# TOPOLOGY OF CONTACTS IN INTERACTING POLYMERS

**By**

**NINA DIRSCHBACHER**

**A thesis submitted to the**
**Graduate School—New Brunswick**
**Rutgers, The State University of New Jersey**
**in partial fulfillment of the requirements**
**for the degree of**
**Master of Science**
**Graduate Program in Physics and Astronomy**
**written under the direction of**
**Prof. Anirvan M. Sengupta**
**and approved by**

_____

_____

_____

**New Brunswick, New Jersey**

**October, 2014**

## ABSTRACT OF THE THESIS

# Topology of contacts in interacting polymers

**By NINA DIRSCHBACHER**

**Thesis Director:**

**Prof. Anirvan M. Sengupta**

Recent developments in the study of chromatin loop domains and gene regulation motivates us to investigate the topology of contacts in interacting polymers. To describe the chromatin loop formation we will use a bead on a string model with excluded volume and attractive polymer-polymer interaction simulated by self avoiding walks. The resulting chain configuration is a balance between these two interactions and can be a coiled state for good solvents or a collapsed one for poor solvent. We can classify the different chain configurations by their topological structures through the genus, a topological invariant of the resulting diagrams. The backtracking algorithm enables us to achieve exact numerical results for the genus distribution of the above mentioned model for small polymer lengths. Since pseudoknots decrease the entropy by restricting the possible configuration, we explore whether we see a suppression of pseudoknots for finite size polymers and intermediate strength of interaction.

# Table of Contents

# List of Figures

# Chapter 1
# Motivation

The aim of this chapter is to emphasize why it is important to examine the specific configuration of polymer chains in biological systems. To do so, I will explain the basic structure of the two biopolymers we are interested in: DNA and RNA. Why the spatial structure is important for gene regulation will be explained in the second part of this chapter.

## 1.1 Biopolymers

DNA and RNA are both polymer chains that fulfill a special folding process in biological systems. In this section, I will show which components are part of this process and how interesting interaction structures can occur.

### 1.1.1 RNA and DNA

RNA has a single string structure composed of the four nucleotides adenine (A), guanine (G), uracil (U) and cytosine (C). In solutions there is an attraction between C and G and between A and U. There is weaker attraction between G and U as well. The original RNA strand has no bonds and thus the formation of the RNA depends on the interaction between the bases and can fold into a lot of different spacial structures. It can lead to loops and inter- or intra-loop interactions which results in *pseudoknot* conformation. Pseudoknots emerge due to crossing interaction chords and are not real knots in space. For illustration, the so-called *kissing hairpin* is shown in Fig. 1.1. By pairing with itself RNA forms a double helix structure with unpaired parts in between. These fragments pair up with other unpaired segments, which is indicated by the red dashed lines in Fig. 1.1. This specific way of interacting generates pseudoknots. I will elaborate on the problem of pseudoknots in chapter 4. They occur mainly in functionally important regions of RNA.

Compared to RNA, DNA has a double helix structure with two backbones that are called polynucleotides, since they consist of different nucleotides. Each nucleotide is composed of one of the following bases: Guanine (G), adenine (A), thymine (T) or cytosine (C). Always two of these DNA bases are compatible, A with T and C with G, to form base pairs. The DNA has a overall negative

charge and is relatively stiff. Nevertheless, a folding process similar to that in RNA occurs, but for different reasons. Here it is not the interaction between bases that lead to a compact structure, since all of them are already in a bonded state. Through interaction with different proteins, it will wrap up to a dense configuration called chromatin. I will explain the DNA folding process in the next section.

Both, RNA and DNA are mainly in their collapsed structured involved in the gene regulation process. Yet the focus in the thesis will be on examining the DNA, i.e. chromatin structure.

### 1.1.2 Chromatin

Chromatin results as a dense packed complex of DNA. The packing process is illustrated in Fig. 1.2. Besides the DNA strand there is another component that plays an important role in the folding: *histones*, positively charged proteins. There are six different core histones: H1, H2A, H2B, H3, H4 and H5. Eight of those (H2A, H2B, H3, and H4, each taken two times), the so called histone octamer, build a globular unity around which the DNA is wrapped. The tight loop of DNA and histones is called a nucleosome. The remaining two histones (H1 and H5) are linker histones. They will tight the nucleosomes together and the resulting structure is a fiber of packed nucleosomes known as chromatin. With the help of other proteins this fiber will be looped and further packaged to the familiar chromosome structure. Stretched human DNA has a length of about two meters and this length is packed up to fit into nucleus of every cell, whose characteristic size is a micron. But more astonishing than that is that this compact complex still provides access for gene expression and the topological structure actually plays a main part in the regulation process.

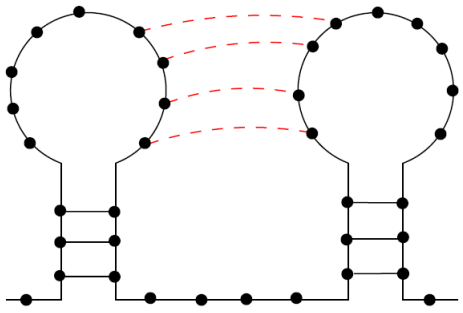


Figure 1.1: The 'kissing hairpin' pseudoknot configuration for a RNA strand (Image from [1]).

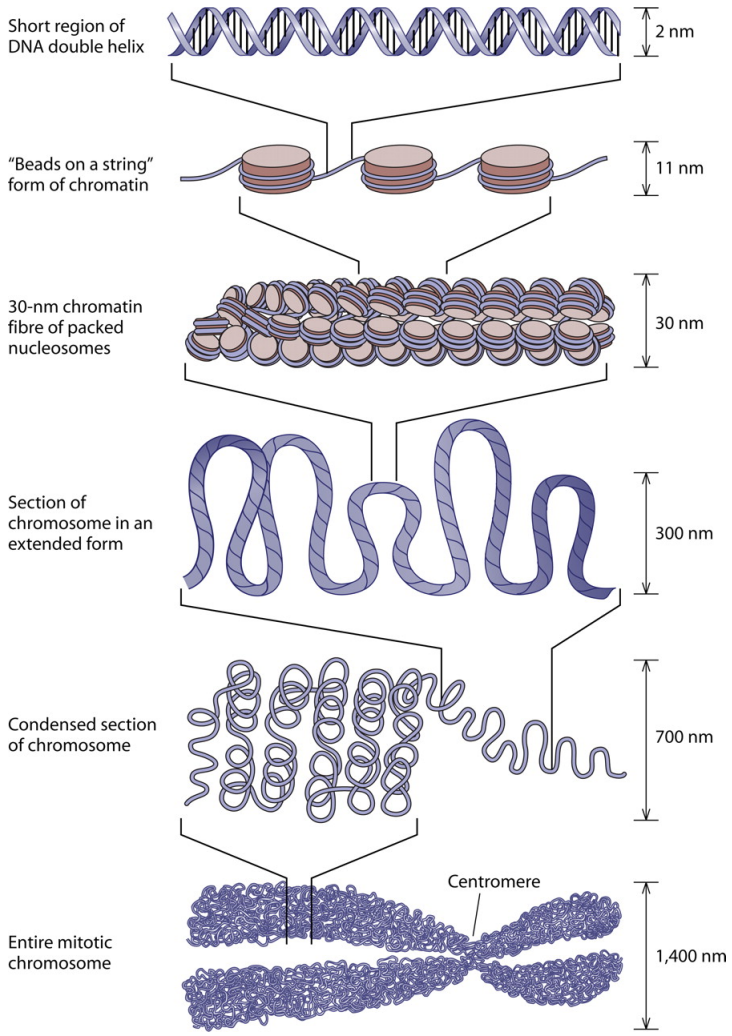



Figure 1.2: Illustration of the chromatin packing process (Image from [2]) .

## 1.2 First evidence of organized chromosome structure

Since the chromosome structure differs a lot depending on the examined cell, one could conclude that the conformation in the nuclei is randomly. That chromosome appears in an organized structure, so called chromosome territories (CTs), was first suggested by Carl Rabl 1885 for animal cell nuclei. Despite light microscopic evidence that favored the territory theory, most results obtained since the 1950s argued that the nuclei is filled with intermingling chromosome fibers and loops. Thus the arrangement of chromosome in territories was even considered to be experimentally disproved. It was not until the 1980s that experiments were performed that gave evidence of structural conformation. One of those experiments was performed by Zorn et al. 1976. They used a laser microbeam to induce UV-damaged DNA within a small part of the nucleus. Depending on the conformation of the of chromosome this procedure would lead to different results (cf. Fig. 1.3). For disorganized

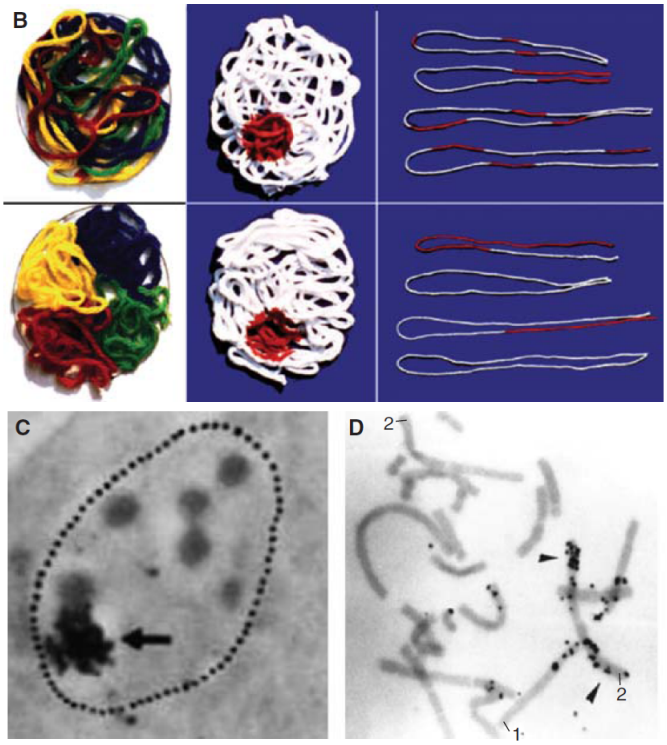



Figure 1.3: Experiments to determine the chromosome structure. B shows the expected result of laser-UV-microbeams experiments. Upper row shows results for disorganized chromatin structure. Bottom row for territorial chromosome arrangement in cell nuclei. Image C and D show the autoradiograph of an Chinese Hamster cell that was partly UV-damaged. One chromosome 1 and one chromosome 2 are intensely marked whereas the other strands are unlabeled. This result supports the theory of a organized chromatin structure (data from [3]).

of chromosome structure the damage would spread throughout different threads. For CTs, on the other hand, the damage would only affect a small subset of threads. Experiments like this gave the first evidence for the of chromosome territory concept. Till the present day the experimental capabilities improved enormously and methods were developed that clearly show that the chromatin in the cell is territorial structured. What is unclear, however, is how exactly this territories are arranged and how the process can be described. Not only for the nuclear range but also for smaller scale chromatin shows a structure that is considered to play main part in regulation process.

The thesis is constructed in the following way: After the general introduction I will go into the state of current research. Experimental methods like FISH and chromatin conformation capture build a good experimental framework to determine the structural genome features. Applications of polymer physics helps to find models with corresponding interactions that can explain higher-order-chromatin structure. I will describe a model of describing polymer chains with a bead on the string model in chapter 3. The equilibrium conformation of a real chain is the balance between attractive and repulsive interaction between these beads. Different conformations of the polymer chains can be topological classified by only one integer number, the genus. I will explain this procedure in chapter 3. Topic of the following chapter is the numerical approach we used to describe this problem. With an exact algorithm to generate self avoiding walks we achieved the genus distribution depending on various parameters. The results will be discussed in the last chapter.

# Chapter 2
# Current research

## 2.1 Experimental methods

Different experimental approaches are used to gain more information about the chromosome structure in nuclei. The determination of the nuclear structure can take place in several scales. Fig. 2.1 illustrates the chromosome structure for resolution in nuclear, chromosome and megabase scale. Depending on the considered range, different interactions and observations become important and different experimental methods are used to determine this features. Two complementary and often used approaches are light microscopy and cell/molecular biology. Light microscopy by fluorescent *in situ* hybridization (FISH) experiments gives more information about the spatial distance as function of genomic linear distance. On the other hand, cell biology by chromosome conformation capture (3C)-based approaches results in contact maps that give indication of the underlying interactions. With the improvements of both methods it is now possible to gain important information about the features of genome architecture. I will explain both methods and the obtained results in this chapter.

