and swollen regions.

As the protein concentration increases, this swollen region shrinks and the polymer size drops sharply. This observation connects the results of the precursor to this study [12], which studied clustering for relatively low values of $x$ and [3], where $x$ was typically larger than 1. In the *Barbieri et al* paper, the system was observed to switch between an open and bridging-induced compacted phase on varying either protein affinity or concentration.

As an extension to this work, it would be interesting to quantify the exponents determining the growth laws of clusters in both the flexible and semi-flexible cases. The simulations could also be performed at a larger scale, although this would be difficult considering our computational resources.

Since the interactions in our model are non-specific, the logical next step is to extend the model to include the sequence specific protein-DNA interactions found in vivo. We develop this extended model in chapter 3.

# Chapter 3

# Transcription Factor Binding Model

The following section is based on the Nucleic Acids Research paper "Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domain" [11] and the Nucleus paper"Simulating topological domains in human chromosomes with a fitting-free mode" [9]

## 3.1 Outline

The aim of the work in this chapter was to design and simulate a model where transcription factors and other proteins would bind to specific sites along a section of chromatin. This idea was a continuation of a previous model [12] (Chapter 2) where transcription factors would act as bridges, linking regions of chromatin which were distant genomically, but were close spatially.

To do this we used data from the ENCODE (Encyclopedia of DNA Elements) project, an online resource for genomic data [49, 82]. This allowed us to identify the regions of DNA where specific proteins would bind. After simulating this bridging process we were able to create contact maps, which show regions of chromatin that are spatially close and also allow us to identify TADs (Topologically Associated Domains). The results of this process were then compared to existing Hi-C contact maps [84], which highlight regions of DNA which are spatially close. This was done in order to see how successfully the model predicted TAD boundaries and other characteristic features of the experimental contact map.

## 3.2 Topological Domains and Contact Maps

The 3D conformation of human chromosomes is an important area of research in genome biology, as this organisation influences gene activity, which in turn has consequences relating to health and general cellular function [16]. A useful representation of a particular chromosome's 3D structure is given by the chromosome's contact map, which involves binning the chromosome up into equally sized sections and then determining which sections are in close spatial contact. This would usually be done at a sufficiently high resolution (Usually on the order of 20+ kbp), so that the output contact map is not too sparse. In the example contact maps in figure 3.2 we can see the brighter square regions along the diagonal which represent TADs - contiguous regions of chromatin where there are a lot of contacts between different parts of the chromatin fibre. There are also a few other typical features visible, such as the high degree of contacts along the diagonal and the sharp boundaries between domains.

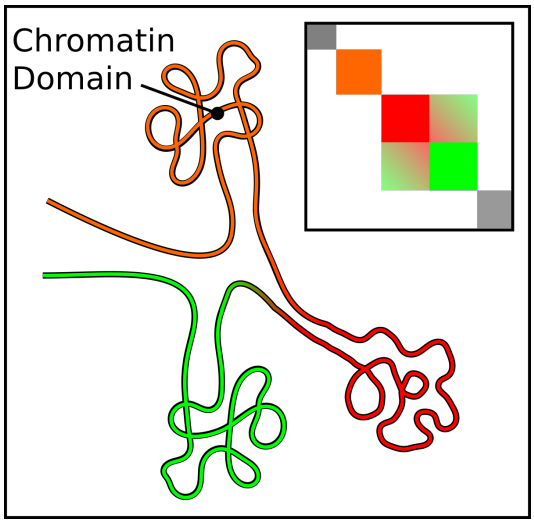



Figure 3.1: A cartoon showing a chromatin fibre folded into TADs. Inset: A possible contact map for this system.

The link between 3D conformation and gene activity also comes from the fact that chromatin folds into local domains (Figure 3.1). For example, genes embedded in a dense and globular domain in "heterochromatin" (inactive chromatin) are likely to have a reduced probability of transcription, while genes in more open regions will have an increased probability of transcription. Transcription factors and polymerases are also known to be localised at domain boundaries, suggesting that in some cases an active gene may act as a boundary [29]. Chromatin domains tend to make few inter-domain contacts, while having many intra-domain contacts. While this is expected from the definition of a domain, it is worth noting the extent of the intra-domain contacts. Even chromatin at the 'edge' of a domain will have a lot of contacts with all other parts of the domain, rather than just having local contacts. These topological domains can also separate active and

inactive regions of the genome [20].

A typical size for a TAD in humans is between 0.1-2 Mbp, though the characteristics of TADs are also dependant on factors like cell type. One measurable characteristic is the probability of two chromatin regions separated by a genomic distance $r$ being in contact $P_c$, which scales with $P_c \propto r^{\alpha}$. For HeLa cells $\alpha = -0.5$ [75], corresponding to larger domains on average than for stem cells where $\alpha = -1.6$ [3].

## 3.3 Fractal and Equilibrium Globule Models

The fractal globule model has been proposed as a general organisational principle for chromatin. The model does not take into account local details of the chromatin and every part of the chromatin fibre is treated in the same manner. *In vivo* this will not be true as some regions will be more active and have a more open conformation, while other inactive regions form denser structures. This model instead seeks to reproduce the average properties of TADs and make more general statements about chromosome architecture. The fractal globule structure consists of a polymer which collapses in a hierarchy of folds: some large scale domains resulting from this folding are highlighted in colour in figure 3.3.

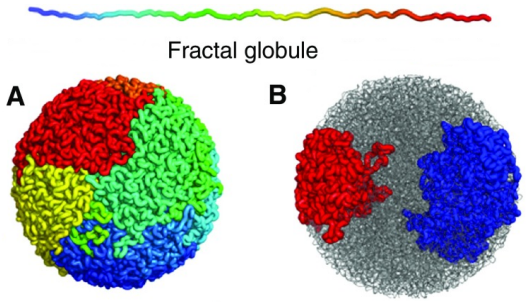


Figure 3.3: Image from [70]. A simulated fractal globule and its domain structure.

These folds can be produced *in silico* by setting a short-ranged, attractive interaction between monomers, and performing a rapid simulations where the polymer is quenched from the swollen phase, without allowing the chain to equilibrate: this typically also results in knot-free structures. As the fractal globule is space-filling, its radius $R$ scales with the number of monomers $N$ with $R \propto N^{\frac{1}{3}}$. The attractive interaction between monomers is an effort to represent molecular cross-links within chromatin. Hierarchical folding may occur thanks to the short simulation run times, which result in mainly local interactions within the chromatin fibre as there is not enough time to 'search' for longer range contacts.

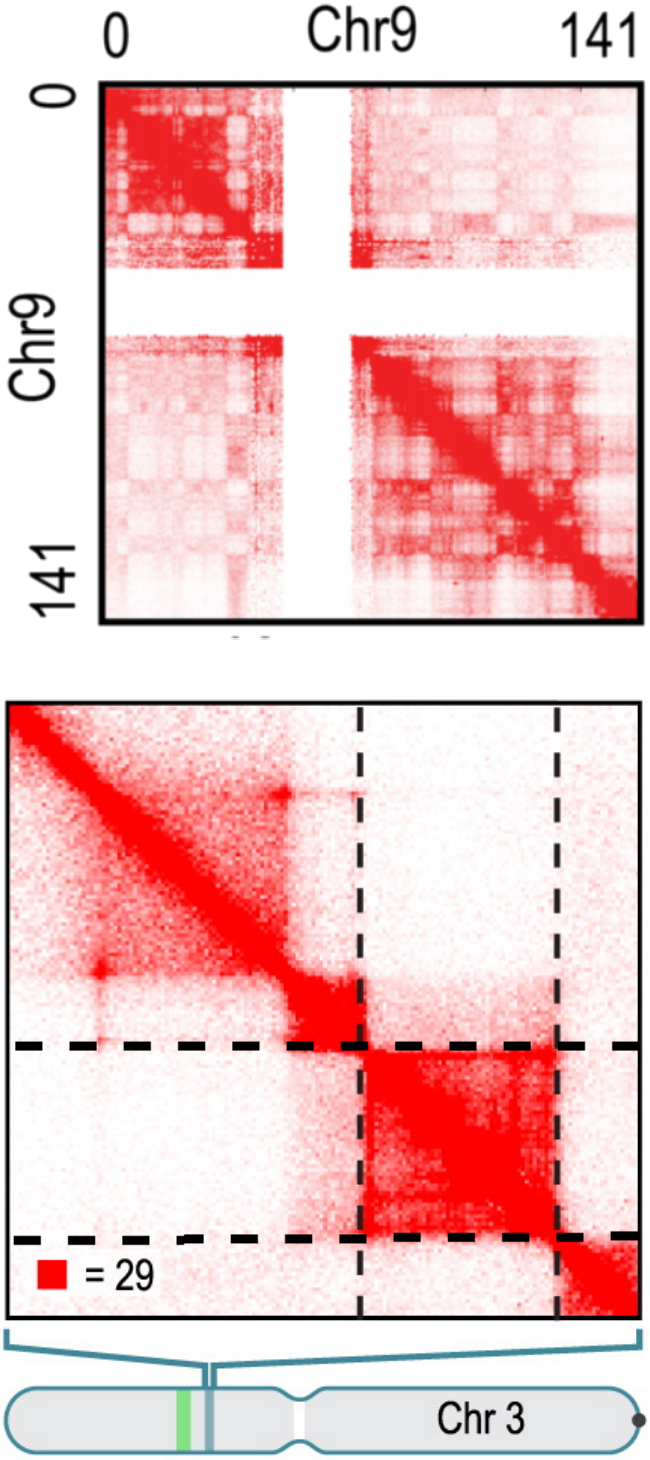



Figure 3.2: (Top) An example contact map, taken from [84], detailing contacts between regions of chromosome 9. The contact-free area near the middle is the centromere of the chromosome (this is the region where duplicated chromosomes are kept together prior to mitosis). (Bottom) A zoomed in view of a contact map for chromosome 3, highlighting a TAD. In both images, brighter red regions indicate more contacts between chromation.

On account of the hierarchical folding, the size of any subsection of the full polymer also has the same scaling behaviour. The "fractal globule" theoretical model for chromatin architecture has $\alpha = -1$ [70] in the scaling relationship mentioned above, which compares favourably with experimental Hi-C data, at least when analysing average contact probability curves, where data from all chromosomes are used at the same time.

As a contrast, we can look at the equilibrium globule (Figure 3.4). This is the equilibrium structure formed when the attractive interaction between monomers is allowed to dominate the repulsive interaction due to the excluded volume of the chain. To acheive this in simulations generally requires neglecting topological constraints as in *Mirny et al* [70]. In this model an individual monomer is considerably more likely to come into contact with a monomer which is a large genomic distance away. In simulations this is reflected in the scaling of $P_c$ for this model, with $\alpha = -\frac{3}{2}$ for short/mid-range interactions ($r \leq N^{\frac{2}{3}}$) and constant for longer range interactions [70]. While the scaling in this model compares well enough with some cell types, it also has a high degree of knotting, which makes it a poor choice for modelling chromatin.

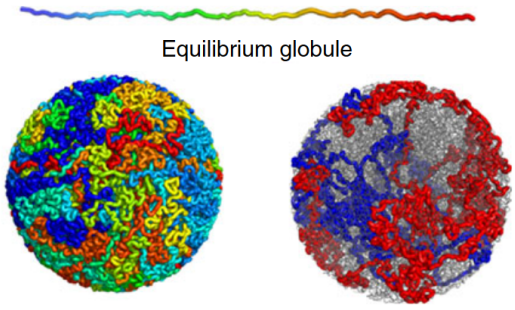



Figure 3.4: Image from [70]. An equilibrium globule and its domain structure.

It is important to note that different cell types have different values for $\alpha$, suggesting the fractal globule model may be a good fit for some cell types, but is not universally reflective of real chromatin. Also, this model does not include local details of the genome (promoters, enhancers etc.), instead looking to reproduce the underlying general details of chromosome architecture.

## 3.4 Experimental Techniques: 3C, Hi-C and others

Experimental data on chromatin contacts were initially provided by a process known as 3C (Chromosome Conformation Capture) and later by methods like 4C, 5C and Hi-C, all of which expand on the original 3C method. All of the

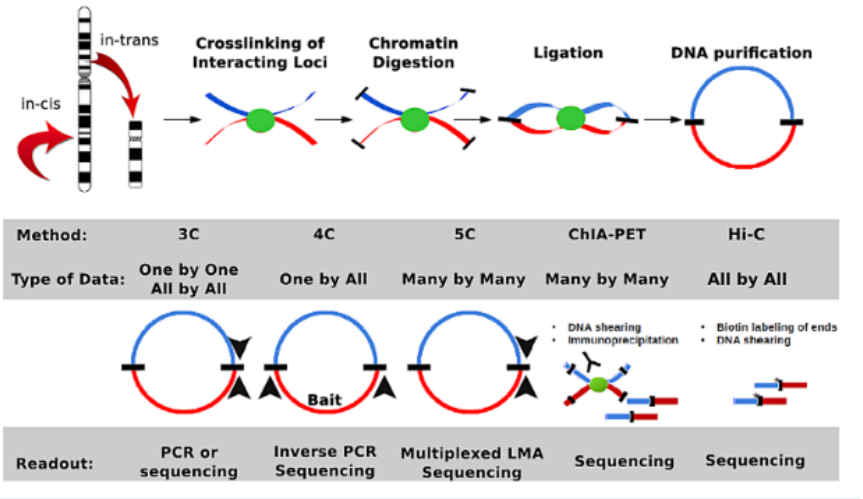



Figure 3.5: Figure from [4]. The methods have similar preparatory stages, but differ greatly in the scope and depth of their analysis.

techniques listed attempt to identify regions of the genome which are spatially proximate in 3D, however they differ in the scope of their analysis. 3C is used to identify interactions only between pairs of specific, pre-determined fragments of the genome, while Hi-C can do this for interaction between all parts of the genome.

All of the methods have the same sample preparation technique, which involves first cross-linking the DNA sample using formaldehyde [67]. This joins spatially close DNA, effectively taking a "snapshot" of the system at the time.

Next, the DNA which is not cross-linked is digested by a restriction enzyme, leaving only the pairs of DNA fragments which were in close spatial proximity. The ends of each pair are then joined (ligated) and the cross links removed, leaving the DNA fragments ready to be analysed [4].

The analysis stage is where the techniques mentioned above diverge. The original 3C method requires specific fragments of interest to be identified before any analysis takes place, as PCR (Polymerase Chain Reaction) methods are used to identify DNA fragments. In PCR, a "DNA primer" which corresponds to a specific sequence is used; if the DNA fragments being tested match up with the DNA primer we get a chain reaction which rapidly increases the number of DNA fragments in the sample. If this happens, then the DNA in our test fragment has been identified!

However, the limitations of this method can be significant. The DNA primer

requirement means that this method is only suitable for regions of the genome where some details are already known, which means 3C experiments are more about testing a particular hypothesis then generating a large dataset which could then be analysed in greater detail. Also, ligated DNA fragments have to be analysed one-by-one, so 3C tends to focus on smaller regions of the genome. Again, this increases the focus on hypothesis testing over data generation.

As an aside, this is not necessarily a bad thing for cases where detailed information on a genomic locus is required. However, for researchers wanting to study entire chromosomes or look for previously unconsidered mechanisms/relationships, 3C may not be the most suitable tool.

4C (Chromosome Conformation Capture on ChIP) was designed to allow study of larger regions of the genome, as chromosomes are known to have both intra-chromosome and long-range inter-chromosome interactions. 4C provides data on how a pre-selected DNA fragment interacts with all other regions of the genome. After the DNA fragments are ligated, only the ones containing the pre-selected fragment are analysed, either by microarray or deep sequencing analysis [103].

5C (Chromosome Conformation Capture Carbon Copy) can be thought of as more like a straightforward upgrade to 3C than 4C is. Like 3C, data can still only be collected for small genomic regions and misses out long range interactions. Where it substantially improves on 3C is the number of different fragments which can be identified, building on 3C "libraries" and using a technique known as LMA (Ligation Mediated Amplification) to simultaneously amplify large numbers of ligated DNA fragments. These can then be analysed via PCR or microarrays, providing a great amount of detail over a small genomic region. Further information on the exact details of the methodology are available in *Dostie et al* [32].

The last of the "C" techniques mentioned above is Hi-C [5], which allows study of both long and shorter range chromosome interactions. The methodology is similar to the previous techniques except sequencing can be done simultaneously for all regions of the chromosome, leading to contact maps of the type we see in *Rao et al*. As it provides data for genome-wide interactions, Hi-C is also a more useful tool than 3/4/5C when there is little pre-existing information about the region being studied.

While Hi-C is an extremely effective tool, it does not completely replace the other methods. 4C and variants such as Capture-C [41] can still be higher resolution than Hi-C, and the extended range of Hi-C brings with it a corresponding in-

crease in sequencing depth which may be avoidable depending on the goals of the experiment. Note: While Capture-C and 4C give similar output (one vs many), the actual experimental methods underpinning the two techniques are distinct from each other. For our purposes, we shall be content with mentioning that Capture-C may be just viewed as a refined version of 4C.

Techniques also exist which allow the characterisation of protein interactions with DNA. These include ChIP-seq (Chromatin Immunoprecipitation Sequencing), which involves enriching the DNA-protein complexes in a system via the use of protein specific antibodies. Later on, the DNA in the complex can be sequenced and its position in the genome identified.

There also futher related experimental techniques available, such as 6C (Combined 3C ChIP Cloning) [94] and ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing) which are also more geared towards gathering data on the protein interactions between DNA, rather than the DNA itself. We do not discuss them further here, as these experiments will not be used for the direct comparison with simulations we discuss in this chapter.

## 3.5 More Simulations With Non-Specific Binding

### 3.5.1 Single Protein Model

Before attempting simulations which incorporate genetic data, we looked at a few more simple models of non-specific binding. Firstly, we looked at a model almost identical to the one used for chromatin in section 2.4, but with the change that stronger binding sites were placed at regular intervals on the fibre (model type A in figure 3.9). Although too simplistic to model real chromatin, this can be thought of as emulating the tight binding of transcription factors to their specific target sites and non-specific binding elsewhere. This initial set-up leads to the clustering seen in section 2.4, along with the formation of chromatin "rosettes" where the strong binding sites are bound tightly to the protein cluster. The chromatin in between the strong binding sites then forms a loop with its base at the protein cluster (Figure 3.6).

The interaction between proteins and the chromatin fibre is also implemented in a similar way to section 2.4, with $\epsilon = 3.5k_BT$ (weak binding) or $\epsilon = 7.1k_BT$

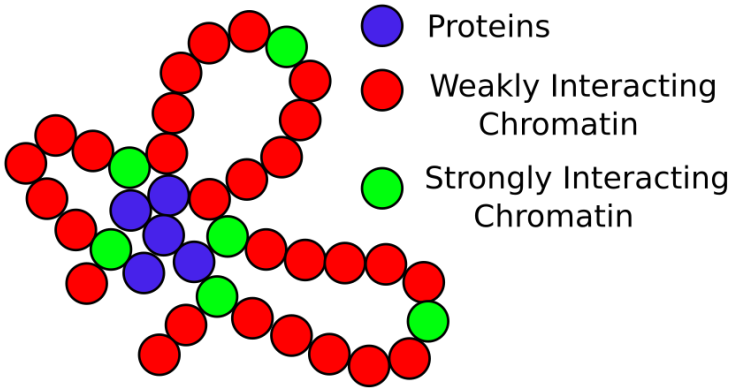



Figure 3.6: A possible configuration of a chromatin "rosette". The strongly interacting sites are likely to make up the base of a loop.

(strong binding) and $r_{cut} = 1.8\sigma$. These choices for $\epsilon$ correspond to values of 4 and 8 in a LAMMPS input script, as similar to Chapter 2 we use a potential which is shifted so to energy is zero at $r_{cut}$.

$$\begin{cases} E = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) & \text{if} \quad r < r_{cut}\sigma \\ 0 \quad \text{otherwise} \end{cases} \tag{3.1}$$

These values for $\epsilon$ allow for transient binding of proteins, which may only bind briefly to a weak binding site before detaching and re-binding at a nearby location. If this happens to be a strong binding site, it may remain bound for long enough to stabilise a loop between two chromatin beads. When this happens, proteins have an increased probability of binding to the same site due to the "Bridging Induced Attraction" mechanism described in chapter 1.

This loop stabilisation could also occur at a weak binding site, but as the protein is more likely to dissociate this happens infrequently. As in the other non-specific binding simulations (section 2.4) there is no attraction between proteins, or between the beads in the fibre - aside from the bonds between adjacent beads.

As figure 3.8 shows, clusters form quickly with their properties having reached steady-state values within $\sim 5 \times 10^4$ simulation time units. Converting to real units follows the same procedure as in section 1.5 (1 simulation time unit is 0.6 ms), implying the simulation is "finished" after $O(10^1)$ seconds.

The average cluster grows to contain ∼12 proteins and ∼6 strong binding sites. Further growth is inhibited for entropic reasons [35], as the entropic cost of bringing together loops (i.e. adding more binding sites to a cluster) scales non-linearly

($\approx n^2$) with the number of loops, while the binding energy scales linearly [64].

As the number of proteins per chromatin bead was relatively low, almost all the proteins end up in a cluster by the end of the simulation run. This puts a partial constraint on cluster size, though larger clusters could still form by the merger of two or more smaller ones. However, this tends not to happen in the simulations as merging two clusters of loops is prevented by the free-energy barrier from loop-loop interactions between clusters. As in figure 3.6, both clusters are likely to have a 'screen' of DNA loops around the proteins. The fact that the DNA is looped rather than linear means it is less likely for the strands to interpenetrate. The amount of chromatin reorganisation and protein dissociation required for cluster merging means it is unfeasable, even during extended time simulation runs. Though a very slow transition cannot ever conclusively be ruled out, it was not observed in the simulations we performed.

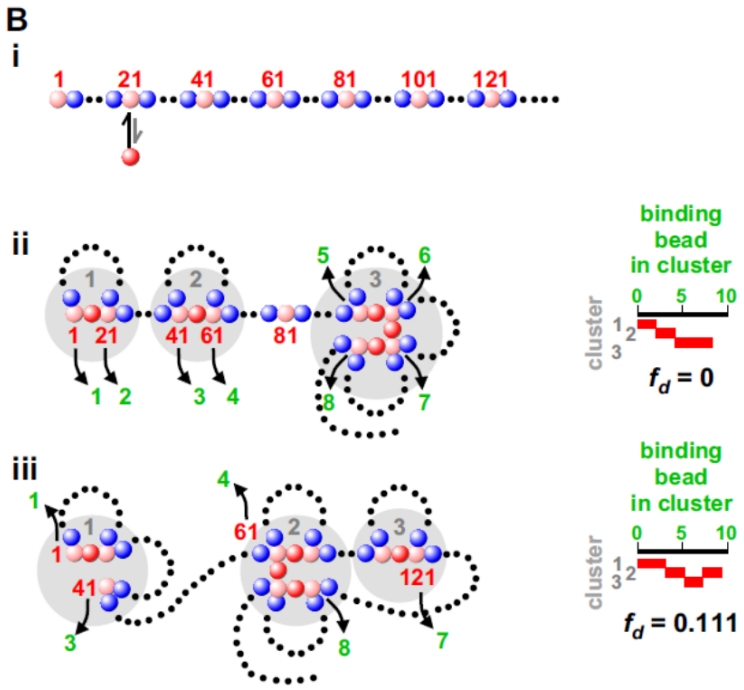



Figure 3.7: An example of a rosettogram where the binding beads are ordered (ii) and slightly disordered (iii). While (i) shows the protein binding sites as being regularly spaced, this is not a requirement - as long as we can order the binding beads (e.g. by genomic distance) we can create a rosettogram.

The rosettogram plot in figure 3.8 shows how the local conformations around a cluster differ. The second cluster forms a perfect rosette - one where successive strongly interacting sites are all located in the cluster. In contrast the first cluster contains a few strong interaction sites, but the fibre then leads away to another

cluster before returning. In fact this happens multiple times for the first cluster!

