DISSERTATION

submitted to the

Combined Faculties for the
Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg
Germany

for the degree of
Doctor of Natural Sciences

put forward by

M.Sc. Gabriell Máté

born in: Huedin

Date of oral examination: 19.12.2013

Graphical and Topological Analysis of the Cell Nucleus

Referees: Prof. Dr. Dieter W. Heermann
Prof. Dr. Michael Hausmann

Abstract

It is well known that our genetic material influences our tendency to develop certain conditions. Finding the causes behind these predispositions assumes the understanding of mechanisms handling and maintaining the genome. While the problem is important from the biological point of view, being one of the basic riddles of life, it also poses interesting questions which may only be answered by physics. Topics include transport, reaction-diffusion, polymer physics, equilibrium and non-equilibrium dynamics and chaos, amongst others. Experimental techniques, like microscopy or molecular biology approaches provide an ever improving insight in the structure of the nucleus, however, computational and modelling approaches are still needed to explain unknown aspects of genetics.

In this thesis we tackle the problem of understanding the structure of the nucleus from the two opposite sides of the experimental "blind-spot". We develop alternative image modelling and analysis tools which are able to capture and recreate the "large scale" density patterns observed in confocal microscopy images of the nucleus. For this, we introduce a generalized Potts model which is extensively analysed also from the statistical mechanics point of view. Furthermore, we apply statistical mechanics and graph theory calculations to study patterns registered with super resolution microscopy techniques. We investigate the effect of irradiation and light stress on the structure of the chromatin, and are able to quantitatively support prior experimental observations regarding structural changes.

Understanding the interaction and classification of proteins, structures which perform vastly different functions on molecular scales, is also important to achieve the final picture. We contribute to this by elaborating a framework to assess topological similarity among these chemicals. Our approach is based on recently developed computational topology algorithms used to calculate fingerprints of the molecules. We discuss three different modifications of the framework and investigate them on real-world datasets. In addition, we recognize that the mentioned fingerprints can be used to calculate the fractal dimension of certain objects, and offer an intuitive explanation for the observed relation.

Zusammenfassung

Wir wissen, dass unser genetisches Material die Tendenz zu bestimmten Entwicklungen beeinflusst. Um die Gründe für diese Veranlagungen herauszufinden bedarf es dem Verständnis der Steuerungsmechanismen, welche das Genom aufrecht erhalten. Dieses Problem ist als eines der grundlegenden Rätsel des Lebens aus biologischer Sicht sehr wichtig. Jedoch stellen sich hierbei auch interessante Fragen, die nur mit Hilfe der Physik beantwortet werden können. Die Themen umfassen unter anderem Transport, Reaktions-Diffusion, Polymerphysik, Gleichgewichts- und Nichtgleichgewichtsdynamik sowie Chaos. Experimentelle Methoden wie Mikroskopie oder molekularbiologische Ansätze bieten zwar immer weiter verbesserte Einsichten in die Struktur des Zellkerns, jedoch sind auch Computermodellierungsansätze noch immer erforderlich um unbekannte Gesichtspunkte der Genetik zu erklären.

In dieser Dissertation behandeln wir das Problem des Verständnisses der Zellkernstruktur von zwei konträren Seiten der experimentellen "Lücke". Wir entwickeln alternative Werkzeuge zur Bildverarbeitung und -analyse, die in der Lage sind, Dichtemuster auf der großen Skale, die durch Konfokalmikroskopie des Zellkerns beobachtet werden, zu erfassen und zu reproduzieren. Darüberhinaus wenden wir Berechnungen der Statistischen Mechanik und der Graphentheorie an, um aus Superauflösungsmikroskopie gewonnene Muster zu analysieren. Wir untersuchen die Auswirkungen von Strahlung und leichter Spannung auf die Chromatinstruktur und sind in der Lage, quantitative Bestätigungen für die experimentellen Beobachtungen struktureller Veränderungen zu erbringen.

Das Verständnis für die Wechselwirkungen und Klassifikation von Proteinen, also Gebilden, welche eine große Bandbreite unterschiedlicher Funktionen auf molekularer Skala ausführen, ist ebenfalls wichtig, um ein Gesamtbild zu erreichen. Wir tragen dazu durch die Ausarbeitung eines Frameworks zur Untersuchung von topolgischen Ähnlichkeiten zwischen diesen chemischen Einheiten bei. Unser Ansatz basiert dabei auf kürzlich entwickelten computertopographischen Algorithmen, die für die Berechnung der Fingerabdrücke dieser Moleküle verwendet werden. Wir diskutieren drei verschiedene Modifikationen für dieses Framework und verwenden es für reale Datensätze. Zusätzlich stellen wir fest, dass die genannten Fingerabdrücke zur Berechnung der fraktalen Dimension bestimmter Objekte verwendet werden können und bieten eine intuitive Erklärung für die beobachteten Verhältnisse.

Publications Related to this Thesis

Certain parts of this thesis have already been published or are currently under peer-review. Papers in preparation are also listed. (Information as of October 7th, 2013)

- C.J. Feinauer, A. Hofmann, S. Goldt, L. Liu, **G. Máté**, D.W. Heermann, Zinc finger proteins and the 3D organization of chromosomes. *Advances in Protein Chemistry and Structural Biology* **(2013)**, *90*, 67-117.

  DOI: 10.1016/B978-0-12-410523-2.00003-1

- **G. Máté**, A Hofmann, N Wenzel and D.W. Heermann, A Topological Similarity Measure for Proteins. *BBA - Biomembranes* **(2013)**, accepted.

- **G. Máté**, and D.W. Heermann, Statistical Comparison of Topological Features of Proteins. *PLoS One* **(2013)**, under peer-review.

- **G. Máté**, and D.W. Heermann, Persistence Intervals of Fractals. *Physica A* **(2013)**, under peer-review.

- **G. Máté**, R. Dickman and D.W. Heermann, A State Dependent Potts Model. **(2013)**, in preparation.

- **G. Máté** and D.W. Heermann, Simulating Microscopy Images of the Cell Nucleus. **(2013)**, in preparation.

- **G. Máté**, L. Shopland and D.W. Heermann, Spatial Correlation of the Nuclear Pore Complexes and Lamin B Receptors. **(2013)**, in preparation.

- **G. Máté**, M. Tark-Dame and D.W. Heermann, Quantifying Effects of Light Stress in Arabidopsis Thaliana. **(2013)**, in preparation.

- Y. Zhang*, **G. Máté***, P. Müller, M. Hausmann and D.W. Heermann, Measuring Structural Changes in Chromatin Induced by Ionizing Radiation. **(2013)**, in preparation.

Additional Publications

- **G. Máté**, J. Benedek and Z. Néda, Spring-Block Model Reveals Region-Like Structures. *PLoS One* **(2011)**, *6*, e16518.

  DOI: 10.1371/journal.pone.0016518

- **G. Máté***, E.Á. Horvát*, E. Káptalan, A. Tunyagi, Z. Néda and T. Roska Periodicity enhancement of two-mode stochastic oscillators in a CNN type architecture. *Cellular Nanoscale Networks and Their Applications (Proceedings)* **(2010)**

  DOI: 10.1109/CNNA.2010.5430275

*equal contribution

- G.E. Pazienza, K. Karacs, E.Á. Horvát and **G. Máté**, Designing efficient CNN algorithms for the Bionic Eyeglass by combining manual and automatic techniques. *Cellular Nanoscale Networks and Their Applications (Proceedings)* **(2010)**

  DOI: 10.1109/CNNA.2010.5430297

- **G. Máté**, Andrś Kovacs and Z. Néda, Hierarchical Settlement Networks. **(2013)**, under peer review.

- **G. Máté**, and Z. Néda, Defining Fuzzy Region Boundaries. **(2013)**, in preparation.

- **G. Máté**, J. Benedek and Z. Néda, Economic Blocks in the European Union. **(2013)**, in preparation.

- **G. Máté**, J. Benedek and Z. Néda, Grouping of Countries Defined by their Import-Export Ties. **(2013)**, in preparation.

Contents

Chapter 1

Introduction

Curiosity of human nature initiated a journey in discovering the universe a few thousand years ago. While in prehistoric times experience was passed to consecutive generations by oral tradition, the development of writing allowed recording observations and spreading knowledge regarding nature with higher fidelity. This not only enabled what we would consider systematic research, but also contributed to the cultural evolution of societies [1].

Although most of the early observations were recorded to facilitate the life of the communities, humans soon developed the desire to understand nature. Behind this desire, however, there always were pragmatic reasons. In the light of the past millennia, the journey is far from being over, and mankind's hope to make new discoveries and the demand for practical solutions keep pushing the frontiers of knowledge. This synergy resulted in some spectacular developments in the recent decades.

Scientific knowledge proved to be an extremely important factor contributing to the development of most of the contemporary societies. Mathematics and physics promoted the invention of computers which infiltrated in the smallest aspects of our lives. High technology materials developed by material scientists and chemists are built in most of the objects we interact with on daily basis. The focus now is on biology, as it is on the path to revolutionize medicine and extend life expectancy.

The advance of different scientific fields helped us understand how cells, basic building blocks of life function. While a malfunction of cells is easily observed by its symptoms, it is often not obvious what causes the abnormal behaviour. In fact, evidence shows that disorders may have a genetic background. Genetic predisposition is observed not only to diseases like cancer, cardiovascular disorders or diabetes but also to viral infections and to conditions like substance dependence, depression or obesity [2–8]. On the other hand, worm-, fly- and mouse-experiments showed that ageing, known to change the structure and function of the cells [9–11], can be slowed by administrating drugs targeting the genetic material of the cells [12].

Understanding the exact role of each gene, the structure of the genome and the mechanisms handling and maintaining it is thus crucial, and it is on of the biggest challenges the scientific community is facing. The question is complex and can only be solved by putting together the pieces of a huge puzzle. However, many of the pieces are still missing and while a faint idea about the global picture is already formed, details may play a very important role.

Although experiments designed to reveal the structure of the cell nucleus – the organelle enclosing the genetic material of the eukaryote cell – have evolved rapidly, certain aspects of the spatial organization of the genome are still not well understood. Despite recently developed clever solutions to overcome diffraction effects, the resolution of light microscopes is still limited and higher resolution microscopy techniques are rather invasive. While, molecular biology approaches offer detailed information on macromolecular scales, relatively little is known about the folding of the famous *double-helix* DNA fibre – a long macromolecule encoding the genetic information – into *chromosomal structures* [13]. However, computational methods offer possible explanations by fitting experimental data with theoretical models, and have the potential to uncover the structures hidden behind the "blind spot" of the microscopes and other experimental approaches [14–16].

Experimental evidence is strongly underpinning the current belief that the DNA fibre is coiled in a hierarchical fashion, forming loops on different scales [17–25]. The first level of folding is organized by *histone protein cores* around which the DNA wraps about 1.6 times forming *nucleosomes*. Nucleosomes are the basic, repeating units of DNA packing and the roughly 10 nm thick fibre with the repeating nucleosome motif is commonly known as the "beads-on-a-string" structure or the 10 *nm chromatin fibre* [26–28]. What happens on the next level of packing is still highly debated. Nevertheless, models suggest that *entropy* has an important contribution [29–33] and considerations about the 10 nm fibre already offer the basis for theoretical and computational modelling.

1.1 Scope

Polymer physics offers excellent tools for studying such fibres both from theoretical and computational perspectives. Polymer models have been applied successfully to reproduce and thus explain a variety of experimental observations [34–38].

On the other hand, it is relatively hard to compare configurations generated by polymer models to confocal microscopy images. *Monte-Carlo computer simulations*, for instance, generate an ensemble of configurations. None of the configurations represents the system by itself, quantities can only be measured in terms of averages over the ensemble. Therefore, a direct comparison is not possible. Furthermore, computer simulations are costly and polymer representation of the chromatin fibre must be *coarse-grained* especially for large genomes [39–41].

Microscopy images are usually analysed within a generic *image processing* framework. Images are segmented to detect different objects, then quantitative measurements may be carried out and morphological properties may be determined using specialized computer algorithms [42–46]. However, this task is relatively difficult when the concept of "object" is not well-defined. This situation is obviously faced in the case of nucleus microscopy images, unless certain specific loci are marked and the markers clearly define the objects interesting for the study at hand.

One of the major focus-points of this thesis is to investigate alternative possibilities for analysing and modelling the spatial density distribution observed in confocal and other high resolution microscopy images of the cell nucleus. Our methods will be based mainly on *statistical mechanics*, *graph theory* and *probability theory*. Some of the used approaches belong to the framework of *graphical models*, a toolbox which takes advantage of an interplay between the aforementioned fields.

Besides investigating and modelling the biophysics of chromatin folding, the organization and functioning of the chromosomes is also tackled on the molecular scales. A

huge amount of proteins – large molecules performing a vast amount of functions, and therefore being extremely important – have been segmented and their structure is stored in public databases [47]. To determine relations among these molecules, the structures are then compared and classified with respect to their shape, sequence, function or other criteria [48–56]. Flexibility turned out to be an extremely important feature of proteins as it may influence binding affinity among other properties [57]. Still, flexibility is often handled in a mechanistic or heuristic fashion by most classifiers.

Breaking with tradition, instead of comparing proteins based on calculated alignments or heuristic features, we propose a method which builds on recently developed *computational topology* algorithms to characterize molecules in terms of *topological invariants* – numbers which stay constant when structures are changed without altering a chemical bond. The proposed framework is therefore an excellent alternative for currently existing methods, handling flexibility in a "native" and mathematically rigorous manner.

During the development of the ideas related to image analysis and topological similarity, several spin-off projects were also exploited: For instance, the model used to describe confocal microscopy images turned out to be very interesting from the statistical mechanics point of view and its thermodynamic properties were extensively studied. Furthermore, we investigated and explained the relation of the used topological invariants to the *fractal dimension* of objects. These spin-offs and the related results will also be presented alongside the projects which inspired them.

1.2 Structure

The thesis is divided into four parts.

First part In the first part, consisting of two chapters, we sketch the biological and experimental background and lay down the theoretical foundations of the used methods.

Second part In the second part we describe our approach to the modelling of the spatial density distribution of the chromatin observed in confocal microscopy images of the cell nuclei. Our method is based on the generalization of the Potts model. This being part of the larger family of Markov random fields, we benefit from the tools developed in the latter framework. Markov random fields provide the necessary tools to learn the parameters of the model directly from microscopy images. After we ensure that the learning works properly, we train our model on real world data.

On the other hand, we investigate the generalized Potts model from the statistical mechanics point of view in the same chapter. We devise a mean field theory which helps us to gain an insight and make predictions regrading the behaviour of the model. We control and support our predictions with computer simulations.

Third part In the third part we describe the methods we applied to analyse the structure of the nucleus and the chromatin either by graph theoretic approaches or by characterizing the images with intensity distributions and their different momenta. We apply scaling of the distributions to fit them to each other and to decide whether they belong to the same family of distributions or not.

In particular, we investigate the effects of the ionizing irradiation over the structure of the chromatin. We find that while the overall local properties do not change, heterochro-

matic regions open after irradiation. We also observe that when enough healing time is allowed to pass before fixation, this decondensation is reversed.

In the same chapter we analyse the spatial relation of the nuclear pore complexes and lamin B receptor proteins. Our most important finding in this analysis shows, while lamin B receptors prefer to be located in spatial proximity of the pore complexes, the position of the nuclear pore complexes are not influenced by the locations of the lamin B receptors.

Fourth part In the last part of this thesis we present methods for assessing similarity between proteins. The methods are based on recently developed computational topology algorithms which are used to calculate persistence intervals, certain topological invariants, of the chemical structures. We introduce three different approaches and apply them on real world data.

Furthermore, we recognize that if persistence intervals indeed characterize the topology, then, they should be able to capture properties like fractality. We derive an intuitive relation between the fractal dimension of object and the structure of persistence diagrams and use the diagrams to estimate the fractal dimension of certain objects with high precision.

While the author of this thesis took part in different collaborations, most of the results presented here constitute his own work produced following the guidance of the principal investigators associated with the different projects. This completely excludes the preparation and the acquisition of the experimental data. Other contributors are acknowledged in the reference boxes.

Chapter 2

Biological and Experimental Background

A Short Survey

In this chapter, we will introduce the biological and experimental concepts important from the point of view of this thesis. Since the thesis is more inclined towards a theoretical work, the needed knowledge in the mentioned two fields is not extensive. In turn, we will spend more time (and text) on the mathematical and physical concepts required to understand our work.