### 2.1.1 FISH

Fluorescent *in situ* hybridization (FISH) allows to determine the rearrangements of chromosomes. It combines fluorescent DNA probes with their target sequence in the chromosome. With help of light microscopy (immunofluorescence) the spatial location of the probes, whose position changes with the chromosome position, can be evaluated and the features of the structures become known.

With FISH it is possible to achieve genome-wide images about the chromatin structure in cells. High resolution FISH even allows results in the megabase (Mb) range. However, it can only characterize a limited number of loci in a small number of cells. FISH is a preferred method for characterizing chromosome variability and dynamics.

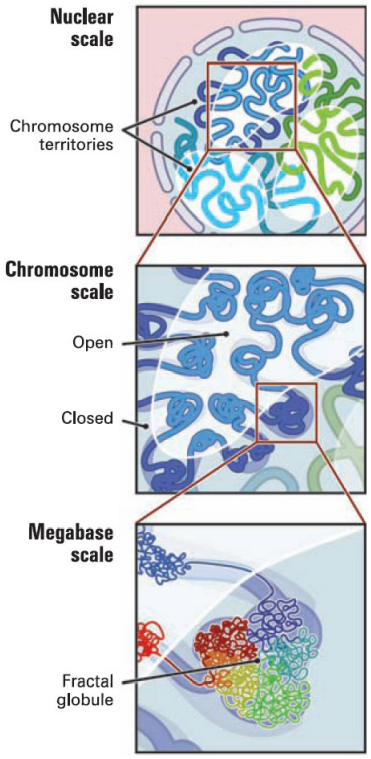



Figure 2.1: Chromosone is shown at different scales. At nuclear scale, chromosome occupies separated territories. Chromosome itself consists of different domains as well. Image from [4].

### 2.1.2 Chromosome conformation capture

Chromosome conformation capture (3C) technique and its variations (4C, 5C, Hi-C) are used to capture the organization of chromosome. It quantitatively measures frequencies of spatial contacts between genomic loci averaged over a large population of fixed cells. This results in so called contact density maps. The process is illustrated in Fig. 2.2. Adjacent chromosome segments are crosslinked and afterwards separated from the non-cross-linked parts. Ligating the fragments and reversing the cross-links results in an array of ligated pairs that can be detected by polymerase chain reaction (PCR). Abundance of pairs is a measure of how often chromosome contact appears.

The range that can be achieved by this method reaches from genome wide results (Hi-C, 4C), through intermediate megabases range (4C, 5C) to fine resolutions in kilobase region (3C). The contact maps are obtained by ensemble-averaged properties of genomic conformations in large populations with typically more than a billion cells. It is not possible to gain information about a single cell.

The two explained experimental methods have different features and limitations and thus they are very complimentary for characterizing the genome features in different ranges. Some of the

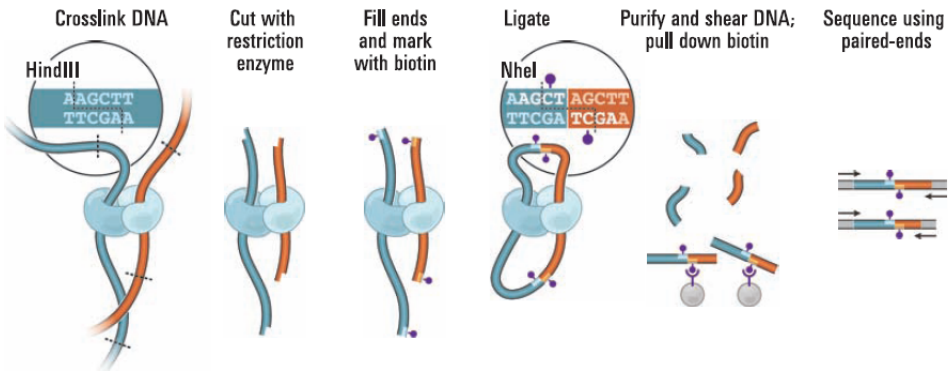



Figure 2.2: 3C method. After crosslinking adjacent chromatin segments with formaldehyde, a restriction enzyme cuts the cross-linked and non-cross-linked parts of DNA apart. DNA fragments are shown in red and blue, the proteins are shown in light blue and cyan. The ends are filled with biotin (purple dot) and ligated. The DNA is purified and sheared and the sequences with paired ends can be detected by PCR. Image from [4].

results obtained by this methods will be mentioned in the following.

## 2.2 Chromosome organization

The explained methods provide a good experimental framework for the study of chromosome organization in the nuclei. In this chapter I will point out some conclusions based on the results FISH and chromosome conformation capture techniques.

### 2.2.1 Chromosome territories

Since the first evidence of chromosome territories the experimental possibilities improved greatly and it is nowadays considered as a known fact that a chromosome inside the nucleus appears in a organized structure. Chromosomes do not mix but instead occupy their own territory. Fig. 2.3 shows a three dimensional false color FISH image of the human chromosome ([5]). One can clearly identify the different CTs. Where these territories touch, intermingling can occur and interaction between loci from different chromosomes is possible. With data from FISH experiments, Cremer and Cremer [3] found evidence for and against CT intermingling (cf. Fig. 2.4).

The same observation was stated by data from chromosome conformation capture method. Contact density maps from Hi-C data in megabase scale indicate that the genome can be separated in two compartments, A and B. Fig. 2.5 shows the map that Dixon et al. [7] achieved. Compartment A corresponds to the orange regions and B to the blue ones. It can be observed that A compartments prefer interaction with other A compartments and in a similar way B compartments prefer interaction among each other. Besides that the interaction frequency in A is significantly higher

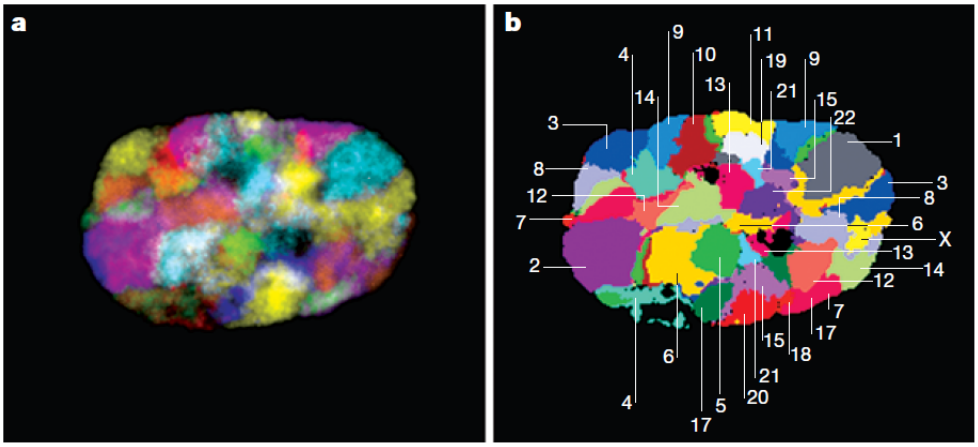



Figure 2.3: False color representation of all chromatin territories by three dimensional FISH. Every chromosome territory can be shown simultaneously in intact interphase nuclei, each in a different color. Image a) shows a red, green and blue image of the 24 labeled chromosomes. In image b) a labeling scheme is used in which each chromosome is labeled with a different set of fluorochromes. Some of the dark regions represent unstained nucleoli (Image from [5]). For more information also see [6].

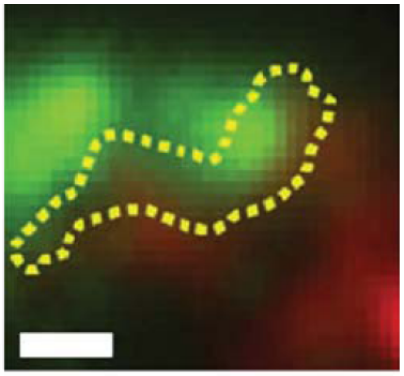


Figure 2.4: Image shows two CTs in different colors. The yellow line marks the are in which intermingling is assumed (data taken from [3]).

compared to B. Lieberman-Aiden et al. [4] brought the different compartments in relation to gene activity. One of them can be identified as euchromatin, a lightly packed, gene rich structure that is transcriptionally active (A), whereas the other compartment, B, is more densely packed.

Intra-chromatin interaction is favored compared to inter-chromatin ones. The compartments occur in alternating pattern and have a size of several megabase. But chromosome itself can be specified in certain domains. This will be subject of the next subsection.

### 2.2.2 Topologically associating domains

With Hi-C and 5C experiments, resolutions at kilobase scale can be achieved and different structural features appear. At this scale, so called topologically associating domains (TAD), that build the

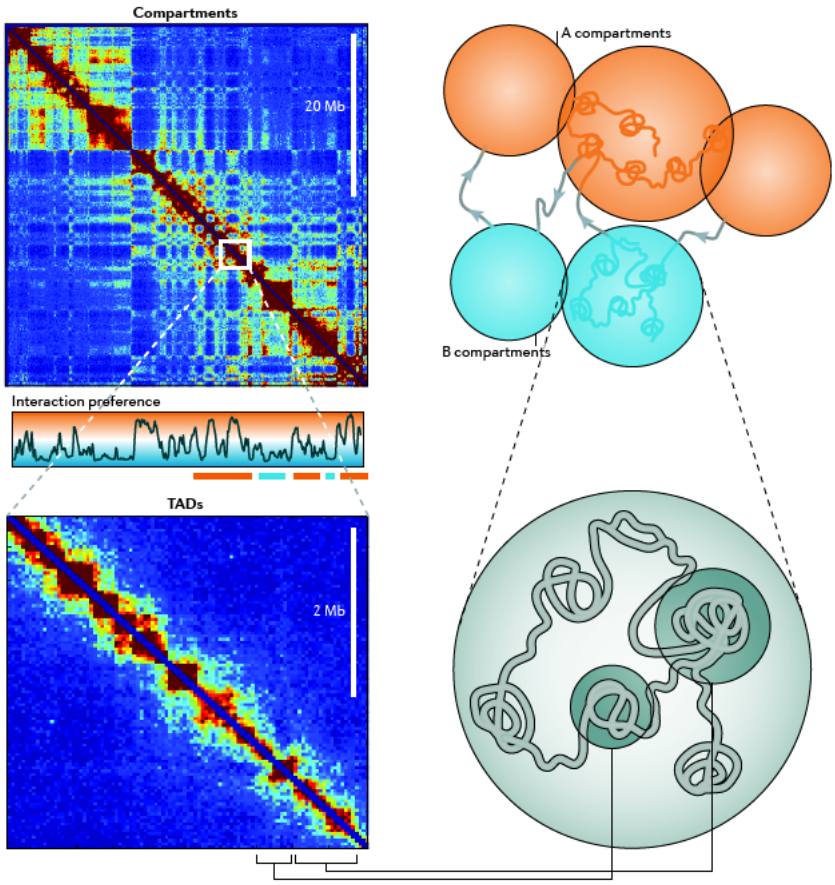



Figure 2.5: Shown is the data from Hi-C determinations for human chromosome 14 for IMR90 cells. Top panel shows compartments at a resolution in megabase scale. Lower panel shows a blow up of a 4 Mb fragment indicating TADs (data taken from [8]).

chromosome, become visible. They differ from the compartments we discussed in the previous section. Nora et al. [9] found that in the 5C interaction map of the mouse X-chromosome at a 4.5 Mb range a series of large structural domains (cf. Fig 2.6) can be seen. Loci in the same TAD tend to interact frequently with each other, whereas the interaction between different TAD is suppressed. This mirrors the importance of the specific topological structure of chromatin in the regulation of gene expression.

The TADs can be hundred of kilobases big. Mouse genomes are composed of over 2.000 TADs, covering over 90% of the genome. It is suggested that they are constructed by chromatin loops. The looping model is also often used to justify long range interaction, an effect that I will explain in the next section.