The contact map for these simulations also show that domains do form in any given simulation run. Since in our model protein clusters are unlikely to merge or grow past 12 proteins, the corresponding chromatin domains also tend to be small. There is no consistency to where domains will form, hence the average (20 run) contact map shows little evidence of any domains.

Other simulations were performed where the location of strong interaction sites was randomised, while keeping the total number of such sites constant. This led to a more ordered chromatin chain with a disorganised fraction, $f_d = 0.06$, meaning only around 1 in 20 strong interaction sites is 'out of order' - as in figure 3.7 (iii). This is perhaps because the randomisation allows for the occasional large gap between interaction sites. This gap could act as a more natural domain boundary between rosettes, as the entropic cost of forming loops increases (logarithmically) with loop size, thereby favouring local loops (hence rosettes) over more non-local structures.

A further alteration to the model can be seen in figure 3.9, where runs of non-binding beads separate sections with strong binding sites placed at regular intervals (type B in figure 3.9). This has a parallel with real chromatin, where active regions alternate with inactive regions.

There is clear evidence of domain formation, with distinct boundaries (the non-interacting regions) between each "pyramid" in the contact map. This result is reflective, at least qualitatively, of observations from simulations of the Caulobacter Crescentus chromosome [55]; the pyramids are also reminiscent of the TADs in the experimental contact maps (Figure 3.2). We can also see evidence of occasional inter-domain contacts.

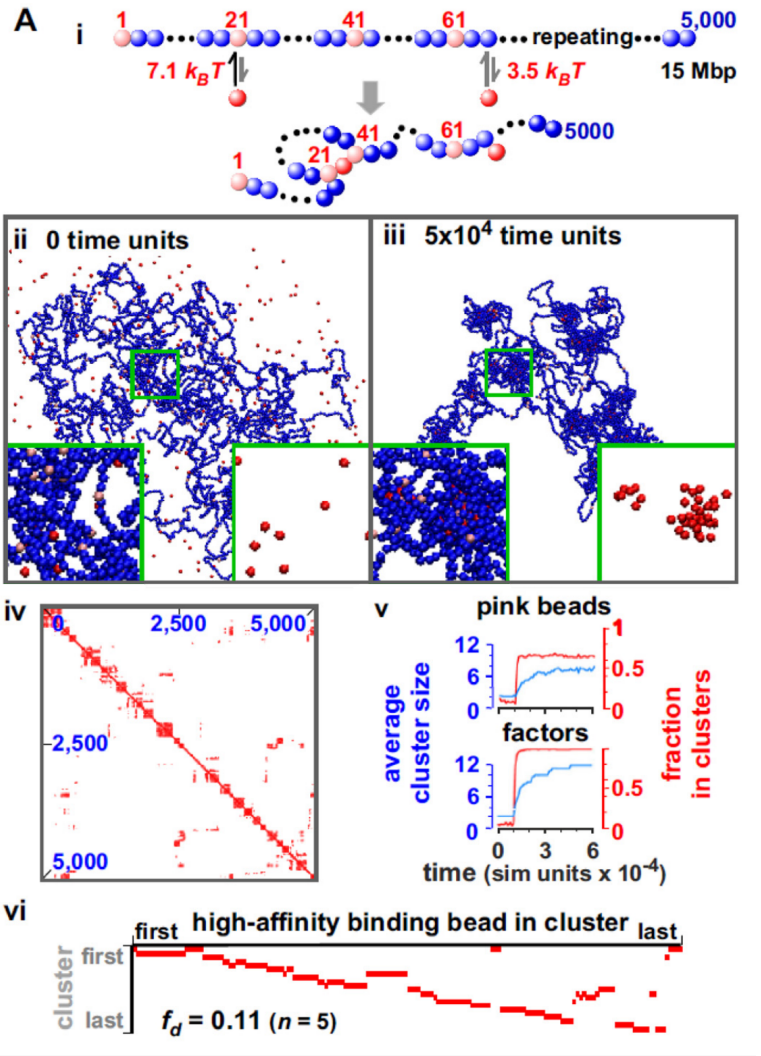



Figure 3.8: Set-up, simulation snapshots and results for the one protein model. i) Interaction strengths between proteins and chromatin. As shown, there are 5000 chromatin beads, which at 3 kbp/30 nm per bead gives 15 Mbp total. As the simulation environment is a cube with side length 3 $\mu$m this corresponds to a volume fraction of $\Theta_c = 0.26\%$, meaning the chromatin is in the dilute regime ($\Theta \ll 1$). The persistence length of the chromatin fibre is 90 nm. There are 250 proteins, also sized 30 nm - giving a volume fraction of $\Theta_p = 0.01$ and $x = \frac{c_p}{c_d} = 0.05$ %. For comparison, this is at the low end of the concentrations used in chapter 2.
ii) Initial conditions for the simulation
iii) Simulation after $5 \times 10^4$ timesteps - protein clustering has begun to take place.
iv) A contact map for a single run of the simulation. Two beads are considered in contact if they are within 150 nm ($5\sigma$) of each other.
v) Properties of the strongly binding chromatin beads and the proteins themselves.
vi) A rosettogram. This plot shows the strongly interacting (high-affinity) beads and which cluster they end up binding to. For example, a horizontal red line means consecutive strongly interacting beads have bound to the same cluster, forming a rosette structure similar to the one in figure 3.6 but where all purple beads are bound to the cluster. $f_d$ is the "disorganised fraction", a measure of how many clusters/rosettes are formed by non-consecutive chromatin beads.

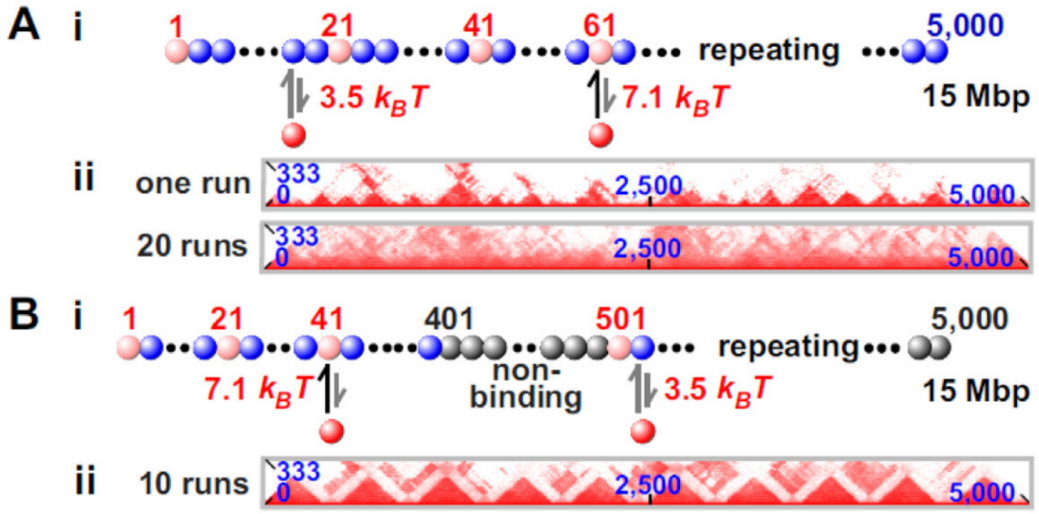



Figure 3.9: Diagonal from contact map for the one protein model with regular binding site spacing (A) and regular spacing with non-interacting regions (B). The difference between a single run and the system average is also illustrated. Simulation type (A) shows domain formation in individual runs but not consistently, while (B) forms domains which are consistent over several simulation runs, along with weak inter-domain contacts. The triangular domains correspond to the regions with strongly binding beads.

### 3.5.2 Two Protein Model

Next, we studied a model where a chromatin fibre interacts with two different types of protein. In this model, the proteins are either red or green with red proteins binding to red sites on the chain and green binding to green. The fibre itself is made up of alternating red/green sections of equal length. This is representative of active and inactive regions of the genome, which have different binding proteins and form separate domains.

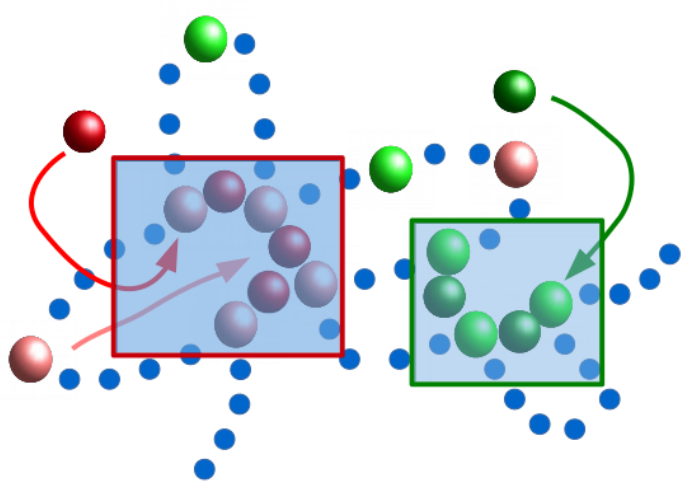


Figure 3.10: Image from [9]. A possible configuration for the two protein model, where dark red/green beads are proteins and light red/green beads are chromatin. This illustrates some of the features seen in simulation such as cluster linking, where the chromatin fibre revisits a cluster it previously interacted with.

As seen in figure 3.11 the protein clusters contain proteins of only one type and mixed clusters have not formed at the end of the simulation. This is not too

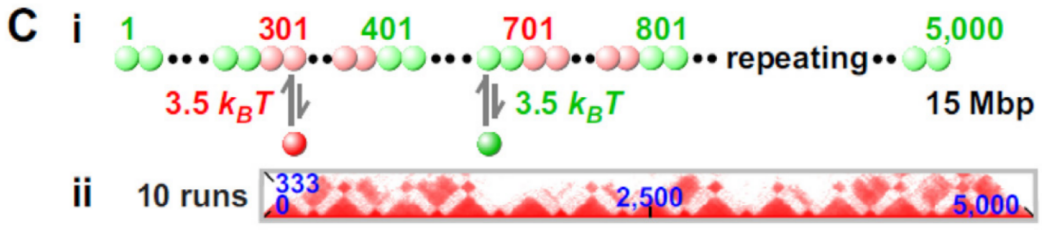



Figure 3.11: Illustration of two protein model and image of diagonal from contact map. Intra-domain contacts are seen in individual runs, but not as consistently as inter-domain contacts. There also appears to be a small effect where inter cluster contacts are more likely towards the end of the chain.

surprising, as the alternating sections of the chromatin fibre mean there is a high entropic cost when forming mixed clusters.

This result is of interest if, as mentioned previously, we consider the different proteins to be analogous to the transcription factors and other proteins associated to different chromatin regions. Experimental Hi-C studies show that many domain boundaries are also boundaries between active and inactive regions, which would also separate their associated transcription factors.

### 3.5.3 Loops and Supercoiling

Chromatin looping and supercoiling are also known to affect domain formation. As mentioned in section 1.6, supercoiled domains are known to share boundaries with TADs [40] so it is of interest to see exactly what influence supercoiling and looping has on domains. It should be noted that this boundary sharing is not one-to-one, as there are many more supercoiling boundaries than TADs. In the case of supercoiling, there is also a strong link to transcriptional activity which is studied in Chapter 4.

The simulation set-up for type D & E in figure 3.12 has linear stretches of beads connecting permanent loops, which are supercoiled in E but not in D. This type of looping can occur *in vivo*, possibly due to CTCF binding to sites around the loop base and stabilising loops [40]. This CTCF stabilisation may also act as a barrier to supercoiling; this idea is implemented in simulation E where supercoiling is conserved within the supercoiled regions, with each region having a linking number of +32. The simulations were also run with a linking number of -32 and as the results were similar. Only the positive supercoiling data is shown in figure 3.12. As in simulations A & B, we have a single type of binding protein - although this time the protein binds to all beads with equal strength.

The results from the averaged contact maps show that domain formation occurs in both models, with more distinct boundaries seen in the supercoiled case. Though it may not necessarily come across from the cartoon in figure 3.12, supercoiled loops have a much higher local density of chromatin even before adding binding proteins. This may explain why the domains are clearer in type E than in type D - the supercoiled regions already have a greater probability of recruiting and stabilising a binding protein in the first place.

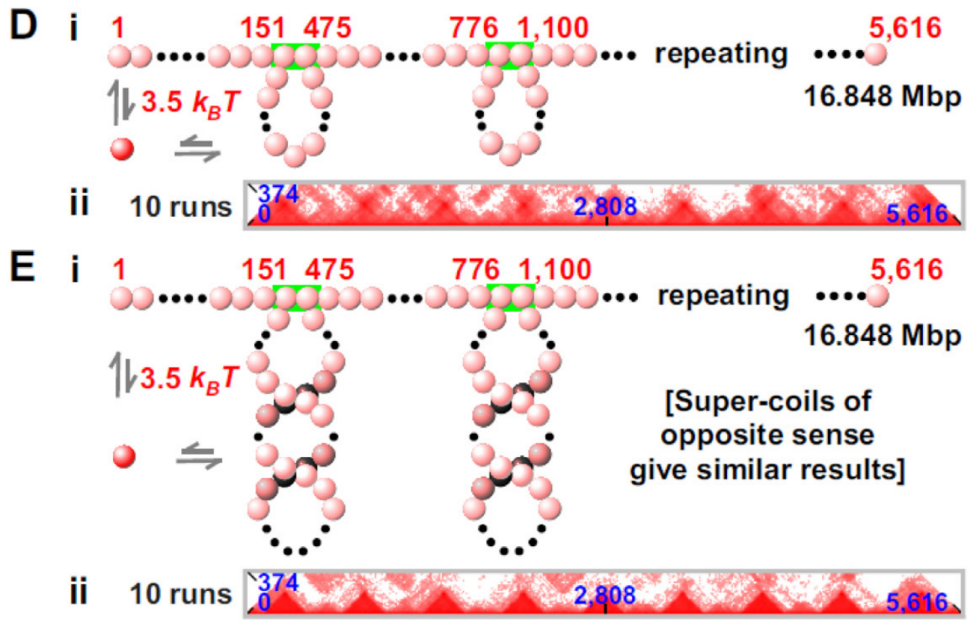



Figure 3.12: Diagonals of a averaged contact map for loop (D) and supercoiled loop (E) simulations. The protein-chromatin interaction rules are also shown. There are slightly more beads than in the previous simulations, 5616 here compared with 5000 previously. This corresponds to a 16.8 Mbp region. We can see considerably clearer domains in the supercoiled case.

Collectively, these models display a few of the potential mechanisms by which topological domains can form - we expect that each of these mechanisms will be active during domain formation *in vivo*.

## 3.6 Domain Properties and Boundary Identification

As mentioned in section 3.3, the probability that two regions of a genome are in contact is related to their genomic distance. Unsurprisingly, greater genomic distances mean a contact is less likely to be made. As previously discussed, this relationship follows a power law where the exponent $\alpha$ is dependent on cell type, ranging from $-0.5$ for HeLa to $-1.6$ for embryonic stem cells. For comparison, recall that the fractal globule model from section 3.3 has $\alpha$ equal to $-1$.

When calculating the power law exponent for the simulations above two regimes are found, representing inter- and intra-domain contacts (Figure 3.13). The intra-domain $\alpha$ values range from $-0.65$ to $-1.05$, falling within the experimentally

observed range. For inter-domain contacts $\alpha$ ranges from $-1.4$ to $-2.06$ though these values should be taken with caution due to the smaller sample size for these longer range contact. These values may also be influenced by the increased likelihood of forming an inter-domain contact for domains near the ends of the chromatin fibre (Figure 3.14). While the number of 'end' domains will always be 2, there are only $\approx 10$ domains in total so these end effects may be significant. For a larger region of chromatin with hundreds or thousands of domains, these effects would not be so noticeable.

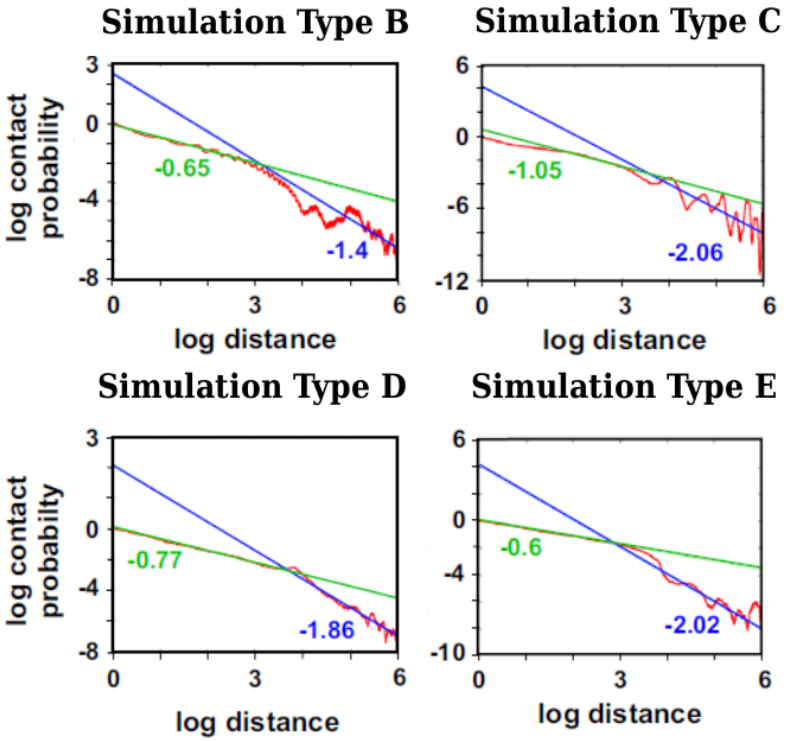



Figure 3.13: $\alpha$ for simulation types B to E. As with experimental measurements of $\alpha$, we see different values characterising short and long range interactions.

Identifying the position of boundaries is not difficult for the simulations presented so far, as the smaller scale and regular structure of the models tend to create distinct, regularly spaced boundaries. However this will not always be the case, as figure 3.2 shows that contact maps may have irregular, ambiguous boundaries.

Because of this it is necessary to have some metric by which boundaries can be identified, along with some degree of automation for larger data sets. Unfortunately, it is not an easy task to programmatically identify every domain boundary with 100% accuracy, so some degree of manual checking is required for the more "borderline" domains.

The base of the domain finding approach we use is the Janus plot (Figure 3.15),

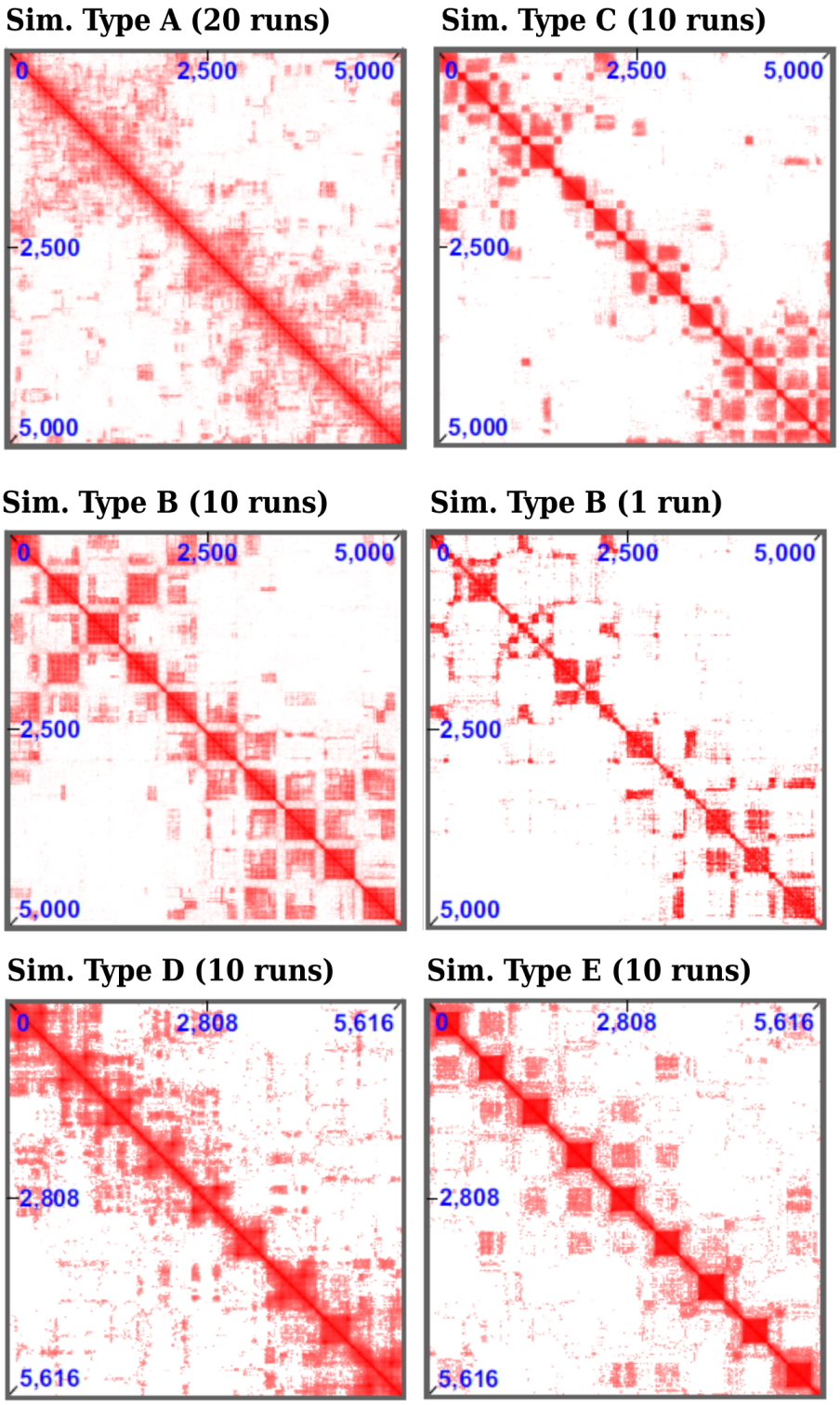



Figure 3.14: Full contact maps for all simulation types. We can also see the difference between a single run and the average, even for just 10 runs. In Hi-C experiments, the contact maps may be an average of thousands of individual cells.

which in its simplest form involves selecting the $i$th bead and measuring the number of contacts it makes to the left ($B_i$) or to the right ($F_i$) in 1D genomic space. For the 2D contact maps this would mean contacts up or down the diagonal. Domain boundaries coincide with the points where there is an abrupt change in the plot i.e. a jump from most contacts being on the left to most contacts being on the right.

This can also be seen in a difference plot ($\Delta_i = F_i - B_i$), for this type of plot boundaries are wherever the signal goes from negative to positive y values. This method is similar to the one used in [29]. One issue with these types of plot is that we can potentially have multiple nearby boundaries due to noise in the signal. This can be avoided by adding the requirement that the signal continues increasing in the positive y direction for a number of beads, though the number chosen will always be to some extent arbitrary.

As long as the number of beads is considerably less than the average domain size, the arbitrariness of the choice should not matter too much. A refinement to this method involves looking at the "insulator signal", which is the derivative of the difference plot. Boundaries should coincide with peaks in this plot. The benefit of this over the standard difference plot is that contacts away from the diagonal may affect where the difference plot crosses the x-axis. This should be avoided by the insulator plot, provided that the number of long-range contacts also does not vary too quickly.

As mentioned above, it is also necessary to have a degree of manual boundary verification. This is required in part due to the difference between Hi-C and simulation data, with simulation data being noisier and with less evenly distributed long-range contacts (Figure 3.18 shows the two contact map types side-by-side). Numerical values which are not significant for Hi-C may be large enough in simulation datasets to give spurious boundaries, so it is important to at least check this has not occurred.

Another potential way of characterising simulation results is the clustering method used by *di Stefano et al* in their steered molecular dynamics simulations of chromosome 19 [28]. This involves looking at the clusters formed by co-localised genes and identifying which subsets of genes cluster together as well as the layout of these clusters. The clusters themselves were found using a k-means clustering algorithm, which allows the number of total clusters to be set in advance. Running with different numbers of clusters generally shows a clear 'best' choice for

the tradeoff between minimising a cost function and having a very high number of clusters.