The two main topics that will be presented here are the structure of the nucleus and the microscopy techniques used to capture image. Microscopy images are the main data source laying at the heart of the analyses we will perform to validate our methods.

2.1 Structure and Function of the Eukaryote Cell Nucleus

The existence of membrane bound organelles, such as the nucleus, distinguish Eukaryotic cells from Prokaryotes [58]. The nucleus (Figure 2.1), usually the largest organelle, plays an important role in storing, protecting and utilizing genetic information [59]. The nuclear envelope forms a boundary between the cell's cytoplasm and the content of the nucleus, a viscous and amorphous mass of material [59]. The nucleus of an interphase cell contains the chromosomes in the decondensed state, in form of highly extended and entangled nucleoprotein fibres, often referred to as *chromatin* [59]. Besides enclosing the chromosomes, the nucleus contains one or more *nucleoli*, organelles playing a role in ribosomal RNA synthesis [59]. The genetic material and proteins interacting with it are suspended in the *nucleoplasm*.

2.1.1 The Nuclear Envelope

The appearance of the nuclear envelope can be considered a landmark in biological evolution [59]. Its presence and role of separating the genetic material from its surroundings is the most important distinction between eukaryotic and prokaryotic cells. The envelope consists of two parallel membranes separated by a distance of 10 – 50 nm [59]. It prevents ions, solutes and macromolecules from travelling freely between the cytoplasm and nucleus.

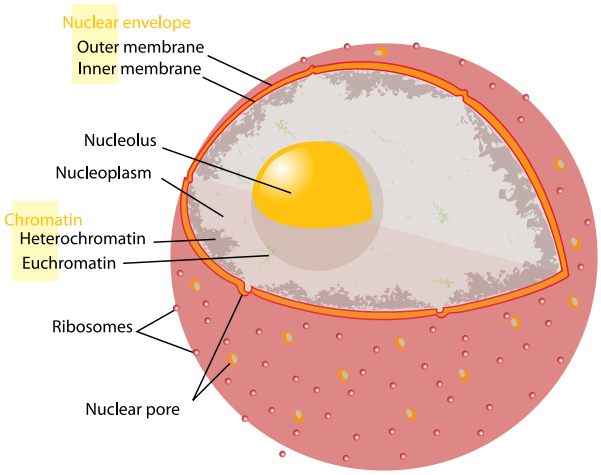

Figure 2.1: Diagram of the nucleus of a human cell.

The two membranes of the envelope fuse at certain locations to embed *nuclear pore complexes* [59], assemblies of different proteins which regulate the exchange of chemicals between the cytoplasm and the nucleus [59]. Mammalian cells contain several thousand of pore complexes on average [59].

The inner part of the nuclear envelope serves as attaching sites of integral membrane proteins [60]. These membrane proteins bind to a filamentous *nuclear lamina* mesh [59]. The lamina mechanically supports the nucleus and creates attachment sites for the chromosomes [59].

2.1.2 The Chromosomes in Interphase

The human cell contains roughly 6400 million base pairs of DNA [59]. This amount is shared among 23 pairs of chromosomes. Each chromosome contains a single DNA molecule. The total cumulative length of these molecules is about 2 m [59]. It is remarkable that a molecule of this length fits into a container with a diameter of about 8 – 10 μm. While during cell division the chromatin condenses to the well-known x-shaped structures, the genetic material is in a diffuse, decondensed state for most part of the cell cycle [61].

What is Known About the Packing of the DNA

The lowest level of packing is governed by groups of *histone proteins* forming *nucleosomes*. A nucleosome core particle (Figure 2.2) consists of 146 base pairs of DNA wrapped 1.67 times around a group of 8 histone proteins [62]. The diameter of a nucleosome is roughly 10 nm [59]. The DNA fibre wrapped around the regularly spaced nucleosomes create a structure known as “beads-on-a-string” with a packing ratio of 5–10 to 1 [63].

Relatively little is known about the higher level structure of the chromatin fibre. When chromatin is studied in vitro, a fibre with a diameter of approximately 30 nm is observed [59]. However, experimental evidence is scarce for the existence of this packing stage of the chromatin inside a living cell nucleus [64] and therefore, the existence of the 30 nm fibre is highly debated. From the physics point of view its existence is also doubtful as it supposes

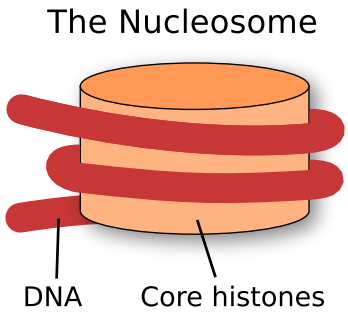

Figure 2.2: Schematic diagram of a nucleosome core particle.

a crystal-like structure. Considering that experiments indicate that the mechanisms inside the nucleus are highly dynamic [65], the stability of such a crystal is questionable.

Regardless of the existence of the 30 nm fibre, experimental evidence supports that the chromatin is organized in a loopy structure [64, 66–68]. Loops form on different scales and it has been shown that entropy plays an important role in the process [68].

On the next scale, different chromosomes occupy different domains inside the nucleus [69]. Experiments show that the arrangement of chromosomes in chromosomal domains is not only a result of geometric constraints [69–72]. Furthermore, highly expressed regions were found to be located significantly closer to the centre of the nucleus while regions with lower expression levels, associated with densely a packed chromatin, were found closer to the envelope [70]. Historically, these denser regions were termed *heterochromatin* however according to a more precise definition, the formation of heterochromatin requires the methylation of histone H3 at lysin 9 (H3K9me) and it is associated with transcriptional silencing and repression of recombination [73].

Alteration in the structure of the nucleus, the condition of the envelope or the organization of the chromatin, has been linked to numerous diseases [74–77]. Therefore it is crucial to understand the structure and the packing of the DNA. Moreover, besides being an important topic from the point of view of biology and medicine, the problem of understanding how the genome functions also poses interesting questions for physicists. As already pointed out, the inside of the nucleus is highly dynamic, creating a stage for a vast amount of physical processes and understanding them may lead to exciting discoveries.

Defects of the DNA

Lesions in the DNA might be caused by many mechanisms and environmental agents. UV and ionizing radiation, the presence of genotoxic chemicals, byproducts of normal metabolic processes are all amongst the possible causes [78]. Besides these factors, the process of DNA replication is prone to produce errors [79]. In fact, it had been estimated that the DNA of individual cells may undergo as many as one million DNA changes per day [61]. Single strand breaks, missing bases, chemically changed bases are all possible forms of lesions. However, the most harmful damages are perhaps the double strand breaks (DSB) [80].

DSBs are usually caused by ionizing radiations and there are two important repair pathways which may be triggered to remedy these errors: homologous recombination and non-homologous end joining [81]. The homologous recombination uses genetic information stored in genetically identical or very similar molecules, most often the sister chromatid, to repair the damage. Non-homologous end-joining, in turn, rejoins two broken ends of DNA

in a sequence independent way. The rejoining, however, is often facilitated by short, 1 – 6 base-pair regions (microhomologies) at the end of the severed DNA fibres [81]. Because in the first phase of the cell cycle (Gap 1 or G1 phase), the chromatin is highly compactified and no redundant genetic information is present, non-homologous end-joining is the main repair pathway in this phase. During the synthesis (S) and Gap 2 (G2) phase homologous recombination is predominant [80].

2.2 Microscopy

Experiments have been an important driving force and the only validation technique in natural sciences, especially when investigating phenomena less accessible to everyday experiences. So it is the case in cell biology, where the dimension of the investigated samples enforce special experimental techniques. Early investigations were possible by the invention of the microscope which dates back to the 16th century [82], although unattested sources report earlier devices. Despite the early developments, improvements of the device were slow until the wave theory of light was developed around the beginning of the 19th century [82]. By the turn of the century, in the early nineteen hundreds, the limiting optical resolving powers had been reached [82].

The resolution of the microscopes is limited by diffraction effects. Because of these effects, a single point light-source will create a circular diffraction image with a bright disk in the centre and concentric alternating bright and dark disks around it when imaged on an optical system with circular elements. This pattern is known as the *Airy diffraction pattern* while the bright disk in the centre is the *Airy disk*. According to the *Reyleigh* criterion, the resolution can be given as the radius of the first dark ring in the Airy pattern and it can be approximated by the *Abbe* equation as

$$r_A = 0.61\frac{\lambda}{A_N},$$

where λ is the wavelength of the light and A_N is the numerical aperture [83]. Two points in the image can be resolved if the distance between the source of the points is larger then the *Airy radius* r_A. If the optical system is more complicated, the pattern generated by a point light-source may differ from the Airy pattern and the observed three dimensional intensity distribution is called the point spread function [84].

In standard light microscopy samples are illuminated by a light-source and images are formed gathering the light scattered by the samples. However, this results in recording light scattered from all the molecules suspended in the sample, and the obtained images might be rather blurry. *Fluorescence microscopy* remedies this by recording the fluorescent light emitted by *fluorophores* – fluorescent molecules [85, 86]. Some fluorophores may be naturally present in the studied samples, others might have been engineered to tag specific regions or attach to given loci along the DNA [87]. This technique, thus, enables the sharper imaging of the investigated structures capturing contour of fluorophores on a very dark background. To trigger fluorescence, fluorphores are excited with high energy photons, usually in the UV spectrum. Since the emitted fluorescent light has a longer wavelength than the light used for excitation, the different wavelength can be separated using spectral emission filters, thus the emitted light can be recorded separately [85].

Technologies to improve image contrast were emerging by the middle of the 20th century. The interest of biologists was aroused by the developing of confocal microscopy. This technique utilizes point-by-point imaging: instead of illuminating the whole specimen, it

illuminates and images only a single point also blocking the light coming from other areas of the specimen by allowing the light to pass only through a pinhole in the focal plane of the objective [88]. By this arrangement the setup is able to produce much sharper images, the resolution depending on the wavelength of used light and is typically about 200 nm [88].

Because the confocal setup filters and blocks any light source which is not in the focus plane of the objective, a quasi three dimensional reconstruction of the sample is possible [88]. The resolution in the axial dimension is limited by diffraction effects and it is the range of a few hundred nm, also depending on the wavelength of the utilised light [83].

In the last decades different techniques were developed in order to further improve the resolution of microscopes. These techniques are collectively know as super-resolution microscopy methods [89–92]. Some of these techniques aim to improve the point spread function either by increasing the numerical aperture or by exploiting interference phenomena [89]. Recorded data are post processed using computerized image processing techniques to obtain the high resolution images.

One of the super-resolution techniques important from the point of view of this thesis is termed localization microscopy. This type of microscopy can be approached with different experimental solutions, however, the idea behind all approaches is to optically isolate fluorophores and avoid the overlapping of Airy disks [92]. In Spectral Position Determination Microscopy (SPDM), a particular type of localization microscopy, this is achieved by using fluorophores which can be bleached in two ways. Besides the regular irreversibly bleached state, these fluorophores can enter a reversibly bleached state with a certain probability. From this reversibly bleached state the fluorophores can stochastically re-enter the fluorescent state in which they emit light for a few milliseconds [92]. Through this process an uncorrelated fleshing of the fluorophores is achieved and if appropriate time intervals are allowed between recording diffraction images, the distance between the molecules recorded in a single image becomes large enough to achieve optical isolation [92]. Applying this technique, a resolution of a few nanometres can be achieved on the surface of the samples, while the best resolution in middle of thicker samples is about 30 nm.

Shortly, this is the knowledge we will need to motivate certain approaches and to interpret our findings. Along the thesis, we will complete the presented information, in case this is needed. Now let us proceed to lay the foundations of our approach.

Chapter 3

Methods and Tools

Theoretical Introduction

In this chapter we will review the theoretical principles laying at the foundation of most of the analysis we carry out in this work. We will use a variety of tools, mainly from the fields of physics (especially statistical mechanics), graph theory and computational topology. We will also describe the framework of *graphical models*. The term graphical model is a "cover name" of many different models stemming from different fields, having the common denominator of utilizing graph theory and probability theory to model complex phenomena. The concept of graphical models is going to be important for one of the centre-piece approaches of this thesis.

3.1 Physics

Over the centuries physics helped us to understand the world surrounding us. The Archimedes principle, heliocentrism, the conservation of matter and energy, the wave theory of light and the atomic theory of matter, X-rays and the discovery of the electron are all among the discoveries that helped developing the present day understanding of the universe.

Around the end of the 19th century, a special field of physics, statistical mechanics emerged. Statistical mechanics aims to predict macroscopic properties of a system by studying the structure and interaction of its microscopic components [93].

3.1.1 A System in the Canonical Ensemble

Starting from considerations regarding a system S containing N particles enclosed in a volume V, immersed in a (infinitely) large heat reservoir S' at a temperature T, it is possible to derive the Boltzmann distribution which is central to many statistical mechanics applications and defines the probability $p(E_i)$ of finding the system in a state i characterized by an energy E_i [94]. The idea behind the derivation of the Boltzmann distribution is to represent the system $S + S'$ with innumerably many identical copies of itself, that is with an *ensemble* of systems and to think about the original system $(S + S')$ as the average behaviour of this ensemble. While fixing the state E_i of S_i, any particular copy is allowed to be in any of the allowed configurations of S' compatible with the state of

S. From other considerations of statistical mechanics we know that any configuration is equally likely [95]. As a consequence, the probability of a state with a given energy E_i will be proportional to the number of configurations of S' compatible with the energy E_i [94]. Assuming thermal equilibrium in $S + S'$, the expansion of the logarithm of the number of these configuration yields the proportionality relation

$$P(E_i) \sim e^{-\beta E_i}, \tag{3.1}$$

where, according to the definition, $\beta = 1/k_B T$, k_B being the Boltzmann constant [95]. The right hand side of the proportionality is known as the *Boltzmann factor*. Normalizing (3.1) we get the Boltzmann distribution

$$P(E_i) = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}}. \tag{3.2}$$

The summation in the denominator is over the possible states/configurations of the system and is called the partition function, being often denoted by Z

$$Z = \sum_j e^{-\beta E_j}. \tag{3.3}$$

It can be shown that the partition function Z characterizes the system in thermodynamic equilibrium as most thermodynamic quantities can be given as functions or derivatives of Z:

$$\begin{aligned} F &= -k_B T \ln Z \\ S &= -\left(\frac{\partial F}{\partial T}\right)_{V,N} \\ P &= -\left(\frac{\partial F}{\partial V}\right)_{T,N} \\ \mu &= \left(\frac{\partial F}{\partial N}\right)_{T,V}, \end{aligned}$$

where F is the free energy, S is the entropy, P is the pressure and μ is the chemical potential [94]. In case of a magnetic system, the magnetic field $\vec{B}$ replaces the volume [94]:

$$\begin{aligned} F &= -k_B T \ln Z \\ S &= -\left(\frac{\partial F}{\partial T}\right)_{\vec{B},N} \\ \vec{M} &= -\left(\frac{\partial F}{\partial \vec{B}}\right)_{T,N} \\ \mu &= \left(\frac{\partial F}{\partial N}\right)_{T,\vec{B}}, \end{aligned}$$

where $\vec{M}$ is the magnetization vector.

The specific heat C and the magnetic susceptibility χ can be calculated as fluctuations of the energy and the magnetization [94]:

$$C = \frac{1}{N k_B T^2}\left(\langle E^2\rangle - \langle E\rangle^2\right) \tag{3.4}$$

$$\chi = \frac{1}{N k_B T}\left(\langle M^2\rangle - \langle M\rangle^2\right) \tag{3.5}$$

$$\tag{3.6}$$

The average value of a physical quantity A can be calculated using the Boltzmann distribution (3.2) as

$$\langle A \rangle = \frac{\sum_i A_i e^{-\beta E_i}}{Z}. \tag{3.7}$$

The framework described here is sometimes referred to as the (T, V, N) ensemble, as these are the quantities that are constant in the system. A more widely used term is *canonical ensemble*.

3.1.2 Phase Transitions and Critical Phenomena

Thermodynamic functions (e.g. χ, C) of systems consisting of interacting microscopic components may exhibit discontinuities under certain circumstances. If the conditions are favourable, microscopic components may interact with each other in a strong, cooperative fashion and their behaviour may be correlated over large domains [96]. Such systems may undergo a *phase transition*.