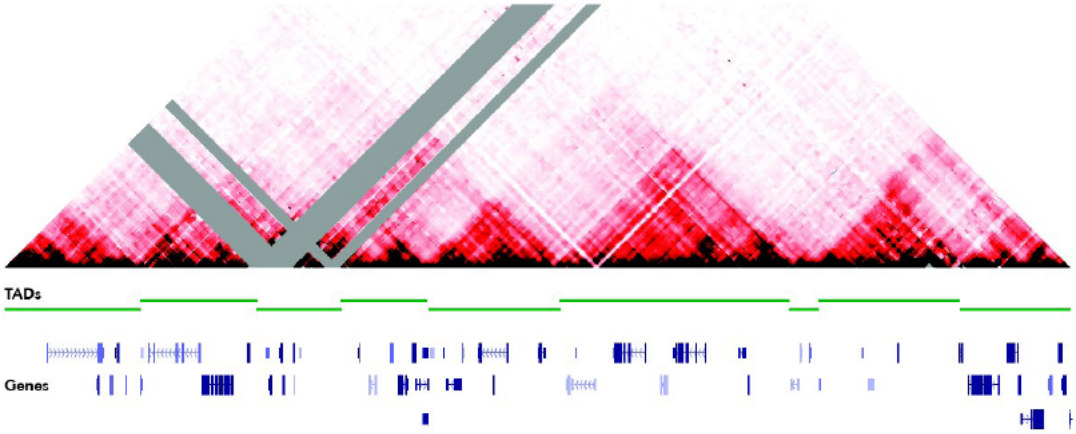



Figure 2.6: 5C interaction map of a 4.5 Mb region on the mouse X chromosome. Map is cut in half on the diagonal. Red indicates the interaction frequency between pairs of loci. Grey represents missing data. Regions in which loci frequently interact correspond to TADs (image from [8]).

### 2.2.3 Long range regulation

Data achieved from chromatin conformation capture experiments show that long range interaction, i. e. interaction between enhancer and promoter with large genomic distance, often occur in gene regulation processes. 3C and 4C interaction profiles in the kilobase range focus on a single locus and the interaction with surrounding chromatin (3C) or genome-wide interactions (4C). Fig. 2.7 shows the interaction profile for 3C data (cf. [10]). The peaks in the interaction frequency indicate long range interaction. The analysis of 5C and Hi-C data achieved interactions maps show the same result. These determination don't refer to one locus but instead generate data that is often illustrated in heat maps.

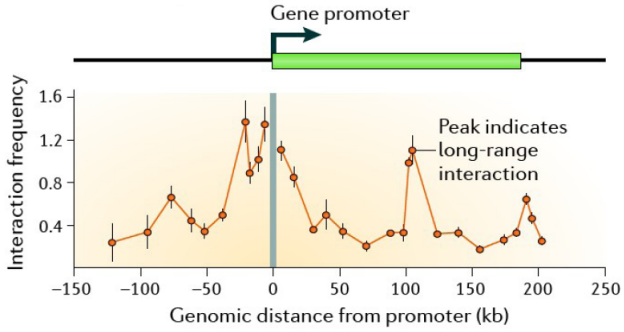



Figure 2.7: 3C data for the CFTR gene in Caco-2 cells. The peaks indicate long range interactions (image from [8]).

Enhancer and promoter are often many basepairs apart and even embedded between other regulatory elements. Wijgerde et al. [11] found with help of FISH experiments that for $\beta$-globin gene, for

example, the enhancer interacts with a gene that is located about 50 kb downstream. Interaction at this range is not intuitive, since one expects the interaction strength to decrease with the distance. Thus the question arises, how enhancer and promoter can interact at such long distances and how can be determined which element regulates which gene at a given time? It was also found that average pattern of interaction at long distance is asymmetric. Most frequent long range interactions are observed with elements that are located upstream of the transcription starting side, although downstream would be possible as well. The reason behind this may lie in some directionality in the looping mechanism.

There are mainly three models that describe the long range gene regulatory: looping, linking and tracking model. For the first two models the spatial chromatin structure is crucial. Recent experimental evidence supports the looping model, which is why I will describe this one more extensive in the next subsection.

Insulating sequences in the DNA interfere with the enhancer-promoter interactions and thus control their interaction. The specific properties of these enhancer blocking elements depend on the model chosen for the interaction description. In the tracking model one assumes that the enhancer controls the promoter by sending a signal through the DNA strand to its target gene. Here the insulators are enhancer-blocking elements that can stop this signal. In the linking model, enhancer and promoter are able to build a linked state because the nucleosomes between them form a coiled structure. The transcription happens through direct contact between enhancer and the promoter. Another idea is that the insulators separate the genome into topologically independent loops. I will describe the looping model in more detail in the next section.

### 2.2.4 Chromatin looping

Loop formations in the kilobase range are a good model to explain long distance promoter-enhancer interactions. Loops can separate genome regions from each other and thus explain independent function. Otherwise they can also bring distantly located sequence elements closer together. They are suggested in all levels of chromatin organization. Local loops at a kilobase scale can describe the long range enhancer-promoter interaction. The looping model is illustrated in Fig. 2.8. Insulators (labeled as EB) separate elements in different loop domains, where intra-loop enhancer-promoter interaction is favored compared to inter-loop interaction. The arrow thickness in Fig. 2.8 indicates the strength of the interaction. Recent studies on the human genome claim that many, if not all gene promoters interact with their enhancer elements through chromatin looping. Sanyal et al. [13] found thousands of significant long-range interactions by using 5C data. Furthermore, loops form in a way that suppresses inter-loop interaction. Thus the loop domain model suggests that contacts

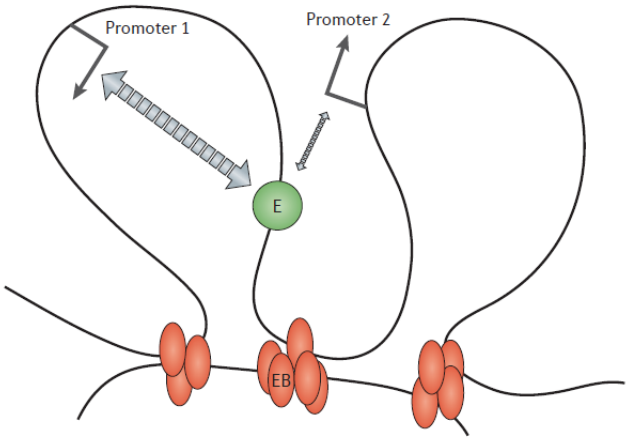



Figure 2.8: Enhancer blocking elements (EB) form different loop domains. Inter-loop enhancer-promoter interaction is suppressed compared to intra-loop interaction (Image from [12]).

in chromatin also avoid 'pseudoknots'.

Now it is important to see whether physical models describe higher-order chromatin contact structure building territorial domains, or CTs, and verify the experimental results. Therefore we start with the method of describing chromosome as a polymer chain. Statistical behavior and spatial structure of polymer chains, depending on interactions that take biological as well as physical constrains into account, help to build a physical framework. I will describe the polymer chain model in the next chapter.

# Chapter 3

# Polymer chains

Polymers in general can have a quite complicated structure since besides linear chains also branched and cross-linked polymers occur. However we are only interested in the linear chains here as they are in accordance with the DNA model. The linear chain can again be described by different models depending on the intent. Three of them are shown in Fig. 3.1. All of them combine monomers to units and describe them by beads centered at one space point. In the *bead-spring model* this units are connected through springs. It is useful to describe the motion of different parts of the chain. The *bead-stick model* represent the connections between the monomers by sticks. This model allows different variations such as beads diameter, stick thickness or restrictions to angles between two adjacent sticks. In the *pearl-necklace model* the pearls are always right next to their adjacent pearls which basically corresponds to a bead-stick model where the stick length equals to the diameter of the pearls (cf. [14]). Since we are interested in the overall conformation of the chain a model that matches the latter one is used in this work. The chain will be considered as a sequence of beads connected by an insignificant flexible string.

For real polymer different effects occur due to interactions between the beads. One is the *excluded*

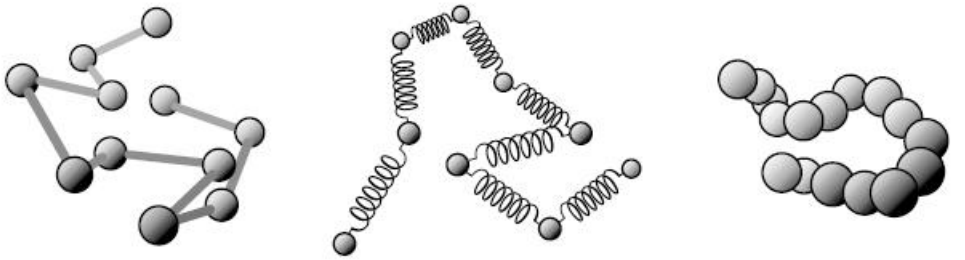



Figure 3.1: Different models for a linear chain: bead-stick model (a), bead-spring model (b), pearl-necklace model (c) (Image from [14]).

*volume*, which leads to a swelling of the chain. Nevertheless in some cases a real chain behaves like an ideal one, a chain without interactions. Therefore both cases will be discussed in the following.

## 3.1 Ideal chain

First we will consider the simplest model of a chain which only takes into account the bonding of the beads on the string but no interaction between them. The position of each bead only depends on the positions of the adjacent ones along the chain and no direction is favored, i.e. every direction has the same probability to be chosen. The mean square end-to-end distance of the chain can be computed by

$$\langle R^2 \rangle = \sum_{i=1}^{N} \langle \mathbf{x}_i^2 \rangle = Nb^2 \ , \tag{3.1}$$

where $b$ is the distance between the monomer units and $N$ the number of monomers. If we take the limit to very long chains, we can use the *Central Limit Theorem* to find the probability distribution of the end-to-end vector. The theorem claims that the iteration over a large number of random variables is normally distributed. We will describe the conformation of the chain structure as a statistical process. Although a DNA strand is still short compared to other applications in thermodynamic, the molecular weight of the chain ($10^7 - 10^9$) reaches a way higher values than the one of single monomers ($\sim 10^2$). Thus we can use statistical methods and assume the thermodynamical large $N$ limit in the following. Instead of talking about a Gaussian distribution as a border case of the ideal chain, we can consider a Gaussian chain. The ideal chain becomes equivalent to the Brownian motion, with the number of beads corresponding to the number of steps and the distance between beginning and end of the chain plays the role of the time. For that reason it can be modeled as a random walk on a lattice. If we extend the ideality to short parts of the chain any two points $\boldsymbol{r}_1,\ \boldsymbol{r}_2$ will follow a Gaussion distribution

$$G(\boldsymbol{r}_1, \boldsymbol{r}_2; n) = \left( \frac{2\pi n b^2}{3} \right)^{-3/2} \exp\left( -\frac{3(\boldsymbol{r}_1 - \boldsymbol{r}_2)^2}{2nb^2} \right) \ , \tag{3.2}$$

where $n$ is the amount of segments between the two points ($n \leq N$) and b is the length of each segment. A Gaussian chain of $n_1$ segments and another one of $n_2$ segments can be joined to a single Gaussian chain of $n_1 + n_2$ segments. In the same way a Gaussian chain can be divided into different parts, as shown by the the following multiplication law which $G(\boldsymbol{r}_1, \boldsymbol{r}_2, n)$ satisfies:

$$G(\boldsymbol{r}_1, \boldsymbol{r}', n_1) G(\boldsymbol{r}', \boldsymbol{r}_2, n_2) = G(\boldsymbol{r}_1, \boldsymbol{r}_2, n_1 + n_2) \ . \tag{3.3}$$

In the limit of $n \to 0$ the $G(\boldsymbol{r}_1, \boldsymbol{r}_2, 0)$ becomes the delta function:

$$G(\boldsymbol{r}_1, \boldsymbol{r}_2, 0) = \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2). \tag{3.4}$$

The free energy $F$ of a system is the sum of energetic and entropic terms

$$F = E - TS \ . \tag{3.5}$$

For the ideal chain only the entropy contributes to the free energy. The energetic term vanishes since there are no interactions between the monomers. With the help of eq. (3.2) we obtain the entropy of the Gaussian chain:

$$S = k_B \ln(G) + const. = -\frac{3k_B}{2\langle R^2 \rangle}(\boldsymbol{r} - \boldsymbol{r}')^2 \ , \tag{3.6}$$

where $\boldsymbol{r}$ and $\boldsymbol{r}'$ are the two ends of the chain. Thus the free energy of the chain is

$$F = \frac{3k_bT}{2\langle R^2 \rangle}(\boldsymbol{r} - \boldsymbol{r}')^2 + const \ . \tag{3.7}$$

In order to minimize the free energy, the chain tries do decrease the magnitude of $\boldsymbol{r} - \boldsymbol{r}'$. The entropy term promotes a shrinking of the chain. Why will we not achieve a vanishing $\boldsymbol{r} - \boldsymbol{r}'$ value? The reason behind this is that is the so called *entropy elasticity*. The name arises from the fact that this elasticity only appears due to the entropic term. The force responsible for a nonzero $\boldsymbol{r} - \boldsymbol{r}'$ value is given by

$$\frac{\partial F}{\partial \boldsymbol{r}} = \frac{3k_bT}{\langle R^2 \rangle}(\boldsymbol{r} - \boldsymbol{r}') \ , \tag{3.8}$$

which is equivalent to a spring force with the spring constant $3k_BT/\langle R^2 \rangle$. It becomes stiffer with increasing temperature. Any chain with a finite size has this elastic behavior (cf. [14]).