This method also leads to genes which are in contact even over mid-to-long genomic distances. This clustering is also used to provide a way to divide up a contact map into larger macro-domains (10 are selected for chr 19).

The goals of the simulation when it comes to boundaries and boundary finding are also different for Hi-C and simulation datasets. While Hi-C experiments seek to return quantitative information about boundaries (and many other things!) the simulations are attempting to test how well results from our underlying models fit the Hi-C data, rather than generating new information based on that data. When performing data analysis on Hi-C data it can be difficult to choose between equally plausible explanations for a feature seen in the data, as there may be no reason to favour one over the other. With our simulation data we can test each potential explanation and see how well they reproduce the data. Of course, the simulations could then show that both models reproduce the data well - though this would still tell you something about how (un)important a particular feature of the model is more generally.

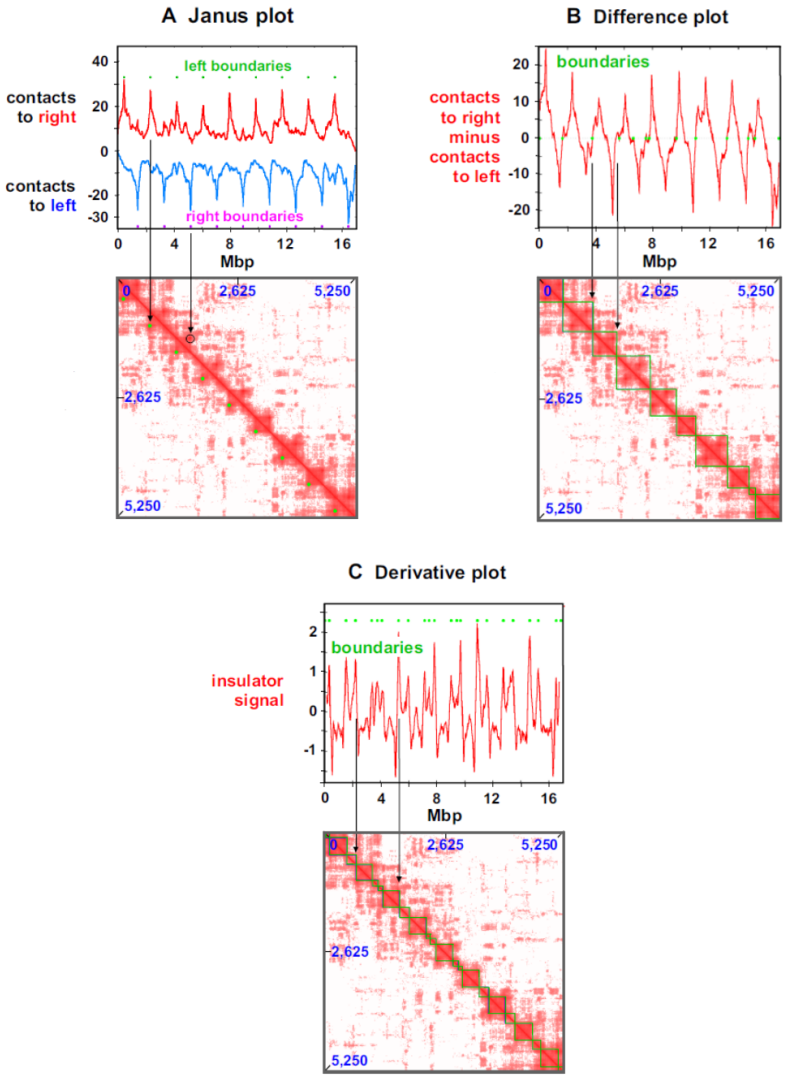



Figure 3.15: Three different methods of boundary identification. (A) Janus Plot showing contacts to right and left of each bead. Peaks in the signal correspond to boundaries. (B) Difference plot for the same data, here boundaries are wherever the signal goes from negative to positive y values. (C) The insulator signal plot, which is the derivative of the difference plot. In this plot type, boundaries can now be found at peaks in the signal.

## 3.7 Adding Genomic Data to the Model

A previous approach to simulating chromatin had involved using experimental Hi-C data as a way to set interaction strengths, which leads to good agreement between simulation and experimental results[89], but may not elucidate the actual mechanisms driving the chromosome organisation. Instead, we opted for a fitting-free model which uses genomic data to determine how different regions of DNA interact with transcription factors. This model uses the set-up for chromatin in section 3.5.1 as a basis, so the beads in the chromatin fibre are representative of 3 kbp of chromatin, have a size of 30 nm and have $L_p = 90$ nm .

In our model we have two protein types, one representative of either transcription factors such as CTCF or polymerases and the other representative of proteins which bind to heterochromatin regions, such as the HP1$\alpha$ protein. These bind to regions of the genome corresponding to euchromatin (active regions) and heterochromatin (inactive regions) respectively. As in the two protein model in section 3.5.2 and the non-specific binding model in chapter 2, these binders are multivalent, meaning they can bind simultaneously to more than one region of DNA and create molecular bridges between distant DNA regions [21].

## 3.8 The Human Genome Project and the Genome Browser

The source of all the genomic data used to set up our simulations is the UCSC Genome Browser [49], an open-access resource providing data for the entire genomes of human and other model organisms. The genome browser is an offshoot of the Human Genome Project and has made data available ever since the first draft of the genome was published in 2000, nearly 10 years after the project began. The project was officially completed in 2003, though there have been several "updated" genomes published since then.

The genome browser website makes a wide range of information available for download [48], with examples including gene locations and expression levels for different cell types.

## 3.9 Using Histone Modifications to Characterise Chromatin States

Any implementation of the model outlined in section 3.7 requires a way of identifying which sections of the genome are "active" or "inactive". This was done by using the Broad ChromHMM dataset from the UCSC genome browser, which characterises sections of the genome based on the properties of individual histones[36, 37]. Histone proteins can be modified after transcription, most commonly leaving specific amino acids either methylated, acetylated or phosphorylated. Then, ChIP-seq methods can be used to identify which modifications are present at each histone.

Promoters, enhancers, transcribed and silenced regions are all associated with specific histone modifications, so these can be inferred from the ChIP-seq data. In [36] a hidden markov model (HMM) is used to make these inferences.

In total the Broad ChromHMM study labels histones as being in one of 15 states. The states we chose to represent in our simulation are in table 3.1. Strong binding refers to a binding energy of $7.1\,k_bT$, weak binding is with a binding energy of $3.5\,k_bT$. Protein type 1 binds to inactive regions (HP1$\alpha$), while protein type 2 binds to active regions (transcription factors or polymerases).

| State | In Simulation | Interaction Style |
|---|---|---|
| 1 - Active Promoter | Yes | Strong - Protein Type 2 |
| 2 - Weak Promoter | No | None |
| 3 - Inactive Promoter | No | None |
| 4/5 - Strong Enhancer | Yes | Strong - Protein Type 2 |
| 6/7 - Weak Enhancer | No | None |
| 8 - Insulator | No | None |
| 9 - Transcriptional Transition | Yes | Weak - Protein Type 2 |
| 10 - Transcriptional Elongation | Yes | Weak - Protein Type 2 |
| 11 - Weak Transcribed | No | None |
| 12 - Repressed | No | None |
| 13 - Heterochromatin | Yes | Weak - Protein Type 1 |
| 14/15 - Repetitive | No | None |

Table 3.1: Possible states from the Broad ChromHMM data, whether they are included in the simulation and their interaction style if they are.

Since the chromatin fibre beads in the simulation represent 3 kbp, or around 15 separate histones, it is possible for a bead to bind both types of protein.

As an example, this could occur when the first histone in a bead is at end of a heterochromatin region, while a enhancer region begins at the final histone in the same bead. The general idea is for a bead to be representative of the features found in that 3 kbp region, rather than indicating that there is 3 kbp of a particular thing at that point.

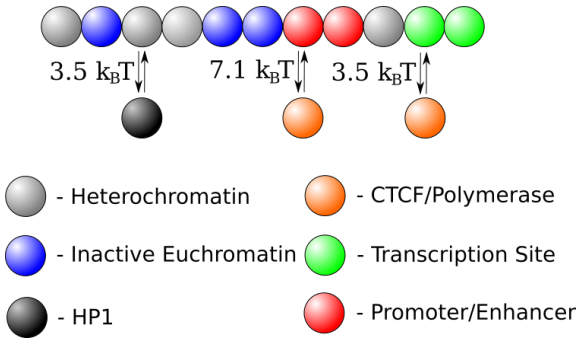



Figure 3.16: Chromosome beads and the strength of their protein interactions.

While transcription factors and polymerases share a bead type, this is not meant to suggest that transcription factors and polymerases are interchangeable in any sense! It would not make sense for a transcription factor to bind at a transcription start site, or a polymerase at a promoter or enhancer site. In both cases they would just be getting in the way of the proteins which were in the "correct" place. However, since our model has separate binding energies for the two sites we can get the different styles of interaction we want with just a single protein type. Since we are only concerned with the overall density of proteins rather than their individual behaviour this model is biophysically appropriate. In the example where a type 2 protein binds to a transcription site, then dissociates and binds to a promoter it should perhaps be thought of as a polymerase coming from the "pool" of proteins, then returning and a transcription factor coming from the pool to replace it.

## 3.10 An Alternative Method to Characterise Heterochromatin

Since the Broad ChromHMM study relies on post-transcriptional histone modifications to characterise histones, it may not work so effectively when identifying heterochromatin. Here, 'post-transcriptional' means 'at a time after significant transcriptional activity', not all histone modifications arise from transcription

alone. Regions of heterochromatin will not have these modifications as these inactive regions are not likely to have been transcribed or gained histone modifications in the first place (Though the modification H3K27me3 can be associated with some types of heterochromatin). Accordingly, any histones with a "low signal" (i.e. no/few modifications) are labelled as heterochromatin.

However, relying on a lack of evidence to characterise something can lead to some issues. For example, single histones could be labelled as heterochromatin even if they are isolated from other heterochromatin regions. Considering the definition of heterochromatin, it does not really make sense to have regions which are labelled heterochromatin but only consist of one or a few histones. More generally, regions could be mischaracterised as heterochromatin simply because the process looks for a lack of distinguishing features, as opposed to their presence.

The alternative method used to characterise heterochromatin is based on the GC content of the chromatin. Regions with a high GC content are associated with more open 3D conformations, while those with low GC content are more likely to be compact and heterochromatic [25]. This allows us to set a threshold for GC content percentage and label all regions below this as heterochromatin. Since the threshold value can be set we also have more flexibility in our model as in principle any region could be labelled as heterochromatin, so we can see to what extent changing this alters our simulations.

This method does also have some downsides, namely the sharpness of the boundary between heterochromatin/non-heterochromatin and potential arbitrariness of threshold value. Picking a threshold value was done on the assumption that the Broad ChromHMM data gave a accurate picture of the amount of heterochromatin globally, even if there were some local inconsistencies. If the Broad ChromHMM data indicated that there were $N_{htr}$ bp of heterochromatin present, the GC content threshold would be set so that our simulations also had $N_{htr}$ bp of heterochromatin. The difference in outcome between the two methods is that using GC content would tend to produce fewer isolated heterochromatin beads. Fortunately, due to the fact that the threshold can be modified we were able to test the robustness of our GC content value by re-running simulations with the threshold increased or decreased. The results from the simulation remained stable for a large range of threshold values, suggesting this method of assigning heterochromatin regions captures the general picture well.

## 3.11 Generating Contact Maps

Contact maps are generated in a fairly straightforward way, the distances between each pair of simulation beads is calculated and if it is below a selected distance the pairs are considered in contact. Unfortunately, the length of experimental cross-links using formaldehyde is not exactly known [104] but it clearly should be short range. For our simulations, beads were considered in contact if they were within 3 bead lengths of each other (qualitatively similar results were found with larger threshold, up to $\sim$ 10 bead lengths). The bin size used in contact maps is also important when seeking to make comparisons with Hi-C data. While the simulation bin size can be as low as the bead size (3 kbp), Hi-C data tends to be binned at a lower resolution than this. For example, the data from *Rao et al* is binned at 20 kbp. This meant our contact maps were binned at 21 kbp as this is the closest multiple of 3 kbp to 20 kbp.

## 3.12 Results - Chromosome 12

The first simulation using genetic data was a 15Mbp region of chromosome 12, ranging from 85 Mbp to 100 Mbp (Figure 3.17). We used data taken from the GM12878 cell line, which is a lymphoblastoid cell type widely used in sequencing projects and is one of the Tier 1 group of cells [96]. The H1 human embryonic stem cell and K562 cell types make up the rest of the Tier 1 cells, which are chosen based on cell availabilty and ease of use.

For this simulation the threshold for GC content was set at 41.8%, there were 3000 Type 1 proteins and 300 Type 2 proteins. Since the proteins in simulation are only representative of real proteins the chosen values for protein number are set the correct order of magnitude, but it if we wanted to set a specific level of protein concentration there would not be a clear 'best' target value. Setting the protein numbers to the values used in simulation means almost all proteins end up binding to the DNA at some point in the run.

We found that separate clusters of Type 1 and Type 2 proteins formed, as in section 3.5.2. These clusters had average sizes of $\sim$14 and $\sim$190 proteins respectively, with the "active" protein clusters showing a good deal of similarity to the rosettes found in previous models. The larger clusters of "inactive" proteins tended to be seen where there were longer runs of heterochromatin or mostly heterochromatin beads, which explains their larger size.

Comparing boundaries with Hi-C experiments we can see the simulation contact map is a decent fit, with 75% of simulation boundaries matching H-C to within 100 kbp. The rosettograms produced for clusters in active regions are also extremely well ordered, with the disorganised fraction $f_d$ equal to 0.02. For the non-specific binding models where $f_d$ was calculated, it was equal to 0.06 and 0.11. Unfortunately we cannot directly compare with an experimental value for $f_d$, but this improvement suggests that the conformation of real genomes might have evolved to minimise tangling and favours the production of the more-ordered rosette structures seen in section 3.5.1.

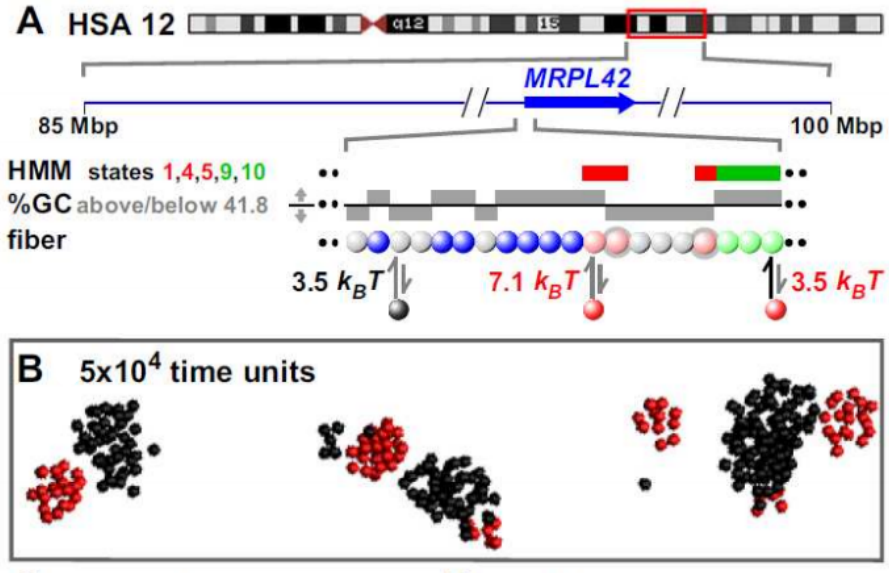



Figure 3.17: (A) The region of chromosome 12 simulated along with bead colourings used in the simulation. These are coloured as follows: Pink - Promoter or Enhancer, Green - Transcription Site, Gray - Heterochromatin, Blue - Inactive Euchromatin, Red - Type 1 Protein, Black - Type 2 Protein. The pink beads here are likely the promoter for the MRPL42 gene pictured.
(B) Protein only screenshots from the simulations themselves. This shows the tendency for similar protein types to cluster together, as also seen in section 3.5.2.

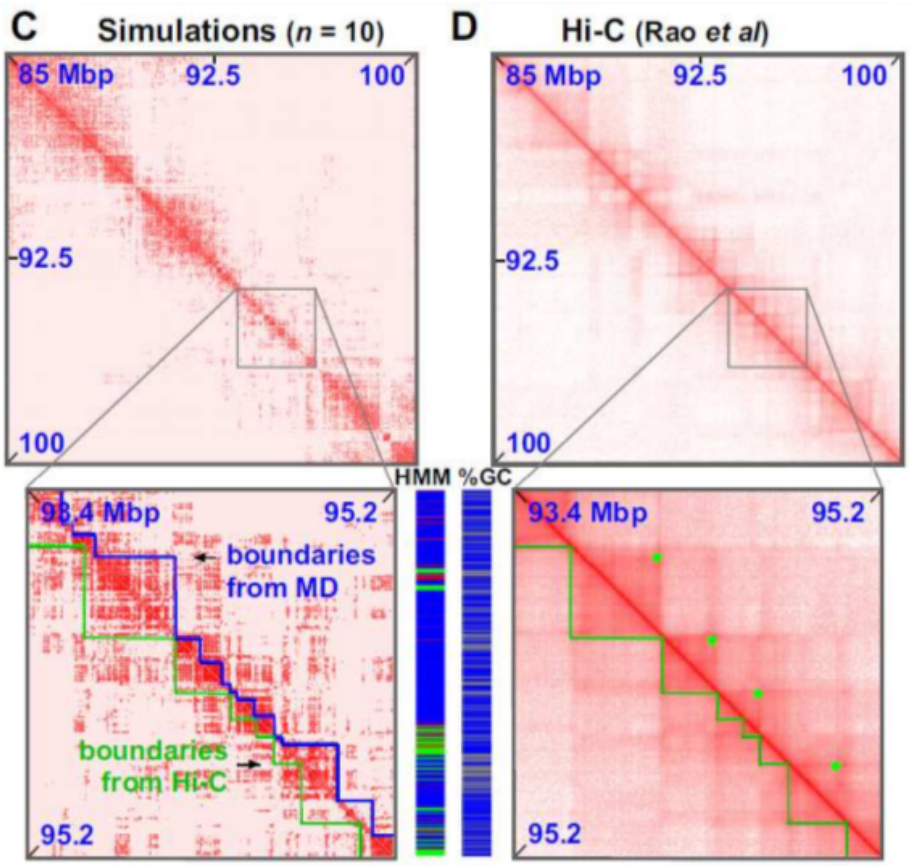



Figure 3.18: A comparison of boundaries found through simulation and Hi-C experiments. Left: Contact Map from simulation, with bin size 7 kbp. Right: Contact map from Hi-C with bin size 10 kbp [84]. The HMM and GC content data are also shown for comparison with domain locations.

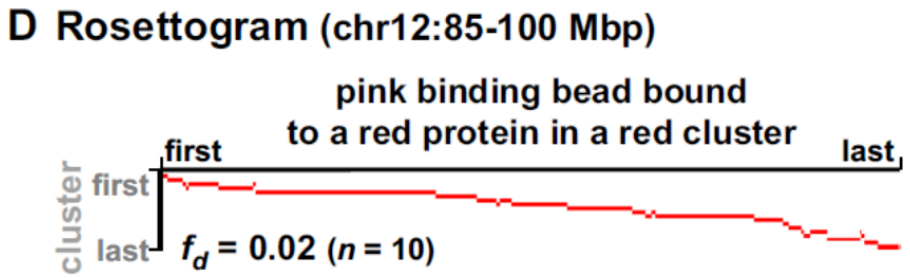



Figure 3.19: A rosettogram for a region of chromosome 12, with disorganised fraction equal to 0.02. The rosettogram is for the active regions of DNA which bind to the red proteins in simulation. The low value for $f_d$ suggests highly ordered, rosette-like structures.

### 3.12.1 Testing Different GC Thresholds

To examine the effect of GC content threshold on contact maps, further simulations were run with the threshold set to 42%, 45% and 48%. The resulting contact maps, shown in figure 3.20, are extremely similar for the first two thresholds - but differ when the threshold is set to 48%. However, this region of the genome is reasonably active and so has a high GC content generally. Because of this, setting the GC threshold high enough means labelling a very large proportion of the beads as heterochromatin, so it is not surprising that this eventually has a visible effect on the contact maps.

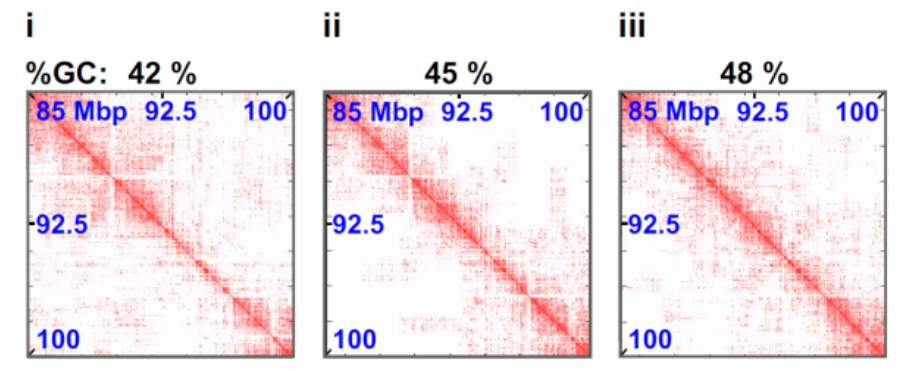



Figure 3.20: Contact maps obtained for the same region of chromosome 12 analysed in figure 3.18, but for different values of the GC threshold before beads are labelled as heterochromatin.

## 3.13 Results - Chromosomes 6 and 14

Further simulations of chromosome sections were performed, this time for 15 Mbp sections of chromosomes 6 and 14. Both regions were selected as they had a mixture of active and inactive binding sites, though as there are lots of regions with this property the choice will always be arbitrary to some extent. The chromatin 6 simulations used data taken from the H1-hESC (Human Embryonic Stem Cell) cell line. While the results and conclusions to be drawn are similar to those of chromosome 12, they help further illustrate both the successes and limitations of the model. The results for chromosome 6 can also be used to justify the choice of using the GC content data to determine heterochromatin regions, rather than the Broad ChromHMM data (Figure 3.21)

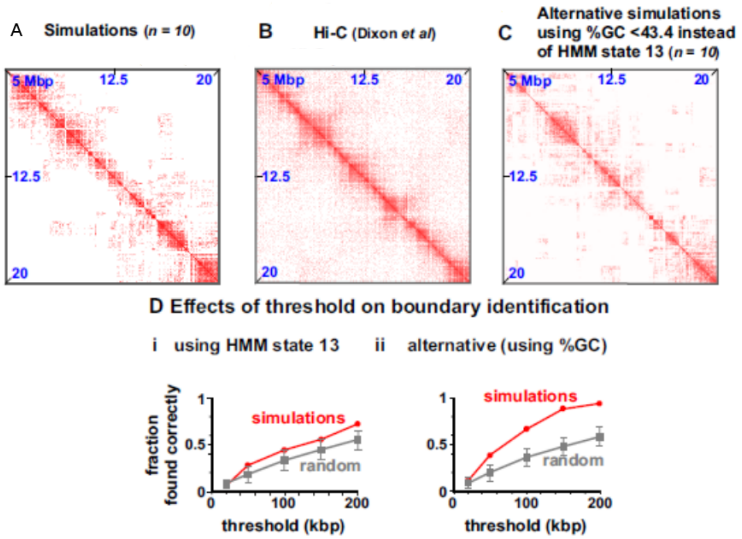



Figure 3.21: Contact maps for (A) Simulations using Broad ChromHMM data, (B) Hi-C for chromosome 6 and (C) GC Content data. (D) A graph showing that fraction of boundaries identified correctly is much greater when heterochromatin beads are identified using GC content data (ii) as opposed to the histone modification data (HMM states) used to identify other regions (i). The threshold here refers to how far a simulation boundary can be from a Hi-C boundary and still be considered correct.
While both methods for determining heterochromatin give better results than setting boundaries randomly, there is a clear improvement when using the GC based method for identifying heterochromatin.