Phase transitions are common in nature, the condensing and freezing of water being perhaps the most often used examples. They can be characterized by looking at the diverging thermodynamic function which is usually a derivative of the relevant thermodynamic potential (the free energy F in the case of the canonical ensemble). If the function is the first derivative of the free energy (for instance the magnetization M), we are dealing with a *first order* phase transition. If the first derivative is continuous, but the second derivative (like the susceptibility χ, this being a derivative of M, thus second order derivative of F) is diverging we speak of a *second order* phase transition or a *critical point*. If the nth order derivative is diverging but all lower order derivatives are continuous the phase transition is an *nth order* phase transition [97, 98]. This is the so-called *Ehrenfest classification*.

A more modern classification however divides the phase transitions in two groups depending on whether the processes involve *latent heat* or not. According to this classification, the phase transitions can either be *first order* or *continuous*. During a first order phase transition the system releases or absorbs latent heat and this class corresponds to the first order phase transitions of the Ehrenfest classification. No latent heat is involved during continuous phase transitions, corresponding to second, third, etc. order phase transitions in the Ehrenfest classification [99].

Systems can usually be described by an order parameter ϕ which characterizes the ordering of the system. The order parameter is usually the first derivative of the relevant thermodynamic potential (e.g. free energy F in the canonical ensemble). Its value is zero in the unordered phase and non zero in the ordered phase [98]. For magnetic systems, M is often chosen as the order parameter.

As a system in an initially ordered state goes through a phase transition, the order parameter drops to zero. Assuming $\phi = M$, ϕ may drop to zero discontinuously for first order phase transitions and continuously for second order phase transitions, according to a *scaling law* [98]

$$\phi \sim (1 - T/T_C)^{\beta}, \tag{3.8}$$

where T_C is the critical temperature. Similarly to the scaling equation (3.8), the specific heat and the susceptibility also scales according to similar relations [98]:

$$C \sim |1 - T/T_C|^{-\alpha}, \tag{3.9}$$

$$\chi \sim |1 - T/T_C|^{-\gamma}. \tag{3.10}$$

α, β and γ are called critical exponents.

There is yet another quantity important in continuous phase transitions: the *correlation length* ξ. In a magnetic spin system, for instance, where σ_i denotes the state at position $\vec{r_i}$, a correlation function $G(\vec{r})$ can be defined as

$$G(\vec{r_i} - \vec{r_j} = \vec{r}) = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle. \tag{3.11}$$

For large enough distances, the correlation function decays exponentially as

$$G(\vec{r}) = e^{-r/\xi}, \tag{3.12}$$

where $r = |\vec{r}|$. Second order phase transitions can be characterized by diverging correlation lengths ξ [100] and the correlation length also scales around the phase transition according to the relation

$$\xi \sim (1 - T/T_C)^{\nu}, \tag{3.13}$$

where ν is the critical exponent for the correlation length [97]. However this is only true for infinitely large systems. For finite systems, the correlation length is limited by the size of the system and $\xi \approx L$, where L is the characteristic system size. On the other hand, critical exponents are not independent [101]. This allows us to perform *finite size scaling analysis* [102] which is especially useful in computer simulations.

3.1.3 Computer Simulations

Unfortunately, the vast majority of the models is analytically intractable. This is due to the exponentially many terms in the partition function Z which makes calculating averages very difficult if not impossible. *Monte-Carlo* computer simulations are popular tools and they may come handy in such cases. These methods are used in many branches of science (physics [103], chemistry [104], applied mathematics [105], biology [106], psychology [107] etc.), but they proved to be useful also in finance [108], risk analysis [109] or different industrial applications [110–112]. To understand Monte-Carlo simulations, however, first we need to discuss some concepts such as *Markov chains*, *ergodicity* and *balance*.

Markov Chains and Markov Processes

In its simplest definition, a *Markov Process* is a stochastic process which does not have memory and its present state depends only on its previous state. A Markov chain is a time series in which the elements are consecutive configurations generated by a Markov process [113].

For a given finite system S theoretically we could enumerate and index all the possible configurations which may be adopted by the system. Assuming that the time series $(S_t)_{t\in\mathbb{N}^+}$ defines a Markov chain and that the system at a given time t has a configuration k while S_{t+1} corresponds to the configuration l, then there must be a finite transition probability π_{kl} from state k to state l. The Markov process which generated $(S_t)_{t\in\mathbb{N}^+}$ is defined by the *transition matrix* π [113].

A Markov chain is thus an ensemble of configurations. However, in order to be useful from the statistical mechanics point of view, a Markov process has to satisfy certain conditions:

- normalization
- ergodicity
- balance

Normalization: Given any initial configuration k, the configurations $\{1, 2, \ldots, k, \ldots, l, \ldots\}$ define a *sample space*, that is:

$$\sum_l \pi_{kl} = 1. \tag{3.14}$$

Ergodicity: Broadly speaking it means, that the systems average is the same, no matter if averaged over time or space. In mathematical terms this can be defined as follows: given two configurations k and l, if $p_k > 0$ and $p_l > 0$, p_k denoting the probability of finding the system in state k, then there must be a positive integer number n such that $(\pi^n)_{kl} > 0$ [113].

Balance: In generic case, satisfying *global balance* is required. This means that the in and out probability flux to and from a given state is equal, that is,

$$\sum_k \pi_{kl} p_k = \sum_k \pi_{lk} p_l, \tag{3.15}$$

where p_k is the probability of state k. However, since we eventually are interested in generating ensembles with Boltzmann distribution, we can replace p_k with the form in Eq. (3.2). This yields

$$\sum_k \pi_{kl} e^{-\beta E_k} = \sum_k \pi_{lk} e^{-\beta E_l}. \tag{3.16}$$

Because of the normalization (3.14), the sum on the right side vanishes

$$\sum_k \pi_{kl} e^{-\beta E_k} = e^{-\beta E_l}. \tag{3.17}$$

This, in turn, means that if global balance holds, the Boltzmann distribution is the eigenfunction/eigenvector of π with eigenvalue 1 [113].

A stronger condition than global balance is *detailed balance*:

$$\pi_{kl} e^{-\beta E_k} = \pi_{lk} e^{-\beta E_l}, \tag{3.18}$$

and it is easy to see that if detailed balance holds, then global balance also holds.

The three enumerated properties are very important from the statistical mechanics point of view as it is possible to show that if a Markov process satisfies these properties than it generates a Boltzmann ensemble, that is, an ensemble characterized by the Boltzmann distribution (3.2) [113].

3.1.4 The Metropolis algorithm

In the picture of the concepts discussed in the previous paragraphs, we can use a Markov chain to generate an ensemble with configurations characterized by energy values following the Boltzmann distribution. Because of ergodicity, we can use the averages calculated over the ensemble to estimate time averages, which are normally measured during experiments. The question is how to chose the transition matrix π.

Note that for larger systems we cannot define π numerically as its dimension grows exponentially with the size of the system. Therefore, we have to propose a transition probability function, a function of the state S_i, the state in which the system is before the transition (initial state), and the state S_f, the state in which the system will be after the transition (final state). The transition probability function $\pi(S_i \to S_f)$ gives the probability of going from the configuration S_i to configuration S_f.

The detailed balance equation (3.18) of the transition probability function $\pi(S_i \to S_f)$ can be rearranged in the following form:

$$\frac{\pi(S_i \to S_f)}{\pi(S_f \to S_i)} = e^{-\beta\left[E(S_f)-E(S_i)\right]}, \tag{3.19}$$

where $E(S_i)$ is the energy of the configuration S_i. The aim is thus to find a transition probability which satisfies the ratio (3.19). The Metropolis idea [114] defines a suitable function $\pi(S_i \to S_f)$ as

$$\pi(S_i \to S_f) = \min\left\{1, e^{-\beta\left[E(S_f)-E(S_i)\right]}\right\}. \tag{3.20}$$

The Markov chain thus can be built by a simple procedure. The *Metropolis algorithm* is summarized in Algorithm 1. The algorithm first initializes the Markov chain S_t at $t = 1$. The initial state can be a random configuration or a predefined state. Then, until $t < t_{max}$, t_{max} being the number of total Monte-Carlo time steps, the following steps are repeated iteratively: First, another configuration S^p is proposed. There are different mechanisms to propose S^p. For instance, S^p can be drawn randomly from the set of possible configurations or it can be a randomly changed version of S_t. Then, the proposed configuration is accepted with a probability dictated by the Metropolis transition function (3.20) $\pi(S_t \to S^p)$. If the proposed configuration is rejected, the current configuration will be assigned to the next state of the Markov chain, that is, $S_{t+1} = S_t$. Last t is increased by one.

Algorithm 1 The Metropolis algorithm

```
1: procedure SAMPLE(β, t_max)
2:     Initialize S_1
3:     for t = 1, t < t_max, t = t + 1 do
4:         propose randomly an S^p configuration
5:         S_{t+1} ← MAKETRANSITION(S_t, S^p, β)
6:     end for
7: end procedure
8: procedure MAKETRANSITION(S_t, S^p, β)
9:     if E(S^p) < E(S_t) then          ▷ E(S^p) is the energy of the configuration S^p
10:        S^f ← S^p
11:    else
12:        x ← rand(0, 1)               ▷ x is a uniformly distributed on the interval [0, 1]
13:        if x < exp{−β [E(S^p) − E(S_t)]} then
14:            S^f ← S^p
15:        else
16:            S^f ← S_t
17:        end if
18:    end if
19:    return S^f
20: end procedure
```

Running the Markov chain long enough, the described procedure guarantees that the energy of the generated samples will be distributed according to the Boltzmann distribution (3.2). However, since the chain was started from a particular configuration, the first

samples might deviate from the distribution. Therefore, when calculating averages, the first samples are usually not considered. After a given *thermalization time* the Markov chain represents a Boltzmann ensemble of the simulated system. This is achieved without knowing the exact from of the distribution, and it is a big advantage as calculating the partition function Z would be very costly even for small systems, since the number of terms in the summation scales exponentially with the size of the system. The trick applied here is to look at ratios of probabilities, thus Z will be cancelled out.

Besides the Metropolis formula (3.20), there are, of course, other possibilities to chose the transition probability. Another popular choice is implemented by the *Glauber dynamics* [115]. In this case, the transition probability has the form of

$$\pi(S_i \to S_f) = \frac{e^{-\beta\left[E(S_f)-E(S_i)\right]}}{1+e^{-\beta\left[E(S_f)-E(S_i)\right]}}, \tag{3.21}$$

which also satisfies the detailed balance condition.

Once the Markov chain is implemented and samples are gathered, averages can be calculated. For instance, in the case of a magnetic spin system, like the *Ising model* [116], the magnetization can be calculated as the average value of the spins. From averages of the magnetization we can calculate the susceptibility χ by using equation (3.5). Similarly, the specific heat C can be estimated through calculating ensemble averages of the energy and using equation (3.4).

By studying how the values of the measured and derived quantities change as the function of the temperature, we can study the behaviour of the system in its ground state or draw conclusions about the aspects of the exhibited phase transitions.

3.1.5 The Ising Model

All concepts introduced so far are rather abstract. To illustrate how simulations work in practice, we use the already mentioned Ising model [116] as a concrete example. The Ising model is a simple *lattice model* [117] of *ferromagnetic* materials [118].

The model consists of *spins* which can either point "up" or "down", arranged according to a lattice structure. The model has been studied in different dimensions and on different lattices [119–128]. Here we present the two dimensional LxL square lattice case as illustrated in Figure 3.1. The energy of the system is defined through the Hamiltonian

$$H = -J_I \sum_{\langle i,j \rangle} \sigma_i \sigma_j - B_I \sum_i \sigma_i, \tag{3.22}$$

where σ_i is the direction of spin i. The "up" direction is encoded by the value 1, the "down" direction is indicated by the value -1. J_I is the strength of the coupling between the spins and B_I is the external magnetic field pointing in the "up" direction.

The model was the first of its kind to properly capture the order-disorder phase transition at the *Curie temperature*, a phenomenon we can also study by computer simulations. In Sections 3.1.4 we already introduced the general theory behind these simulations. Here we will elaborate on the concrete steps we need to take to measure the average magnetization $\langle M \rangle$, the specific heat C and the magnetic susceptibility χ of the Ising model with no external field ($B_I = 0$).

Configurations of the system are numerically stored as LxL matrices with values ± 1. Each entry in the matrix corresponding to a spin in the system. The total number of spins is $N = L^2$. A value 1 represents a spin pointing "up", a value -1 represents a spin pointing

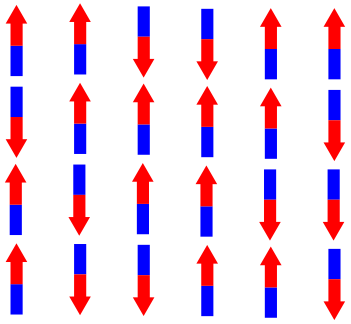

Figure 3.1: The two dimensional Ising model on a square lattice.

"down". The energy $E(S_l)$ of the configuration S_l (represented by a matrix) is calculated by the Hamiltonian (3.22), thus σ_i takes the value of the entry with linear index i in the matrix representation of S_l. If the magnetic field B_I is zero, the lowest energy is achieved when all spins point in the same direction, either "up" or "down". Thus the ground state is degenerated in this case. If B_I is non-zero, the ground state will be determined by the sign of B_I. If the field is positive, the ground state will be reached when all spins point upwards. If the field is negative, all spins point down in the ground state. As already stated, we will simulate the $B_I = 0$ case.

At finite (non-zero) temperatures, the entropic effect of the temperature will manifest itself and, depending on the rapport of T with J_I, the system will undergo a second order phase transition [129]. Since T divides J_I in the Boltzmann factor (3.1), the temperature and the coupling are not two independent parameters of the simulation. Therefore, it is convenient to set J_I to a constant value, and vary only the temperature. For convenience we will set the coupling J_I to 1.

As indicated earlier, we chose the average magnetization $\langle M \rangle$ as the order parameter of the system. To simulate the system at a desired temperature T, we fix the inverse temperature $\beta = 1/k_B T$ and sample K' configurations using the Metropolis algorithm. However, to ensure a faster equilibration and more representative samples, the algorithm is fine-tuned for spin-systems. In the updated version of the algorithm, described in Algorithm 2, new configurations are proposed by flipping a randomly selected spin and saving every Nth state of the Markov chain ($n_{trials} = N$ in Algorithm 2), thus allowing the chance for each spin to be flipped between two consecutive states of the chain. The algorithm can be summarized as follows: An initial state S_1 is chosen and t is set to 1. Then, in every iteration the algorithm tries to flip a random spin. The flipped configuration is accepted with the Metropolis transition probability (3.20). After N flip-trials the configuration of the spins is saved in S_{t+1} and t is increased by one. The N trials between two consecutive saved states define a *Monte-Carlo step*. Algorithm 2, setting t_{max} to K'.

Since the Markov chain in the Metropolis algorithm is started from an arbitrary configuration which might not be characteristic for the simulation temperature T (e.g. a random configuration for $T = 0$), an initial thermalization is required to assure that equilibrium is reached. Equilibrium in this case means that the Markov chain generates samples with energy values distributed according to the Boltzmann distribution. To achieve this, the first k samples will be discarded, keeping only $K = K' - k$ samples. To keep indexing of samples simpler, we re-index the kept samples as $S_1 = S_{k+1}$, $S_2 = S_{k+2}$, ...

Average values of the magnetization and energy can be calculated using these samples

Algorithm 2 The Metropolis algorithm

procedure SAMPLE($\beta, t_{max}, n_{trials}$)
  Initialize S_1
  for $t = 1$, $t < t_{max}$, $t = t + 1$ **do**
    $S^i \leftarrow S_t$
    for $n = 1$, $n \leq n_{trials}$, $n = n + 1$ **do**
      propose randomly an S^p configuration
      $S^i \leftarrow$ MAKETRANSITION(S^i, S^p, β)
    end for
    $S_{t+1} \leftarrow S^i$
  end for
end procedure
procedure MAKETRANSITION(S^i, S^p, β)
  if $E(S^p) < E(S^i)$ **then** ▷ $E(S^p)$ is the energy of the configuration S^p
    $S^f \leftarrow S^p$
  else
    $x \leftarrow rand(0, 1)$
    if $x < \exp\{-\beta\left[E(S^p) - E(S^i)\right]\}$ **then**
      $S^f \leftarrow S^p$
    else
      $S^f \leftarrow S^i$
    end if
  end if
  return S^f
end procedure

in the following way:

$$\langle M \rangle = \frac{1}{NK} \sum_{t=1}^{K} \sum_{i} \sigma_i(S_t), \tag{3.23}$$

$$\langle M^2 \rangle = \frac{1}{NK} \sum_{t=1}^{K} \sum_{i} \sigma_i^2(S_t), \tag{3.24}$$

$$\langle E \rangle = \frac{1}{K} E(S_t), \tag{3.25}$$

$$\langle E^2 \rangle = \frac{1}{K} E(S_t)^2, \tag{3.26}$$

where $\sigma_i(S_t)$ is the state of spin i in configuration S_t and $E(S_t)$ is the energy of the system in configuration S_t calculated using the Hamiltonian (3.22). Then, the specific heat C and the susceptibility χ can be calculated using formulas (3.4) and (3.5).