## 3.2 Real chain

For a real chain we will, in contrast to the ideal ones, consider different interactions between the monomers. A repulsive part is originated in the excluded volume effect, which I will discuss for a simple example in the next section.

### 3.2.1 Excluded volume

In this example we consider two spheres, A and B, shown in Fig 3.2. It is not possible for the center of B to get in the volume $\nu_e$ around A. This results in a change of entropy. To calculate this change lets consider that the volume excluded by each sphere rises from zero to $\nu_e$. The space left for the

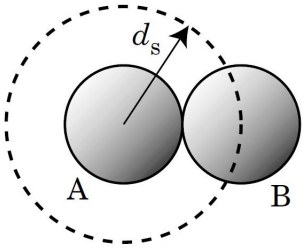



Figure 3.2: Demonstration of the excluded volume. It is not possible for the center of sphere B to approach A closer than $d_S$

other sphere decreases from $V$ to $V - \nu_e$. The resulting entropy change is

$$\Delta S = k_{\mathrm{B}} \ln\left(\frac{V-\nu_e}{V}\right) = -k_{\mathrm{B}} \sum_{k=0}^{\infty} \frac{\left(\frac{\nu_e}{V}\right)^{k+1}}{k+1} = -k_{\mathrm{B}}\left(\frac{\nu_e}{V} + \mathcal{O}\left((\frac{\nu_e}{V})^2\right)\right)$$
$$\Delta S \cong -k_{\mathrm{B}} \frac{\nu_e}{V} \tag{3.9}$$

where we assumed that $\nu_e \ll V$. For $N$ spheres there are $N^2/2$ interacting pairs and thus the entropy change due to the excluded volume is

$$\Delta S \cong \frac{N^2}{2} \frac{\nu_e}{V} \; . \tag{3.10}$$

The change of the free energy is

$$\frac{\Delta F}{k_{\mathrm{B}} T} \propto b^3 \frac{N^2}{R^3} \; , \tag{3.11}$$

where we used that $\nu_e \cong b^3$. Although the change in $F$ originates from entropic effects we will regard it as a repulsive interaction between the monomer units.

### 3.2.2 Coil and globule state

A real polymer chain in a solvent can have two macroscopic states: the *globule* or the *coil* state. The globule is a dense structure from which almost all solvent molecules are suppressed by the monomers of the chain and the density fluctuation is less than the density itself. Polymer-polymer bonds are favored which leads to this collapsed form. The coil state on the other hand pulsates macroscopically. The density fluctuations are about the same order as the density itself ([15]). The coil is an open structure and repulsive interactions dominate. Which state the chain in a solvent adopts depends on the balance between polymer-polymer and polymer-solvent interactions and thus on the solvent quality. The quality can be changed by varying the temperature or the concentration

of the solvent. For a good solvent (high $T$ values) the chain is in the coil state whereas for the poor solvent (low $T$ values) it takes on the globular state. At the so called *theta temperature* $T = \Theta$ the real chain behaves like an ideal chain. This temperature is the borderline between good and poor solvent.

The equilibrium confirmation is the balance between attractive and repulsive interactions that minimizes the free energy Eq. (3.5). For the coiled state we can assume a low monomer density for long chains and thus expand the free energy in powers of the density $n$ (virial expansion):

$$E = NkT(Bn + Cn^2 + \dots) \,, \tag{3.12}$$

where $B$ , $C$ , $\dots$ are the virial coefficients which are responsible for the two, three, etc. body interactions. They only depend on the potential of the interaction between the units, i.e

$$\begin{gathered} B = \frac{1}{2} \int \mathrm{d}^3 r \, ( \, 1 - \exp\left(U(r)/kT\right)) \;, \\ C > 0 \,, \end{gathered} \tag{3.13}$$

where the potential $U(r)$ can vary greatly depending on the solvent. From Eq. (3.13) however we can identify the rough form of B. For high temperatures ($T > \Theta$) the exponential factor will be small and $B$ is positive. This is the region of a good solvent. The interaction represented by the $B$-term are repulsive and higher corrections terms to the free energy can be neglected. The chain is in the coil state. With decreasing temperature, the value of $B$ decreases and for one point $B$ has to vanish. This temperature is the $\Theta$-temperature. The whole energy term vanishes and the chain adopts the ideal coil conformation. In the poor solvent ($T < \Theta$), $B$ has negative values and thus the attractive interaction part dominates and the chain starts to shrink. Due to the fact that it can't shrink to zero and that the density $n$ becomes larger we have to take higher terms from the virial expansion into account, i.e. repulsive three body interactions. $B < 0$ is the region where the chain has a globule conformation. The calculation of the free energy can not be approximated with the virial expansion any more since the monomer density is large in this region and higher orders can't be neglected.

A microscopic description of the approximation is possible by the dimensionless Hamiltonian (cf. [16]):

$$\begin{aligned} H[c(\tau)] = \frac{1}{2} \int_0^N \mathrm{d}\tau \, |\dot{c}(\tau)|^2 + \frac{1}{2!} g_0 \int_0^N \mathrm{d}\tau \, \int_0^N d\tau' \delta[c(\tau) - c(\tau')] \\ + \frac{1}{3!} w_0 \int_0^N \mathrm{d}\tau \, \int_0^N d\tau' \int_0^N d\tau'' \delta[c(\tau) - c(\tau')] \delta[c(\tau) - c(\tau'')] \end{aligned} \,, \tag{3.14}$$

that only takes two-and three body interactions into account. The coupling constants for the interactions are given by $g_0$ and $w_o$, $c(\tau)$ represents a continuous chain conformation in $d$-dimensional space, $\tau$ is the contour variable, $\tau \in [0, N]$ and $\dot{c} \equiv dc/d\tau$. The fixed end-vector partition function is defined as

$$G(\boldsymbol{R}, N; g, w) = \int_{c(0)}^{c(N)} \mathcal{D}[c(\tau)] \exp\{-H[c(\tau)]\} \; . \tag{3.15}$$

This model describes a chain with a certain spatial structure. Thus it can be used to describe the coil state but it is unusable close to the transition point and for the globule region. As described above, $w_0$ will always be assumed to be positive since it represents repulsive interaction. This model has been used to calculate mean-field approximations (cf. [16]). However it can also be extended to a field theoretical description that provides a systematically determination and can be used to study the two different states as well as the transition between them (Amaral [17], Widyan [18], Flores et al. [19]).

### 3.2.3 Dimension of a real chain

In mean field theory the equilibrium size of a chain has a exponential dependence of the length $N$

$$R \propto N^{\nu} \tag{3.16}$$

and can be obtained by minimizing the free energy. Aim in this section is to determain the exponential $\nu$ for the different states. We will perform this calculation in three dimension but similar argumentations are valid for two dimensions as well. In section 3.1 we already discussed the case of the chain at the $\Theta$ temperature. In this state the end-to-end distance of the chain increases linearly with the length of the chain

$$R \propto N^{1/2} \; , \quad \nu = 1/2 \qquad \textbf{ideal state} \; . \tag{3.17}$$

For the globular stat ($T < \Theta$) we can assume the chain to be tightly packed in a cube analogous to sphere packing. Therefor the mean square end-to-end distance is equivalent to the radius of circumscribing sphere $R_{sp}$. If $a = N^{1/3} \cdot 2b$ is the length of the sides of the cube then $R = \frac{\sqrt{3}}{2}a$. Therefor we achieve:

$$R \propto N^{1/3} \; , \quad \nu = 1/3 \qquad \textbf{globule state}. \tag{3.18}$$

For the coil state ($T > \Theta$) we will use Florys approximation to determine the exponential behavior. Flory was the first one deriving this scaling behavior by taking excluding volume effects into account.

For the energetic term we substitute the virial expansion (3.12) and neglect every order higher than two in the density. The entropic term is given by Eq. (3.6). Therefor the free energy results in

$$F = k_B T \left( \frac{3N^2}{R^3} B + \frac{3}{2} \frac{R^2}{Nb^2} \right) . \tag{3.19}$$

To achieve a relation between $R$ and $N$ we will minimize $F$ with respect to $R$:

$$\begin{gathered} \frac{\partial F}{\partial R} = -\frac{BN^2}{R^4} + \frac{3R}{Nb^2} \equiv 0 \\ R \propto N^{3/5} , \quad \nu = 3/5 \qquad \textbf{coil state}. \end{gathered} \tag{3.20}$$

The equilibrium size is a balance between repulsive excluded volume effects and compressing forces due to the loss of entropic energy. The exponent $\nu$ is called the *Flory exponent*. Because of the simple form the theory is often referred to as a prototype of meanfield theory. It is not trivial to go beyond the mean field approximation. In renormalization group theory one could estimate the exact value of $\nu = 0,591...$ which is remarkably close to the Flory exponent. A generalized form of the exponent

$$\nu(d) = \frac{3}{d+2} \tag{3.21}$$

predicts exact values for $d < 4$.

A numerical model for describing a real chain is a self avoiding walk, in the same way as the ideal chain was equivalent to a random walk. For a SAW it is not possible to occupy a site more often than once what mirrors the repulsive part of the interactions in the chain. Moreover we will add attractive interaction between the beads to the model. I will amplify the method further in chapter 4.

### 3.2.4 Coil-globule transition

We can identify the different density behavior of the two states as different phases of the system. The transition between them is consequently equivalent to a phase transition. In statistical mechanics however a phase transition is only defined for $N \to \infty$ which is not present when dealing with DNA. Nevertheless there is a transition between the coil and globule state that has a finite width $\Delta T$. In this region it is not possible to distinguish between the two states. It separates regions with different phases, i.e. regions in which one of the phases gives the magnificent contribution to the partition function. The transition is called phase transition if $\Delta T$ vanishes for $N \to \infty$ (cf. [15]).

# Chapter 4

# Topological classification

If we want to analyze the preference of certain polymer structures it is helpful to separate the configurations in different classes and specify them by values that can be compared to other models and experimental results. One in the literature often used way to do is to separate the diagrams depending on the way the interaction in the chain occur. I will explain this method in the following section.

## 4.1 Planar diagrams

DNA and RNA chains can be described by different structures. The primary structure displays the sequence of nucleotides in the order in which they are linked together. There is no information about the spatial structure of the polymer. The next higher order is the secondary structure. Here is not only the three dimensional form but also non crossing interactions between the monomers considered. After all there is the tertiary structure which takes all possible interactions among the monomers into account. As soon as interactions cross, so called *pseudoknots* emerge. They are no true knots in space but generated by the polymer's self-contacts. Diagrams with no crossing arcs are also called *planar-diagrams*. The disadvantage of energy-based methods is that they deal almost only with secondary structures. Pseudoknots can't be included in systematical way. However, it is possible to find a topological classification of polymer chains. Let us consider a self avoiding

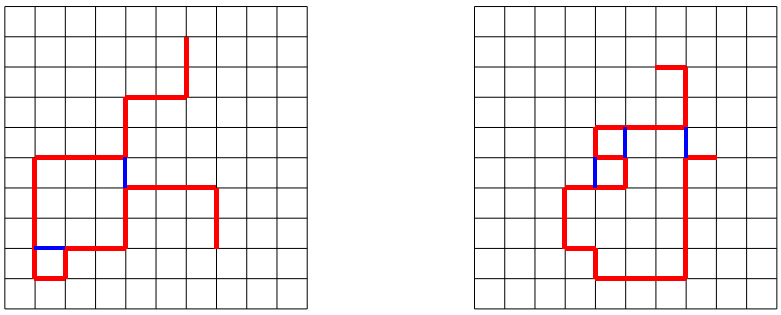


Figure 4.1: Two example for self avoiding walks on a lattice which will produce a planar (left walk) and a non planar (right walk) diagram. The blue lines indicate interactions between adjacent sites.