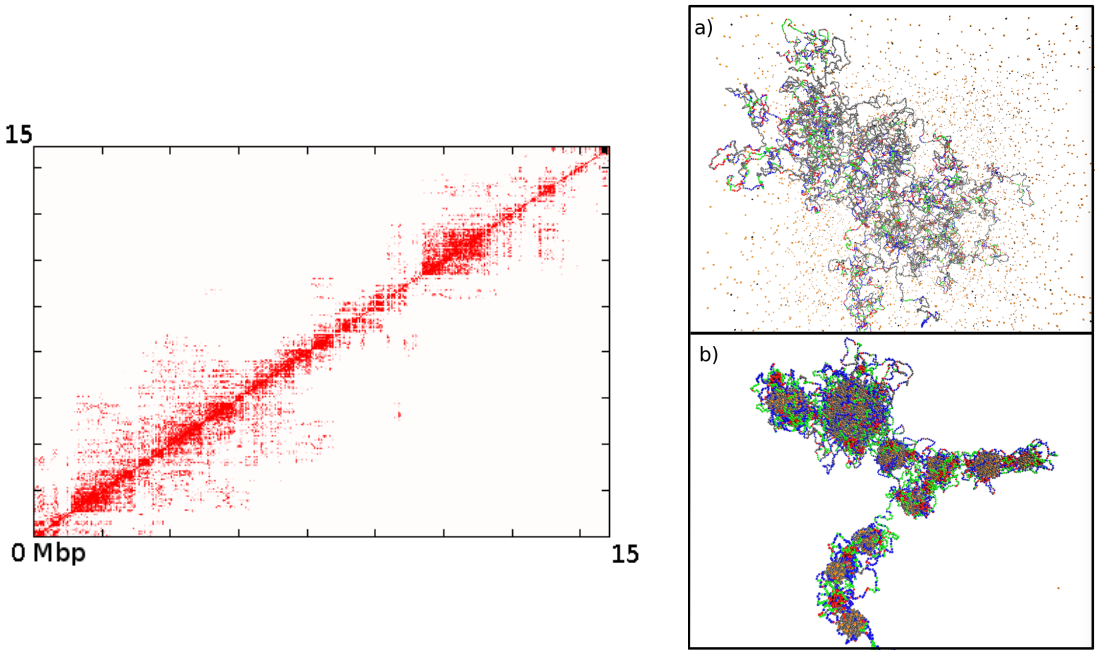



Figure 3.22: Left: Average contact map for simulations of chromosome 14. Right: A snapshot of a simulation run for chromosome 14, with (a) taken towards the beginning of the run and (b) towards the end.

## 3.14 Results - Full Chromosome Simulation

After successfully modelling chromosome sections we attempted to simulate the entirety of chromosome 19, using the same model as previously. Chromosome 19 was chosen for computational reasons, as it is one of the shortest chromosomes. Larger chromosomes should give similar results, but would take considerably longer to simulate - chromosome 1 is almost 5 times as long and simulation times scale as $\sim$(chromosome length)$^2$. This simulation had an even higher degree of success when comparing boundaries with Hi-C, getting around 85% correct to within 100 kbp.

For this chromosome, we also characterised the beads which were found at boundaries. Boundary elements should in theory be more accessible to polymerases and other transcriptional machinery as they are by definition on the periphery of a domain. Because of this, we would expect to find more active and less inactive beads at the boundary compared to any other region of the domain, with this placement driven by protein binding mechanisms. This was verified to some extent (Figure 3.24), and boundaries were found to contain a greater then average number of active or non-interacting beads.

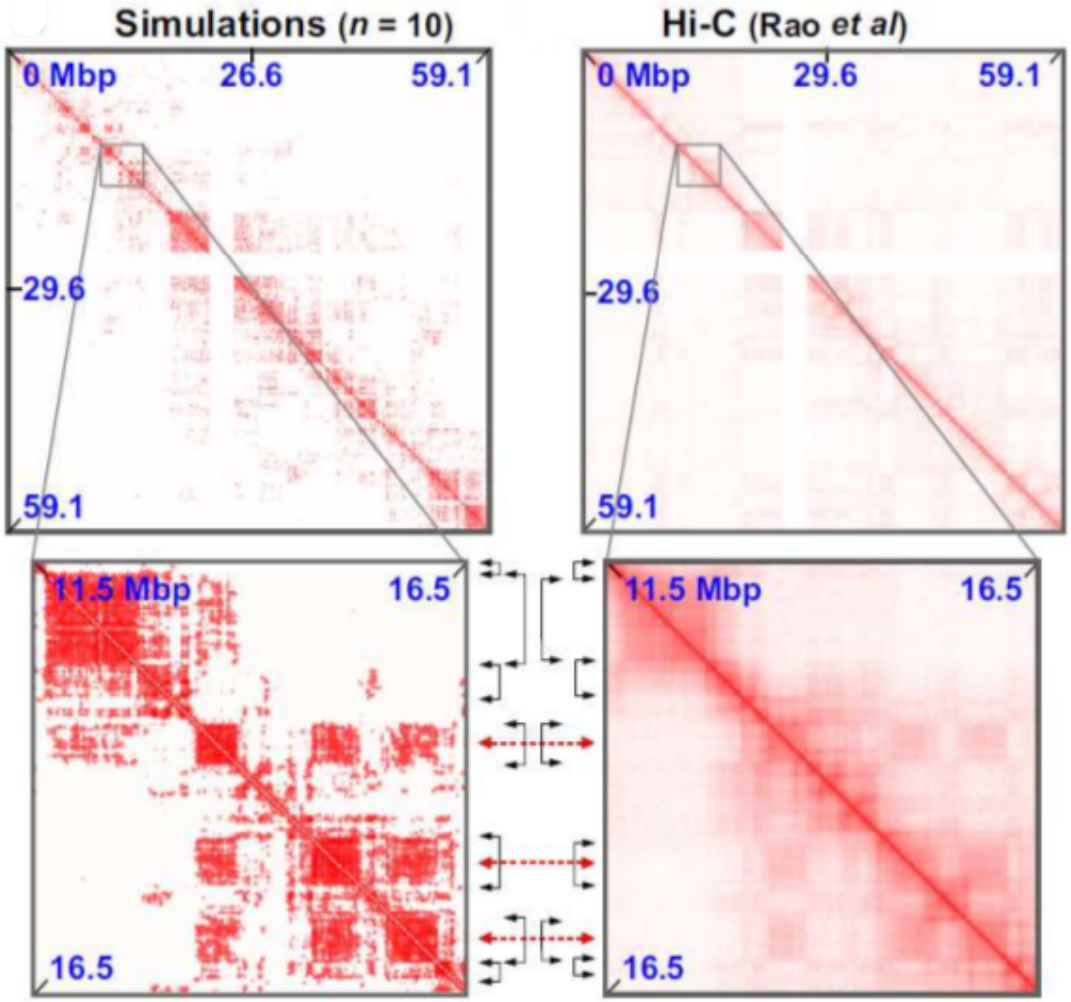



Figure 3.23: Contact maps for chromosome 19 from both simulation (left) and experiment (right). From the zoomed region, we can see the simulations reproduce the Hi-C results with good accuracy. The simulated contact maps also have fewer long-range, non-domain contacts - something which was true in general when comparing simulation and experimental results. This may be a consequence of the simulation contact maps being made up of considerably fewer samples than the experimental Hi-C maps. Some of the longer range, weaker intensity contacts seen in Hi-C may not occur regularly enough in simulation to be detected with a sample size of 10. This could also come about since the polymer is more dilute in simulation than in the cell.

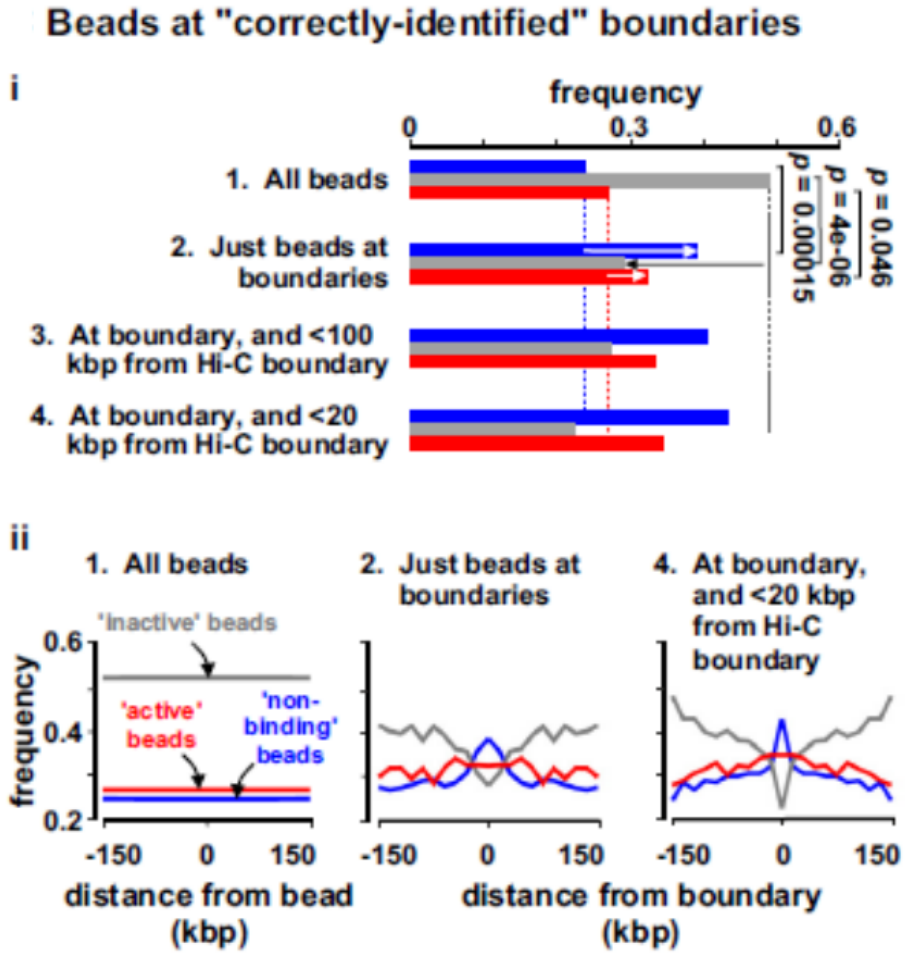



Figure 3.24: The proportion of beads found at or near to boundaries. Red "active" beads are promoters, enhancers or transcriptionally active areas, blue beads are non-interacting and grey "inactive" beads are heterochromatin. There is a clear reduction in inactive beads at boundaries, and a corresponding increase in active and non-interacting beads. P-values for the distributions were calculated assuming a Poisson distribution.

## 3.15 Summary

In the above simulations, we see bridging-induced attraction cause transcription factor proteins to cluster together, even when the only attractive interaction in the system is between proteins and binding sites on the chromatin fibre. When this interaction is expanded to encompass multiple protein and binding site types, we observe the proteins forming clusters which are segregated by type - as long as their binding sites are separated genomically.

This clustering was also found to create chromatin loops, which are organised into topological domains and are observed for simple test models, chromosome fragments and even entire chromosomes.

Even though the full chromosome model only contains two proteins types and three interactions, it still replicates boundaries from Hi-C with 85% accuracy, while also placing appropriate (active) sequences at the boundaries themselves. This level of agreement, towards the favourable visual comparison between simulated and experimental contact maps (e.g, Figures 3.18 and 3.23) is quite remarkable given the relative simplicity of the model.

We also find that active regions favour folding into more ordered rosette-like structures with mainly local loops, as opposed to the more compact, globular structures seen in inactive regions. This occurs due to entropic limitations on the number of loops in a rosette, along with the fact that active binding sites are more sparsely distributed than inactive sites, which tend to be in longer consecutive runs which favour localised binding.

## 3.16 Running Further Simulations

As with the other chapters, you can view videos of selected simulations at `http://www2.ph.ed.ac.uk/~s0841882/chapter2.html` or `http://www.jjthesis.co.uk/chapter2.html` and download a software package to run the above simulations, or variants of them. This allows for simulations to be run using a model which includes supercoiling, for different chromosomes & chromosome fragments or just generally using different parameters (protein no., interaction strength etc.). Some data analysis tools are also provided to help make contact maps and look at other variables of interest.

## 3.17 Future Work

Another popular model for genome organisation which involves bridging CTCF binding sites by loop extrusion, with these loops then forming the TADs seen in contact maps [38]. This model also helps account for the observation that CTCF bridging depends on the directionality of the CTCF binding sites [43]. There are still some issues with using just this CTCF model to determine TADs, as it requires a high processivity motor protein to extrude the loops - which protein this would be is undetermined at present. Experiments with CTCF knockouts also do not show major effects on domain formation, so it would be surprising if this extrusion behaviour were the only factor. However, combining this extrusion model with the bridging-induced attraction model here could potentially improve on the results we obtain here.

# Chapter 4

# Supercoiling-Dependent Transcription

This chapter is based on the Physical Review Letters paper "Stochastic Model of Supercoiling-Dependent Transcription" [8].

## 4.1 Outline

The project detailed in this chapter is an investigation of the link between transcription and supercoiling. In both eukaryotes and prokaryotes, transcription is known to affect the local supercoiling density by causing positive supercoiling to build up ahead of the transcribing polymerase, while an equivalent amount of negative supercoiling is built up behind. Alongside this, local supercoiling density also influences transcription probability as negatively supercoiled areas are more unwound and this allows easier access for polymerases. In this chapter we describe a numerical model which incorporates these ideas, and allows us to characterise how the different regimes of the model depend on the model's parameters.

As we shall see, by changing the amount of transcriptionally induced supercoiling flux we can drive a sharp transition from a regime where gene transcription occurs randomly (low flux), to one where transcription is strongly correlated and regulated by supercoiling (high flux). In this regime we also observe transcriptional bursts, supercoiling waves and upregulation of divergent gene pairs – all these have counterparts in experimental observations.

## 4.2 Supercoiling and DNA

While supercoiling was briefly mentioned in 1.6, it is useful here to discuss it in more detail, alongside its relation to DNA. Since double stranded DNA consists of two intertwined chains, a good starting point is the quantity known as the linking number (Lk) and how it applies to pairs of closed curves.

We can calculate the linking number for an untwisted pair of curves by looking at the points where the two curves cross over each other. We can then see which of these crossings are right-handed (+1) and which are left-handed (-1) (Figure 4.1), the linking number is equal to half the sum of these values.

In most cases only the absolute value $|Lk|$ is considered, as $Lk$ is dependent on the orientation of the curves. This does not matter for abstract curves where the orientation is arbitrary, but if we want to relate the curves to something which does have a specific orientation (like DNA!) it is useful to consider $Lk$ as a value which can be negative as well as positive.

To apply these ideas to a DNA molecule we can think of the two curves as being the phosphate backbones of the DNA. It is also useful to introduce the quantities twist ($Tw$) and writhe ($Wr$). Twist consists of all the internal crossings made by the DNA duplex (i.e., the number of times the magenta curve crosses over the cyan curve in figure 4.2), while writhe counts the number

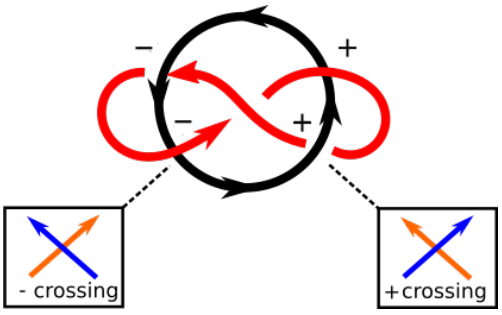



Figure 4.1: A structure known as the whitehead link, with crossings and handedness shown. The linking number is 0 as all the right-handed crossing have a left-handed partner. It's also worth noting that the crossing in the middle is counted twice (once as +, once as -) as we move around the red curve, so does not contribute to the linking number.

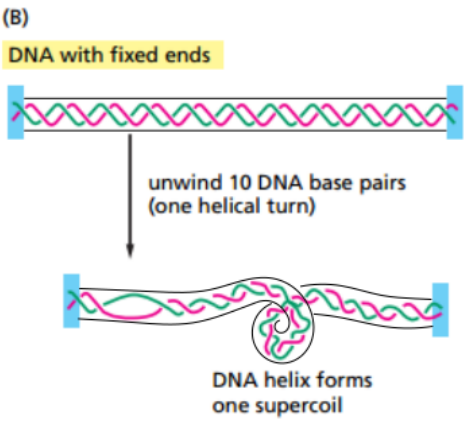
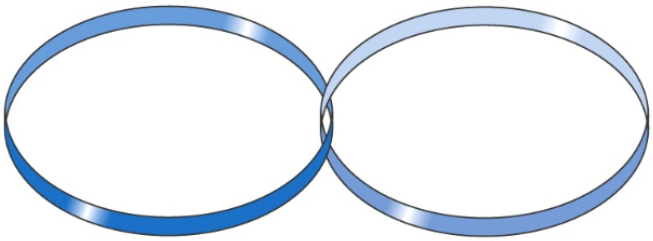
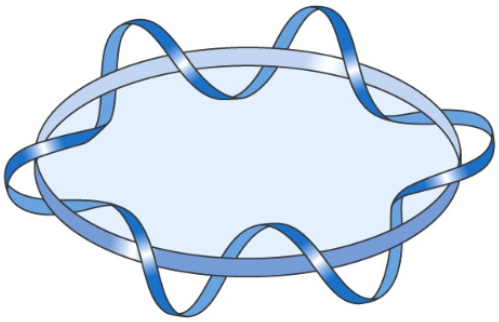



Figure 4.2: Image from [1]. While the number of crossings and linking number remains the same, we have decreased the twist of the molecule and increased writhe.



$Lk = 1$
**(a)**



$Lk = 6$
**(b)**

Figure 4.3: Some more examples, this time with a non-zero linking number. For (b), if we "travel" along both curves in a clockwise direction we can see that the lower curve in any crossing is always going from right to left when seen from the upper curve's perspective.

of the self-crossings of the centreline (i.e., the backbone of the double-stranded DNA) in 3D. This is illustrated in figure 4.2 where a DNA strand goes from a twist of $-10$ and zero writhe, to a twist of $-9$ and writhe of $-1$. Twist and writhe sum up to give the linking number, $Lk \equiv Wr + Tw$.

Although that the proof that the linking number is the twist plus the writhe requires some sophisticated maths, it is reasonably intuitive that Lk is conserved for DNA molecules which are looped or have "fixed" ends. This means that we can consider two DNA conformations with the same linking number but different twist and writhe as topologically equivalent, so a DNA molecule could transition between these states. An example of this is shown in figure 4.4.

Twist, writhe and linking number do not have to be integer values either and while it is obvious that a chain can have a half twist, it is less clear what a non-integer writhe looks like. While we know a loop with $Wr = 2.5$ would look like

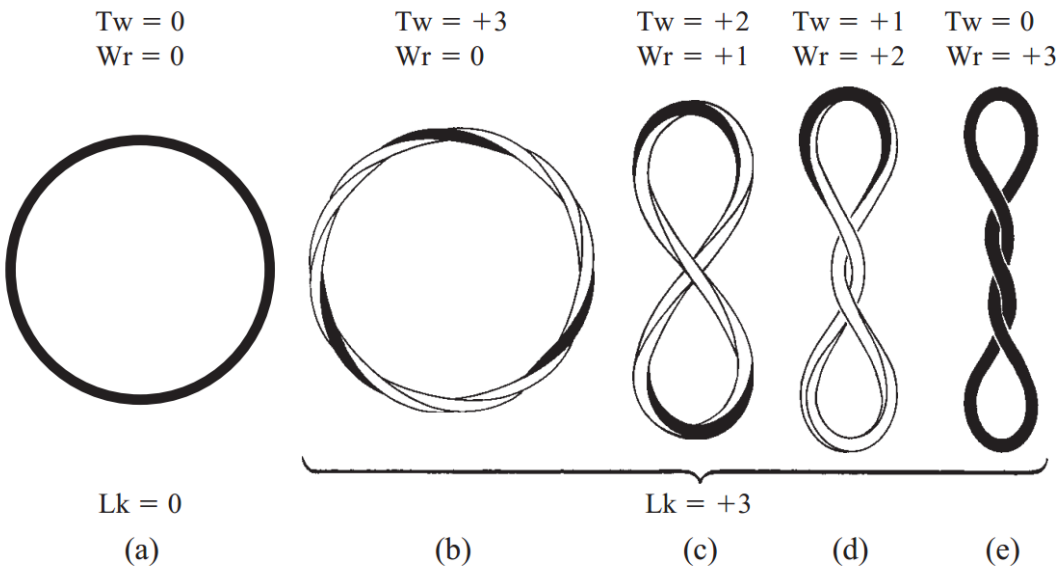



Figure 4.4: Image from [14]. The topological equivalence of twist and writhe is shown by loops (b-e).

a hybrid of (d) and (e) in figure 4.4, this is quite a difficult concept to represent diagrammatically!

Since DNA is naturally coiled it helps to use more relevant values for twist, setting $Tw = 0$ to mean a perfectly straight "ladder" of base pairs would mean any DNA conformations we are actually likely to find will have extremely large values for twist. Instead, we can say DNA has a twist of zero in its relaxed state and measure twist from this new baseline.

DNA within prokaryote and eukaryote genomes is actually a little underwound compared to relaxed DNA, with a around 1 helical turn "missing" for every 20 [63]. This has been shown experimentally[7], with a 7000 bp DNA loop having a linking number of $-40$ – a significant amount even for as small a loop as this! We can see in figure 4.5 that most of the contribution to the linking number is due to writhe ($Wr = -36\,;\, Tw = -4$). While the assignment of twist and writhe values is arbitrary in terms of topology, energetic considerations have a significant effect. The twist/writhe distribution will attempt to minimise the free energy from twisting ($Tw$) and bending ($Wr$) the DNA, which explains why we see certain configurations more than others. While a DNA loop with a twist of $-30$ and writhe of $-10$ would be topologically equivalent to the observed DNA loop, the energy cost of untwisting the DNA to such an extent is high enough that this does not happen.

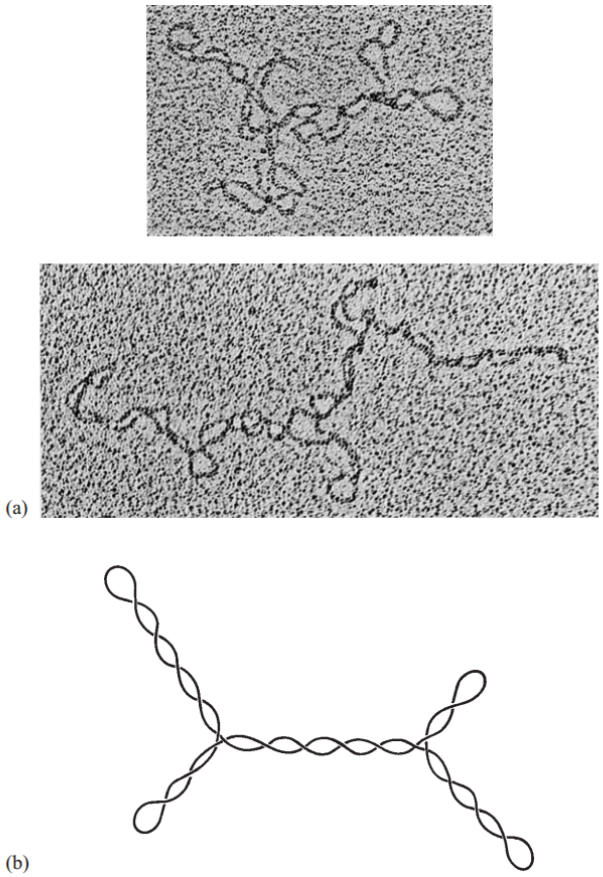



Figure 4.5: Image from [7]. (a) Electron micrographs of negatively supercoiled DNA from *E-Coli* bacteria. (b) Cartoon schematic of the DNA from (a).