Performing the simulation and gathering statistics for different temperatures T we can plot the order parameter $\langle M \rangle$ as the function of the temperature. For convenience, during the simulation we set $k_B = 1$. Therefore, the temperature is expressed in units of k_B. Furthermore, we usually plot the averages as a function of $\beta = 1/k_B T$ instead of T. As a consequence of using $k_B = 1$, considering also that $J_I = 1$, β is expressed in units of J_I/k_B.

The plot in Figure 3.2 shows the average magnetization $\langle M \rangle$ (which is also used as order parameter) for a 100x100 Ising model after $k' = 5000$ thermalization Monte-Carlo steps, with the average calculated over $K = 10^5$ samples. As this plot illustrates this, the average magnetization is close to zero in the high temperature (small β) range. As the temperature decreases (increasing β) and reaches the *Curie temperature* T_C, the system reaches its critical point and undergoes a phase transition. According to the exact solution, $T_C = 2.2692$ [129]. This corresponds to $\beta = 0.4407$, which is around the value where we observe a jump in the order parameter.

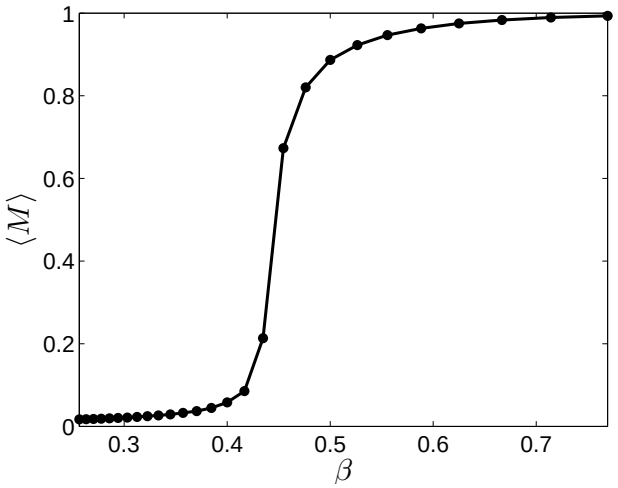

Figure 3.2: The average magnetization $\langle M \rangle$ used as order parameter, plotted as the function of the inverse temperature β for the two dimensional Ising model. Note the phase transition occurring around $\beta = 0.4407$.

Figure 3.3 presents the diverging specific heat C and the magnetic susceptibility χ. Although the plots show a jump in these two quantities, the divergence is not real as this could only happen in infinitely large system. In order to observe real divergence, the correlation length ξ should be infinitely large. Here, however, the correlation length is cut by the finite size of the lattice.

The computer simulation method presented through the case study of the Ising model can be easily adapted to many models. It constitutes a generic framework for simulations in different sciences, as we shall see this later.

3.2 Elements of Graph Theory

Graph theory [130] emerged from the field of mathematics as a much needed tool to treat problems involving pairwise connections among different entities. The famous paper by Euler regarding the problem of *seven bridges of Königsberg* is often considered as the first graph-theoretic approach to solve a problem [131].

From the point of view of the present thesis, graphs are twofold important. First, they appear as a backbone on which probabilistic models are built (see *graphical models* in section 3.3 and our applications in Chapter 4), secondly, they offer a framework for the analysis of certain data (Chapter 5). Therefore, we consider it important to highlight the basics of graph theory, at this stage confining ourselves only to definitions.

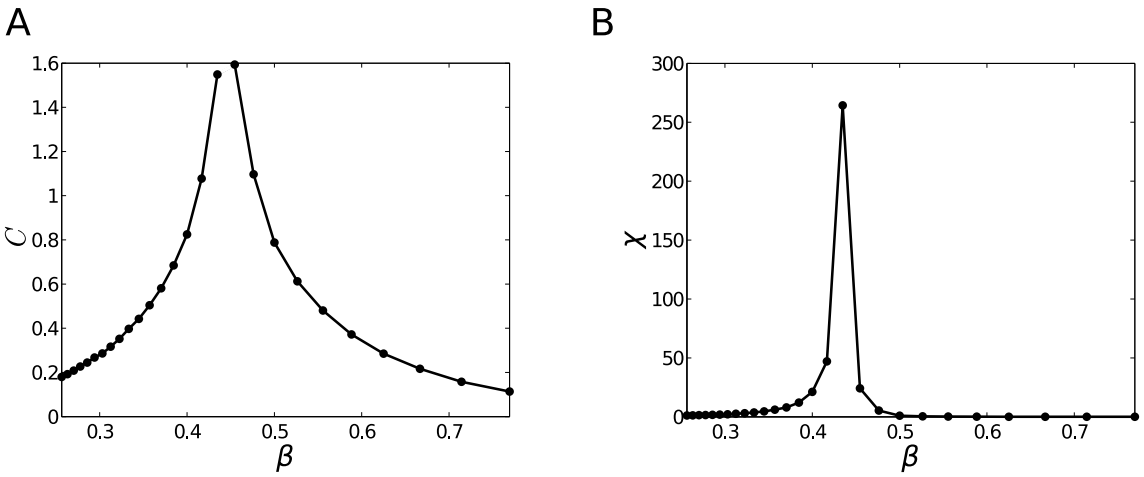

Figure 3.3: The specific heat C (in Panel A) and the magnetic susceptibility χ (Panel B) as the function of the inverse temperature β for the two dimensional Ising model.

A graph G is defined as the tuple $G(V, E)$, where E is the set of edges between vertices (or nodes) contained by the set V. The notion emphasises the data-structure-like construction of graphs: given a set of vertices V relations among nodes are stored in edges $(a, b), a \in V, b \in V$, that is, $E = \{(a, b)|a \in V, b \in V\}$ which implies $E \subseteq L \times L$. A graph is *directed*, if the set of edges consists of ordered pairs of vertices. Unless otherwise stated, we will work with *undirected* graphs, that is, $(a, b) = (b, a)$ for any edge $(a, b) \in E$.

Small graphs are easily illustrated with a diagram similar to the one in Figure 3.4, nodes being represented by the black points while edges are marked by the lines connecting the nodes. However, it is difficult to embed more complicated graphs in the two dimensional space. In fact, the problem of embedding proves to be relatively difficult and, besides the aspects of visualization, it finds many applications which are much beyond the scope of this section [132]. In the following we introduce the nomenclature we will use and some

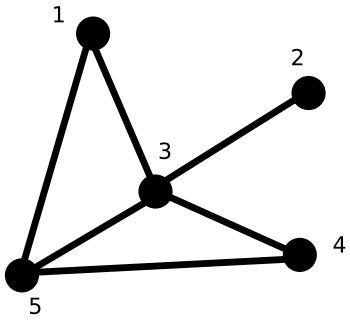

Figure 3.4: A sample graph $G(V, E)$ with vertices $V = \{1, 2, 3, 4, 5\}$ and edges $E = \{(1, 3), (1, 5), (2, 3), (3, 4), (3, 5), (4, 5)\}$.

basic concepts which might occur several times in the following stages.

A vertex v is said to be *incident* with an edge e if $v \in e$. The two vertices incident with an edge e are the *end-vertices* or *ends* of e, while e *joins* its two ends. Two vertices x and y are *adjacent* if they are joined by an edge. Two edges are adjacent if they have

a common end. If all vertices in a graph G are pairwise adjacent, then G is a *complete graph*.

The graph $G'(V', E')$ is a *subgraph* of $G(V, E)$, if $V' \subseteq V$ and $E' \subseteq E$.

The set of vertices adjacent with vertex x are the *neighbours* of x and they are denoted by $n(x)$. Similarly, the set of vertices adjacent with any of the vertices from a set $U \subseteq V$ are denoted by $n(U)$ and are called the neighbours of U.

The *degree* of a vertex x is defined as the number of the neighbours of x, that is, $deg(x) = |n(x)|$. The *average degree* $deg(G)$ of a graph G can be computed as

$$deg(G) := \frac{1}{|V|} \sum_{v \in V} deg(v), \tag{3.27}$$

where $|V|$ denotes the number of vertices in V.

It is possible to construct the *degree distribution* of a graph as

$$h(d) = \frac{1}{|V|} \sum_{v \in V} \delta_{d,deg(v)}, \tag{3.28}$$

where δ is the Kronecker symbol. Then, of course, $deg(G)$ must be the mean value of the degree distribution h. Indeed, calculating the mean of h, we get

$$\sum_{d=1}^{\infty} d\, h(d) = \frac{1}{|V|} \sum_{v \in V} \sum_{d=1}^{\infty} d\, \delta_{d,deg(v)} = \frac{1}{|V|} \sum_{v \in V} deg(v) =: deg(G). \tag{3.29}$$

A *path* is a nonempty graph $P(V, E)$, so that, $V = \{x_1, x_2, \ldots, x_k\}$ and $E = \{(x_1, x_2), (x_2, x_3), \ldots, (x_{k-1}, x_k)\}$. The path P *connects* x_1 with x_k. x_1 and x_k are the *ends* of the path P. The number of the edges in E, that is, $|E|$ is the *length* $l(P)$ of the path P. A graph $G(V, E)$ is *connected*, if there is a path between any two vertices from V.

In addition to the enumerated definitions, we introduce the concept of *weighted graph*. A weighted graph $G(V, E)$ is a graph in which a real number $w(e)$ is assigned to each edge e. $w(e)$ is the *weight* or *length* of e. In order to have a unified framework, we can assign weights 1 to all edges e' of an unweighted graph $G'(V'E')$ so that $w(e') = 1,\ \forall e' \in E'$. Then we can redefine the length of a path P as

$$l(P) = \sum_{e \in \mathscr{E}(P)} w(e), \tag{3.30}$$

where $\mathscr{E}(P)$ denotes the set of edges of P.

3.3 Probabilistic Graphical Models

Probabilistic graphical models [133] are models taking advantage of an interplay between graph theory and probability theory. These models represent the dependencies of random variables with graph structures, thus creating a relatively transparent framework for modelling complex data. The versatility of graphical models is already indicated by its popularity and the numerous applications from different fields ranging from computer vision to statistical mechanics [134–139].

There are two main types of graphical models: *directed* and *undirected*. As the names suggest they are based on directed and undirected graphs, respectively. While directed graphical models are very useful tools especially in machine learning and artificial intelligence, undirected graphical models are more popular in physics applications. As we will

see later, this is not a question of preference, but has rather well grounded reasons. For the sake of completeness, however, we will introduce directed graphical models through the means of a simple example. We will invest more effort in describing undirected graphical models and emphasising their connections to physics.

Directed graphical models [140], also known as *Bayesian networks* or *belief networks* represent a set of random variables and their conditional dependencies via directed graphs. One of the classical textbook examples, as presented in [141], demonstrates how states and events like rain fall, the functioning of the sprinkle, having a wet pavement, the slipperiness of the pavement and the current season are related/conditioned upon each other. The enumerated events and conditions constitute a system we might want to model. Thus, we will represent each element by a random variable (season – X_1, rain fall – X_2, sprinkler – X_3, wet pavement x_4, slippery pavement – X_5). Each variable is binary (for instance $X_3 = 1$ means the sprinkler is on, $X_3 = 0$ means the sprinkler is off), except X_1 which has four states. This defines a state space of 64 states and it seems fairly complicated to assign probabilities for each of the states.

To simplify the model, we apply common sense, and illustrate the dependencies of the variables by a directed graph, for instance, by the one shown in Figure 3.5. Doing so helps

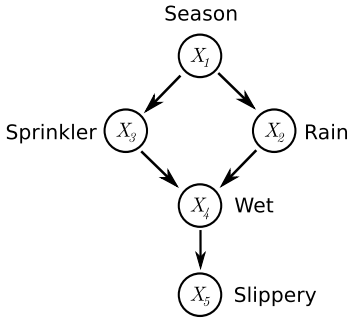

Figure 3.5: A simple Bayesian network, representing the dependencies of five random variables.

us to understand the dependencies among the random variables. While this seems easy in a simple model like the present one, it gets complicated for complex models and thinking in terms of directed graphs facilitates the modelling process. The graph clearly indicates the factorization, and, based on the *chain rule* [142], we can rewrite the joint probability distribution as

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4), \tag{3.31}$$

where $P(x_1, x_2, x_3, x_4, x_5)$ is a shortened notation for $P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5)$, and $P(X_2|X_1)$ denotes the conditional probability of the event $X_2 = x_2$ given the event $X_1 = x_1$. The product on the right-hand side of (3.31) requires only 24 parameters, significantly less compared to 64 parameters required by the joint distribution. Thus, applying a Bayesian network, we simplified the initial model represented by the joint probability distribution $P(x_1, x_2, x_3, x_4, x_5)$.

3.3.1 Undirected Graphical Models

Undirected graphical models, also known as *Markov random fields* [143], similarly to Bayesian networks, represent the relationships of random variables with a graph structure. However, as the name suggest, undirected graphical models rely on undirected

graphs, graph edges indicating stochastic interactions between random variables represented by nodes.

According to the definition, Markov random fields characterized by a graph $G(X, E)$ describe a joint probability distribution over a set of random variables $X_1, X_2, \ldots, X_n$ as products of *compatibility functions* ψ:

$$P(\vec{x}) = P(x_1, x_2, \ldots, x_n) = \frac{1}{Z} \prod_{c \in \mathscr{C}} \psi_c(\vec{x_c}), \tag{3.32}$$

where the product is over *cliques* $\vec{x_c}$ in the graph G and Z is a normalization constant. Cliques are complete subgraphs of a graph, an example is given in Figure 3.6. The com-

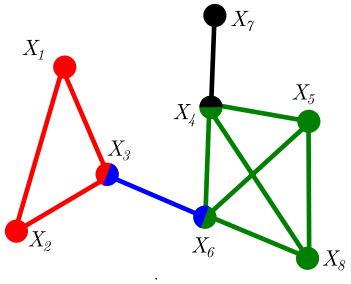

Figure 3.6: A simple Markov random field illustrating cliques with different colours. The cliques are $\{(X_1, X_2, X_3), (X_3, X_6), (X_4, X_5, X_6, X_8), (X_4, X_7)\}$.

patibility functions ψ assess the affinity between their arguments. The higher the value of $\psi_{(X_i, X_j, \ldots)}(x_i, x_j, \ldots)$ is, the more compatible the values $x_i, x_j, \ldots$ are.

If $P(\vec{x})$ is strictly positive (it is nowhere zero), according to the *Hammersley-Clifford theorem* the joint distribution can be represented as a *Gibbs measure* [143]

$$P(\vec{x}) = \frac{1}{Z} e^{-\mathcal{E}(\vec{x})}, \tag{3.33}$$

where $\mathcal{E}(\vec{x})$ is called the *energy function*. Since the right-hand side of Equation (3.32) and that of Equation (3.33) must be equal, the energy must have the form

$$\mathcal{E}(\vec{x}) = -\sum_c \log \psi_c(\vec{x_c}) = -\sum_c \phi_c(\vec{x_c}), \tag{3.34}$$

where the functions ϕ are called *potential functions*.

3.3.2 Relation to Physics

Equations (3.33) and (3.34) reveal a clear relation between Markov random fields and the formalism of the canonical ensemble. They resemble equations (3.2) and (3.22). In fact, Markov random fields emerged from lattice models of physics, like the Ising model, as a more generic and widely applicable framework. Therefore, it is not surprising, that $\mathcal{E}$ is called energy and the Z normalising constant in equation (3.32) is termed as partition function. However, these quantities do not necessarily have physical meaning. The case is similar with the potential functions. In most of the cases they have nothing in common with physical potentials.

It is clear now, that the Ising model described in Section 3.1.5 is a special Markov random field, with a graph defined by a two dimensional square lattice and potential functions $\sigma_i \sigma_j$, where i and j are neighbouring spins.

Ising models, however, were studied not only on regular lattices, but also on graphs with different properties [144–146]. There are also many modifications of the applied energy function and it has become popular in image segmentation applications [147–149].