(a) Planar diagram (b) Non-Planar diagram

Figure 4.2: Planar and non-planar diagram corresponding to the self avoiding walk shown in Fig. 4.1 (Image from [20]).

walk on a lattice with nearest neighbor interactions. One can consider a lot of different interaction models. In the one that we will use later attractive interaction only appears between neighboring sites that are not adjacent in the chain. To illustrate the method we will use this assumption here. The represented ideas however are general valid. Two examples of SAWs are illustrated in Fig. 4.1. The red line indicates the path and the blue ones the interaction with neighboring sites. Now we can pull the ends of the walk apart and will receive the diagrams shown in Fig. 4.2. The solid line represents the walk and the dashed lines the interaction between the sites. The left one shows the diagram corresponding to the left walk in Fig. 4.1. It is planar since the interaction lines do not cross. In image (b) is the diagram corresponding to the right lattice illustrated. There the chords do cross and a pseudoknot appears. Pseudoknots lower the entropy since they restrict the possible configurational structure. This conformations can only be energetic favorable when the energetic loss through the bond is bigger than the gain in energy through entropic causes.

The configurations with pseudoknots can be very complex structures and they differ from walk to walk. Thus to classify the structure it is preferable to use another variable that can be associated to every topological configuration: the *genus*.

## 4.2 The genus

The genus $g$ is a topological invariant of a diagram and is related to the Euler characteristic $\xi = 1-2g$ ([21]). It is a characteristic for the pairing in the polymer chain and helps to predict the full tertiary structure in a systematic and comparative way. The topological complexity of a pseudoknot can be described by just one integer number. To determine the genus for a given diagram we will use the 't Hoofts double line formalism, developed in connection with large-N matrix theory. In QCD, for example, this means that every gluon comes with $N$ colors and every gluon propagator is represented as a double line. For the limit $N \to \infty$, planar diagrams give the main contribution ([22]). The diagrams shown in Fig. 4.2 for instance will change to a form shown in Fig. 4.3. In the double line formalism, every propagator, i.e. every double line, receives a factor $\frac{1}{N}$ ('t Hoofts rule) and every closed loop results in a factor of $N$, since there are $N$ possible choices of color. Thus, every diagram

(a) Planar diagram (b) Non-planar diagram

Figure 4.3: Planar and non-planar diagram in t'Hoofts double line formalism (Image from [20]).

contributes with a factor

$$\frac{1}{N^{P-L}} \,, \tag{4.1}$$

where $P$ is the number of propagators and $L$ the number of closed loops. According to 't Hooft, powers of $\frac{1}{N^2}$, i.e. factors $\frac{1}{N^{2g}}$, give information about the topology. Thus we can classify the topological structure of the diagrams by the integer $g$, the genus:

$$g = \frac{P-L}{2} \,. \tag{4.2}$$

For a planar diagram, like the one shown in Fig. 4.3 (a) the genus is $g = 0$, since it has two closed loops and two double lines. The genus of the right diagram results in $g = 1$, since it has one closed loop and three double lines.

Now we know a method to determine the genus for a given diagram. But what does this number actually mean? For the geometrical interpretation one can think about a sphere with $g$ holes. Zero holes correspond to a sphere, a sphere with one hole is equivalent to a torus, one with two holes is equivalent to a double torus and so on. The genus of a diagram is the minimum number of holes the sphere must have in order to draw the diagram on it without intersecting chords. Thus it specifies on which kind of surface the diagram can be drawn without crossing lines. An often used way to represent the interactions is by using *circular diagram*. They are obtained by closing the ends of the diagram to a circle. An example of different circular diagrams and the corresponding genus is shown in Fig. 4.4.

A diagram with no crossing arcs corresponds to $g = 0$ and can be drawn on a sphere. As soon as lines do cross at least a torus is the necessary surface. Anyhow, the genus is not equivalent to the number of crossing lines.

Even separating the diagrams in only two classes, i.e. planar and non-planar ones, is of enormous practical use and helps for specifying chromatin intra - and inter-loop interactions. The same process can be done for RNA. The kissing hairpin from chapter 1, for instance, has a circular diagram shown in Fig. 4.5. The inter-loop interaction generates a pseudoknot and the genus of the structure is $g = 1$.

We will use this topological classification to study the tertiary structure of a self interacting

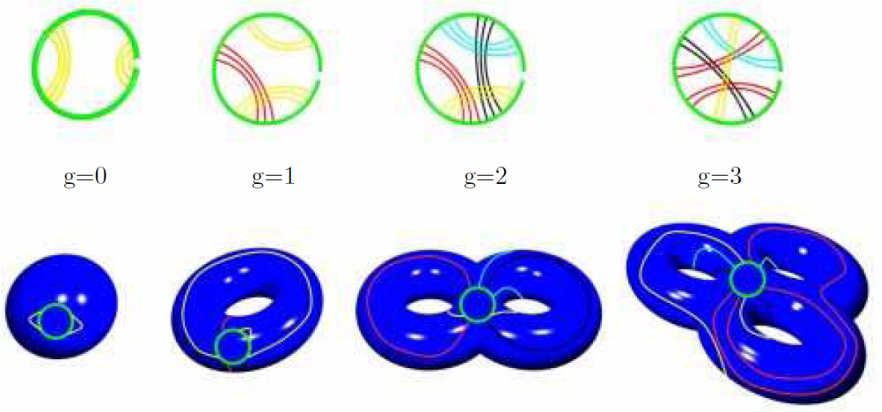



Figure 4.4: Circular diagrams and corresponding genus (Image from [20]).

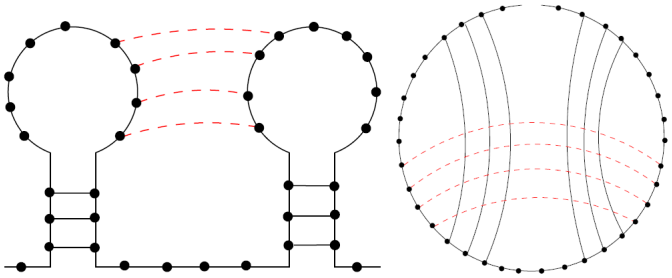


Figure 4.5: Kissing hairpin and the corresponding circular diagram (Image from [1]).

polymer chain. The procedure will be like described above. After generating a self avoiding walk we will map it to the circular diagram regarding the interactions in the polymer. From there we can conclude which genus, and therefore which topological structure preferred appears.

# Chapter 5
# Numerical method

After we discussed possible polymer models in chapter 3, I will now introduce the model we will use for the analyzes of the spatial structure. We will use a simple interaction model to describe the interaction in the chains. The model can be extended and different variations are possible. Because of the method we chose for generating the SAWs we achieve exact results but also limit our determination to a small number of sites. I will explain the method in the second section.

## 5.1 The model

As I mentioned before, a large chain can be considered as a statistical system and since chromatin is long compared to single monomer dimensions, we will use statistical methods to describe the chromatin chain. In chapter 3 we saw that a real chain can be understood by considering a self avoiding walk. The repulsive effect of excluded volume follows from the generation of self avoiding walks. Moreover we will add attraction between adjacent sites that are not neighbors in the chain. We will use short range interaction because the interaction density in a real chain is expected to be low, since not every possible pair will interact.

The strength of the interaction, and with it the energy value, will vary depending on the solvent. It is assumed that every bond in the walk has the same energy. We will add a statistical weight, comparable to the Boltzman factor, to every walk and sum over all numbers of macrostates, i.e. all possible configurations. Thus we can set up the canonical partition function $Z(E)$ for all the walks with specified length:

$$Z(E) = \sum_{i=1}^{n_w} \exp\left(\sum_{j=1}^{m_i} -\beta\ E\right) = \sum_{i=1}^{n_w} \mathrm{e}^{-\beta E m_i}\ , \tag{5.1}$$

where $\beta = 1/k_{\mathrm{B}}T$, $n_w$ the number of possible walks and $m_i$ is the number of bonds the $i$th walk has. Eq. (5.1) can be written as

$$Z(w) = \sum_i w^{m_i}\ , \tag{5.2}$$

where we introduced the quantity $w$ which stands for the interaction strength and will be used as one of the parameters to classify the walks. In the good solvent ($T > \Theta$) repulsive interactions are

predominant which corresponds to a small attraction, i. e. small $w$, in our model. The state of the system is equivalent to the coiled state. On the contrary for the bad solvent ($T < \Theta$) bonds are favored since it lowers the energy of the system. The attraction predominates (big $w$) and the results will be equivalent to those of a chain in the collapsed chain. At the $\Theta$ temperature, repulsive and attractive interaction are in balance and the chain should behaves like an ideal chain. For our model this means the walk should have the properties of a random walk. For varying $w$ from high values to lower ones we should be able to see the indications of coil-globule transition for the system.

To generate statistical average values we took all the possible configurations into account. I will describe the algorithm that we used in the next section. From the relation between the contacts in the walk we can determine the genus of the diagram and therewith conclude to favored structures of the chain. We know that inter-loop interactions in chromatin are suppressed. In order to describe the chromatin we need a model that mirrors this feature and thus shows preferred planar structures. Determining the appearance of planar or non-planar structures is necessary to see if the chosen interaction model can be used to describe the chromatin structure.

## 5.2 Backtracking algorithm

To generate the self avoiding walks, we used the so called *backtracking algorithm*. This is an exact method, where every possible configuration with $N$ steps and the corresponding properties is taken into account. The walker will take the first $N$ steps in a defined direction. After the last step is gone, he will go one step back and from there check the adjacent site in a certain direction. If attempting this site does not lead to an intersection in the walk, the walker will occupy it and take the next step. If the site is occupied, the walker will go the step back and try another direction. The order in which the walker is going to check the directions is given and thus recursively all possible walks are generated.

Fig. 5.1 shows part of a self avoiding walk. The smaller red point indicates the growing tip. For the next step, every adjacent site will be checked. Has the walker been on this site before, it is not possible to visit this site again, since this leads to intersection. Thus the blue lines in Fig. 5.1 indicate admissible growth. Next step can only be fulfilled to an unoccupied site (green line). The walker takes track of the already gone path. This can also be represented by a tree diagram. Every node shows the so far gone steps and every child represents the walks that are possible from the parent node by adding one step. The leaves are the possible walks with $N$ steps. An example for three steps is shown in Fig. 5.2.

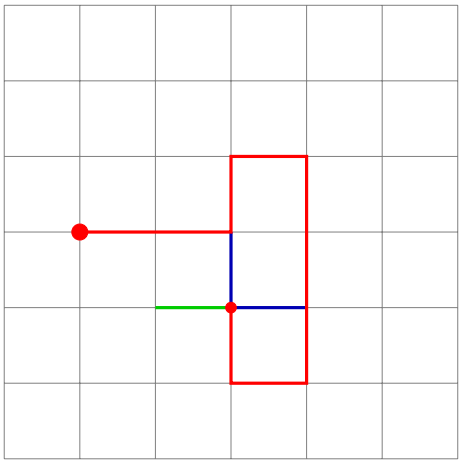


Figure 5.1: Self avoiding walk. The smaller point indicates the growing tip. The green path indicates admissible, the blue one inadmissible growth.