### 4.2.1 Supercoiling and Transcription

Transcription has a significant effect on supercoiling, and vice versa. In the transcription process, a protein known as polymerase moves along the DNA and "reads" it, in order to produce a copy in the form of messenger RNA. This messenger RNA will later be translated into a specific protein, depending on which gene was transcribed.

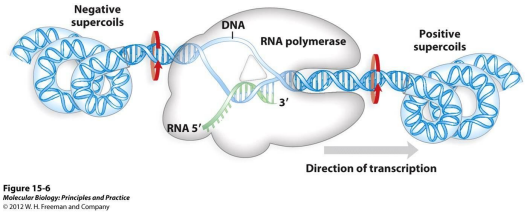



Figure 4.6: Image from [23]. An RNA Polymerase transcribing a gene, leading to a positively supercoiled region ahead of the transcription direction and a negatively supercoiled region behind.

While the polymerase transcribes DNA, the DNA is split into two regions - ahead and behind the direction of transcription (Figure 4.6). In the crowded cel-

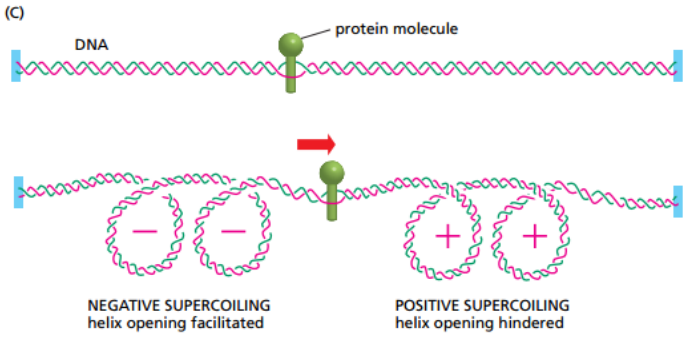



Figure 4.7: Image from [1]. An illustration showing the twin supercoiled domain model.

lular environment the polymerase is thought to be unable to rotate freely during transcription. This leads to a build up of helical turns ahead of transcription and a deficit of helical turns behind. This happens rapidly, with the linking number decreased behind the polymerase and increased ahead of it by 1 for every 10 bp transcribed in B-DNA. A and Z-DNA have slightly looser windings, with 11 and 12 base pairs per helical turn respectively.

Though linking number is still conserved for the whole DNA section or loop, the two sections will be significantly over- or under-wound. After transcription is finished this difference will eventually disappear, but not instantaneously. For this reason it is useful to define a quantity $\sigma$, representing the degree of local supercoiling compared to the equilibrium state. In 4.1 $Lk_0$ is the level of supercoiling in an equilibrated system ($\sigma = 0$ everywhere) and $Lk$ is a local measure of linking number. Values for $Lk$ can be positive or negative and $Lk$ is conserved for the system as a whole.

$$\sigma = \frac{Lk - Lk_0}{Lk_0} \tag{4.1}$$

The effect of transcription on supercoiling is based on some experimental observations [58] and is known as the "twin supercoiled domain model". This refers to the twin domains formed by the polymerase which have supercoiling of equal magnitude but opposite sign. An illustration of this model is seen in figures 4.6 and 4.7.

Negatively supercoiled regions facilitate helix opening, while positively supercoiled regions hinder it. The more open a region of DNA is, the more accessible it is to polymerases and other proteins. This may increase the likelihood of tran-

scription, or even make it possible at all by allowing the appropriate transcription factors to bind [45, 72].

The ubiquitous presence of supercoiling in transcription and overall cell function requires the cell to employ some topological enzymes to control and/or relax it. Topological enzymes are present in both prokaryotes and eukaryotes, and there are many copies and types of these in a single cell. All of these enzymes act in different ways but have the same general function - to add or remove supercoiling from regions of the DNA. These can generally be classified as either type I A/B/C or type II A/B, which change the linking number by $\pm 1$ and $\pm 2$ respectively. In bacteria, supercoiling is regulated in part by an enzyme known as gyrase. The sole function of gyrase is to make a break in the DNA and pass another strand through this break before resealing it, effectively reducing the linking number by 2 (Figure 4.9). Without this enzyme bacteria will eventually die, suggesting a level of structural openness is required for the cell to perform basic functions. Due to this, several anti-bacterial drugs (such as quinolones) work to inhibit gyrase in order to destroy bacteria. Since gyrase is not found in human cells, this should specifically target bacterial cells.

It could also be implied from the above and other studies [39] that positively supercoiled regions block transcription, which could be useful in some contexts. However in bacteria a large amount of the genome is functional, so anything which blocks transcription is likely to be unhelpful.

Type 1A topoisomerases have a significantly different mechanism of action to their type II counterparts, making only a single strand break in the DNA as opposed to the double strand break in type II. They then rotate the free strand $360^\circ$, modifying the linking number by $\pm 1$ (see figure4.8).

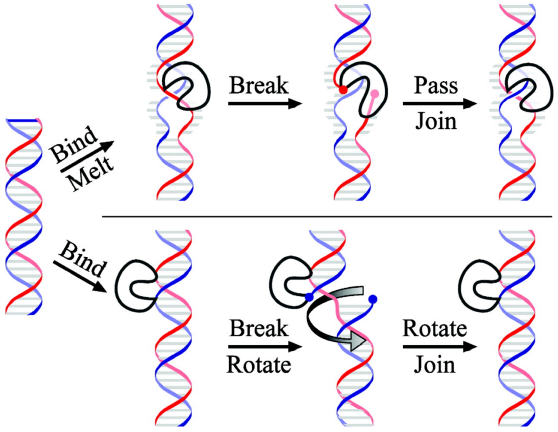



Figure 4.8: Image from [27]. The mechanism of action for some type I enzymes, where a single strand is cut, rotated and rejoined in order to change the linking number by 1.

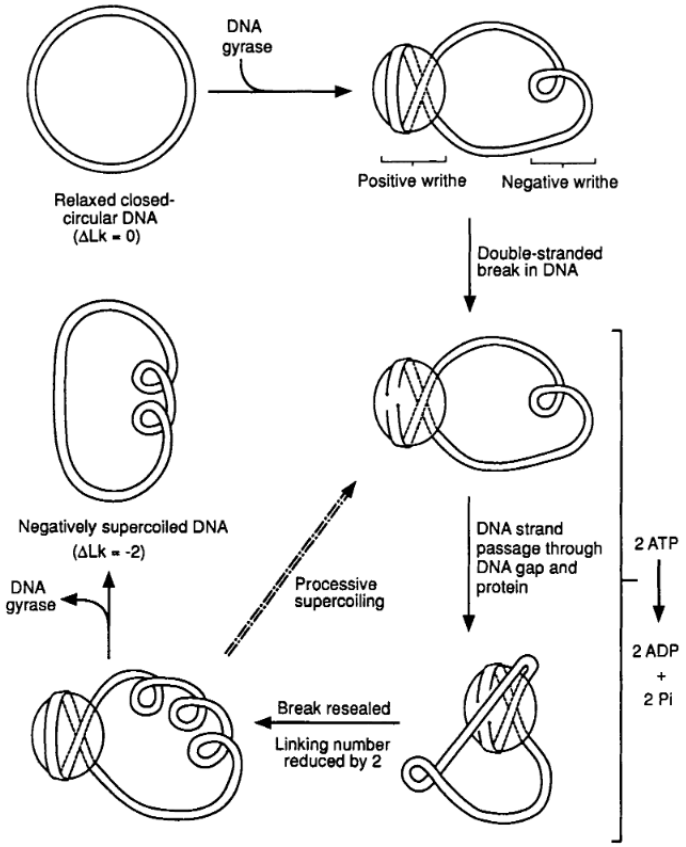



Figure 4.9: Image from [85]. The proposed strand-passing mechanism which allows gyrase and similar type II enzymes to reduce a DNA molecule's linking number by 2.

## 4.3 A Numerical Model For Supercoiling

In this section we move from molecular dynamics to a lattice-based modelling approach. Our DNA is modelled as a 1D lattice and position on the DNA specified as $x$, with each lattice site having a length ($l$) of 15 bp. We assume that the DNA lattice contains an arbitrary number ($n$) of genes, with each gene having a length equal to $\lambda$. For the simulations in this section $\lambda$ was constant, meaning all the genes were the same length - but this is not required by the model and could be changed for future simulations. Each gene has a promoter at a particular position on the DNA ($y_i$ for $i = 1 \,..\, n$). There are also an arbitrary number of polymerases ($N$), which are able to transcribe the genes.

Gene transcription is modelled as a stochastic process; at each timestep, each of the free polymerases in the system has a probability ($k_{on}$) to bind to a promoter and begin transcription. When a polymerase binds to a promoter it moves along the gene with a velocity $v$, until it reaches the end of the gene and is re-added to the pool of free polymerases. As an example, the position along the DNA of a polymerase transcribing the $i$th gene is $x_j = y_i + vt_j$ where $t_i$ is the time since the polymerase bound to gene $i$.

Each point of the lattice has an associated value of $\sigma$, representing a measure of local supercoiling (see equation 4.1). Our model uses the following diffusive dynamics for $\sigma$:

$$\frac{\partial \sigma(x,t)}{\partial t} = \frac{\partial}{\partial x}\Big[D\frac{\partial \sigma(x,t)}{\partial x} - J_{tr}(x,t)\Big] \tag{4.2}$$

The $\frac{\partial J_{tr}(x,t)}{\partial x}$ part of the equation represents the flux of supercoiling across the polymerase, while the other term is diffusive.

$$\begin{aligned} J_{tr}(x,t) &= \sum_{i=1}^{N} J_i(t_i)\delta(x - x_i(t_i))\xi_i(t) \\ |J_i(t_i)| &= J_0\Big(1 + \frac{vt_i}{l}\Big) \end{aligned} \tag{4.3}$$

Equation 4.2 represents the change in local supercoiling through the DNA. Setting $J_{tr} = 0$ would allow the system to eventually relax to a state where supercoiling is evenly distributed throughout the DNA. The diffusion constant ($D$) sets the timescale for this process, and within naked DNA experiments have measured a value of $D \sim 0.1kbp^2/s$ or less [59].

Equation 4.3 is perhaps not so nice mathematically, but can be related to the transcription process fairly straightforwardly. Overall, $J_{tr}$ is the supercoiling flux generated by a transcribing polymerase and has its sign dependant on the direction of transcription. The function $\xi_i(t)$ is used as a filter for whether a gene is currently transcribing, being equal to 1 if it is and 0 if not. Similarly, $\delta(x-x_i(t_i))$ is a filter which tracks the current location of an active polymerase. Finally, $J_i(t_i)$ is the magnitude of the transcriptional flux, where $J_0$ is a constant which sets the flux generated per bp. transcribed and the $\frac{vt_i}{l}$ term reflects supercoiling being "pushed forward" by the polymerase.

We also used a simpler version of this equation which uses static polymerases, which attach to the gene promoters and generate supercoiling at that location. This was done as a first attempt at characterising this system, as well as to provide a more solid foundation for analytical work. Some exact results for this model, as well as mean field and scaling results for the travelling polymerase model are detailed in appendix B.

One illustrative result is the steady state solution of our static polymerase equation for one (switched-on!) gene located at $x = 0$:

$$\frac{\partial \sigma(x,t)}{\partial t} = \frac{\partial}{\partial x}\Big[D\frac{\partial \sigma(x,t)}{\partial x} - J_0\delta(x)\Big] \tag{4.4}$$

We use the boundary condition $\sigma(0,t) = 0$ , along with the condition that there is no flux of supercoiling out of the system, so the overall level of supercoiling is fixed. If we attempt to solve our equation on an infinite domain, this implies $\frac{\partial \sigma}{\partial x} = 0$ for $x \to \pm\infty$ (Setting this condition fixes the overall level of supercoiling). The fact we are considering a steady state also means we have $\frac{\partial \sigma(x,t)}{\partial t} = 0$.

This gives us:

$$\frac{\partial^2}{\partial x^2}D\sigma - \frac{\partial}{\partial x}J_0\delta(x) = 0 \tag{4.5}$$

$$\frac{\partial}{\partial x}\Big[\frac{\partial}{\partial x}D\sigma - J_0\delta(x)\Big] = 0 \tag{4.6}$$

$$D\frac{\partial \sigma}{\partial x} - J_0\delta(x) = c \tag{4.7}$$

$$\text{Since } \frac{\partial \sigma}{\partial x} = 0 \text{ for } x \to \pm\infty \implies c = 0 \tag{4.8}$$

Integrating again gives $D\sigma - J_0 H(x) = c'$

Where H(x) is a Heaviside function with $\begin{cases} H(x) = 1 \text{ for } x > 0 \\ H(x) = \frac{1}{2} \text{ for } x = 0 \\ H(x) = 0 \text{ for } x < 0 \end{cases}$

Substituting in the boundary condition $\sigma(x = 0, t) = 0$ and $H(0) = \frac{1}{2}$ gives $c' = -\frac{J_0}{2}$

$$\text{And finally, } \sigma = \frac{J_0}{D}\left(H(x) - \frac{1}{2}\right) = \frac{J_0}{2D} sgn(x) \tag{4.9}$$

$$\text{where } sgn(x) = 2H(x) - 1, \quad i.e. \begin{cases} sgn(x) = 1 \text{ for } x > 0 \\ sgn(x) = 0 \text{ for } x = 0 \\ sgn(x) = -1 \text{ for } x < 0 \end{cases}$$

While this result does not tell us anything too surprising, it is nice to see the basic idea of creating positive supercoiling ahead of transcription and negative supercoiling behind coming out of the initial equations. The result also shows positive and negatively supercoiled regions separated by the gene, though in a system with periodic boundary conditions these regions would meet and cancel out.

### 4.3.1 Technical Details and Limitations of the Model

It is worth justifying some of the ideas behind the above model, as well as the limitations when moving from a continuous equation to a lattice representation. We start from an approximation of a free-energy density for DNA with supercoiling $\sigma$, shown below.

$$f = \frac{A\sigma^2}{2} \tag{4.10}$$

Here, $A$ is a positive constant which sets the scaling between $f$ and $\sigma^2$ and has the same form found in *Marko et al* [65] ($[A] = [k_B T * L * C] = kgm^4 s^{-2}$). In *Marko et al*, the constant $A$ is a determined by the polymers persistence length for bending (please note, in *Marko et al* this is also denoted by $A$) and twisting ($C$) as well as a constant $\alpha$ determined by how a change in linking number is partitioned between twist and writhe ($\alpha$ is calculated via simulation). The length of the polymer is given as $L$. A free energy has also been determined experimentally as $f \approx$

$10.0 k_B T N \sigma^2$, where $N$ is the number of base pairs [98]. Since our model considers DNA loops and linear DNA with fixed ends, the overall level of supercoiling is fixed and this means we can use "Model B" [17, 46] dynamics for the system.

In the equation below $x$ is the position along the DNA, $M$ is the mobility associated to supercoiling density and $t$ is time. This gives us an effective diffusion coefficient $D = MA$.

$$\frac{\partial \sigma(x,t)}{\partial t} = M\nabla^2 \frac{\partial f}{\partial \sigma} = MA\nabla^2 \sigma(x,t) \equiv D\nabla^2 \sigma(x,t) \tag{4.11}$$

However, this free energy is only appropriate for small values of $\sigma$ [65]. While there are improvements that could be made to the functional, there are also other issues which arise at large $|\sigma|$ values. One issue is to do with transcription probability, as repeated underwinding of a local DNA region will not cause the transcription probability to increase indefinitely. In fact, the opposite happens for large enough values of negative supercoiling. We also would not expect supercoiling to be created if transcription occurs in an area with linking number $\approx 0$ ($\sigma \approx -1$), as the polymerase does not have to unwind anything.

Because these extra issues would not be resolved for other free energy functionals, it is more sensible to stick with our harmonic approximation but remaining aware of its issues for large $|\sigma|$. In simulations, these issues are manifest more as local inaccuracies in regions with $\sigma > 1$. While the analytical results may not be valid for these regions, the general principles behind the model and thus the simulation should still apply.

While a continuous representation of our model uses $\delta(x - x_i)$ to implement supercoiling flux only at appropriate positions, it needs to be altered to take into account the fact that polymerases have a finite size $\sim \Delta x = 15$ bp . Because of this we use a regularised form for analytical calculations, with $\delta = \frac{exp(-x^2/(4l^2))}{2l\sqrt{\pi}}$. Here the support of the function ($l$) is set as the size of a lattice site, $\Delta x$. In some of the earlier simulation runs the delta function was replaced with a kroenecker delta $\delta_{x,0}$, meaning regularisation occurs with $l \approx \Delta x$.

A final detail to be aware of is the fact that a polymerase can bind as soon as the promoter is free, even if the previous polymerase is only a single lattice site away. While this is not impossible, it would be more difficult for a second polymerase to bind with the first one in close proximity. However there is no simple way to

characterise this, outside of imposing arbitrary restrictions or alterations to $k_{on}$.

One possible method to get around this issue would be to couple the 1D simulations to a 3D molecular dynamics code, however doing this in an efficient way so as to be able to follow the supercoiling dynamics for a comparable amount of time as in the 1D model would be extremely challenging in practice.

While equation 4.3 has perhaps more terms than would be expected, the actual process being modelled is not so complicated. We can think of the polymerase as adding twist to the lattice site in front of its position while decreasing it at the lattice site behind. It also "pushes" the supercoiling forward as it moves along the DNA as shown in figure 4.10.

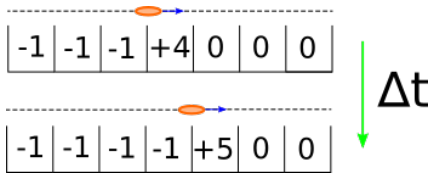



Figure 4.10: An example (without diffusion!) of the polymerase building up supercoiling as it moves along a gene.

The term $\xi_i(t)$ is set to zero or 1 depending if the polymerase is active or inactive and in the "pool" of polymerases. This pairs with the function $\delta(x - x_i(t_i))$ which means that flux is only applied at the current location of the polymerase, $x_i(t_i)$. The final term $J_i(t_i)$ determines both the flux applied per timestep and causes the supercoiling to rack up in front of the polymerase. The flux applied at a single timestep increases at each timstep by $\frac{J_0 v \Delta t}{l}$, i.e. the flux at $t_i$ is $J_0(1 + \frac{v t_i}{l})$. This gives the effect seen in figure 4.10.

Based on the finding that negative supercoiling facilitates polymerase and transcription factor binding we assume that the probability for a polymerase to bind to a particular gene depends on the value of $\sigma$ at that gene's promoter $\sigma_p$. Since the form of the polymerase binding probability distribution is unknown and there is no obviously superior choice, a linear coupling to supercoiling was used with $k_{on} = k_0 \max(1 - \alpha\sigma_p, 0)$. The max function is simply because it does not make physical sense for the binding rate to be less than 0. Additionally, $k_0$ is the polymerase binding rate for $J_0 = 0$, while $\alpha$ is the sensitivity to $\sigma_p$.

The linear coupling used here still leads to highly non-linear dynamics, as supercoiling created by transcription favours the transcription of upstream genes (against the direction of transcription), while hindering the transcription of downstream genes (with the direction of transcription).

We can also create dimensionless parameters out of combinations of the ones used

in the model. There are three main parameters which are of interest:

$$\begin{aligned} \Phi &= \Big(\frac{k_{on}N}{n}\Big)\tau \\ \Theta &= \Big(\frac{k_{on}N}{n}\Big)\frac{\lambda^2}{D} \\ \frac{\bar{J}}{D} &= J_0\Big[1+\frac{\lambda}{2l}\Big] \end{aligned} \tag{4.12}$$

$\Phi$ is the product of transcription initiation rate and transcription time and gives a measure of how often an average gene is being actively transcribed. Clearly, boosting $k_{on}$, the number of polymerases $N$ or transcription time $\tau$ should increase this value. Since the polymerases follow a uniform probability distribution when selecting which gene to attempt to bind to, increasing the number of genes makes any specific gene less likely to be selected.

$\Theta$ is a measure of how quickly supercoiling diffuses away between transcription events. Since supercoiling flux is generated at a constant rate, the length of the gene $\lambda$ sets the magnitude of the flux, and $D$ controls its diffusivity.

Finally, $\frac{\bar{J}}{D}$ is a measure of the supercoiling generated at the promoter site while the gene is active, with $\bar{J}$ being the average supercoiling flux during transcription. The value $\frac{\lambda}{2l}$ measures how many lattice sites away the midpoint of the gene is. Since flux increases linearly, the flux at this midway point is equal to the average value for the whole gene. From dimensional analysis we would also expect that $\bar{J} \approx v\lambda$ .

### 4.3.2 Relating Simulations and Theory to Measured Quantities

The parameters defined above can be related to observable quantities in both prokaryotes and eukaryote. We can derive an estimate for average supercoiling ($\bar{\sigma_p}$) at a promoter (the full derivation can be found in appendix B) as

$$\bar{\sigma_p} \approx -\Big[\frac{\Phi}{\Phi+1}\Big]\frac{\bar{J}}{D}. \tag{4.13}$$

This result allows us to relate supercoiling density and parameters in the simulation with expermentally observed values. In bacteria the baseline value of super-

coiling is $\sigma_p \approx -0.05$, with experiments suggesting that $\sigma_p \approx -0.01$ is enough to affect polymerase binding [86]. Transcription rates in bacteria are of order $\sim$10 RNA molecules per minute ($k_{on} \sim 0.16$), with a typical gene size ($\lambda$) of around 1 kbp and polymerase transcription velocity $v \sim 100$ bp/s. For *E. coli* specifically, there are approximately 3000 polymerases per cell [53] and approximately 5000 genes so we can get a rough estimate for $\Phi \approx 0.42$. As mentioned previously $D \approx 0.1 kbp^2/s$. Plugging these numbers into equation 4.13 gives $\sigma_p \approx -0.3$. This difference from the baseline suggests that supercoiling can be relevant to transcription in prokaryotes.

Depending on the parameters of the simulation, the value of $\sigma_p$ can be significantly affected. For example setting $\Phi = 10$ and $\frac{\bar{J}}{D} = 1.0$ gives $\sigma_p \approx -1.0$ at the promoter, again suggesting the behaviour here will differ from an "average" region of the DNA. By using genes which remain a constant size and polymerases which transcribe at a constant rate (i.e. for $\Phi$ fixed), we can see how varying the values of $\frac{\bar{J}}{D}$ affects the simulation results (see section 4.5).

We can repeat the above calculation for eukaryotes, though the results should be seen as more of an order of magnitude estimate as some parameters are not yet known for chromatin – for example the supercoiling diffusion rate $D$.

Transcription in eukaryotes is considerably slower in terms of polymerase transcription velocity ($v \approx 25$ bp/s), and genes are longer ($\lambda = 1.6$ kbp in yeast; 10 kbp in humans). The number of transcripts produced (around 1 per hour in humans; 10 per hour in yeast) is lowered further by the need for several transcription factors to co-localise at a promoter before transcription can be initiated.

Using the above numbers gives $\sigma_p = -0.03$ for yeast and $\sigma_p = -0.13$ for humans. This suggests supercoiling could potentially be relevant to transcription in eukaryotes also, though it's important to take into account the caveats mentioned previously!

## 4.4 Mutual Information and Conditional Entropy

In the results section we require a way to characterise the correlations between genes, namely to what extent the transcription of one gene affects the transcription of another in the same system. To do this we use quantities from information theory known as mutual information and conditional entropy [22]. Since these

are not widely used in physics, they are defined here.

Both quantities are defined in terms of a time series, in our case this will be the index of the gene transcribed across the time period of our simulation. We can refer to this series as $i_q$, where $i_1$ is the index of the gene transcribed in the first transcription event, $i_2$ the second, etc.