On the other hand, statistical mechanics also benefit from the synergy created by the interdisciplinary approaches to applications of these models. Methodologies developed in the generic framework of Markov random fields are applicable for the Ising model and its generalized versions, as we will see in later parts of this thesis.

3.4 Elements of Algebraic and Computational Topology

As some of our approaches have a theoretical background relying on computational and algebraic topology, this section will be dedicated to the introduction of the most important theoretical concepts from the point of view of this thesis. Most of this introduction is based on references [150] and [151].

Algebraic topology is the field of mathematics which barrows methods from abstract algebra to study *topological spaces*. A topological space is, in fact, a set of points with neighbouring relations. More exactly, for each $a \in A$, let $N(a)$ denote a non-empty collection of subsets of A, called the *neighbourhoods* of a. Neighbourhoods are required to satisfy the following conditions:

1. if $n(a) \in N(a)$, that is, if $n(a)$ is a neighbourhood of a, then $a \in n(a)$.
2. If $n(a)$ and $n'(a)$ are neighbourhoods of a, then $n''(a) = n(a) \cap n'(a)$ is also a neighbourhood of a.
3. If $n(a)$ is a neighbourhood of a and $n'(a) \subseteq n(a)$, then $n'(a)$ is also a neighbourhood of a.
4. If $n(a)$ is a neighbourhood of a and $n'(a) = \{b | b \in n(a)\}$, then $n'(a)$ is also a neighbourhood of a. $n'(a)$ is called the *interior* of $n(a)$.

If the conditions are satisfied than the structure consisting of A and $N(A)$ is called a topological space [152].

3.4.1 Simplical Complexes

Let $u_0, u_1, \ldots, u_k$ denote points in the $\mathbb{R}^d$ space. A point $x = \sum_{i=1}^{k} \lambda_i u_i$, where $\lambda_i \in \mathbb{R}$, is an *affine combination* of u_i if $\sum_{i=1}^{k} \lambda_i = 1$. The *affine hull* is the set of affine combinations. If the vectors $u_i - u_0$ and $u_j - u_0$, $\forall i, j \in \{1, 2, \ldots, k\}$ are linearly independent, then th vectors $u_0, u_1, \ldots, u_k$ are *affinely independent*.

An affine combination $x \sum \lambda_i u_i$ is a *convex combination* if $\lambda_i \geq 0, \forall i$. The set of convex combinations is called the *convex hull*. A *k-simplex* σ is the convex hull of $k+1$ affinely independent points ($\sigma = conv\{u_0, u_1, \ldots, u_k\}$). The dimension of σ is k. The first four k-simplexes correspond to well-known geometric objects: a 0-simplex is a *vertex*, a 1-simplex is an *edge*, a 2-simplex is a *triangle* and a 3-simplex is a *tetrahedron*.

If $\sigma = conv\{u_0, u_1, \ldots, u_k\}$ is a k-simplex, then the convex hull τ of any non-empty subset of the $k+1$ points is a *face* of σ (denoted as $\tau \leq \sigma$). τ is a *proper face* if the subset is not the entire set ($\tau < \sigma$). The *boundary* of σ is the union of all of its proper faces.

A *simplical complex* is a finite collection K of simplexes, such that, if $\sigma \in K$ and $\tau \leq \sigma$, then $\tau \in K$, furthermore, if $\sigma' \in K$ and $\sigma'' = \sigma \cap \sigma'$, then $\sigma'' = \emptyset$ or, alternatively, $\sigma'' \leq \sigma$ and $\sigma'' \leq \sigma'$.

The *nerve* of a set A is the collection of all non-empty subcollections whose sets have a non-empty common intersection, that is

$$NrvA = \left\{ B \subseteq A | \bigcap B \neq \emptyset \right\}. \tag{3.35}$$

Let X be a finite set of points in $\mathbb{R}^d$. Let $B_x(r) = x + r\mathbb{B}^d$ denote a closed ball with radius r, centred around point x. The *Čech complex* of X and r is the nerve of the collection of the balls $B_x(r)$, $x \in X$, but the balls in the complex are represented by their centres. Mathematically speaking,

$$\check{C}ech(X, r) = \left\{ \sigma \subseteq X | \bigcap_{x \in \sigma} B_x(r) \neq \emptyset \right\}. \tag{3.36}$$

A complex related to the Čech complex is the *Vietoris-Rips* complex. The latter contains all simplexes whose edges are in the complex:

$$VR(X, r) = \{ \sigma \subseteq X | \operatorname{diam}(\sigma) \leq 2r \}, \tag{3.37}$$

where $\operatorname{diam}(\sigma)$ is the diameter of simplex σ. Evidently, the edges in the Čech and Vietoris-Rips complexes are the same for the same radius r, furthermore, $\check{C}ech(X, r) \subseteq VR(X, r)$ as the latter contains every simplex faced by the edges in the complex.

The *Voronoi cell* V_x of a point $x \in X$ is the set of points for which x is the closest, that is,

$$V_x = \left\{ y \in \mathbb{R}^d | \|x - y\| \leq \|x' - y\|, \forall x' \in X \setminus \{x\} \right\}. \tag{3.38}$$

The collection of the Voronoi cells V_x as x runs over the elements of X is called the *Voronoi diagram*.

Connecting the centres x and x' ($x, x' \in X$) of those Voronoi cells V_x and $V_{x'}$ which meet in a common piece of their boundary, we get the *Delaunay complex*. A geometric realizations of the Delaunay complex is known as the *Delaunay triangulation*. Because of its construction, the Delaunay triangulation is the dual of the Voronoi diagram. In fact, the Delaunay complex is the nerve (more precisely is isomorphic to the nerve) of the Voronoi diagram:

$$DT(X) = \left\{ \sigma \subseteq X | \bigcap_{x \in \sigma} V_x \neq \emptyset \right\}. \tag{3.39}$$

Denoting with $R_x(r)$ the intersection of the ball $B_x(r)$ with the Voronoi cell V_x, that is, $R_x(r) = B_x(r) \cap V_x$, we define the *alpha complex* as the nerve of the collection of $R_x(r)$ as x runs over the elements of X:

$$Alpha(X, r) = \left\{ \sigma \subseteq X | \bigcap_{x \in \sigma} R_x(r) \neq \emptyset \right\}. \tag{3.40}$$

Obviously, because $R_x(r) \subseteq V_x$, the alpha complex is a subcomplex of the Delaunay complex.

3.4.2 Homology Groups

Let L be a simplical complex and k a dimension. A *k-chain* is defined as a formal sum of k-simplexes in L: $\gamma = \sum c_i \sigma_i$, where c_i are coefficients. k-chains are added componentwise: $\gamma + \gamma' = \sum (c_i + c'_i)\sigma_i$, where $\gamma' = \sum c'_i \sigma_i$. Coefficients are often only from the set

$\{0,1\}$ and they satisfy $1+1=0$ (*modulo 2 coefficients*). The k-chains, together with the addition operation form the *group of k-chains* denoted as $(\mathbf{C}_k,+)$ or simply $\mathbf{C}_k$.

We can redefine the boundary of a k-simplex as similar formal sum of its $k-1$ dimensional faces:

$$\partial_k\sigma=\sum_{j=0}^{k}\sigma_{-u_i}, \tag{3.41}$$

where the σ_{-u_i} denotes the face in which u_i is omitted.

For a k-chain, the boundary is the sum of the boundaries of its simplexes,

$$\partial_k\gamma=\sum c_i\partial_k\sigma_i, \tag{3.42}$$

that is, the boundary maps a k-chain to a $(k-1)$-chain. We denote this by

$$\mathbf{C}_k\longrightarrow\mathbf{C}_{k-1}. \tag{3.43}$$

Based on the definition, $\partial_k(\gamma+\gamma')=\partial_k\gamma+\partial_k\gamma'$, which is the defining property of *homomorphism*, a map between groups commuting with the group operation. Therefore ∂_k is referred to as the *boundary map* or *boundary homomorphism*. The *chain complex* is the sequence of chain groups connected by the boundary map:

$$\ldots\xrightarrow{\partial_{k+2}}\mathbf{C}_{k+1}\xrightarrow{\partial_{k+1}}\mathbf{C}_k\xrightarrow{\partial_k}\mathbf{C}_{k-1}\xrightarrow{\partial_{k-1}}\ldots. \tag{3.44}$$

A *k-cycle* is defined as p-chain with empty boundary: $\partial\gamma=0$. Since ∂ commutes with the addition operator, the k-cycles and the addition form the *k-th cycle group*, denoted as $(\mathbf{Z}_k(L),+)$, or simply $\mathbf{Z}_k(L)$.

A *k-boundary* is a k-chain that is the boundary of a $(k+1)$-chain: $\gamma=\partial\kappa$, where $\kappa\in\mathbf{C}_{k+1}$. Since ∂ commutes with the addition operator, the k-boundaries and the addition form the *k-th boundary group*, denoted as $(\mathbf{B}_k(L),+)$, or simply $\mathbf{B}_k(L)$ (not to be confused with $B_x(r)$, the ball with radius r around point x).

The *k-th homology group* is the k-th cycle group is the k-th cycle group module the k-th boundary group, that is, $\mathbf{H}_k(L)=\mathbf{Z}_k(L)/\mathbf{B}_k(L)$. The *$k$-th Betti number* ϑ_k is the rank of this group: $\vartheta_k=\operatorname{rank}\mathbf{H}_k(L)$.

Elements of $\mathbf{H}_k$ are eobtained by adding all k-boundaries to a given k-cycle. The elements are called *homology classes* and they are *homologous*.

3.4.3 Persistent Homology

Let L be a simplical complex and $f:L\to\mathbb{R}$ a monotonic (non-increasing) function, that is, $f(\sigma)\leq f(\tau)\Leftrightarrow\sigma\leq\tau$. This implies that $L(l)=f^{-1}(-\infty,l]$ is a subcomplex of L, $\forall l\in\mathbb{R}$. We can find $n+1$ values for l denoting them as by $l_0,l_1,\ldots,l_n$, such that

$$\emptyset\subseteq L_0\subseteq L_1\subseteq\ldots\subseteq L_n\subseteq L, \tag{3.45}$$

where $L_i=L(l_i)$. This sequence is called a *filtration*. Exemples for this are the Čech and the alpha complexes with the radius r corresponding the parameter l.

Since for any $i\leq j$ there is an inclusion map from the underlying space L_i to that of L_j, an induced homomorphism $f_k^{i,j}:\mathbf{H}_k(L_i)\to\mathbf{H}_k(L_j)$ must exist for each dimension k. The filtration thus corresponds to a sequence of homology groups connected by homomorphisms:

$$0=\mathbf{H}_k(L_0)\to\mathbf{H}_k(L_1)\to\ldots\to\mathbf{H}_k(L_n)=\mathbf{H}_k(L). \tag{3.46}$$

Homology classes may "born", "stay alive" or "die" as we go from L_{i-1} to L_i. We group classes according to intervals of l values over which they were "alive".

The *k-th persistent homology groups* are the images of the homomorphisms

$$\mathbf{H}_k^{i,j} = \mathrm{im} f_k^{i,j}, \tag{3.47}$$

for $0 \leq i \leq j \leq n$. The *k-th persistent Betti numbers* $\vartheta_k^{i,j}$ are the ranks of these groups: $\vartheta_k^{i,j} = \mathrm{rank}\, \mathbf{H}_k^{i,j}$.

Persistent Diagrams We visualize the collection of persistent Betti numbers by plotting *persistent diagrams* in the following way: We can calculate the number μ_k^{ij} of independent k-dimensional classes that were alive over the interval $[l_i, l_j]$ as

$$\mu_k^{ij} = (\vartheta_k^{i,j-1} - \vartheta_k^{i,j}) - (\vartheta_k^{i-1,j-1} - \vartheta_k^{i-1,j}). \tag{3.48}$$

We will represent the classes with a horizontal bars on a graph. The abscissa will correspond to l. Each class will be represented by a bar, therefore, we will draw μ_k^{ij} bars with starting point at l_i and end-point at l_j. The horizontal order of the bars is arbitrary.

The persistent diagram prepared in this manner can be considered to be a barcode or fingerprint of the topology of L.

Chapter 4

Analyzing and Modeling Images

References

- G. Máté, R. Dickman and D.W. Heermann, *A State Dependent Potts Model*, in preparation (2013).
- G. Máté and D.W. Heermann, *Simulating Microscopy Images of the Cell Nucleus*, in preparation (2013).

Chapter Summary

Confocal microscopy images provide a non-invasive insight into the structure of the nucleus. Since the resolution of the microscopes are limited by diffraction effects, these microscopes are unable to resolve the chromatin fibre. However, the data provided by images recorded with these techniques is still very valuable as it captures the spatial density distributions of the chromatin in a quasi three dimensional manner. Despite of the accurate information regarding the three dimensional arrangements, modelling approaches which capture these distributions are scarce.

In this chapter, we introduce a generalized Potts model, which is able to recreate the density patterns observed in the microscopy images. The generalization consists in the introduction of interactions between all spin states. A magnetic field, acting differently upon different states is also introduced. In the most generic case, the model is parametrized by the interactions between the states and the strengths of the field for each spin state separately. We benefit from the relation of the Potts model and the more general class of Markov random fields, and use learning techniques developed in the latter framework to learn the parameters of our model using microscopy images as training data. Although learning parameters is common in image processing techniques using Markov random fields, it is less often the case in physics related applications. Therefore, we introduce the concept of learning through a simple example in which we estimate the temperature of the Ising model. Then, we describe the learning procedure used to train the generalized Potts model. After testing the performance of the learning we apply the approach on nucleus images and demonstrate that the framework correctly captures the density features of the images.

As the generalized model is interesting from the statistical mechanics point of view, we also study the thermodynamic properties of the model on a constrained parameter space. We engage in theoretical calculations which predict a shift in the transition point as we vary the model parameters. The developed mean field theory also indicates a highly degenerate ground state for parameter combinations which correspond to an antiferromagnetic model. We support our theoretical findings with computer simulations. Moreover, Markov Chain Monte-Carlo calculations reveal a variety of interesting phases exhibited by the model.

4.1 The Modified Potts Model for Confocal Microscopy Images

References

This section is adapted from our manuscript, which is intended to be submitted for publication,

- G. Máté and D.W. Heermann, *Simulating Microscopy Images of the Cell Nucleus*, in preparation (2013).

We would like to thank Lindsay S. Shopland for kindly providing the microscopy images.

A problem faced by modern biology and medicine is the understanding of the organization of the chromatin fibre in the nucleus [153–155]. The structure of the fibre is know to have a major impact on different precesses occurring in the nucleus. These process include the repair of the damaged DNA, replication, silencing [156–160], etc. Understanding the packing is thus crucial.

In recent years high throughput experiments boosted studies which deal with modelling this packing [14, 20, 161–163]. Most approaches are based on polymer models [164] and are able to explain many findings of the experiments. For instance, the levelling off of the mean square physical distance between positions on the chromatin fibre when measured as a function of the distance along the fibre can be explained with a polymer model which is allowed to loop randomly [165].

Another experimental approach which allows non invasive investigation of the cell nucleus is conventional confocal microscopy. Although this technique lacks a resolution which would allow the observation of the fibre, it is a relatively good and reliable source of information. Despite this, modelling approaches based on confocal images are scarce. Although the images do not provide direct information about the fibre structure, they might serve as a perfect source of data to model the statistical mechanics of the spatial density distribution of the chromatin in the nucleus.

In this study, instead of concentrating on the chromatin fibre, we aim to develop a method which models this density distribution. As we want to use confocal microscopy images as the starting point of our approach, we need to consider two factors. First we need to understand how to exploit the data and how image processing algorithms may assist us. Second, we need to find a suitable model stemming from statistical mechanics.

The aim of image processing algorithms vary from segmenting the image to enhancing it. Often, these algorithms are based on detecting contours, edges and thus objects in pictures. However, defining objects in images of cell nucleus is not possible with conventional approaches. The human eye and brain is trained to easily recognize objects on natural images, neurons in the brain interact and form representations of objects within fractions of the second [166], but this is only possible because we spend the first years of our life training our brain to do so. Image processing algorithms can be trained in a similar manner. We can devise different models which are able to represent the same object and we can train the models, that is, learn the parameters of the models on real world images [167].

On the other hand, we need to consider the physics of the nucleus. In physics and

biology (and not just) one encounters many systems presenting self-organizing behaviour. Self organization results from the interactions of units and leads to properties observed on larger scales. Examples range from the fractal structures of bacteria colonies often modelled with a diffusion limited aggregation process to the magnetic domains of certain solids described with the well-known Potts model [168–175].