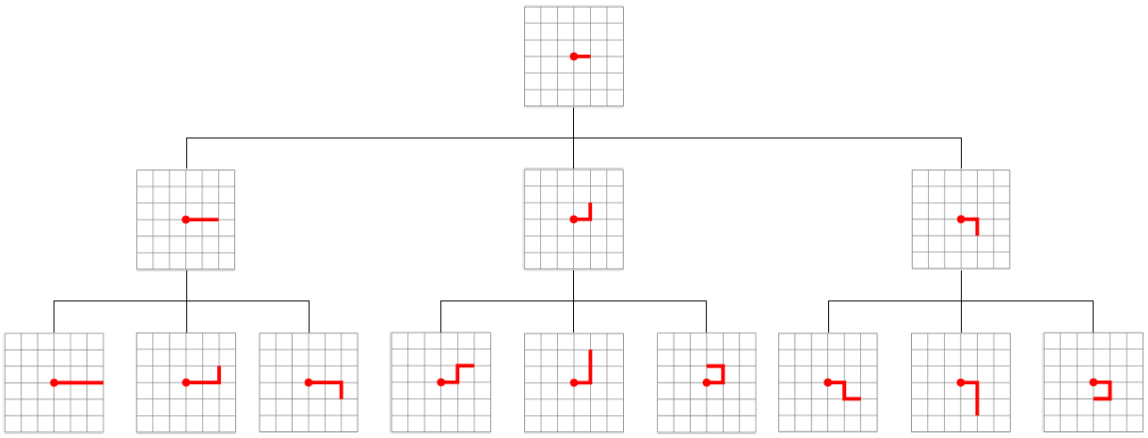


Figure 5.2: Tree diagram for SAW with three steps.

The backtracking algorithm is a brute force method that becomes very complex for higher dimensions or more steps. Compared to statistical Monte Carlo simulations it is not possible to go to the asymptotic behavior for the exact calculation. But where extrapolation methods for the backtracking algorithm bring an uncertainty, statistical fluctuations in Monte Carlo simulations leads to statistical uncertainty.

The highest number of steps we used for our determinations was twelve steps. Schram [23], for instance, optimized the exact enumeration by using the length-double algorithm. It generates self avoiding walks with length $2N$ by connecting two SAWs of length $N$. Besides this they exploit the symmetry of the cubic lattice and parallelized the process. This enabled them the exact enumeration of walks with length $2N = 36$ sites for a cubic lattice in three dimensions. Thus in this thesis we set the foundation for exact determination of topological classification where improvements are possible to achieve results for longer chains.

# Chapter 6
# Results

In this chapter I will discuss the results achieved by using the just explained model. Firstly, I will explain the different statistical ensembles we can use to interpret our results. Afterwards we will analyze at which point the coil-globule transition appears, since we expect different results depending on the polymer state. After discussing the scaling behavior of our polymer model, we will analyze the topological features in the last part of this chapter.

## 6.1 Microcanonical and canonical ensemble

From statistical mechanics we know that there are different ways to treat a statistical system depending on the available variables, i.e. three different ensembles. Every consideration will lead to the same results in the thermodynamical limit ($N \to \infty$). For us two of those ensembles are of interest: microcanonical and canonical ensemble.

For the canonical ensemble energy exchange with the environment is possible which means that the microstates of the system can differ in their energy. Every microstate appears with a probability that is weighted by the Boltzmann factor. We used this method for determining the partition function in chapter 5.

The microcanonical ensemble is isolated, i.e. no energy exchange with the environment. This means that every microstate has the same energy and thus appears with the same probability. The weight for every microstate is the same, i.e. the Boltzmann factor becomes one.

For us this differentiation is of relevance since it allows us to analyze our results in two different ways. We can use the microcanonical ensemble, which is equivalent to $w = 1$, and will achieve a genus distribution only depending on the number of interactions in the walk. This universal view allows us to compare general features of models that do not depend on the specific value of the variable $w$. On the other hand, the canonical ensemble, i.e. the weighting process of the microstates, allows us to differentiate between the coil and the globule state. The genus distribution however depends on two variables: the number of interactions and the attraction strength. In the following it will be specified which point of view is used.

## 6.2 Transition point

Before we are going to discuss different structures and scaling behaviors of our polymer model it is necessary to know in which region we can distinguish between the coil and globule state and where the transition occurs. Thus for this chapter we will use the $w$-dependence of the system to specify the different states.

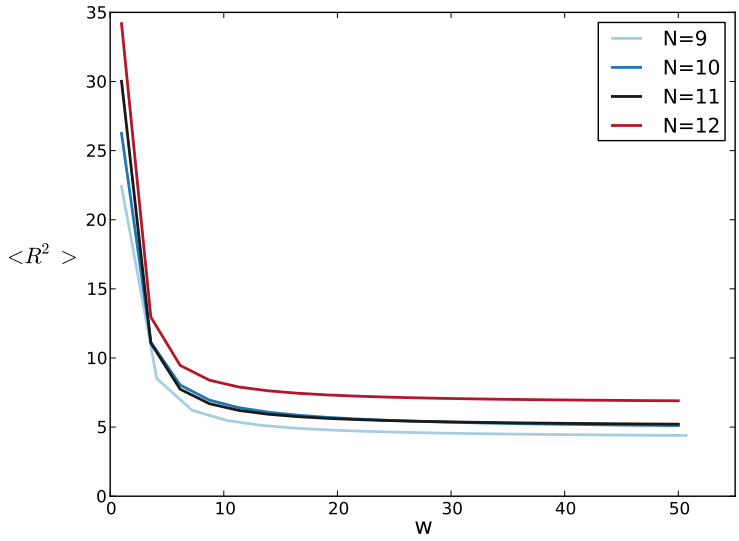
Fig. 6.1 shows the change of the mean square end-to-end vector $\langle R^2 \rangle$ with increasing interaction strength $w$ for different length $N$ of the chain in two and three dimensions. All graphs show the same behavior as the end-to-end distance will decrease rapidly for small interaction strength and reach a plateau after $w \approx 10$. Thus the chain forms the open, coiled structure corresponding to a large $\langle R^2 \rangle$ for very small $w$. Once the dimension shrank and the plateau is reached, the polymer collapsed to the compact globule state.

The steep drop of the end-to-end vector clearly indicates a transition. To examine if this transition really is a phase transition the value of $N$ has to be a lot higher than the ones used here. However, we now know for which interaction values the polymer has an open structure and for which it adopts a collapsed state. We can use this information to study the spatial dimension of the system for the different phases.

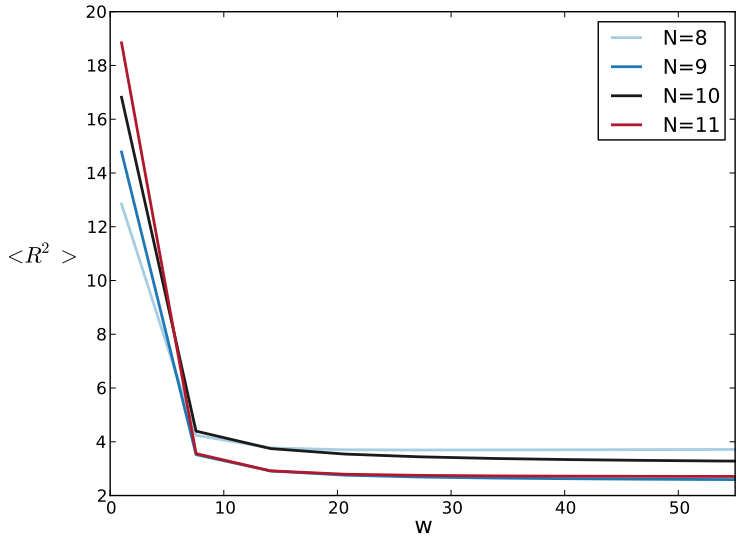
## 6.3 Scaling behavior

In chapter 3 we already found the relation between the mean square end-to-end distance and the length of the chain in the different states. In this section we will determine the scaling behavior of our numerical model for different interaction strengths. Fig. 6.2 shows $\langle R^2 \rangle$ plotted against the length of the chain in logarithmic scale for different interaction strengths. A logarithmic scale is used to ascertain the slope of the fitting line and thus the Flory exponent. We can see that for weak interactions, the coil state respectively, the points are positioned on a straight line with almost no fluctuation. The value for the slope of the lines is very accurate and in the region of Florys exponent ($2\nu \approx 1.2$). Otherwise, for the globule state ($w = 20,\ w = 30$) the values of the end-to-end distance are highly fluctuation. Although the alterations get smaller for longer chains, the fitting line is imprecise and we would have to take higher $N$ values into account to achieve a reliable result. Nonetheless the deviation from the expected value, $2\nu = 0.66$, is relatively small.

We have no comparative value for the slope in the transition region but fluctuations are less compared to the globule one. The results indicate that our model, at least for short chains, describes the coil state a lot more accurate than the globule one.




(a) Two dimension




(b) Three dimension

Figure 6.1: End-to-end distance $\langle R^2 \rangle$ for different length of the chain plotted versus the interaction strength $w$ in two and three dimensions. On the steep decrease of $\langle R^2 \rangle$ we can see that as $w$ increases, the chain transitions from the coil to the globule state.

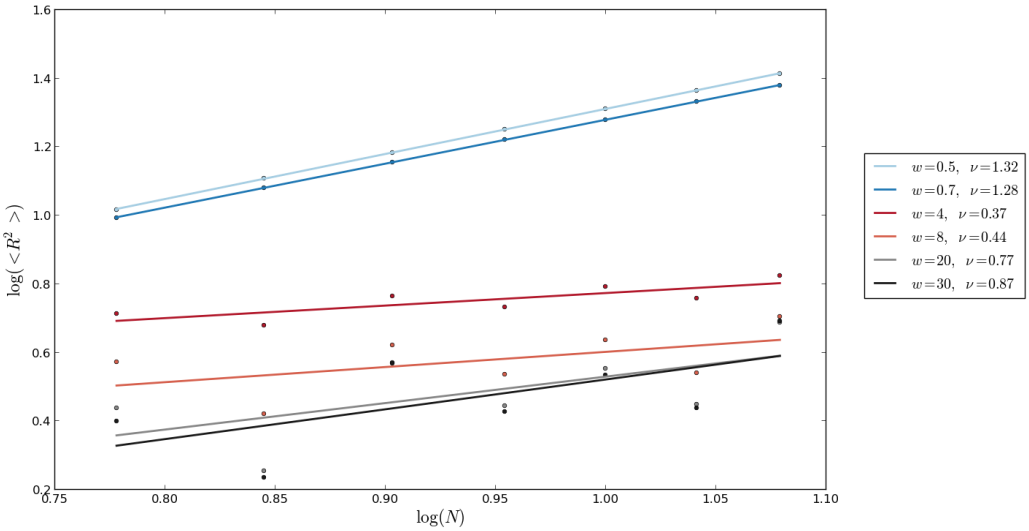



Figure 6.2: End-to-end distance $\langle R^2 \rangle$ plotted against the length of the chain $N$ for different interaction strengths with the corresponding determined Flory exponent.

## 6.4 Topological properties

In this section we will discuss the topological properties of our polymer model. Aim is to work out if there are preferred conformations which the polymer chain of our model adopts depending on the amount of interactions in the chain. We will differentiate them by their genus. Thus for every possible configuration we determined the genus depending on the number of interaction in the chain. In the first part we will do this for the canonical point of view. Later we will discuss some properties of the chain for the $w = 1$ case.

### 6.4.1 Canonical point of view

In this paragraph we used the canonical point of view and weighted every microstate with the Boltzmann factor. This enables us to determine different features for the coil and globule state of the chain, since we expect different genus contribution depending on the interaction strength $w$. Fig. 6.3 shows the percental occurrence of configurations with different genera plotted vs the number of interacting pairs $m$ in a chain with $N = 12$ steps and for small $w$. It is observable that configurations with only a few interactions are predominant and the corresponding genus is very low. This matches with our expectations. For weak interaction the chain is in the coil state. The open structure does not favor interaction between monomers and thus a non planar configuration is unlikely.