### 4.4.1 Conditional Entropy

The conditional entropy $S$ of a time series $i_q$ is defined as

$$S(\{i_q\}) = -\sum_{i,j} p(i,j) \log\left[p(i|j)\right], \tag{4.14}$$

where $p(i,j)$ is the probability of observing the transcription of gene $i$ followed by $j$. Note that the time series format does not put any constraint on how long after transcription of $i$ this occurs, it just requires it to be the next transcription event. $p(i|j)$ is the conditional probability of gene $i$ being transcribed next, given that gene $j$ was the last one transcribed. In general $P(i,j) \neq P(j,i)$; in a system with two genes (Figure 4.11) we expect transcribing gene $i_1$ to direct positive supercoiling to $i_2$, reducing the probability of transcription. However transcribing $i_2$ will have the reverse effect on $i_1$ with the negative supercoiling increasing the probability of transcription.

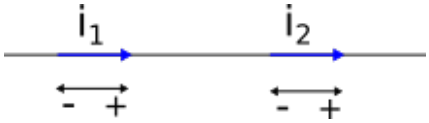


Figure 4.11: Two genes $i_1$ and $i_2$. Blue arrows indicate transcription direction, black arrows indicate supercoiling flux.

The conditional entropy is maximised at $log(n)$ if transcription events are uncorrelated ($\bar{J} = 0$) and minimised at 0 for a maximally correlated process, for example when a single gene is repeatedly transcribed.

### 4.4.2 Mutual Information

The mutual Information $I$ is defined as

$$I(\{i_q\}) = \sum_{i,j} p(i,j) \log\left[\frac{p(i,j)}{(p(i)\,p(j)}\right] \tag{4.15}$$

where $p(i)$ is the overall probability that gene $i$ is activated. The mutual information is equal to 0 if $p(i,j) = p(i)\,p(j)$, in our system this would correspond to gene transcription being random ($\bar{J} = 0$). Its value therefore measures the divergence of the joint probability distribution for successive transcription events from that of a random process. In statistical mechanics systems it is often found that the mutual information peaks at or close to phase transitions, where correlations are maximal.

## 4.5 Results - Randomly Positioned Uni-Directional Genes

The simulations here use parameters relevant to bacterial DNA, as well as peroidic boundary conditions in the 1D model to simulate a DNA loop. While we also consider this model for parameters relevant to eukaryotes, it is not clear if the periodic boundary condition would still apply when loops are held together by architectural proteins. The first set of simulations use genes which are randomly placed along the DNA, though a initial choice for a placement would be rejected if a part of the new gene is within 1 kbp of an existing gene. All the genes in this simulation transcribe in the same direction (left to right in figures).

To obtain the results shown in this section, multiple runs were performed with $\frac{\bar{J}}{D}$ varied from 0 to 3.5.

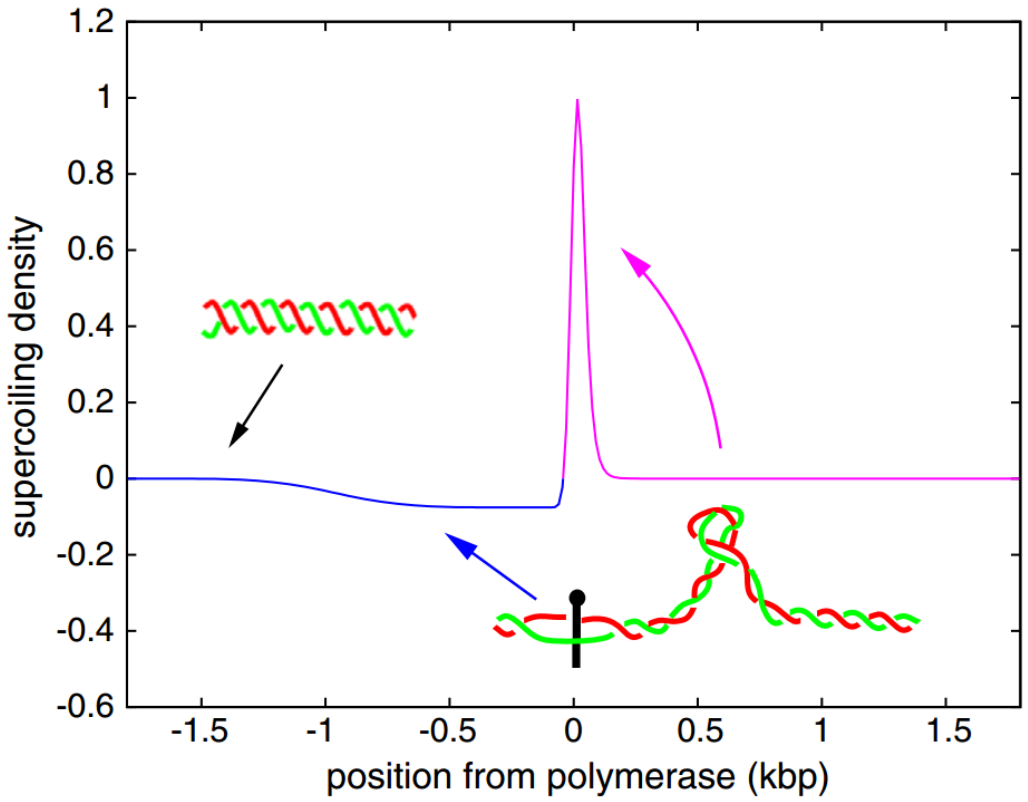



Figure 4.12: A snapshot of a simulation with $\frac{\bar{J}}{D} = 1.7$, showing supercoiling density close to a gene which is being transcribed. The graph shows the build up of positive supercoiling just ahead of the polymerase and a negative supercoiling "wake" behind it.

We see two behavioural regimes as the value of $\frac{\bar{J}}{D}$ is modified. The first is the "relaxed" regime, which occurs for small values of $\frac{\bar{J}}{D}$. In this regime the levels of supercoiling generated by transcription are low enough that they do not significantly affect the transcription probability of neighbouring genes. In this case gene transcription can be modelled as a Poisson process, with every gene being read on average the same number of times.

As $\frac{\bar{J}}{D}$ increases, we reach a point where supercoiling does have an effect on transcriptional dynamics, which we call the supercoiling-regulated regime. In this regime we observe transcriptional bursts, which are when a gene is repeatedly transcribed. This occurs due to the negative supercoiling "wake" close to the promoter being large enough to increase transcription probability, leading to a positive feedback loop between transcriptional activity and probability.

We also observe transcription waves (Figure 4.13 (b)), where the positive supercoiling generated by transcription silences a downstream and promotes transcription of an upstream gene. Transcription of this upstream gene will also produce positive supercoiling affecting the initially transcribed gene, leading to the transcriptional waves observed.

The characteristics of the supercoiling-regulated regime are seen whether genes are randomly positioned or positioned a fixed distance apart. However there is an extra characteristic observed for randomly positioned genes, where average transcription probability is dependent on the distance from upstream neighbours. If this distance is large, the diffusion of the positive supercoiling generated from transcription will result in less of an effect on transcription probability for the downstream gene (Figure 4.13 (c)).

The transition between the two regimes can be seen in figure 4.13 (d) and figure 4.14 (a), which both indicate that the transition is gradual as $\frac{\bar{J}}{D}$ is increased.

Decreasing the order of magnitude of $k_0$ results in a sharper peak in the mutual information, as well as the peak (or maximum value) occurring at lower $\bar{J}$. However the general behaviour of the system is more or less the same. Transcription rate is also mainly dependant on $\frac{\bar{J}}{D}$, rather than $k_0$ – though $k_0$ does have an effect at small $\frac{\bar{J}}{D}$.

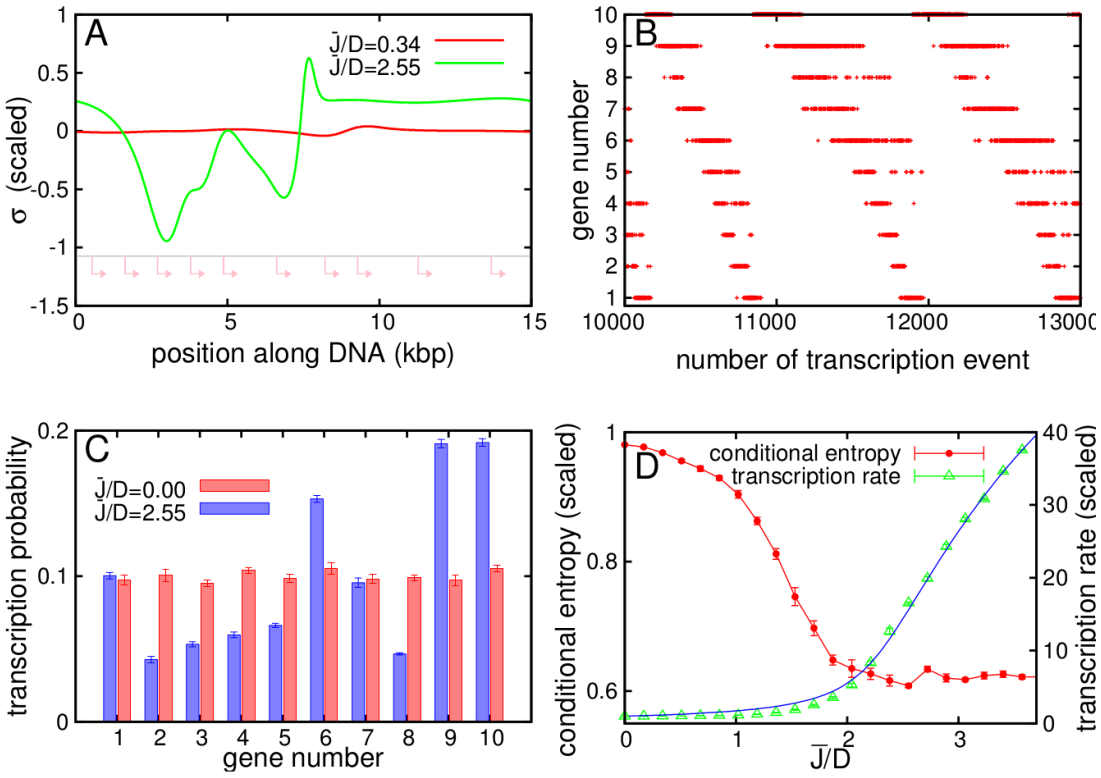



Figure 4.13: Simulations for a 15 kbp DNA loop, red arrows indicate genes and transcription direction.
(A): A snapshot of two separate simulations with $\frac{\bar{J}}{D} = 0.34$ (relaxed) and $\frac{\bar{J}}{D} = 2.55$ (supercoiling-regulated).
(B) Part of the time series showing the order of transcribed genes; transcription waves can also be seen.
(C) Average transcription probability for $\frac{\bar{J}}{D} = 0$ (relaxed) and $\frac{\bar{J}}{D} = 2.55$ (supercoiling-regulated).
(D) Conditional entropy and transcription rate for varying $\frac{\bar{J}}{D}$. The blue line indicates the transcription rate derived from the analytical theory in appendix B.

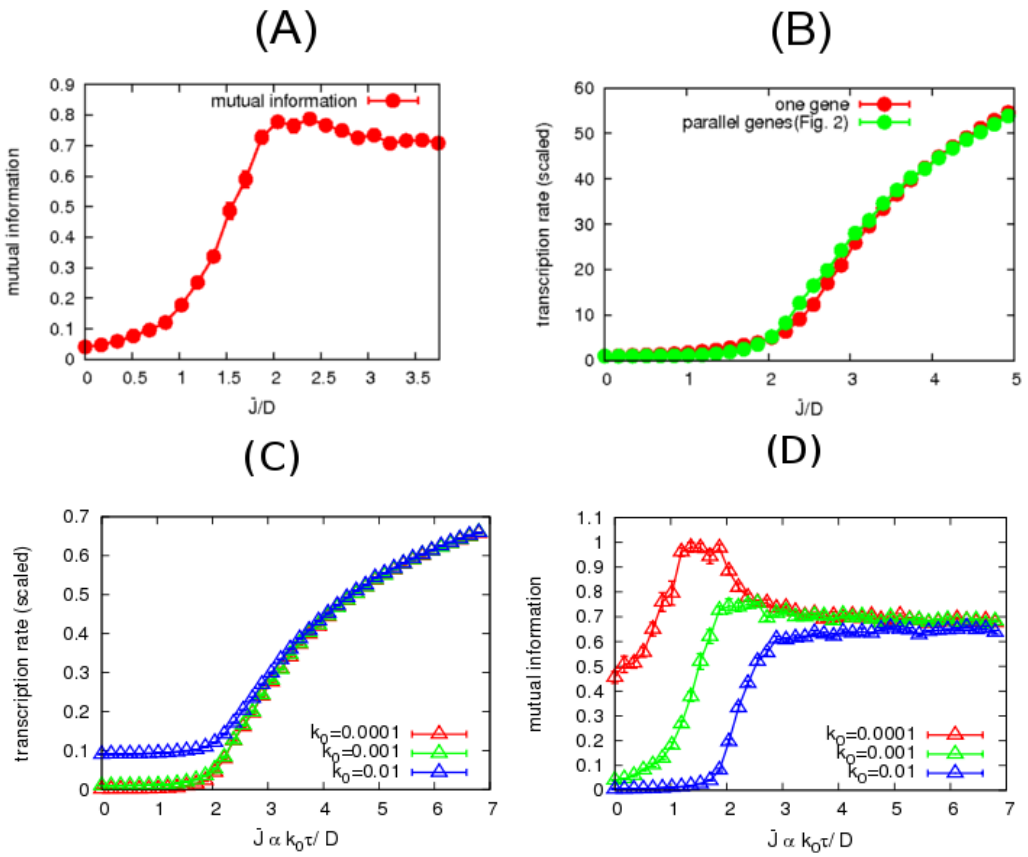



Figure 4.14: Simulation results for the system in figure 4.13.
(A) Mutual information for varying $\frac{\bar{J}}{D}$.
(B) Overall transcription rate for the system in figure 4.13 and a single-gene model. The overall transcription rate is normalised with the expected value at $\frac{\bar{J}}{D} = 0$ for both cases.
(C & D) Transcription rates and mutual information for different values of $k_0$. The x-axis values are plotted in terms of $\bar{J}\alpha k_0\tau/D$ for comparison with the results for the mean field model in appendix B. All data points for C & D are an average of 7 simulation runs.

## 4.6 Results - Bidirectional Genes

As genes can be encoded in either the forward or reverse strand of the DNA double helix, the model in section 4.5 was modified to give each gene a transcription direction. As with the unidirectional genes there exist both a relaxed and supercoiling-regulated regime, however the characteristics of the supercoiling-regulated regime now depend on both gene position and direction. Transcription was observed to be boosted for pairs of genes which point in opposite directions but away from each other (divergent). Conversely, transcription is decreased for genes which are facing each other (convergent). Similar mechanics apply for parallel genes as in section 4.5.

In the supercoiling regulated regime we often find that a divergent pair dominates in terms of transcription frequency, with only this pair being transcribed at the simulation end. This causes the build up of a large amount of negative supercoiling around the gene pair, while positive supercoiling is distributed evenly across the other genes (see figure 4.15 (a)). This forms a positive feedback loop, as a pair being transcribed means they are more likely to be transcribed again.

For systems with multiple divergent pairs, other factors must also be considered. For example, a pair with a short distance between genes is more likely to dominate at the simulation end, as would a pair which comes at the end of a run of parallel genes. Fluctuations are also significant enough that even when a pair possesses both of these properties it does not guarantee that the pair will dominate.

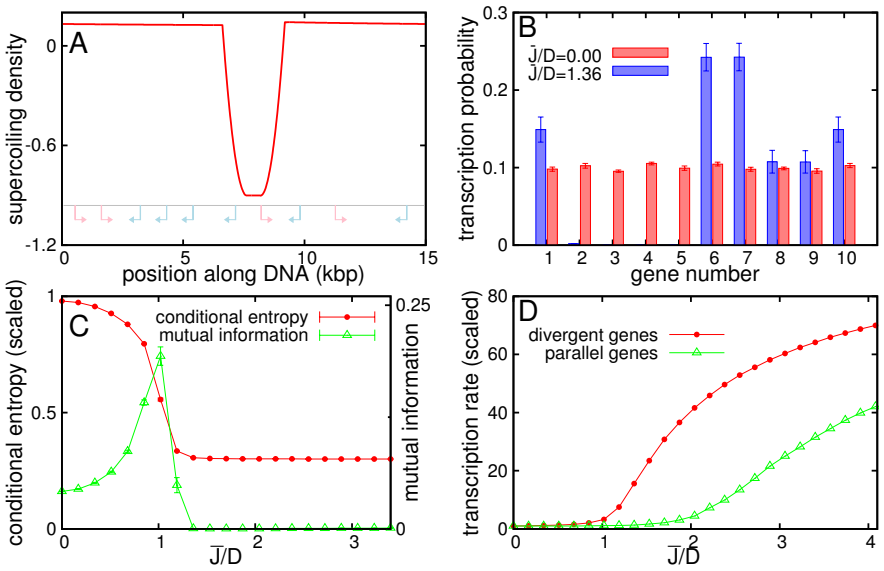



Figure 4.15: (A) A snapshot from the end of a simulation with $\frac{\bar{J}}{D} = 1.36$ where only two genes are being transcribed. Since supercoiling is conserved, the positive supercoiling is distributed across the other genes and prevents their transcription.
(B) Average transcription probabilities across multiple simulation runs for $\frac{\bar{J}}{D} = 0$ (Red) and $\frac{\bar{J}}{D} = 1.36$ (Blue). The divergent pair of genes 6 & 7 is transcribed more regularly due to the closeness of the genes, as well as the negative supercoiling generated by the parallel genes 3,4 and 5.
(C) Conditional entropy & mutual information, scaled by $\log(n)$.
(D) Transcription rate for different gene configurations, scaled by $k_0 N$. Parallel genes corresponds to the set-up in figure 4.13, while divergent genes is an arrangement where the first 5 genes transcribe upstream and the final 5 downstream, as seen in figure 4.16 (B). This creates a divergent pair at genes 5 & 6.

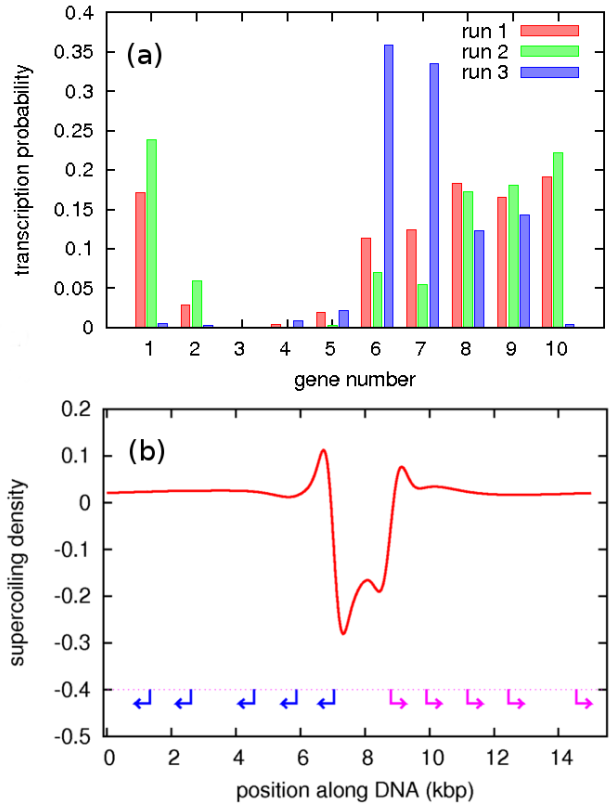



Figure 4.16: (a) Transcription rate for individual simulation runs, for the simulation set up in figure 4.15. (b) A snapshot of a simulation with a single divergent gene pair.

## 4.7 Results - Topoisomerases

If the positive feedback loop for divergent pairs described above were to exist in isolation, there would be serious negative consequences for the cell as a whole. Clearly, transcribing only a limited set of genes to the exclusion of everything else would not be a good thing, especially in bacteria where a large proportion of the genome is functional.

The way a cell can break up this feedback loop is through topoisomerases (see section 4.2, which can add or remove supercoiling in various ways. This occurs at a rate of 0.1-1 supercoil/s in cells [92], with the rate depending on whether the topoisomerase is type I or type II, along with local cellular conditions. This now means $\sigma$ is no longer conserved, as a topoisomerase modifying supercoiling in a local region does not automatically imply that the reverse process will be occuring elsewhere.

Incorporating this effect in the model involved adding a non-conserved reaction term to equation 4.2, giving:

$$\frac{\partial \sigma(x,t)}{\partial t} = \frac{\partial}{\partial x}\Big[D\frac{\partial \sigma(x,t)}{\partial x} - J_{tr}(x,t)\Big] - k_{topo}\sigma \tag{4.16}$$

$k_{topo}$ is a relaxation rate, which can be associated with a length scale $l_{topo}$ $\sqrt{D/k_{topo}}$. This is the distance around the topoisomerase for which supercoiling-mediated interaction will be screened.

We can see the effect of this in figure 4.17, with a significant down regulation of transcription for larger values of $k_{topo}$, along with a rise in conditional entropy - showing the loss of correlations in the transcription process.

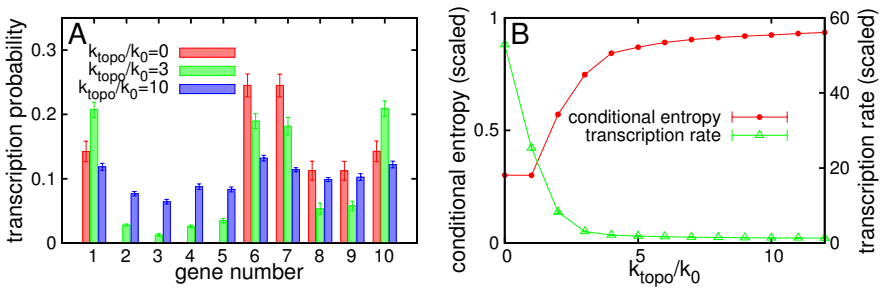



Figure 4.17: Simulations with $\frac{\bar{J}}{D} = 2.55$, meaning the system is in the supercoiling-regulated regime for $k_{topo} = 0$. (A) The transcription probability for different $k_{topo}$. The increase in transcription from being in a divergent pair is almost completely lost for $k_{topo}/k_0 = 10$. (B) Conditional Entropy and Transcription Rate.

It is important to note that our model only includes topoisomerases which act to decrease supercoiling towards the equilibrium value (i.e. $\sigma = 0$). There are also other enzymes which can act to increase negative supercoiling, for example gyrases in bacterial cells.

## 4.8 Videos & Downloads

Videos of the simulations and downloads of the code used to run them are available at `http://www2.ph.ed.ac.uk/~s0841882/chapter3.html` or `http://www.jjthesis.co.uk/chapter3.html`.

## 4.9 Summary

This section details a model which combines a continuum description for the evolution of supercoiling with stochastic transcriptional dynamics, showing how the two processes affect one another. We see two regimes, a low-flux relaxed and high-flux supercoiling-regulated regime, along with the crossover region between them. In the simulations with parallel genes, we also observe features seen experimentally in both prokaryotes and eukaryotes, such as transcriptional bursts[18, 19, 42]. We also found that that genes were regulated depending on their separation from upstream neighbours.

When gene directionality was included we found divergent gene pairs to be highly transcribed, corresponding with the observation that divergent genes are often as-

sociated with essential genes which tend to be highly expressed [99]. Bidirectional promoters are also found in mammalian genomes at a higher rate than expected, so this result could be of interest here also!

Our model also shows topoisomerases can act to downregulate transcription, something which is observed experimentally [34, 74].

A new prediction of the current model is that of transcription waves. Whilst they have not yet been observed experimentally, it is conceivable that they could be recreated in the lab by using DNA plasmids and DNA editing techniques, which allow in principle to position genes on such plasmids in a controlled way.