It turns out that the latter model generalizes very well and its applications include not only the description of magnetization but also grain growth and network analysis amongst others [176, 177]. The Potts model is also a popular model in image processing, especially in image segmentation [178–180].

4.1.1 The Potts model

The Potts model is one of the most investigated spin models of statistical mechanics. The model is relatively well understood, therefore it offers a very good starting point for our study.

The model consists of spin variables (scalers) arranged according to a given graph which might be a regular lattice or an arbitrary network. Spins interact along the edges of the graph and the energy stemming from the interactions is defined by the Hamiltonain

$$H = -J \sum_{<i,j>} \delta\sigma_i\sigma_j - h \sum_i \delta_{\sigma_i a}, \tag{4.1}$$

where J is the interaction parameter, the notation $< i, j >$ means a summation along the edges of the graph, σ_k is the state of spin k and δ is the kronecker symbol, h is the magnetic field and a represents one of the q possible states. The probability of a given configuration of the Potts model in the canonical ensemble is given by the Boltzmann distribution:

$$P(\{S\}) = \frac{e^{-\beta H}}{Z}, \tag{4.2}$$

where Z is the partition function and $\beta = 1/k_B T$, T being the temperature of the system and k_B the Boltzmann factor.

In an attempt to model nucleus densities with Potts models, different density levels were assigned to spin variables and images were studied in the framework of the Potts model. It has been shown that most of the configurations present in the density patterns found in microscopy images (like the one in figure 4.1) are in the vicinity of the phase transition emerging in a system defined by the Hamiltonian (4.1) [181]. In this region the model is very sensitive to the value of the temperature-parameter and thus the model in its classical form is not suitable to the accurate study of the images.

4.1.2 The Modified Potts Model

Because of the phase transition, the Potts model generates samples which are either mostly random (above the critical point) or mostly ordered (below the critical point). Therefore generating configurations with smoother spatial transitions as the ones observed in microscope images is impossible. We either observe random densities or blocks of completely uniform densities.

The domain structure of the latter is caused by the kronecker simbol in the Hamiltonian (4.1) which drives the energy-function causing the aggregation of similar states. This leads us to the idea of introducing interactions between different states in somewhat similar fashion to the so-called *XY-model* [182] but in a more “open-minded” way. While the

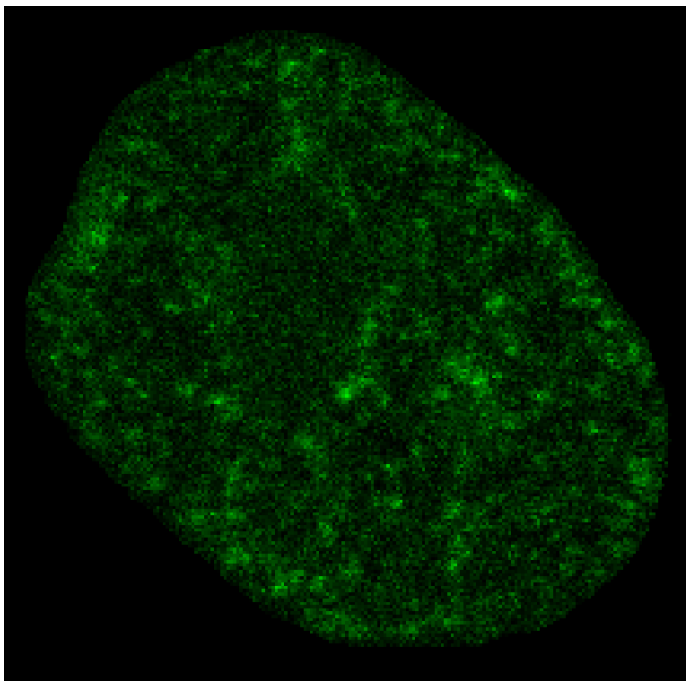

Figure 4.1: Sample of an ES cell.

XY model applies scalar products of the spin directions as interaction strength, we allow arbitrary interaction values. That is, we define the model by the Hamiltonian

$$H = -\sum_{<i,j>} J_{\sigma_i \sigma_j} - \sum_i h_{\sigma_i}, \tag{4.3}$$

where $J_{\sigma_i \sigma_j}$ is a qxq symmetric matrix and h is a q dimensional vector. We define the model without further constraints on the parameters and when modelling the images, we simply learn them from the images we model.

4.1.3 Learning

As mentioned, the Potts model has been extensively used in image processing for segmentation purposes. In fact, the model is the member of a larger model-family called Markov Random Fields (MRFs) [134], a particular class of the model-group referred to as graphical models. Graphical models aim to use the results of graph theory and probability theory to build powerful frameworks for statistical analysis, image processing, computer vision, machine learning, etc. As the Potts model is a special case of the MRFs, methods and techniques developed in the mentioned fields can be applied.

The main problem in image processing applications is finding the value of the temperature- and magnetic field-like parameters which would generate distributions represented by some given sample-configuration. The difficulty in inference problems is caused by the calculation of the partition function which contains exponentially many terms and quickly becomes intractable. As a consequence, the likelihood function cannot be calculated. Therefore, maximum likelihood estimation is often problematic. Luckily, there are plenty of alternatives.

Traditional Approach in the Ising Case

To demonstrate how estimating parameters is possible we start with a simple example, the $q = 2$ Potts model, also known as the Ising model. In this case parameters can be inferred based on counting spins with certain neighborhood configurations [183]. In the Ising model a spin can point either up or down (usually encoded by the values of $+1$ and -1) and the Hamiltonian is given by

$$H = -J_{Ising} \sum_{<i,j>} \sigma_i \sigma_j, \tag{4.4}$$

where the $< i, j >$ notation means neighboring i and j spins, that is the summation is over edges of the network defined by the lattice. σ_i denotes the state of the spin i, and it has a value of either $+1$ or -1.

From the Boltzmann distribution (4.2), the probability of spin k being in the state γ, given the rest of the configuration can be given as

$$\begin{aligned} P(\sigma_k = \gamma | \{S\} \setminus S_k) = \\ &= \frac{e^{-\beta\left[-J_{Ising}\sum_{(i,j)\in\mathbb{E}\setminus edg(S_k)}\sigma_i\sigma_j - J_{Ising}\gamma\sum_{j\in n(k)}\sigma_j\right]}}{\sum_{\gamma=\pm1} e^{-\beta\left[-J_{Ising}\sum_{(i,j)\in\mathbb{E}\setminus edg(S_k)}\sigma_i\sigma_j - J_{Ising}\gamma\sum_{j\in n(k)}\sigma_j\right]}} \\ &= \frac{e^{-\beta\left[-J_{Ising}\gamma\sum_{j\in n(k)}\sigma_j\right]}}{\sum_{\gamma=\pm1} e^{-\beta\left[-J_{Ising}\gamma\sum_{j\in n(k)}\sigma_j\right]}}, \end{aligned}$$

where $\mathbb{E}$ is the set of pairs of indexes pointing to interacting spins, in the present case, the nearest neighbor spins, $edg(S_k)$ is the set of pairs of indexes pointing to interacting spins, one of the spins always being spin S_k and $n(k)$ is the set of indexes pointing to the neighbors of spin S_k. Since J_{Ising} and β are not independent of each other, we can absorb β in J_{Ising}. Thus, finally we get

$$P(\sigma_k = \gamma | \{S\} \setminus S_k) = \frac{e^{J_{Ising}\gamma\sum_{j\in n(k)}\sigma_j}}{e^{-J_{Ising}\sum_{j\in n(k)}\sigma_j} + e^{J_{Ising}\sum_{j\in n(k)}\sigma_j}}. \tag{4.5}$$

On the other hand, because of the local Markov property, the state of the spin k depends only on the state of its neighbors. Therefore,

$$P(\sigma_k = \gamma | \{S\} \setminus S_k) = P\left[\sigma_k = \gamma | n(S_k)\right], \tag{4.6}$$

where $n(S_k)$ is the set of spins interacting with spin S_k (the neighbors of S_k). Since in the conditional probability (4.5) the neighborhood of S_k appears only as a summation of the spin values in the neighborhood, we can characterize $n(S_k)$ with this sum. Assuming only nearest neighbor interactions, this sum can take only five values: -4, -2, 0, 2 and 4. Therefore, we can give the condition in Eq. (4.6) as

$$P\left[\sigma_k = \gamma | n(S_k)\right] = P\left[\sigma_k = \gamma | \sum \sigma_{n(k)} = \alpha\right], \tag{4.7}$$

where α can take one of the five mentioned values. Moreover, since spins are indistinguishable, we can estimate this probability as

$$P\left[\sigma_k = \gamma | \sum \sigma_{n(k)} = \alpha\right] = \frac{N_\alpha^\gamma}{\sum_\gamma N_\alpha^\gamma},$$

where N_α^γ is the number of spins pointing in the direction of γ with its neighbors summing up to α. As a consequence, we can calculate the ratio of the following conditional probabilities:

$$\frac{P\left[\sigma_k = \gamma | \sum \sigma_{n(k)} = \alpha\right]}{P\left[\sigma_k = -\gamma | \sum \sigma_{n(k)} = \alpha\right]} = \frac{N_\alpha^\gamma}{N_\alpha^{-\gamma}}. \tag{4.8}$$

Replacing the left side of equation (4.8) with the closed forms of the conditional probabilities obtain in equation (4.5), we obtain the following ratio:

$$\frac{e^{J_{Ising}\gamma\alpha}}{e^{-J_{Ising}\gamma\alpha}} = \frac{N_\alpha^\gamma}{N_\alpha^{-\gamma}}. \tag{4.9}$$

Using equation (4.9) we can estimate the J_{Ising} coupling as

$$J_{Ising} = \frac{1}{2\alpha\gamma} \log\left(\frac{N_\alpha^\gamma}{N_\alpha^{-\gamma}}\right). \tag{4.10}$$

Stochastic Approximation Procedure

While the method described in subsection 4.1.3 does not generalize to more complicated models, plenty of approaches have been developed for estimating parameters in Markov Random Fields.

The common starting point of most approaches is the derivative of the log-likelihood. This derivative for the model defined in Eq. (4.4) can be given as

$$\begin{aligned}\frac{\partial \log P(\theta|\{S_0\})}{\partial\theta} &= \frac{1}{\partial\theta}\left[\sum_{<i,j>} J_{\sigma_i\sigma_j} - \sum_i h_{\sigma_i} - \log Z\right] \\ &= \phi_\theta(S_0) - \frac{\partial \log Z}{\partial\theta},\end{aligned} \tag{4.11}$$

where $\theta = \{J_{ab}|a,b \in \{1,2,\ldots,q\}\} \cup \{h_a|a \in \{1,2,\ldots,q\}\}$ is the vector of parameters, S_0 is the observed data for which we want to estimate the θ parameters and $\phi_\theta(S)$ is the so-called potential corresponding to parameter θ. For instance, the potential for a given J_{ab} parameter can be given as

$$\phi_{J_{ab}} = \sum_{<i,j>} \delta_{a\sigma_i}\delta_{b\sigma_j}. \tag{4.12}$$

Similarly

$$\phi_{h_a} = \sum_i \delta_{a\sigma_i}. \tag{4.13}$$

The last term in the derivative (4.11) defines an average over the model distribution:

$$\frac{\log Z}{\partial\theta} = \frac{\sum_S \phi_\theta(S) e^{-\sum_{<i,j>} J_{\sigma_i\sigma_j} - \sum_i h_{\sigma_i}}}{Z} = \sum_S \phi_\theta(S)P(S) = \langle\phi_\theta\rangle_m, \tag{4.14}$$

which, as we already discussed this, is intractable.

Different methods handle this problem with different approaches. Some methods calculate a pseudo-likelihood [184], others, like Markov chain Monte-Carlo (MCMC) maximum likelihood estimation , use importance sampling to estimate the partition function [185].

In the present study we used a method called stochastic approximation procedure (SAP) [186], which make use of MCMC sampling to estimate the model average $\langle \phi_\theta \rangle_m$. SAP had been found to work very well on Markov random fields.

Let us define a Markov process characterized by a transition probability $\pi(S_t \rightarrow S_{t+1})$ which satisfies the detailed balance condition

$$P(S_t)\pi(S_t \rightarrow S_{t+1}) = P(S_{t+1})\pi(S_{t+1} \rightarrow S_t). \tag{4.15}$$

We initialize M Markov chains with a random initial configuration $\{S^1_{t=1}, S^2_{t=1}, \ldots, S^M_{t=1}\}$. The θ parameter vector is also initialized with random values. We simulate the Markov chains and after each Monte-Carlo step we calculate the $\langle \phi_\theta(S^t) \rangle_{MC}$ average of the potentials over the Monte-Carlo samples. Based on these averages we update the parameters according to the update rule

$$\theta = \theta + \eta \left[\phi_\theta(S_0) - \langle \phi_\theta(S^t) \rangle_{MC} \right],$$

where η is the learning rate and it is decreased after each update. In case we have a set $\{S^1_0, S^1_0, \ldots, S^N_0\}$ of observed data sample, we can replace the $\phi_\theta(S_0)$ with the $\langle \phi_\theta(S_0) \rangle$ average calculated over the observed samples. In this case, the update rule is modified and it writes as

$$\theta = \theta + \eta \left[\langle \phi_\theta(S_0) \rangle - \langle \phi_\theta(S^t) \rangle_{MC} \right]. \tag{4.16}$$

To achieve a better learning performance, we first set all J interactions to zero and learn the h magnetic-field-like parameters which govern the intensity distribution of the images. Then, using the learned fields, we restart the Markov chains and learn the J couplings. Furthermore we only need to learn the upper triangular part of J since J is supposed to be symmetric.

The stochastic approximation procedure is summarized in algorithm 3.

In this work we used the Metropolis algorithm to calculate the $\pi(S_t \rightarrow S_{t+1})$ transitions.

4.1.4 Testing

Before training our model in real-world data, we perform a few tests in order to make sure that the method will produce the desired results.

Three different test-scenarios were devised, all based on samples generated with known parameters. In the first scenario we analyze how the change of the number of allowed states affect the mean squared error of the learning. In the second test we investigate how the number of learning samples affect the learning error. In the final test we check how varying inhomogeneity of the interactions effects the learning.

Changing the Number of States

In this test we generate samples with known parameters using the original q-state Potts model and investigate how q influences the efficiency of the learning. Changing q influences the number of parameters of the model as the model has $q(q+1)/2$ interaction parameters in J and q parameters in h.

Figure 4.2 plots the mean squared error as the function of the number of Monte-Carlo updates. The figure illustrates how the estimated parameters approach the real ones as the learning progresses, indicating a converging learning process.