Fig. 6.4 shows the same situation for values of $w$ in the transition region. This is the region

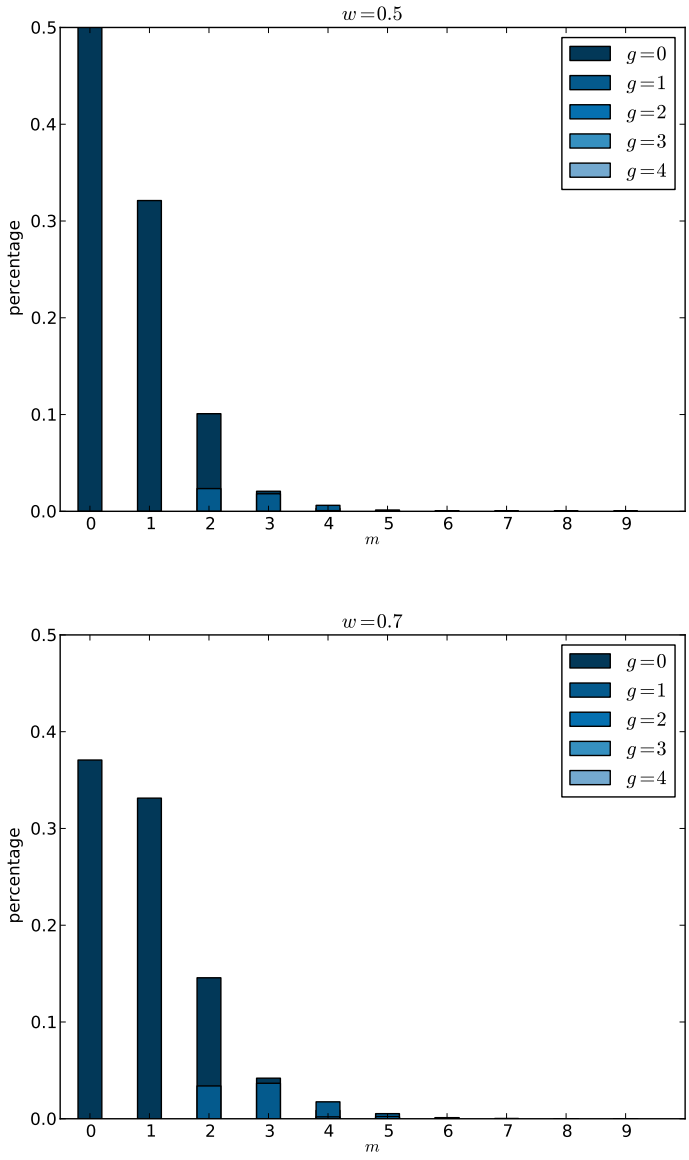



Figure 6.3: Genus occurrence versus number of interactions $m$ for $w = 0.5$ and $w = 0.7$. Configurations with higher genus are clearly suppressed.

where the diagrams differ a lot depending on the value for $w$. Firstly the appearance of different number of interaction pairs and genus seems to be evenly distributed. As $w$ goes closer to values in the globule region, this changes and higher genus as well as configurations with more interactions are favored. At the same time for all of them the most emerged number of interaction pairs is $m = 8$. Although the specific percental occurrence changes with $w$, the maximum of the contribution is the same.

If we higher the strength of the interaction further the chain adopts the globule state. The

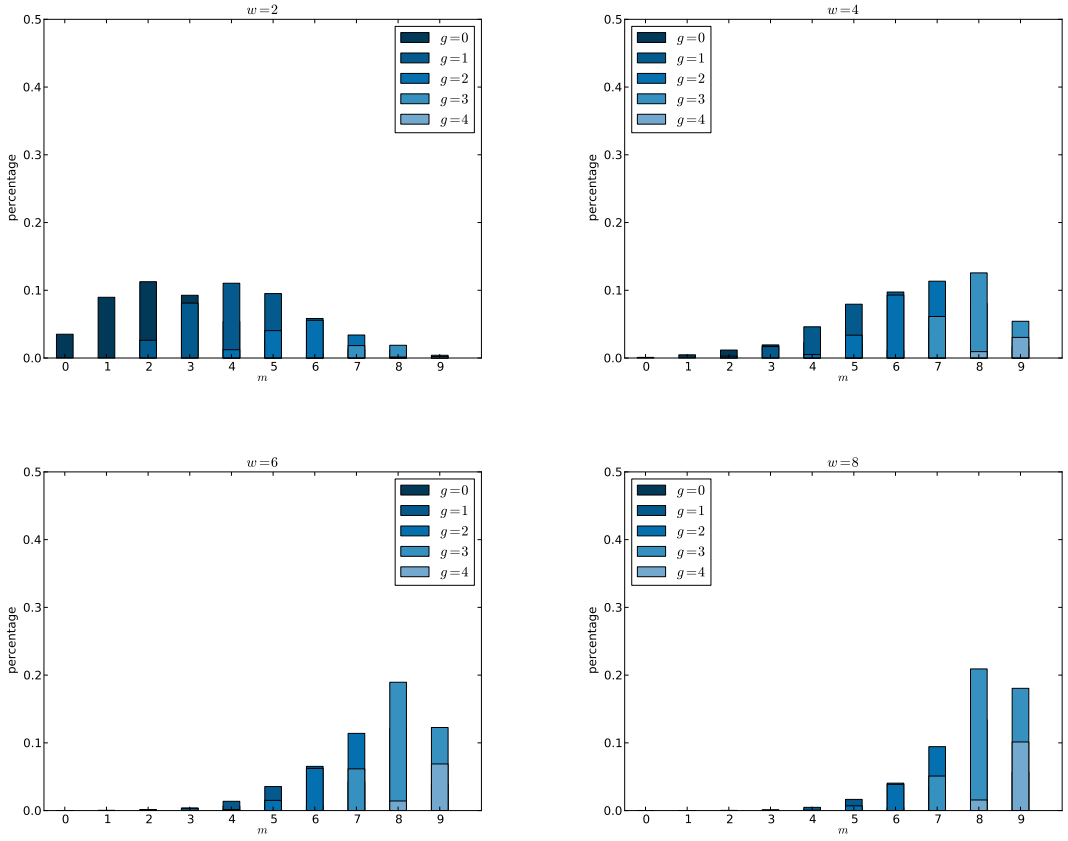



Figure 6.4: Genus occurrence versus $m$ for a chain with length $N = 12$ and different interaction strength $w$ in the transition region. After the genus distribution gets more evenly, every plot has the same maximum value independent on $w$.

corresponding diagrams are shown in Fig. 6.5. Here configurations with higher genus and more interaction pairs are favored. Again this is in accordance with our expectations. Since the form of the chain in the globule state is very crumpled, that is to say very dense, almost every site has at least one interaction partner and pseudoknot configurations are very likely.

Vernizzi et al. [21] achieved similar results. They used short range interaction model as well. However, they used a spin model, where every site has a certain spin and only adjacent sites with spins that point to each other can interact. Since they used Monte Carlo simulation they used chains that are a a lot longer ($N = 500$) which makes it hard to compare. For $N = 20$ however they obtained a averaged higher genus for the globule state and a low one for the coiled one.

### 6.4.2 Microcanonical point of view

Next we will discuss the features of our model in the microcanonical picture. No microstate is weighted, $w = 1$, and the genus distribution depends solely on the number of interactions in the

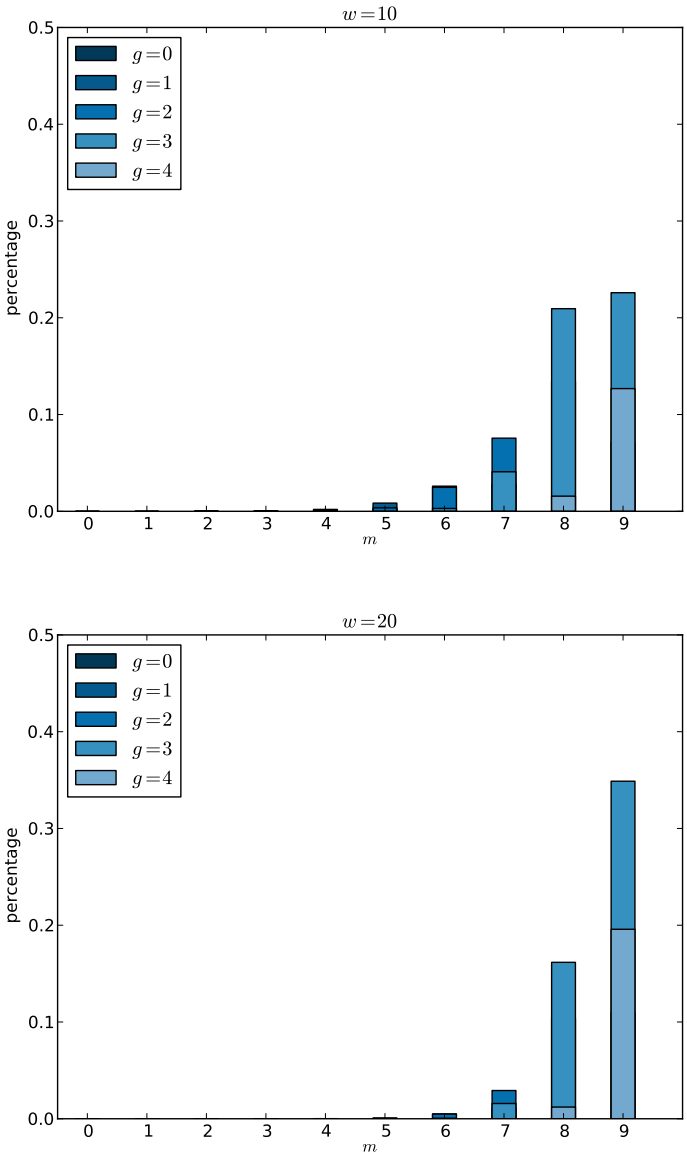



Figure 6.5: Genus occurrence versus $m$ for a chain with length $N = 13$ and interaction strength in region of globule state. Configurations with higher genus are favored.

polymer. The corresponding plot is shown in Fig. 6.6. It clearly shows the suppression of higher genus in the configuration. Even for a bigger number of interaction the highest genus value is $g = 2$. There are two ways to interpret this result. Firstly, one can think about a system with a great quantity of chains and observe it at one moment in time. Most of the chains will have a configuration with low genus, only a few will show pseudoknot formations. Otherwise one could look at one chain and study it's conformation over a long time range. The chain mainly will have a form with low genus, even a planar form. From time to time it will adopt a conformation with more

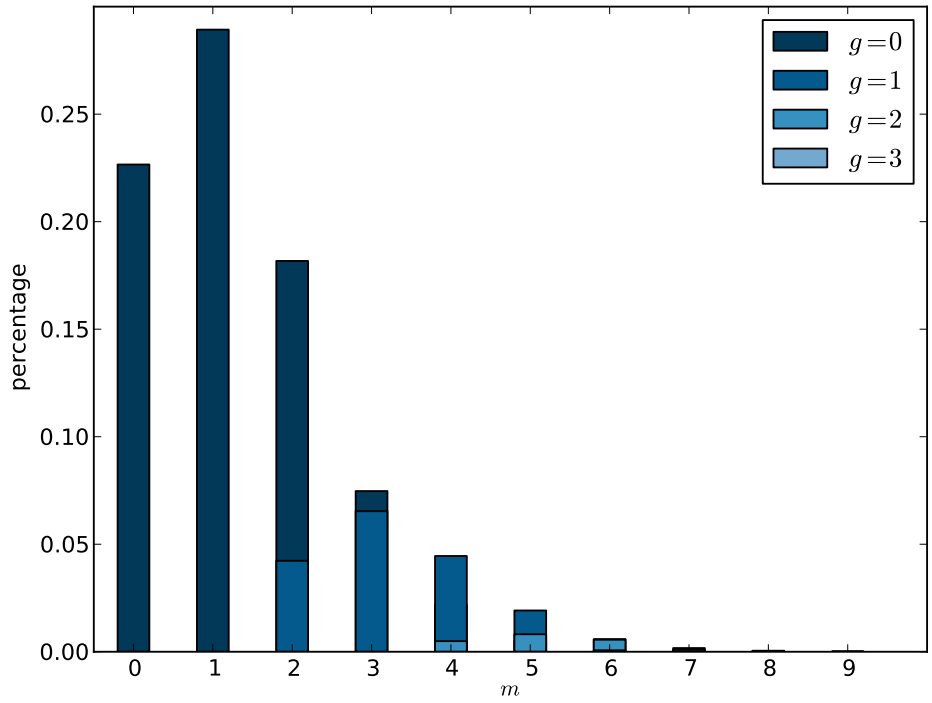



Figure 6.6: Occurrence of the different genera depending on the number of interactions in the chain for a 13 sites long chain with $w = 1$. Higher genus configurations are clearly suppressed.

pseudoknots, but the appearance of this forms are comparable rare and strongly suppressed.