Further work based around this model could involve adapting it to include the effects of further topological enzymes, or designing a version more specific to eukaryote genomes. This would involve incorporating the effect of nucleosomes on the supercoiling density of the system, as these can act as barriers to supercoiling. An additional avenue which would be very interesting to explore is to model polymerase movement in 3 dimensions: this would be important as it would allow discrimination between twist and writhe, which is impossible to achieve within our diffusive 1D model.

# Chapter 5

# Conclusions

In this thesis we discussed a number of problems in biophysics relating to DNA and chromosomes. This is a vast area of research and this thesis of course provides only a small selection of topics. For the majority of the thesis, we have discussed numerical results obtained from molecular dynamics simulations of biopolymers, although the last results chapter contains both analytical results and simulations performed with a simple stochastic 1D model. The overall main message of the thesis is that such physically inspired models can help us understand some aspects of the behaviour of DNA and chromosomes, in a way which is complementary to what is done with experiments – which is normally the method of choice to analyse these systems.

In chapter 2, we studied a simple coarse grained model for protein-DNA bridging, which uncovered a generic mechanism for cluster formation which we call "bridging-induced attraction". For this mechanism to arise requires proteins to be able to bridge the chromatin via multivalent interactions, something which is observed experimentally in the HP1$\alpha$ protein in humans. This mechanism should be at work in both bacterial DNA and eukaryotic chromatin, even in the absence of specific interactions between the genome and its associating proteins.

The bridging-induced attraction is a simple thermodynamic feedback loop. When proteins bind, they may bridge distant sites along the DNA, which increase the genomic concentration locally. This increase in concentration will stimulate the binding of other proteins, as binding is more likely wherever there is more DNA. In turn this increases concentration further, leading to a positive feedback loop which yields clustering.

The simple protein-DNA bridging model studied here provides a generic mechanism for cluster formation among bacterial DNA and chromatin. Even when the interaction is completely non-specific, there is a qualitative similarity between the results of the simulations here and the observations of experimental studies.

Our simulations show a clear link between protein and polymer clustering. They also show that depending on the concentration of proteins, we either get clustering, or, for sufficiently large concentration, full collapse of the polymer. Additionally, we find that flexibility plays a significantly qualitative role, as the clusters formed with semi-flexible DNA are rod-like, whereas those formed with flexible chromatin are quasi-spherical.

As an extension to the work in chapter 2, it would be interesting to quantify the exponents determining the growth laws of clusters in both the flexible and semi-flexible cases. The simulations could also be performed at a larger scale, although this would be difficult considering our computational capabilities.

In chapter 3, we built on this model and attempted to apply it to a eukaryotic chromosome. As a first step, we included specific binding, and multiple transcription factors. This was motivated by the observation that in human chromosomes different types of proteins bind to active and inactive region of the genome. The specific binding leads to clustering via the bridging-induced attraction mechanism, but now clusters grow only up to a certain size due to the steric interactions between loops which join the sites where specific binding takes place. The clusters formed were also observed to be mostly of only one protein type, either proteins which bind to active regions or those which bind to inactive regions. This suggests a possible pathway for the separation of active and inactive regions of chromatin.

By using bioinformatic data to assign the sites of interactions between active and inactive proteins and regions of human chromosomes, we were able to create contact maps, which determine which genomic regions are likely to be in contact with each other. The simulation results compare favourably, although qualitatively, with experimental contact maps.

Another popular model for genome organisation, which we did not consider in Chapter 3, is that of loop extrusion. This model is appealing because it helps accont for the observation that bridging by the CTCF regulatory proteins depends on the directionality of the CTCF binding sites, which is not captured by our simple model. It would therefore be of interest, as an extension of the work in

chapter 3, to combine this extrusion model with the bridging-induced attraction model which we have presented in this thesis.

Finally, in chapter 4 we presented a simple numerical model which links transcriptional activity to the local supercoiling of DNA. This model combines a continuum description of supercoiling with a stochastic description of transcription, where supercoiling is able to diffuse across the DNA and polymerases bind to genes with a probability based on the level of local supercoiling at a gene promoter. A polymerase binding to a promoter and beginning transcription then causes a flux of supercoiling across the polymerase – representing positive supercoiling being pushed in the direction of transcription, and a negative supercoiling wake being left behind.

In general, we were able to characterise two regimes present in our simulations. One of these is the "relaxed" regime where flux generated by transcription is low and genes transcribe randomly. The other is the "supercoiling-regulated" regime, where supercoiling flux is high and gene transcription is dependent on the gene's positioning and the transcription history of the system. When this model is applied to a genome where all genes are transcribed in the same direction, we observe features also seen experimentally in both prokaryotes and eukaryotes, such as transcriptional bursts. We also found that a gene's separation from its upstream neighbours can either promote or suppress transcription, depending on whether the neighbour is distant or close by.

When considering a system where genes can transcribe in opposite directions, we found a significant increase in transcriptional activity for bi-directional gene pairs – something which is also seen experimentally in yeast genomes. Bidirectional promoters are also found in mammalian genomes at a higher rate than expected, so this result could be of interest here also. We also extended the model to incude the effects of topological enzymes, which led to the expected down-regulation of transcription.

A possible extension to the work in this chapter could be to include the effects of futher topological enzymes, or redesigning the model to incorporate more details of eukaryote genomes. This would involve including the effect of nucleosomes on the supercoiling density of the system, as these can act as barriers to supercoiling. An additional avenue which would be very interesting to explore is to model polymerase movement in 3 dimensions: this would be important as it would allow discrimination between twist and writhe, which is impossible to achieve

within our diffusive 1D model.

As a whole, the work in this thesis demonstrates the power of simple to understand, large-scale models. While the interaction rules of the simulations presented here are often not hugely complex, the results they lead to display considerably more complex behaviours. The fact that these models also compare favourably to experimental results shows the power of physics based methods. This is also despite the levels of coarse-graining applied to the system in order to make it computationally tractable, suggesting general physical principles may have as much influence as more detailed 'biological' interactions in these systems. The projects I have worked on also sought to create models which go beyond being 'toy model' descriptions of a system. While these are often very interesting in their own right, more and more biophysics publications are making efforts to match up with real data and provide results which are more directly applicable to real world problems. It is my hope that the work here can also be said to have carried on this trend!

# Appendix A

# Additional Derivations

## A.1 Angular Potentials

In the following $\theta_{ijk}$ is the angle between the three points $i$,$j$ and $k$; $\mathbf{r_{ij}}$ is a vector from point $i$ to point $j$ and $r_{ij}$ is the scalar length of vector $\mathbf{r_{ij}}$. Operations using a vector are implicitly over the $x$,$y$ and $z$ component, i.e $\frac{\partial}{\partial \mathbf{r}} = \sum_{a=x,y,z} \frac{\partial}{\partial r_a}$.

The cosine potential for the angle interaction in LAMMPS is:

$$E = K[1 + cos(\theta_{ijk})] \tag{A.1}$$

with

$$\theta_{ijk} = cos^{-1}\Big(\frac{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}{r_{ji} r_{jk}}\Big) \tag{A.2}$$

For $\alpha = i, j, k$ the force $\mathbf{F}_\alpha$ is,

$$\mathbf{F}_\alpha = -\frac{\partial E(\theta_{ijk})}{\partial \mathbf{r}_\alpha} \tag{A.3}$$

$$-\frac{\partial E(\theta_{ijk})}{\partial \mathbf{r}_\alpha} = -\frac{\partial E(\theta_{ijk})}{\partial \theta_{ijk}} \frac{\partial \theta_{ijk}}{\partial cos(\theta_{ijk})} \frac{\partial cos(\theta_{ijk})}{\partial \mathbf{r}_\alpha} \tag{A.4}$$

$$-\frac{\partial E(\theta_{ijk})}{\partial \mathbf{r}_\alpha} = K sin(\theta_{ijk}) \cdot \frac{1}{sin\theta_{ijk}} \frac{\partial cos(\theta_{ijk})}{\partial \mathbf{r}_\alpha} \tag{A.5}$$

$$\frac{\partial cos(\theta_{ijk})}{\partial \mathbf{r}_\alpha} = \frac{\partial}{\partial \mathbf{r}_\alpha}\left(\frac{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}{r_{ji} r_{jk}}\right) = (\delta_{\alpha j} - \delta_{\alpha i})\frac{\mathbf{r_{jk}}}{r_{ij} r_{jk}} + (\delta_{\alpha j} - \delta_{\alpha k})\frac{\mathbf{r_{ji}}}{r_{ij} r_{jk}} - cos(\theta_{ijk})\left((\delta_{\alpha j} - \delta_{\alpha i})\frac{\mathbf{r_{ji}}}{r_{ij}^2} + (\delta_{\alpha j} - \delta_{\alpha k})\frac{\mathbf{r_{jk}}}{r_{jk}^2}\right) \tag{A.6}$$

Where equation A.6 uses the results

$$\frac{\partial}{\partial \mathbf{r}_\alpha}\mathbf{r_{ji}} = \frac{\partial}{\partial \mathbf{r}_\alpha}(\mathbf{r_j} - \mathbf{r_i}) = \delta_{\alpha j} - \delta_{\alpha i} \tag{A.7}$$

$$\frac{\partial}{\partial \mathbf{r}_\alpha}\frac{1}{r_{ij}} = \frac{\mathbf{r_{ij}}}{r_{ij}^3} \tag{A.8}$$

Writing out the results for $\alpha = i, j, k$ in equation A.5 gives:

$$\begin{aligned}
\mathbf{F_i} =& \frac{K cos(\theta)}{r_{ij}^2}\mathbf{r_{ji}} - \frac{K}{\mathbf{r_{ij}} \cdot \mathbf{r_{kj}}}\mathbf{r_{jk}} \\
\mathbf{F_j} =& \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{ji}} - \frac{K cos(\theta)}{|\mathbf{r_{jk}}|^2}\mathbf{r_{jk}} + \frac{K}{\mathbf{r_{ji}} \cdot \mathbf{r_{jk}}}\mathbf{r_{jk}} - \frac{K cos(\theta)}{|\mathbf{r_{ji}}|^2}\mathbf{r_{ji}} \\
\mathbf{F_k} =& \frac{K cos(\theta)}{r_{jk}^2}\mathbf{r_{jk}} - \frac{K}{\mathbf{r_{ij}} \cdot \mathbf{r_{jk}}}\mathbf{r_{ji}}
\end{aligned} \tag{A.9}$$

Which are the forces used in the LAMMPS source code and given in section 1.4.

# Appendix B

# Analytical Results For Static Polymerase Model

The work in this appendix comes from the supplementary information submitted with the paper "A stochastic model of supercoiling-dependent transcription" [8].

The material in this appendix was written by Davide Marenduzzo, but is included as the main text references some of the results derived here.

## B.1 Static polymerase models: exact results

In this section we obtain some exact results and scaling relations; we will work within the static polymerase model, but in the next section we will also apply them to the travelling polymerase model.

We begin by considering the static polymerase model, where there is a single gene. We start from the equation for $\sigma(x,t)$, and imagine that the gene is on:

$$\frac{\partial \sigma(x,t)}{\partial t} = \frac{\partial}{\partial x}\left[D\frac{\partial \sigma(x,t)}{\partial x} - J_0\delta(x)\right], \tag{B.1}$$

where we use the boundary condition that $\sigma(0,t) = 0$, and consider no flux boundaries (so that the overall supercoiling is fixed; we solve the equations on an infinite domain, so this implies $\frac{\partial \sigma}{\partial x} = 0$ at $x \to \pm\infty$). In steady state ($\frac{\partial \sigma(x,t)}{\partial t} = 0$),

the solution of Eq. (B.1) is given by

$$\sigma(x) = \frac{J_0}{2D}\mathrm{sgn}(x), \tag{B.2}$$

where $\mathrm{sgn}(x)$ is the sign function, so $\sigma = \frac{J_0}{2D}$ for positive $x$, and $\sigma = -\frac{J_0}{2D}$ for negative $x$. This solution shows that the typical value of the supercoiling density is $|\sigma| \sim J_0/D$ (however it is only accurate for a gene which is always on).

It is also of interest to examine how the solution evolves in time to yield Eq. (B.2) at steady state. To address this, we consider an initial condition with $\sigma \equiv 0$, and we imagine that the gene is switched on at time $t = 0$. Then, while the gene is switched on, the Laplace transform of $\sigma(x,t)$, which we shall call $\hat{\sigma}(x,s)$, with

$$\hat{\sigma}(x,s) \equiv \int_0^\infty dt \exp(-st)\sigma(x,t), \tag{B.3}$$

satisfies the following equation

$$D\frac{\partial^2\hat{\sigma}}{\partial x^2} - s\hat{\sigma} = -J_0\frac{\delta'(x)}{s}, \tag{B.4}$$

where $\delta'(x)$ represents the derivative of the Dirac delta function.

One way to solve Eq. (B.4) is to observe that the Green's function, i.e. the solution of

$$D\frac{\partial^2 g(x,x')}{\partial x^2} - sg(x,x') = \delta(x-x'), \tag{B.5}$$

which decays to 0 at $|x-x'| \to \infty$, is given by

$$g(x,x') = \frac{\exp\left(-\sqrt{\frac{s}{D}}|x-x'|\right)}{2\sqrt{Ds}}. \tag{B.6}$$

Then, the solution of Eq. (B.4) is

$$\begin{aligned}\hat{\sigma}(x,s) &= \int_{-\infty}^{+\infty} dx'\, g(x,x')\left[-J_0\frac{\delta'(x')}{s}\right] \\ &= \frac{J_0}{2Ds}\exp\left(-\sqrt{\frac{s}{D}}|x|\right)\mathrm{sgn}(x).\end{aligned} \tag{B.7}$$

In real space, the solution is found by inverse Laplace transform; at time $t = \tau$, when transcription stops in our model, it is given by

$$\sigma(x,\tau) = \frac{J_0}{2D}\mathrm{erfc}\left(\frac{|x|}{2\sqrt{D\tau}}\right)\mathrm{sgn}(x), \tag{B.8}$$

where erfc is the complement of the error function. This solution tends to Eq. (B.2) when $\tau \to \infty$; it also shows that, while the gene is on, again the typical value of supercoiling density *in the neighbourhood of the promoter* is $\sim J_0/D$.

After the gene is switched off the supercoiling density satisfies the diffusion equation,

$$\frac{\partial \sigma(x,t)}{\partial t} = D\frac{\partial^2}{\partial x^2}\sigma(x,t), \tag{B.9}$$

with the initial condition that $\sigma(x,\tau)$ is as given by Eq. (B.8). The solution can be written as

$$\sigma(x,t) = \int_{-\infty}^{\infty} dx' \, \frac{\exp\left[-\frac{(x-x')^2}{4Dt}\right]}{\sqrt{4\pi Dt}} \frac{J_0}{2D} \mathrm{erfc}\left(\frac{|x'|}{2\sqrt{D\tau}}\right) \mathrm{sgn}(x'), \tag{B.10}$$

where for simplicity we have shifted time so that the gene switches off at time $t = 0$ and the solution holds for $t \geq 0$. Eq. (B.10) can be used to infer that $\sigma(0,t) \equiv 0$ (for the static polymerase model), and $\sigma(x,t) \sim t^{-3/2}$ for large $t$ and for $x \neq 0$.

## B.2 Static and travelling polymerase models: mean field theory, and scaling

We now use the results obtained from the last section to build a simple mean field theory for our model.

We start from the observation that, within the static polymerase model, the on rate for RNA polymerase, $k_{\mathrm{on}}$, depends on the extent of negative supercoiling upstream of the promoter (at $x_0 < 0$), according to the formula (see main text),

$$k_{\mathrm{on}} = k_0 \left[1 - \alpha\sigma(x_0,t)\right], \tag{B.11}$$

where, since this is always positive, we do not need the max function as in the main text.

We propose a simple mean field theory, where the value of $\sigma(x_0,t)$ is replaced with the average supercoiling profile over the whole simulation, $\bar{\sigma}(x_0)$. An equation for $\bar{\sigma}$ can be written down by finding the steady state solution of Eq. (B.1) when the flux is replaced by its average $J_0\delta(x)\frac{k_{\mathrm{on}}\tau}{k_{\mathrm{on}}\tau+1}$, where $\frac{k_{\mathrm{on}}\tau}{k_{\mathrm{on}}\tau+1}$ is the fraction of

time that the gene is on (this last formula can be obtained by realising that the polymerase has an on rate equal to $k_{\rm on}$ and an effective off rate equal to $1/\tau$). If we do this, we find that

$$\bar{\sigma}(x_0) = -\frac{k_{\rm on}\tau}{k_{\rm on}\tau + 1}\frac{J_0}{2D}. \tag{B.12}$$

We should note that this solution, as the previous ones, works for open, no flux, boundary conditions (our simulations instead have periodic boundary conditions, but the scaling of $\bar{\sigma}$ does not change).

We can now plug in this expression for $\bar{\sigma}$ in Eq. (B.11), to get a self-consistent equation, similar in spirit to a mean field theory,

$$k_{\rm on} = k_0\left[1 - \alpha\bar{\sigma}(k_{\rm on})\right] \sim k_0\left[1 + \alpha\frac{k_{\rm on}\tau}{k_{\rm on}\tau + 1}\frac{J_0}{2D}\right]. \tag{B.13}$$

Eq. (B.13) has a solution which depends smoothly on $\frac{J_0}{D}$: in other words, there should be no discontinuity in the transcription rate (proportional to $k_{\rm on}$, see below) as a function of $J_0$. Another way to understand this is to realise that Eq. (B.13) is essentially equivalent to the mean field equation for the magnetisation versus temperature in the Ising model in the presence of a non-zero magnetic field (the $k_0$ term): it is well known that this equation in this case describes a smooth crossover and no thermodynamic phase transition.

While we have derived our mean field equation, Eqs.(B.12) and (B.13) for the static polymerase model, numerically we found that Eq. (B.12) also applies well for the travelling polymerase model, with $J_0$ replaced by $\bar{J}$, the average supercoiling flux during transcription. Specifically, for the travelling polymerase model, the average supercoiling density at the promoter, which we call $\bar{\sigma}_{\rm p}$, is given by

$$\bar{\sigma}_{\rm p} = -\frac{k_{\rm on}\tau}{k_{\rm on}\tau + 1}\frac{\bar{J}}{2D} = -\frac{\Phi}{\Phi + 1}\frac{\bar{J}}{2D}, \tag{B.14}$$

where $\Phi = k_{\rm on}\tau$ is one of the dimensionless numbers introduced in the main text, for $N = n = 1$. Eq. (B.14) is used in the main text to estimate the supercoiling densities at promoters in bacteria, yeast and human cells.

By plugging Eq. (B.14) into Eq. (B.11), we can find an explicit expression for $k_{\rm on}$

in our mean field theory, which is given by

$$\begin{aligned} k_{\rm on}\tau &= \frac{h+\sqrt{h^2+4k_0\tau}}{2} \\ h &= k_0\tau\left(1+\frac{\alpha\bar{J}}{2D}\right)-1. \end{aligned} \tag{B.15}$$

The overall transcription rate $k_{\rm t}$ (of the single gene considered up to now in the simplified theory) can be estimated as follows,

$$k_{\rm t} = \frac{k_{\rm on}}{1+k_{\rm on}\tau} \tag{B.16}$$

where the correction $\frac{1}{1+k_{\rm on}\tau}$ takes into account the fact that the maximum transcription yield per gene is equal to $1/\tau$, when the polymerase is transcribing the gene at all times. Figure B.1 shows some examples of the overall transcription rate $k_{\rm t}$, for different values of $k_0\tau$. As anticipated when analysing the static polymerase model, for any $k_0 \neq 0$, there is no discontinuity in the transcription rate, so that the switch between uniform and supercoiling-regulated regime is a crossover. The only limit in which this would become a true nonequilibrium transition is if $k_0 \to 0$, while keeping the product $\bar{J}\alpha k_0\tau/D$ constant. Eqs. (B.15) and (B.16) also highlight a useful criterion to determine when supercoiling starts to significantly affect transcriptional rate (hence transcription): this occurs when

$$\frac{\bar{J}\alpha k_0\tau}{2D} \sim 1. \tag{B.17}$$

In other words, the value of $\bar{J}/D$ (which is the parameter varied in the main text) at which we should expect the crossover between the uniform and the supercoiling-dominated regime is equal to $2/(\alpha k_0\tau)$.

Note that, as is the case in general for mean field approximations, the assumption that $k_{\rm on}$ depends on the *average* supercoiling profile, $\bar{\sigma}$, is only appropriate when the supercoiling profile does not vary too much in time, so that the instantaneous profile for $\sigma$ is close to the average one. This is the case when there is not enough time for the supercoiling to diffuse away in between transcription events. The physical dimensionless parameter determining when this is the case, in the travelling polymerase model, is $\Theta = \frac{k_{\rm on}\lambda^2}{D}$. If $\Theta$ is small, then diffusion is fast and while the gene is off the supercoiling is much smaller than the average value, and our mean field theory is not valid.

Fortunately, even when $\Theta$ is relatively small (Figure B.1, where the minimum

value of $\Theta$ is $\sim$0.44) our numerical results suggest that the value of $\sigma$ at the promoter, $\sigma_{\rm p}$, at the moment when the gene is switched on (which is the relevant value to use in Eq. (B.11)), depends on $k_{\rm on}$ linearly for small $k_{\rm on}$, so that the same qualitative considerations apply as in our simplified mean field theory (i.e., the system displays a crossover rather than a phase transition as $\bar{J}/D$ is increased). We can further perform a simulation to find the value of $\sigma_{\rm p}$ as a function of $k_{\rm on}$ (kept constant for each simulation, see figure B.1 and its caption). We can then fit the resulting data with the following functional form,

$$|\sigma_{\rm p}| = \frac{a k_{\rm on}}{b k_{\rm on} + 1} \tag{B.18}$$

where $a$ and $b$ are positive constants determined via fitting (Figure B.1). At this point, we can follow the procedure described above, where Eq. (B.18) is plugged into Eq. (B.11) to yield a semianalytical estimate for $k_{\rm on}$: this is an improvement with respect to the mean field estimate, Eq. (B.15). In a system with one polymerase and one gene, the rate $k_{\rm on}$ determined self-consistently via Eq. (B.11) gives the overall transcription rate $k_{\rm t}$ by using Eq. (B.16). For a system with $N$ polymerases and $n$ genes, substituting $k_{\rm on}$ with $k_{\rm on}N/n$ we obtain the predicted transcription rate per gene. This rate is a good approximation of the transcription rate per gene in the case of genes oriented along the same direction.

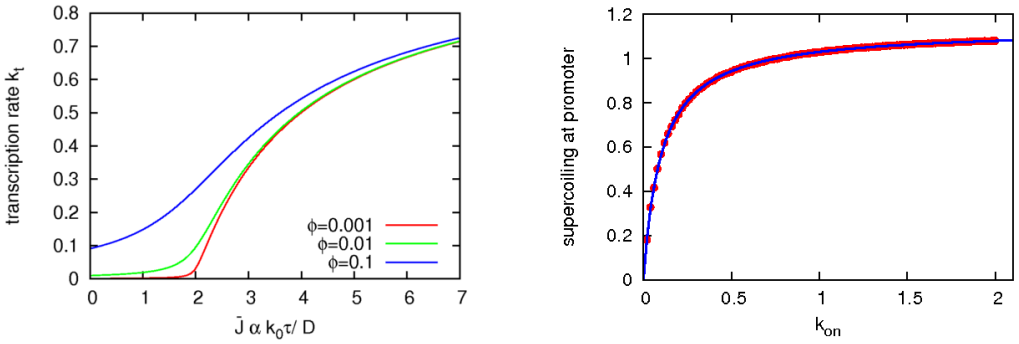



Figure B.1: Left: Plot of the transcription rate, found by using Eq. (B.15) and Eq. (B.16), for $\alpha = 100$ (as in the main text), and different values of $k_{\rm on}\tau$ (see legend).
Right: (A) Plot of the local supercoiling density (absolute value) at the promoter as a function of $k_{\rm on}$ for a single gene, on a lattice of size $1000\,\Delta x$ (with periodic boundary condition). To make this plot we run our simulations with $\alpha = 0$ so that $k_{\rm on}$ can be fixed as an input. The fit is to Eq. (B.18), and the resulting parameters are $a \sim 11.18 \pm 0.02$ and $b = 9.85 \pm 0.02$. This simulation was performed with $\bar{J}/D = 2.55$; in order to get the transcription rate as a function of $\bar{J}$ we further assumed a linear dependence of $\sigma_{\rm p}$ on $\bar{J}$ overall (as in Eq. B.12).