Algorithm 3 Estimating the θ parameters with the SAP

procedure ESTIMATE($\{S_0^1, S_0^1, \ldots, S_0^N\}, \eta_0$)
  for $a, b \in \{\overline{1..q}\}$ **do**
    $J_{ab} \leftarrow 0$
    $h_a \leftarrow rand$
  end for
  for $m \in \{\overline{1..M}\}$ **do**
    $S_1^m \leftarrow rand$
  end for
  $t \leftarrow 1$
  $\eta \leftarrow \eta_0$
  while h NOT converged **do**
    for $m \in \{\overline{1..M}\}$ **do**
      Sample S_{t+1}^m using $\pi(S_t \rightarrow S_{t+1})$
    end for
    for $a \in \{\overline{1..q}\}$ **do**
      $h_a \leftarrow h_a + \eta \left[\frac{1}{N} \sum_{n=1}^{N} \sum_i \delta_{a\sigma_i}(S_0^n) - \frac{1}{M} \sum_{m=1}^{M} \sum_i \delta_{a\sigma_i}(S_t^m) \right]$
    end for
    Decrease η
  end while
  for $a, b \in \{\overline{1..q}\}$ **do**
    $J_{ab} \leftarrow rand$
  end for
  for $m \in \{\overline{1..M}\}$ **do**
    $S_1^m \leftarrow rand$
  end for
  $t \leftarrow 1$
  $\eta \leftarrow \eta_0$
  while h NOT converged **do**
    for $m \in \{\overline{1..M}\}$ **do**
      Sample S_{t+1}^m using $\pi(S_t \rightarrow S_{t+1})$
    end for
    for $a \in \{\overline{1..q}\}$ **do**
      for $b \in \{\overline{a..q}\}$ **do**
        $J_{ab} \leftarrow J_{ab} +$
        $\eta \left[\frac{1}{N} \sum_{n=1}^{N} \sum_{<i,j>} \delta_{a\sigma_i}(S_0^n) \delta_{b\sigma_j}(S_0^n) - \frac{1}{M} \sum_{m=1}^{M} \sum_{<i,j>} \delta_{a\sigma_i}(S_t^m) \delta_{b\sigma_j}(S_t^m) \right]$
      end for
    end for
    Decrease η
  end while
end procedure

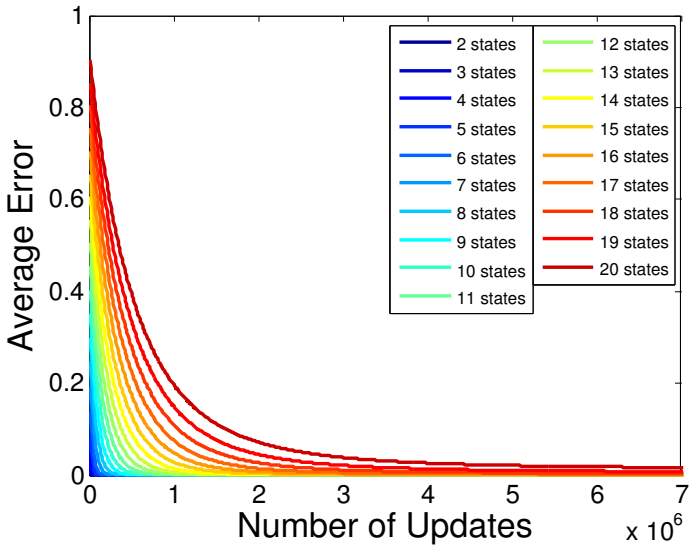

Figure 4.2: The mean squared error of the learned parameters for samples generated with different number of allowed states as a function of the number of Monte-Carlo updates of the learning process. The error decreases as the learning progresses.

Figure 4.3 plots the standard error at the last Monte-Carlo step as the function of the number of allowed states of the Potts model (i.e. q). We observe that the error increases as we allow more states. This is a consequence of the decreasing statistics for increasing number of potentials.

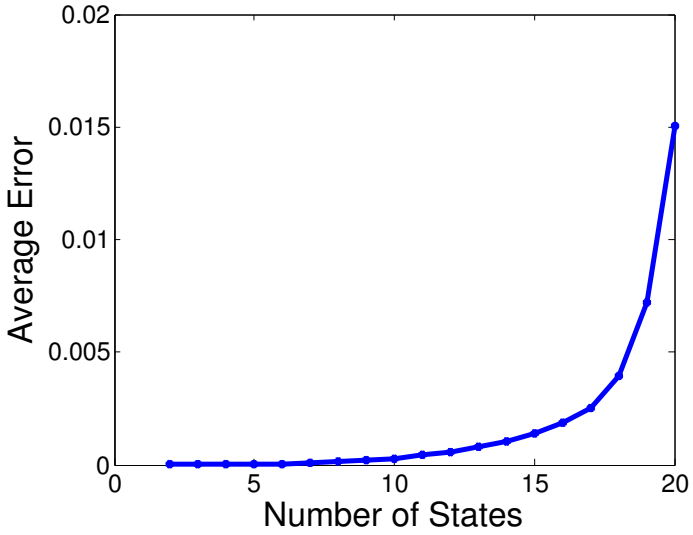

Figure 4.3: The mean squared error of the parameters as a function of the number of allowed states q.

Changing the Number of Samples

In this test we generate samples with known parameters for the $q = 10$ state Potts model (keeping q) constant. We analyze how the number of learning samples influences the

learning error.

In figure 4.4 the mean squared error is plotted as a function of the number of Monte-Carlo updates. The error is again presenting a decreasing trend as the learning progresses. This means that the learning converges in all cases.

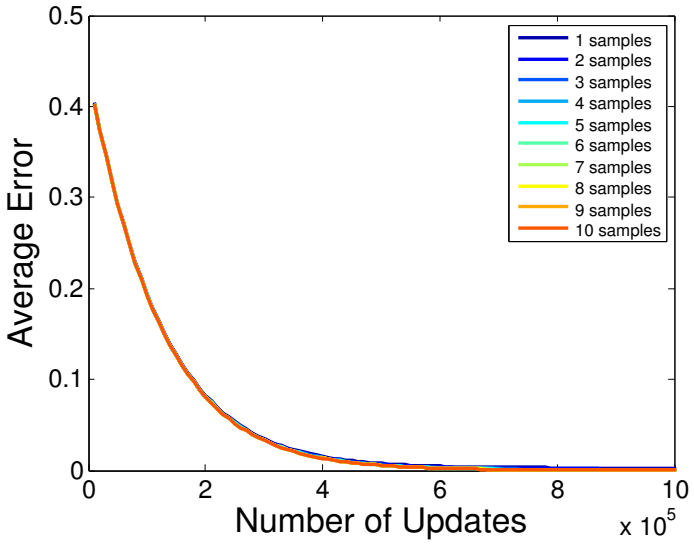

Figure 4.4: Mean squared error of the learning as a function of the number of Monte-Carlo updates. The number of parameters is kept constant while the number of learning samples varies.

Figure 4.5 plots the error at the last Monte-Carlo update as the number of the input learning samples (S_0). The figure demonstrates that the error decreases as we feed more and more learning samples to the algorithm. Increasing the number of learning samples increases the statistics over the potentials, thus assuring a more precise update of the parameters. Assuring enough input data, we can counter-balance the increased learning error when the number of parameters is large.

Changing the Homogeneity of the Parameters

This test-case studies how the learning behaves when samples are generated within the framework of the generalized Potts model. The number of states is kept constant at $q = 10$, the number of learning samples is also constant (using 20 samples). Instead we change here the structure of J when generating the samples. While in the previous test-cases J was strictly diagonal and h constant, here we step-by-step increase the off-diagonal values of J and study how does this increase effect the learning efficiency. The increase is constrained so that values on the different diagonal lines parallel to the main diagonal are kept constant.

Figure 4.6 presents the mean squared error for this test case as the function of the Monte-Carlo updates. We again see a converging learning procedure. Note that the decrease of the error is non uniform.

Figure 4.7 plots the mean squared error for the present test case at the last Monte-Carlo step. It is interesting to note that at an inhomogeneity level of 4 the error exhibits a peak. This means that learning is difficult in this case. This is probably caused by a combination of parameters which define a system close to a phase transition. This may prevent the system to easily explore the configuration-space. However, since in real-world scenario we

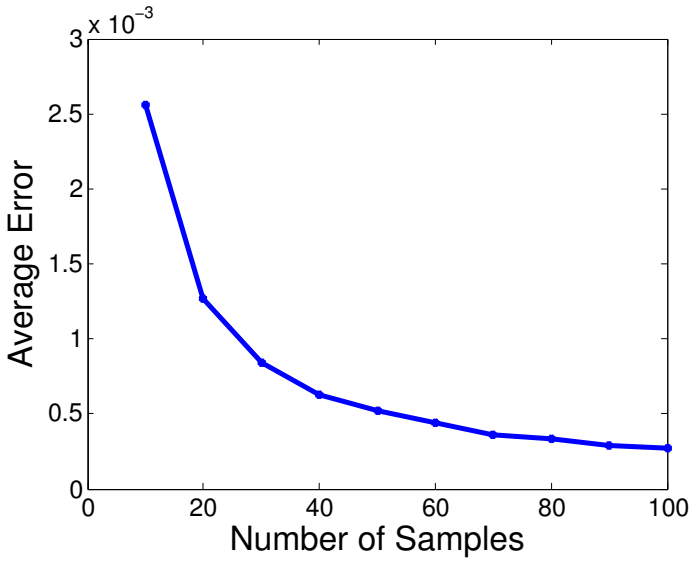

Figure 4.5: Mean squared error of the learning at the last Monto-Carlo update as a function of the number of input samples. The error show a decreasing trend as the number of samples is increased.

allow a total inhomogeneity of J we are far away from the region where learning might be less efficient.

4.1.5 Applying the Method on Real-World Images

We apply the learning procedure on conventional confocal microscopy images of mice cells. Samples were prepared so that activate genes were labeled with antibodies raised against the trimethylation of histon H3 on lysine 4 (H3K4-Me_3). An example image sampled from an ES cell is presented in figure 4.1. The model is trained on data obtained from two different cell lines: mouse embryonic stem (ES) cells and 3T3 fibroblasts.

Fifty 100x100 images were randomly sampled from 30 cells and discretized to 20 different density levels ($q = 20$), assuming that fluorescence intensity is proportional to the density. We use these samples as training data for the learning algorithm. The J and h parameters were learned with the described SAP. Figure 4.8 presents the progression of a learning process of the J interaction matrix. The plot illustrates how after a transition period each entry from J converges to a stable value.

In addition to the interaction parameters in J, it is also possible to learn the boundary conditions. This is useful as it is known that the structure of the chromatin might differ close to the nuclear envelope. Boundaries can be introduced in the model as an extra state, and assigning an extra row and column in J for this state.

After learning the interaction and field parameters, we can use the Monte-Carlo process used for the learning to generate samples. Figure 4.9 displays such a configuration generated using parameters learned from 3T3 fibroblast images. Visually comparing it to a real 3T3 sample (figure 4.10), they appear very similar.

Of course, we can also measure similarities, for instance the intensity distributions. Figures 4.11 and 4.12 present the intensity distributions calculated on ES and 3T3 cells and on the corresponding generated configurations, respectively. We observe a very good agreement of these distributions.

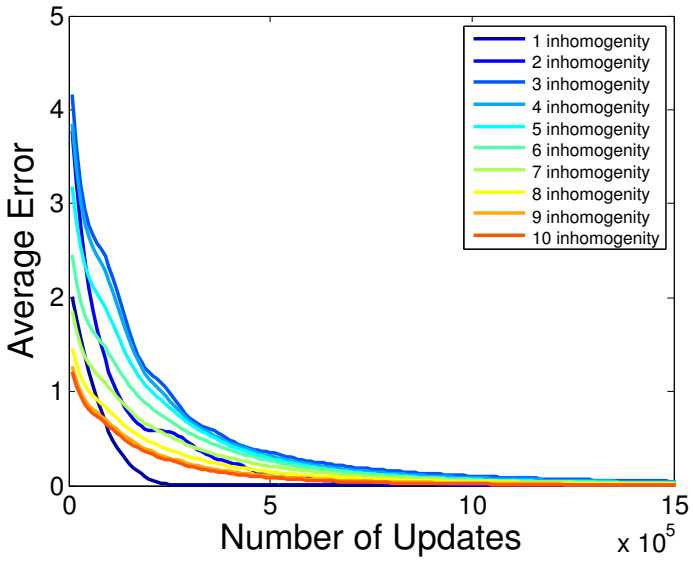

Figure 4.6: The figure illustrates the mean squared error as the function of the Monte-Carlo updates. In this test, both the number of states and the number of learning samples was kept constant while varying the structure of J by adding non-zero off-diagonal elements. The parameter inhomogeneity with a value of x in the legend means that the first $x-1$ diagonals from the main diagonal of J are non-zero.

4.1.6 Discussion and Conclusions

We presented a novel approach for modelling confocal microscopy images. The approach is based on a modification of one of the most known and successful models of statistical mechanics, the Potts model. While in the original version the Potts model considers interactions only among spins pointing in the same direction (or being in the same state), in the modified version we introduced interactions between all spin-states. This allows the model to exhibit configurations which present a local mixture of the states. While in the original model a mixed configurations obligatorily meant a disordered state, this is not the case any more in the modified model as these states may arise from the complex interactions of the spins.

When density patterns are modelled in the framework of the Potts model and states of the model represent density levels, the modified model enables a smooth transition between densities. The original Potts model would generates patterns which are either random, corresponding to a high-temperature case, or, below the critical temperature, ordered same-state blocks would be observed. While the behaviour of the latter model is suitable for image segmentation when objects need to be identified by such same-state blocks, this behaviour was not desired and thus eliminated in the present approach by the applied modification.

We introduced a stochastic approximation procedure which learns the model parameters in two phases. First the magnetic-field like parameters are learned. Then, with the obtained field parameters implemented in the model, the method learns the interactions in a second state. We tested our approach on artificially generated data. We demonstrated that the learning process always converges to the real generating parameters.

We applied the learning successfully on real world data obtained from mouse ES and 3T3 cells. Besides learning an "interaction" between pixel intensities, we also learned

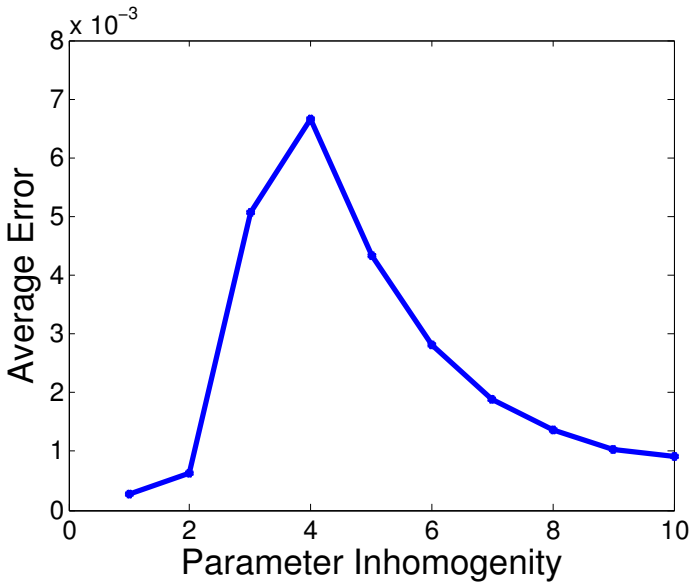

Figure 4.7: The mean squared error at the last Monte-Carlo update of the test case 4.1.4. The horizontal axis represents the parameter inhomogeneity and its value represents the number of non-zero diagonals in the upper (or lower) triangular part of J. It is interesting to note that at inhomogeneity 4 we observe a peak of the standard error. This means that learning is difficult in this case.

the boundary conditions found in the images, thus making sure that the structure of the generated images resembles the original ones also close to the boundary.

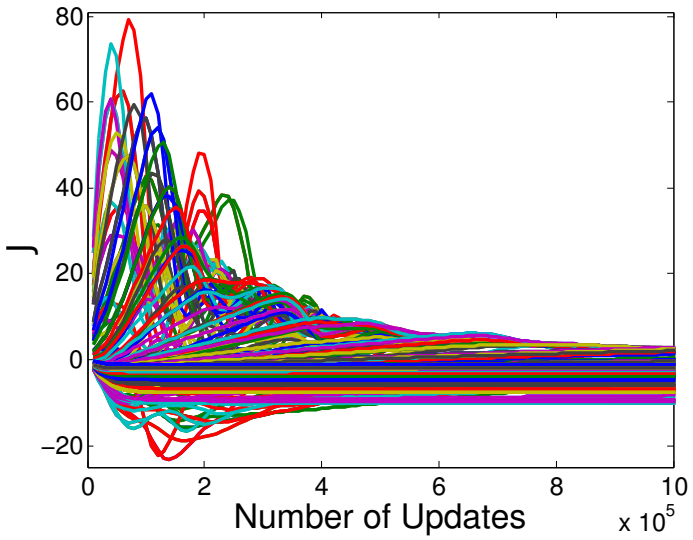

Figure 4.8: Example of the learning process of J. Each curve corresponds to an entry from the interaction matrix.

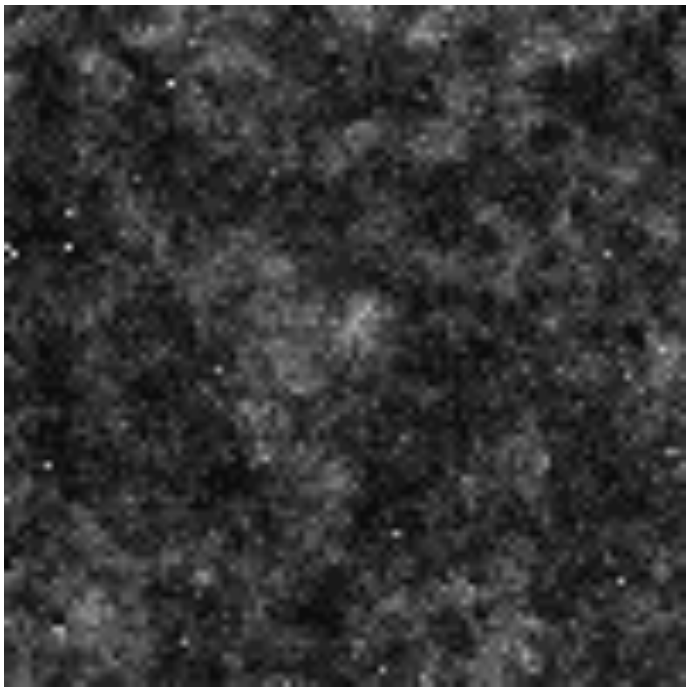

Figure 4.9: Configuration generated using parameters learned on 3T3 fibroblast images.