A polymer chain, disregarded of the solvent, will most of the time have a configuration with low genus, that is to say with almost now pseudoknots or even a planar structure. Configurations with more complex structure and higher genus appear, but they are rare. This is a reasonable observation since structures without pseudoknots are energetic more favored. Pseudoknots restrict the possibles configuration of the chain and thus lower the entropy. The same result was observed by Mukhopadhyay et al. [24]. They used a model where the sites were temporary fluctuating between active and inactive states, whereby only two active states can interact and multiple bonds are disallowed. In the determination of long-range enhancer-promoter interaction by chromatin looping they observed that configurations with many contacts are relatively rare.

The results in the last two sections are in accordance with the expected properties of a self interacting polymer chain and thus not surprising. In the next section we will explore if our model yields to configurations with comparatively small genus distributions.

### 6.4.3 Comparison to null model

From experimental results we know that inter-domain interaction in chromosome is suppressed (cf. chapter 2). In order to describe this feature, the chain conformation achieved with our interaction model should preferable have planar structure. To see if our model has this property we will compare the genus distribution to one achieved from random placement of bonds (null model). We used Monte Carlo simulations to create rainbow diagrams where we placed the interacting pairs randomly on the chain. The choice of the interacting pairs is not related to their spatial distance or the chain configuration. With this comparison we want to find out if our model is particularly suitable to describe chromosome structure, i.e. if it features low genus contributions. Fig. 6.7 shows the genus distribution for both models depending on the amount of interaction. We used the microcanonical point of view, which means that we did not weight the different eigenstates ($w = 1$).

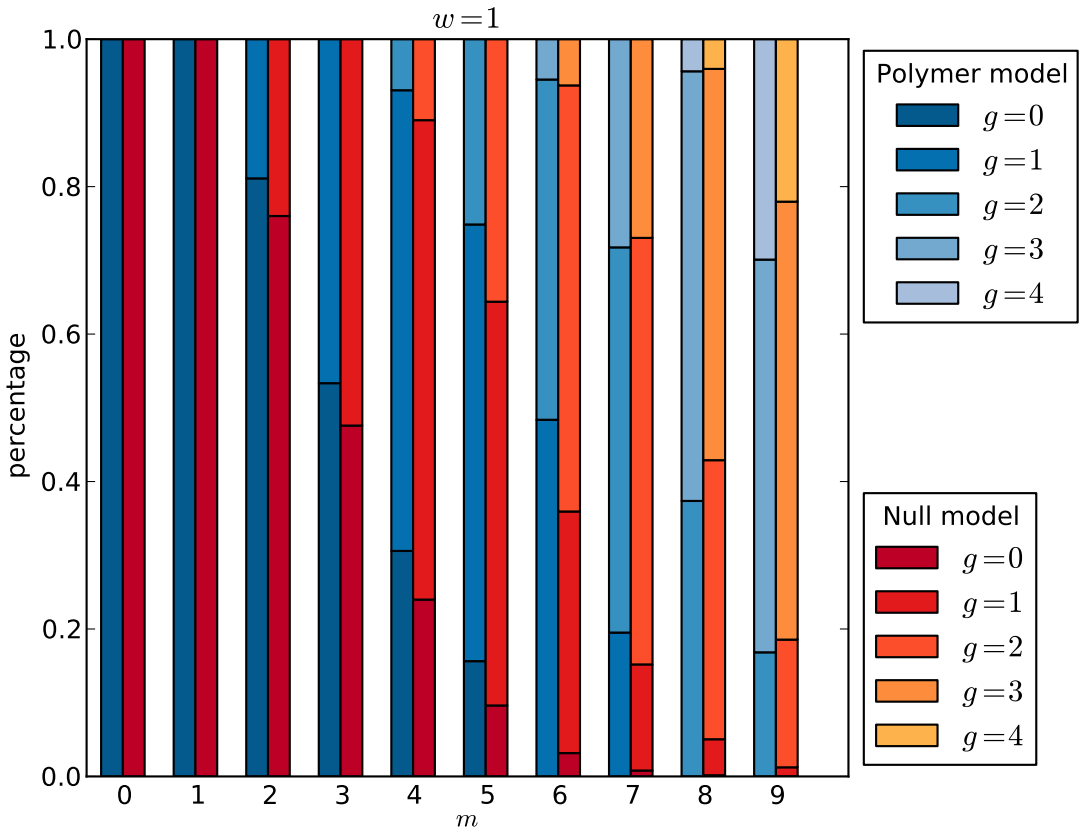



Figure 6.7: Comparison of the genus distribution depending on the number of interactions for the interacting polymer model and Monte Carlo simulations with randomly distributed bonds. Polymer model shows moderate high genus suppression for intermediate interaction strength $m$.

As we can see from Fig. 6.7, the interacting polymer model, in comparison with the null model,

favors lower genus for small number of interacting pairs. This however changes for rising $m$. From $m = 6$ on, our model has in average higher genus than the Monte Carlo simulated ones. The high genus expression in the polymer model is moderate for intermediate interaction strength $m$, but still lower than expected. There are multiple ways to improve our model. We will discuss this in the next section.

### 6.4.4 Discussion

In our preliminary comparison with the random bond placements the high genus suppression is moderate for intermediate contact numbers ($m \approx 4, 5$). There are a lot of different reasons why the suppression is lower than expected. One is that the used polymer chains are relatively short. There are however options to improve this issue. Schram accomplished exact determinations with $N = 36$ long chains. Although the average genus increases with increasing $N$ (cf. Vernizzi et al. [21]), the pseudoknots due to interaction between pairs that are far separated along the chain decreases. The main contribution to the genus for longer chains is obtained by short crossing, i.e. interaction crossing from pairs that are not far separated in the chain (Kabakcioglu and Stella [25]). By separating a long chain into two halves, Kabakcioglu and Stella [25] found that the number of pseudoknots, that are formed between the two halves of the chain diverges logarithmically with $N$, whereas the total number of pseudoknots increases linearly in $N$. To explore this behavior, chains with $N = 12$ steps are too short.

Another point is that we consider multiple bonds for one site. The lack of bond-saturation leads to a more complex interacting structure and features crossing chords. The number of pseudoknots is higher than in models where every site can only form one bond. Mukhopadhyay et al. [24] and Vernizzi et al. [21] used both one-bond interaction models for their numerical analysis. It is difficult to compare our results with the one obtained in their method since the results are determined by Monte Carlo simulations and the chains are accordingly long. It is however reasonable to expect lower genus configurations for one-bond models.

Furthermore we did not take elasticity effects into account. Stiff polymer, like DNA, have a rodlike shape and are not as flexible as our bead on the string model. This constrains the spacial structure of the chain and results in a more open confirmation what also reduces the number of pseudoknot. A more realistic model has to include stacking energies as well. All this points require further study and an improvement of the used model to build a more adequate description for chromosome.

# Chapter 7
# Conclusion

In this thesis we explored the topology of a polymer chain with nearest neighbor interaction. The topology of chromatin plays an important role in the gene regulation process and experimental results allow us to test the validity of physical models. With the application of polymer physics, one can build a physical framework for the description of chromatin. Aim is to find an interacting polymer chain model that features intra-domain interaction and thus confirms the chromatin looping model, a model that can explain long-range enhancer-promoter interaction as well as topological associated domains. We used the backtracking algorithm to generate SAWs at a cubic lattice and added attractive interaction, i.e. interaction between sites that are nearest neighbors in the lattice, but not adjacent in the chain. Using the backtracking algorithm enables us to achieve exact results.

With this model we could explore the coil-globule transition and the spatial structure of polymers with short lengths. For weak attractive interaction between the sites the chain adopts an open structure. Through increasing the attraction, the end-to-end distance in the chain becomes smaller and the chain collapses to the globule state. The topological structure was specified by the genus. We could show that the preferred configurations in the coil state are those with low genus, i.e. planar structures. Conformations that lead to pseudoknots are unfavored. However, as the interaction strength rises and the chain adopts the globule state, conformations with higher genus are favored. If we regard the chain independent of the different states, the polymer in our model mainly adopts contributions with low genus. Higher genus conformations appear, but they are relatively rare. This results are in accordance with the results obtained in other studies ([24], [21]) and our expectations. By regarding the genus distribution for every value $m$, we found that high genus suppression in our polymer model, in comparison to one achieved by considering interactions randomly placed in the chain, without regarding the spacial structure of the polymer, is moderate, especially for intermediate interaction strength.

There are different reasons that lead to a lower than expected high genus suppression, such as chain length, no bond-saturation, no elasticity, etc. Still, there are multiple ways to improve our model which enables us to study the topological structure further. The basic mechanism, i.e. classifying the polymer structure by their topological invariant, the genus, depending on the number

of interactions in the chain, can be used and gives us lots of possibilities for further investigations.

# Bibliography

[1] Graziano Vernizzi and Henri Orland. Large-n random matrices for rna folding. *ACTA PHYSICA POLONICA B*, 36(9), 2005.

[2] An Jansen and Kevin J. Verstrepen. Nucleosome positioning in saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, 75, 2011.

[3] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harb Perspect Biol; 2:a003889*, 2010.

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

[5] Michael R. Speicher and Negel P. Carter. The new cytogenetics: Blurring the boundaries with molecular biology. *Nature Reviews Genetics*, 6:782–792, 2005.

[6] Andreas Bolzer et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, 3:e157, 2005.

[7] Jesse R. Dixon et al. Topological domains in mammalion genomes identified by analysis of chromatin interactions. *Nature*, 4854, 2012.

[8] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *J. Phys. G: Nucl. Part. Phys.*, 24:1061–1076, 1998.

[9] Elphege P. Nora et al. Spatial partitioning of the regulatory landscape of the x-inactivation center. *Nature*, 485, 2012.

[10] Nele Gheldof et al. Cell-type-specific long-range looping interactions identify distant regulatory elements of cftr gene. *Nucl. Acids Res.*, 38, 2010.

[11] Mark Wijgerde, Frank Grosveld, and Fraser Peter. Transcription complex stability and chromatin dynamics in vivo. *Nature*, 377, 1995.

[12] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7, 09 2006.

[13] Amartaya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489, 09 2012.

[14] Iwao Teraoka. *Polymer Solutions: An Introduction to Physical Properties*. John Wiley & Sons, Inc., 2002.

[15] I. M. Lifshitz, A. Yu. Grosberg, and A. R. Khokhlov. Some problems of statistical physics of polymer chains with volume interaction. *Reviews of Modern Physics*, 50(3), 1978.

[16] A L Kholodenko and Karl F Freed. Coil-globule transition: comparison of field theoretic and conformational space formulations. *J. Phys A: Math. Gen.*, 17:2703–2727, 1984.

[17] Marcia G do Amaral. $\phi_2^6$ theory on the lattice - an effective potential study. *Nature Reviews Genetics*, 14, 2013.

[18] Hatem Widyan. Bubble formation in $\phi^6$ potential. *arXiv:quant-ph/0902.0885*, 12 2008.

[19] Gabriel H. Flores, Rudnei O. Ramos, and N.F. Svaiter. Tunneling and nucleation rate in the $(\frac{\lambda}{4!}\phi^4 + \frac{\sigma}{6!}\phi^6)_3$. *ar$\chi$iv*, 02 1999.

[20] Michael Bon, Graziano Vernizzi, Henri Orland, and A. Zee. Topological classification of rna structures. 24:1061–1076, 1998.

[21] Graziano Vernizzi, Paolo Ribeca, Henri Orland, and A. Zee. The topology of pseudoknotted homopolymers. *arXiv:q-bio/0508042v2*, 2005.

[22] Gerard 'T Hoof. Large n. *arXiv:hep-th/0204069v1*, 04 2002.

[23] Raoul Schram. Exact enumeration of self-avoiding walks. Master's thesis, University of Utrecht, September 2011.

[24] Swagatam Mukhopadhyay, Paul Schedl, Vasily M. Studitsky, and Anirvan M. Sengupta. Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators. *PNAS*, 108, 12 20118.

[25] A. Kabakcioglu and A. L. Stella. Pseudoknots in a homopolymer. *arXiv:cond-mat/0312515v1*, 12 2003.