# List of Figures

# Bibliography


[1] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell - 5th Edition*. Garland Science, 2008.

[2] Aumann, F., F. Lankas, M. Caudron, and J. Langowski. "Monte Carlo simulation of chromatin stretching." *Phys. Rev. E* 73: (2006) 041,927. https://link.aps.org/doi/10.1103/PhysRevE.73.041927.

[3] Barbieri, M., M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi. "Complexity of chromatin folding is captured by the strings and binders switch model." *Proceedings of the National Academy of Sciences* 109, 40: (2012) 16,173–16,178. http://www.pnas.org/content/109/40/16173.abstract.

[4] Barutcu, A. R., A. J. Fritz, S. K. Zaidi, A. J. van Wijnen, J. B. Lian, J. L. Stein, J. A. Nickerson, A. N. Imbalzano, and G. S. Stein. "C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization." *Journal of Cellular Physiology* 231, 1: (2016) 31–35. http://dx.doi.org/10.1002/jcp.25062.

[5] van Berkum, N. L., E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander. "Hi-C: A Method to Study the Three-dimensional Architecture of Genomes." *J Vis Exp* 1869. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149993/. 1869[PII],20461051[pmid].

[6] Biggin, M. D. "Animal Transcription Networks as Highly Connected, Quantitative Continua." *Developmental Cell* 21, 4: (2011) 611–626. http://dx.doi.org/10.1016/j.devcel.2011.09.008.

[7] Boles, C., J. H. White, and N. R. Cozzarelli. "Structure of plectonemically supercoiled DNA." *Journal of molecular biology* 213: (1990) 931–51.

[8] Brackley, C. A., J. Johnson, A. Bentivoglio, S. Corless, N. Gilbert, G. Gonnella, and D. Marenduzzo. "Stochastic Model of Supercoiling-Dependent Transcription." *Phys. Rev. Lett.* 117: (2016) 018,101. https://link.aps.org/doi/10.1103/PhysRevLett.117.018101.

[9] Brackley, C. A., D. Michieletto, F. Mouvet, J. Johnson, S. Kelly, P. R. Cook, and D. Marenduzzo. "Simulating topological domains in human chromosomes with a fitting-free model." *Nucleus* 7, 5: (2016) 453–461. http://dx.doi.org/10.1080/19491034.2016.1239684. PMID: 27841970.

[10] Brackley, C. A., A. N. Morozov, and D. Marenduzzo. "Models for twistable elastic polymers in Brownian dynamics, and their implementation for LAMMPS." *The Journal of chemical physics* 140 13: (2014) 135,103.

[11] Brackley, C. A., J. Johnson, S. Kelly, P. R. Cook, and D. Marenduzzo. "Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains." *Nucleic Acids Res* 44, 8: (2016) 3503–3512. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856988/. 27060145[pmid].

[12] Brackley, C. A., S. Taylor, A. Papantonis, P. R. Cook, and D. Marenduzzo. "Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization." *Proceedings of the National Academy of Sciences* 110, 38: (2013) E3605–E3611. http://www.pnas.org/content/110/38/E3605.abstract.

[13] Brinkers, S., H. R. C. Dietrich, F. H. de Groote, I. T. Young, and B. Rieger. "The persistence length of double stranded DNA determined using dark field tethered particle motion." *The Journal of Chemical Physics* 130, 21: (2009) 215,105. https://doi.org/10.1063/1.3142699.

[14] Calladine, C., and H. Drew. *Understanding DNA*. Academic Press, 1992.

[15] Carey, N.

[16] Cavalli, G., and T. Misteli. "Functional implications of genome topology." *Nat Struct Mol Biol* 20, 3: (2013) 290–299. http://dx.doi.org/10.1038/nsmb.2474.

[17] Chaikin, P. M., and T. C. Lubensky. *Principles of Condensed Matter Physics*. Cambridge University Press, 1995.

[18] Chong, S., C. Chen, H. Ge, and X. S. Xie. "Mechanism of Transcriptional Bursting in Bacteria." *Cell* 158, 2: (2014) 314 – 326. http://www.sciencedirect.com/science/article/pii/S0092867414007399.

[19] Chubb, J. R., T. Trcek, S. M. Shenoy, and R. H. Singer. "Transcriptional Pulsing of a Developmental Gene." *Current Biology* 16, 10: (2006) 1018 – 1025. http://www.sciencedirect.com/science/article/pii/S0960982206014266.

[20] Ciabrelli, F., and G. Cavalli. "Chromatin-Driven Behavior of Topologically Associating Domains." *Journal of Molecular Biology* 427, 3: (2015) 608 – 625. http://www.sciencedirect.com/science/article/pii/S0022283614005129. Functional Relevance and Dynamics of Nuclear Organization.

[21] Cournac, A., and J. Plumbridge. "DNA Looping in Prokaryotes: Experimental and Theoretical Approaches." *Journal of Bacteriology* 195, 6: (2013) 1109–1119. http://jb.asm.org/content/195/6/1109.abstract.

[22] Cover, T. M., and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[23] Cox, M., J. Doudna, and M. OĎonnell. *Molecular Biology: Principles and Practice*. W. H. Freeman, 2011.

[24] Dame, R. T., M. C. Noom, and G. J. L. Wuite. "Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation." *Nature* 444, 7117: (2006) 387–390. http://dx.doi.org/10.1038/nature05283.

[25] Dekker, J. "GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p." *Genome Biology* 8, 6: (2007) R116. http://dx.doi.org/10.1186/gb-2007-8-6-r116.

[26] Dekker, J., K. Rippe, M. Dekker, and N. Kleckner. "Capturing Chromosome Conformation." *Science* 295, 5558: (2002) 1306–1311. http://science.sciencemag.org/content/295/5558/1306.

[27] Dekker, N. H., V. V. Rybenkov, M. Duguet, N. J. Crisona, N. R. Cozzarelli, D. Bensimon, and V. Croquette. "The mechanism of type IA topoisomerases." *Proceedings of the National Academy of Sciences* 99, 19: (2002) 12,126–12,131. http://www.pnas.org/content/99/19/12126.abstract.

[28] Di Stefano, M., A. Rosa, V. Belcastro, D. di Bernardo, and C. Micheletti. "Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19." *PLOS Computational Biology* 9, 3: (2013) 1–13. https://doi.org/10.1371/journal.pcbi.1003019.

[29] Dixon, J. R., 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, 7398: (2012) 376–380. http://dx.doi.org/10.1038/nature11082.

[30] Doi, M., and S. F. Edwards. *The Theory of Polymer Dynamics*. Oxford University Press, USA, 1986. http://www.worldcat.org/isbn/0198520336.

[31] Dorman, C. J. "H-NS: a universal regulator for a dynamic genome." *Nat Rev Micro* 2, 5: (2004) 391–400. http://dx.doi.org/10.1038/nrmicro883.

[32] Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements." *Genome Res* 16, 10: (2006) 1299–1309. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1581439/. 16954542[pmid].

[33] Doye, J. P. K., T. E. Ouldridge, A. A. Louis, F. Romano, P. Sulc, C. Matek, B. E. K. Snodin, L. Rovigatti, J. S. Schreck, R. M. Harrison, and W. P. J. Smith. "Coarse-graining DNA for simulations of DNA nanotechnology." *Phys. Chem. Chem. Phys.* 15: (2013) 20,395–20,414. http://dx.doi.org/10.1039/C3CP53545B.

[34] Dunaway, M., and E. A. Ostrander. "Local domains of supercoiling activate a eukaryotic promoter in vivo." *Nature* 361, 6414: (1993) 746–748. http://dx.doi.org/10.1038/361746a0.

[35] Duplantier, B. "Statistical mechanics of polymer networks of any topology." *Journal of Statistical Physics* 54, 3: (1989) 581–680. https://doi.org/10.1007/BF01019770.

[36] Ernst, J. "Mapping and analysis of chromatin state dynamics in nine human cell types." *Nature* 473: (2011) 43–49. http://www.nature.com/nature/journal/v473/n7345/full/nature09906.html.

[37] Ernst, J., and M. Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." *Nature Biotechnology* 28: (2010) 817–825. http://www.nature.com/nbt/journal/v28/n8/full/nbt.1662.html.

[38] Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny. "Formation of Chromosomal Domains by Loop Extrusion." *Cell Rep* 15, 9: (2016) 2038–2049. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4889513/. 27210764[pmid].

[39] Gartenberg, M. R., and J. C. Wang. "Positive supercoiling of DNA greatly diminishes mRNA synthesis in yeast." *Proceedings of the National Academy of Sciences* 89, 23: (1992) 11,461–11,465. http://www.pnas.org/content/89/23/11461.abstract.

[40] Gilbert, N., and J. Allan. "Supercoiling in {DNA} and chromatin." *Current Opinion in Genetics & Development* 25: (2014) 15 – 21. http://www.sciencedirect.com/science/article/pii/S0959437X13001500. Genome architecture and expression.

[41] Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing." *Nat Biotech* 27, 2: (2009) 182–189. http://dx.doi.org/10.1038/nbt.1523.

[42] Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox. "Real-Time Kinetics of Gene Activity in Individual Bacteria." *Cell* 123, 6: (2005) 1025 – 1036. http://www.sciencedirect.com/science/article/pii/S0092867405010378.

[43] Guo, Y., Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, and Q. Wu. "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function." *Cell* 162, 4: (2015) 900–910. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4642453/. 26276636[pmid].

[44] Hagerman, P. J. "Flexibility of DNA." *Annual Review of Biophysics and Biophysical Chemistry* 17, 1: (1988) 265–286. https://doi.org/10.1146/annurev.bb.17.060188.001405. PMID: 3293588.

[45] Hatfield, G. W., and C. J. Benham. "DNA Topology-Mediated Control of Global Gene Expression in Escherichia coli." *Annual Review of Genetics* 36, 1: (2002) 175–203. https://doi.org/10.1146/annurev.genet.36.032902.111815. PMID: 12429691.

[46] Hohenberg, P. C., and B. I. Halperin. "Theory of dynamic critical phenomena." *Rev. Mod. Phys.* 49: (1977) 435–479. https://link.aps.org/doi/10.1103/RevModPhys.49.435.

[47] Johnson, J., C. A. Brackley, P. R. Cook, and D. Marenduzzo. "A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction." *Journal of Physics: Condensed Matter* 27, 6: (2015) 064,119. http://stacks.iop.org/0953-8984/27/i=6/a=064119.

[48] Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. "The UCSC Table Browser data retrieval tool." *Nucleic Acids Research* 32: (2004) D493–D496. http://dx.doi.org/10.1093/nar/gkh103.

[49] Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, and A. M. Zahler. "The Human Genome Browser at UCSC." *Genome Research* 12, 6: (2002) 996–1006. http://genome.cshlp.org/content/12/6/996.abstract.

[50] Konrad, M. W., and J. I. Bolonick. "Molecular Dynamics Simulation of DNA Stretching Is Consistent with the Tension Observed for Extension and Strand Separation and Predicts a Novel Ladder Structure." *Journal of the American Chemical Society* 118, 45: (1996) 10,989–10,994. http://dx.doi.org/10.1021/ja961751x.

[51] Kubo, R. "The fluctuation-dissipation theorem." *Reports on Progress in Physics* 29, 1: (1966) 255. http://stacks.iop.org/0034-4885/29/i=1/a=306.

[52] Kunze, K.-K., and R. R. Netz. "Salt-Induced DNA-Histone Complexation." *Phys. Rev. Lett.* 85: (2000) 4389–4392. http://link.aps.org/doi/10.1103/PhysRevLett.85.4389.

[53] Kühner, S. "Proteome organization in a genome-reduced bacterium.", . http://bionumbers.hms.harvard.edu//bionumber.aspx?id=107820&ver=0.

[54] Laughton, C. A., and S. A. Harris. "The atomistic simulation of DNA." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1, 4: (2011) 590–600. http://dx.doi.org/10.1002/wcms.46.

[55] Le, T. B. K., M. V. Imakaev, L. A. Mirny, and M. T. Laub. "High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome." *Science* 342, 6159: (2013) 731–734. http://science.sciencemag.org/content/342/6159/731.

[56] Lebrun, A., and R. Lavery. "Modelling Extreme Stretching of DNA." *Nucleic Acids Research* 24, 12: (1996) 2260–2267. http://dx.doi.org/10.1093/nar/24.12.2260.

[57] Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326, 5950: (2009) 289–293. http://science.sciencemag.org/content/326/5950/289.

[58] Liu, L. F., and J. C. Wang. "Supercoiling of the DNA template during transcription." *Proceedings of the National Academy of Sciences* 84, 20: (1987) 7024–7027. http://www.pnas.org/content/84/20/7024.abstract.

[59] van Loenhout, M. T. J., M. V. de Grunt, and C. Dekker. "Dynamics of DNA Supercoils." *Science* 338, 6103: (2012) 94–97. http://science.sciencemag.org/content/338/6103/94.

[60] Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. "Crystal structure of the nucleosomecore particle at 2.8 A resolution." *Nature* 389. http://www.nature.com/nature/journal/v389/n6648/full/389251a0.html.

[61] Ma, K., . https://commons.wikimedia.org/w/index.php?curid=22965076.

[62] Maeshima, K., R. Imai, S. Tamura, and T. Nozaki. "Chromatin as dynamic 10-nm fibers." *Chromasoma* .

[63] Marenduzzo, D. "Course Notes.", . http://www2.ph.ed.ac.uk/~dmarendu/GeometryAndPhysicsSoftCM/.

[64] Marenduzzo, D., and E. Orlandini. "Topological and entropic repulsion in biopolymers." *Journal of Statistical Mechanics: Theory and Experiment* 2009, 09: (2009) L09,002. http://stacks.iop.org/1742-5468/2009/i=09/a=L09002.

[65] Marko, J., and E. Siggia. "Statistical mechanics of supercoiled DNA." *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 52, 3: (1995) 2912–2938.

[66] Micheletti, C., and N. M. Toan. "Inferring the effective thickness of polyelectrolytes from stretching measurements at various ionic strengths: applications to DNA and RNA." *Journal of Physics: Condensed Matter* 18. http://iopscience.iop.org/article/10.1088/0953-8984/18/14/S11.

[67] Miele, A., and J. Dekker. "Mapping cis- and trans- chromatin interaction networks using Chromosome Conformation Capture (3C)." *Methods Mol Biol* 464: (2009) 10.1007/978–1–60,327–461–6_7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874836/. 18951182[pmid].

[68] Milo, R., and R. Phillips. "Bionumbers e-book.", . http://book.bionumbers.org/how-many-proteins-are-in-a-cell/.

[69] ———. "How big is the "average" protein?", . http://book.bionumbers.org/how-big-is-the-average-protein/.

[70] Mirny, L. A. "The fractal globule as a model of chromatin architecture in the cell." *Chromosome Research* 19, 1: (2011) 37–51. https://doi.org/10.1007/s10577-010-9177-0.

[71] Misteli, T. "Chromosome territories: The arrangement of chromosomes in the nucleus." *Nature Education* 1, 1.

[72] Mizutani, M., T. Ohta, H. Watanabe, H. Handa, and S. Hirose. "Negative supercoiling of DNA facilitates an interaction between transcription factor IID and the fibroin gene promoter." *Proceedings of the National Academy of Sciences* 88, 3: (1991) 718–722. http://www.pnas.org/content/88/3/718.abstract.

[73] Mylonas, E., and D. I. Svergun. "Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering." *Journal of Applied Crystallography* 40, s1: (2007) s245–s249. https://doi.org/10.1107/S002188980700252X.

[74] Naughton, C., S. Corless, and N. Gilbert. "Divergent RNA transcription." *Transcription* 4, 4: (2013) 162–166. http://dx.doi.org/10.4161/trns.25554.

[75] Naumova, N., M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker. "Organization of the Mitotic Chromosome." *Science* 342, 6161: (2013) 948–953. http://science.sciencemag.org/content/342/6161/948.

[76] Neidhardt, F., and R. Curtiss. *Escherichia Coli and Salmonella : Cellular and Molecular Biology*. American Society of Microbiology, 1996. http://bionumbers.hms.harvard.edu/bionumber.aspx?id=104954&ver=16.

[77] Nicodemi, M., and A. Prisco. "Thermodynamic Pathways to Genome Spatial Organization in the Cell Nucleus." *Biophysical Journal* 96, 6: (2009) 2168–2177. http://dx.doi.org/10.1016/j.bpj.2008.12.3919.

[78] Nishino, Y., M. Eltsov, Y. Joti, K. Ito, H. Takata, Y. Takahashi, S. Hihara, A. S. Frangakis, N. Imamoto, T. Ishikawa, and K. Maeshima. "Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure." *The EMBO Journal* 31, 7: (2012) 1644–1653. http://emboj.embopress.org/content/31/7/1644.

[79] Ou, H. D., S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O'Shea. "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells." *Science* 357, 6349. http://science.sciencemag.org/content/357/6349/eaag0025.

[80] Papantonis, A., and P. R. Cook. "Transcription Factories: Genome Organization and Gene Regulation." *Chemical Reviews* 113, 11: (2013) 8683–8705. http://dx.doi.org/10.1021/cr300513p. PMID: 23597155.

[81] Pluth, J., and D. Sridharan. "Chromosome Scaffold." In *Brenner's Encyclopedia of Genetics (Second Edition)*, edited by Stanley Maloy, and Kelly Hughes, San Diego: Academic Press, 2013, 576 – 578. Second edition edition. https://www.sciencedirect.com/science/article/pii/B9780123749840002473.

[82] Project, E. "ENCODE: Encyclopedia of DNA Elements.", . https://www.encodeproject.org/.

[83] Raap, A. K. "Advances in fluorescence in situ hybridization." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 400, 1: (1998) 287 – 298. http://www.sciencedirect.com/science/article/pii/S0027510798000293.

[84] Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E.-L. Aiden. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159, 7: (2014) 1665–1680. http://dx.doi.org/10.1016/j.cell.2014.11.021.

[85] Reece, R. J., and A. Maxwell. "DNA Gyrase: Structure and Function." *Critical Reviews in Biochemistry and Molecular Biology* 26, 3-4: (1991) 335–375. http://dx.doi.org/10.3109/10409239109114072. PMID: 1657531.

[86] Rhee, K. Y., M. Opel, E. Ito, S.-p. Hung, S. M. Arfin, and G. W. Hatfield. "Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli." *Proceedings of the National Academy of Sciences* 96, 25: (1999) 14,294–14,299. http://www.pnas.org/content/96/25/14294.abstract.

[87] Ringrose, L., S. Chabanis, P.-O. Angrand, C. Woodroofe, and A. Stewart. "Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances." *The EMBO Journal* 18, 23: (1999) 6630–6641. http://dx.doi.org/10.1093/emboj/18.23.6630.

[88] Rosa, A., and R. Everaers. "Structure and Dynamics of Interphase Chromosomes." *PLOS Computational Biology* 4, 8: (2008) 1–10. https://doi.org/10.1371/journal.pcbi.1000153.

[89] Rousseau, M., J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette. "Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling." *BMC Bioinformatics* 12, 1: (2011) 414. http://dx.doi.org/10.1186/1471-2105-12-414.

[90] Schiessel, H. "The physics of chromatin." *Journal of Physics: Condensed Matter* 15, 19: (2003) R699. http://stacks.iop.org/0953-8984/15/i=19/a=203.

[91] Schiessel, H., W. M. Gelbart, and R. Bruinsma. "DNA Folding: Structural and Mechanical Properties of the Two-Angle Model for Chromatin." *Biophysical Journal* 80: (2001) 1940–1956.

[92] Terekhova, K., K. H. Gunn, J. F. Marko, and A. Mondragón. "Bacterial topoisomerase I and topoisomerase III relax supercoiled DNA via distinct pathways." *Nucleic Acids Research* 40, 20: (2012) 10,432–10,440. http://dx.doi.org/10.1093/nar/gks780.

[93] Thoma, F., T. Koller, and A. Klug. "Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin." *The Journal of Cell Biology* 83, 2: (1979) 403–427. http://jcb.rupress.org/content/83/2/403.

[94] Tiwari, V. K., and S. B. Baylin. "Combined 3C-ChIP-Cloning (6C) Assay: A Tool to Unravel Protein-Mediated Genome Architecture." *Cold Spring Harb Protoc* 2009, 3: (2009) pdb.prot5168–pdb.prot5168. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3586525/. 20147103[pmid].

[95] Valle, F., M. Favre, P. De Los Rios, A. Rosa, and G. Dietler. "Scaling Exponents and Probability Distributions of DNA End-to-End Distance." *Phys. Rev. Lett.* 95: (2005) 158,105. http://link.aps.org/doi/10.1103/PhysRevLett.95.158105.

[96] Various. "Encode Project Website.", . https://www.genome.gov/26524238/.

[97] Vinograd, J., J. Lebowitz, R. Radloff, R. Watson, and P. Laipis. "The twisted circular form of polyoma viral DNA." *Proc Natl Acad Sci U S A*

53, 5: (1965) 1104–1111. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC301380/. 4287964[pmid].

[98] Vologodskii, A. V., , and N. R. Cozzarelli. "Conformational and Thermodynamic Properties of Supercoiled DNA." *Annual Review of Biophysics and Biomolecular Structure* 23, 1: (1994) 609–643. https://doi.org/10.1146/annurev.bb.23.060194.003141. PMID: 7919794.

[99] Wang, G.-Z., M. J. Lercher, and L. D. Hurst. "Transcriptional Coupling of Neighboring Genes and Gene Expression Noise: Evidence that Gene Orientation and Noncoding Transcripts Are Modulators of Noise." *Genome Biology and Evolution* 3: (2011) 320–331. +http://dx.doi.org/10.1093/gbe/evr025.

[100] Wang, W., G.-W. Li, C. Chen, X. S. Xie, and X. Zhuang. "Chromosome Organization by a Nucleoid-Associated Protein in Live Bacteria." *Science* 333, 6048: (2011) 1445–1449. http://science.sciencemag.org/content/333/6048/1445.

[101] Wani, A. H., A. N. Boettiger, P. Schorderet, A. Ergun, C. Münger, R. I. Sadreyev, X. Zhuang, R. E. Kingston, and N. J. Francis. "Chromatin topology is coupled to Polycomb group protein subnuclear organization." *Nature Comms.* 7: (2016) 10,291. http://dx.doi.org/10.1038/ncomms10291. Article.

[102] Watson, J. *Molecular Biology of the Gene, 6th Edition.* Pearson, 2007.

[103] van de Werken, H., P. de Vree, E. Splinter, S. Holwerda, P. Klous, E. de Wit, and W. de Laat. "4C technology: protocols and data analysis." *Methods Enzymol.* .

[104] Zeng, P.-Y., C. R. Vakoc, Z.-C. Chen, G. A. Blobel, and S. L. Berger. "In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation." *BioTechniques* 41, 6: (2006) 694–698.

[105] Zinchenko, A. A., T. Sakaue, S. Araki, K. Yoshikawa, and D. Baigl. "Single-chain compaction of long duplex DNA by cationic nanoparticles: modes of interaction and comparison with chromatin." *The Journal of Physical Chemistry B* 111, 11: (2007) 3019–3031.

[106] Zinchenko, A. A., K. Yoshikawa, and D. Baigl. "Compaction of Single-Chain DNA by Histone-Inspired Nanoparticles." *Phys. Rev. Lett.* 95: (2005) 228,101. https://link.aps.org/doi/10.1103/PhysRevLett.95.228101.