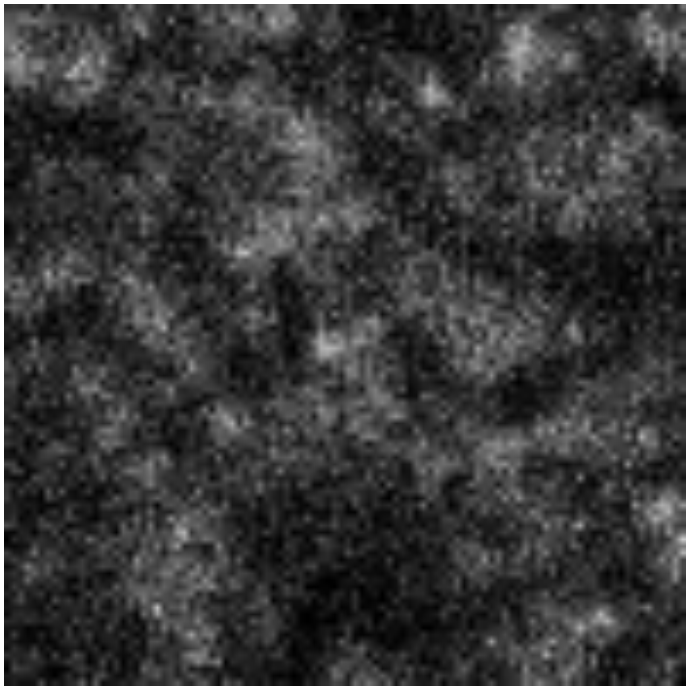

Figure 4.10: A slice from a 3T3 fibroblast cell nucleus.

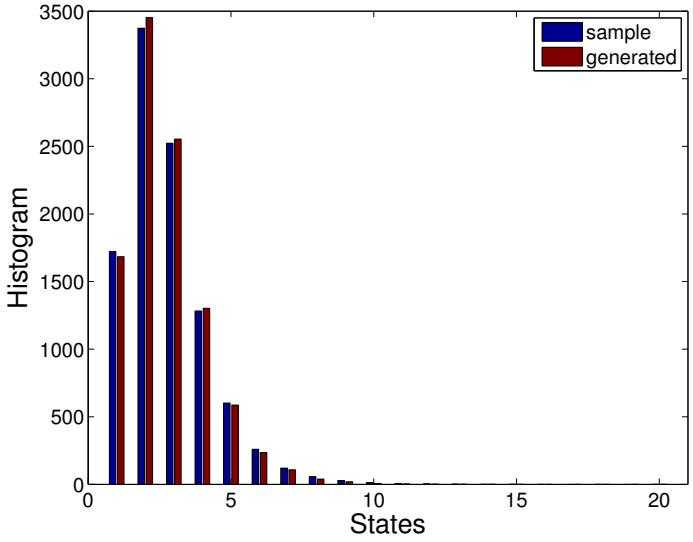

Figure 4.11: Intensity distributions of ES cells and configurations generated with parameters learned from ES cell images.

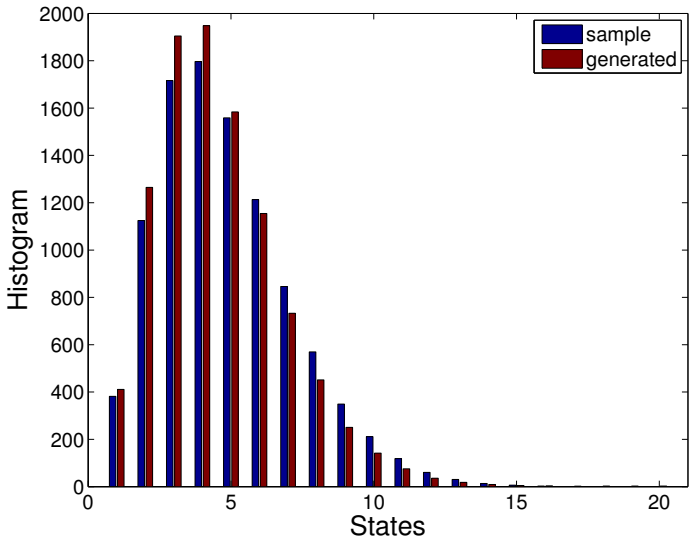

Figure 4.12: Intensity distributions of 3T3 cells and configurations generated with parameters learned from 3T3 cell images.

4.2 Thermodynamic Properties of the Generalized Potts Model

References

This section is adapted from our manuscript, which is intended to be submitted for publication,

- G. Máté, R. Dickman and D.W. Heermann, *A State Dependent Potts Model*, in preparation (2013).

Since its first appearance, the Potts model [187] gained popularity not just among physicists but also scientists from other fields [188–194]. It was successfully applied in grain growth simulations, image segmentation, modelling social and ecological phenomena and data, etc. [176, 178, 195, 196]. Inasmuch as the model is a member of the broader Markov Random Field family [197], the general tools developed in the latter framework can be easily applied to a given instance of the model. This also enables the useful generalization of the model and as it was shown already in the early developments (Ising, vector Potts, XY), the model generalizes very well. In this study we investigate a promising modification of the Potts model, which contains the aforementioned generalizations as special cases.

The very well known *q-state Potts model* [187] is described by the Hamiltonian

$$H_c(S) = -J^c \sum_{<i,j>} \delta_{\sigma_i \sigma_j}, \tag{4.17}$$

where S is a configuration, σ_i is the state of the spin with linear index i in S, and J^c is the coupling constant. σ_i can have q states corresponding to q different discrete spin orientations. Let us denote these states by $\alpha_1, \alpha_2, \ldots, \alpha_q$. Although studying the model one may get insight in the behaviour of ferromagnetic systems [198], the fact that the interactions are present only between similar states (similar spin directions) is a very strong limitation of the model. The *XY model* [182] lifts this limitation by introducing interactions between non similar states through the scaler product of the spins:

$$H_{XY}(S) = -J^{XY} \sum_{<i,j>} \vec{\sigma_i}\vec{\sigma_j}. \tag{4.18}$$

One can rewrite the Hamiltonians (4.17) and (4.18) by discretizing the directions and placing all the possible interaction scalers ($\delta_{\sigma_i\sigma_j}$, $\vec{\sigma_i}\vec{\sigma_j}$) to the appropriate positions of a matrix

$$H_c(S) = -\sum_{<i,j>} J^c_{\sigma_i\sigma_j},$$
$$H_{XY}(S) = -\sum_{<i,j>} J^{XY}_{\sigma_i\sigma_j},$$

where $J^c_{\sigma_i\sigma_j} = \delta_{\sigma_i\sigma_j}$ and $J^{XY}\sigma_i\sigma_j = \vec{\sigma_i}\vec{\sigma_j}$. As a matter of fact, nothing constraints us from implementing an arbitrary function of $\vec{\sigma_i}$ and $\vec{\sigma_j}$, provided that the given function is symmetric with respect to its variables. Therefore, we will use an arbitrary q dimensional symmetric square matrix J as a generalized coupling between the spins and write the H Hamiltonian of the generalized model as

$$H(S) = -\sum_{<i,j>} J_{\sigma_i\sigma_j}. \tag{4.19}$$

Of course this induces the broadening of the parameter-space but also makes the model more flexible and capable of exhibiting a variety of interesting phases as this will be shown in the following. Since this work aims to introduce the possibilities offered by this generalization and not to exhaust the study of it, we will constrain ourselves in the numerical studies to two dimensional systems and cases when the value of the parameters on the individual diagonals of J are the same. We are, of course, aware that going from two to three dimensions can drastically influence the behaviour of the system. We also believe that, when it comes to apply the model for instance to biological data, the real strength of the formalism is the case when J is only required to be symmetric. However, in this case we would have to deal with $(q+q^2)/2$ parameters for a given value of q and this is inappropriate when studying the model. Nevertheless we will make additional comments, if possible, on the cases when all the parameters are independent.

4.2.1 The Ising case

For the Ising ($q = 2$) case [199] J must be a 2×2 symmetric diagonal matrix with equal values on the diagonal:

$$J_{\text{Ising}} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}. \tag{4.20}$$

Using the proposed generalization, we have non-zero off-diagonal elements, and thus

$$J = \begin{pmatrix} a & b \\ b & a \end{pmatrix}. \tag{4.21}$$

One can rewrite J in the following way:

$$J = \frac{(a-b)}{a} \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + b = \frac{(a-b)}{a} J_{\text{Ising}} + b. \tag{4.22}$$

As the system in the canonical ensemble can be characterized by the Boltzmann distribution, which gives the probability of a given configuration S as

$$p(S) = \frac{1}{Z} e^{-\beta H(S)}, \tag{4.23}$$

where $\beta = 1/k_B T$ is the inverse temperature, k_B being the Boltzmann constant and Z is the partition function. $H(S)$ is the energy of a configuraiton S calculated with the Hamiltonian (4.19). It is obvious that the $(a-b)/a$ pre-factor in Equation (4.22) in fact multiplies β. Therefore, the model will behave like the original Ising model on a temperature $Ta/(a-b)$. While the $+b$ constant on the right-hand side of Equation (4.22) shifts the energy scale, it does not change the distribution.

Note that a particular combination of a and b parameters is not independent from β. β, therefore, can be absorbed in J and the system can be parametrized only with the value of a and b. For instance, a system with $a = 1$, $b = 0.5$ at an inverse temperature $\beta = 2$ will be equivalent with a system parametrized with $a = 2$, $b = 1$ at inverse temperature $\beta = 1$. The absorption of β in J is equivalent with fixing $\beta = 1$ and modifying only the a and b values.

Note also that for $b > a$ we get an antiferromagnetic Ising model. Furthermore, the case when we have different values on the diagonal, that is

$$J' = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}, \tag{4.24}$$

corresponds to a situation in which the model would include a special external field, but this field acts on pairs of spins of the preferred state. In this sense the field influences neighbourhoods. One can understand this by rewriting the Hamiltonian (4.19) in the following way:

$$H(S) = -\sum_{<i,j>} J'_{\sigma_i\sigma_j} = -\sum_{<i,j>} J_{\sigma_i\sigma_j} - h\sum_{<i,j>} \delta_{\sigma_i\sigma_2}\delta_{\sigma_i\sigma_j}, \tag{4.25}$$

where $h = a_2 - a_1$. If h is positive, the model prefers neighbourhoods with state σ_2, otherwise σ_1 will be favoured.

4.2.2 The $q = 3$ case

As the classical Ising model was extensively studied in vast amount of studies (see for instance references [200–204]) we proceed to the $q = 3$ case. Similarly to (4.22) we can rewrite J as follows:

$$J' = \begin{pmatrix} a' & b' & c' \\ b' & a' & b' \\ c' & b' & a' \end{pmatrix} = \begin{pmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{pmatrix} + c' = J + c', \tag{4.26}$$

using the notations $a = a' - c'$ and $b = b' - c'$. Compared to the reduced interaction matrix J, c' only rescales the energy. Therefore we study the system defined by J (obeying the tacit constraints $c' < a'$ and $c' < b'$).

The Bragg-Williams Approximation

Since a $b > a$ situation resembles an antiferromagnetic model, we divide the system into two sublattices according to a chessboard pattern, the black sites constituting one lattice (denoted by A) while the white sites the other (denoted by B). We then rewrite the Hamiltonian in the following form:

$$H = (b-a)\sum_e \sum_{k=1}^{3} \delta_{\sigma_{e_A}\alpha_k}\delta_{\sigma_{e_B}\alpha_k} + b\sum_e (\delta_{\sigma_{e_A}\alpha_1}\delta_{\sigma_{e_A}\alpha_3} + \delta_{\sigma_{e_A}\alpha_3}\delta_{\sigma_{e_A}\alpha_1}) - 4N_s b, \tag{4.27}$$

where the summation with index e is over the edges between neighboring spins on different sublattices, and N_s is the number of spins in a sublattice (i.e., $N_s = N/2$).

At high enough temperatures we can replace the kronecker functions by their averages, this is known as the *Bragg-Williams approximation* [205]:

$$\sum_e \delta_{\sigma_{e_A}\alpha_k}\delta_{\sigma_{e_B}\alpha_k} = 4N n_k^A n_k^B, \tag{4.28}$$

where $n_k^L = N_k^L/N_s$, N_k^L being the relative number of spins found in state α_k on lattice L. Therefore the Hamiltonian will have the following form:

$$H = 4N\left[(b-a)\sum_{k=1}^{3} n_k^A n_k^B + b n_1^A n_3^B + b n_3^A n_1^B - b\right]. \tag{4.29}$$

The number of microstates on lattice L can be given as

$$\Omega^L = \frac{N_s!}{N_1^L! N_2^L! N_3^L!} \tag{4.30}$$

and the total number of microstates is $\Omega = \Omega^A \Omega^B$. The entropy is then

$$S = S^A + S^B = -N \sum_L \sum_{k=1}^{3} n_k^L \ln n_k^L, \tag{4.31}$$

Based on these, the free energy writes as

$$\begin{aligned} F \quad &= H - TS = \\ 4N \quad &\left\{ (b-a) \sum_{k=1}^{3} \left[n_k^A n_k^B - T \frac{n_k^A \ln n_k^A + n_k^B \ln n_k^B}{4(b-a)} \right] + b n_1^A n_3^B + b n_3^A n_1^B - b \right\}. \end{aligned} \tag{4.32}$$

The first term in the curly bracket corresponds to the free energy of the original Potts model [206] at a shifted temperature. The shift depends on the $(b-a)$ difference. Although we do not believe that in reality the relation is as simple as this approximation predicts it, nevertheless we conclude that larger b values require larger a values in order to reach criticality, that is, the critical point is shifting according to a relation between a and b.

Note that the last terms in the expression (4.32) of the free energy are symmetric in states α_1 and α_3, but do not include state α_2. Whenever states α_1 and α_3 appear in a configuration, the energy is increased. Since state α_2 does not cause an energy increase by this term, α_2 will be preferred against α_1 and α_3. This is only the case if the value of parameter b is relevant compared to the value of a. In the $b \ll a$ limit we are approaching the Potts case, thus the extra terms are vanishing. However, we shell see state α_2 dominating the system if $b \gg a$, which corresponds to an antiferromagnetic interaction among neighbouring states α_1 and α_2 and neighbouring states α_3 and α_2.

If we think about the Bragg-Williams approximation as an expansion of the free energy, it is important to note, that the coefficients of the terms in the approximation depend on the a and b model parameters and possibly vanish for certain combinations of these. This may be also the case for higher order terms in a non-linear expansion. This observation will be important when interpreting the Landau expansion; as a matter of fact, we sketched the Bragg-Williams approximation specifically to emphasize this point.

A Landau Expansion of the Model

As we already noted before, depending on the model parameters, the system is able to exhibit ferromagnetic as well as antiferromagnetic behavior. Therefore we need order parameters which are able to capture this propriety. Since the n_k^L quantities are not independent ($n_1^L + n_2^L + n_3^L = 1$), following reference [207], we define the following independent combinations:

$$\begin{aligned} \psi_{1L} &= n_2^L \\ \psi_{2l} &= n_1^L - n_3^L, \end{aligned}$$

where L can be either A or B. To observe both ferromagnetism and antiferromagnetism it is useful to consider the following order parametrs:

$$\begin{aligned} \psi_1 &= \psi_{1A} + \psi_{1B} \\ \psi_2 &= \psi_{2A} + \psi_{2B} \\ \phi_1 &= \psi_{1A} - \psi_{1B} \\ \phi_2 &= \psi_{2A} - \psi_{2B} \end{aligned}$$

In case of ferromagnetic behavior ψ_{iL} is independent of L, therefore ϕ_1 and ϕ_2 are zero while in case of antiferromagnetic behavior the net magnetization is zero, that is $\sum_L \psi_{iL}$ vanishes, that is ψ_1 and ψ_2 are both zero.

The system exhibits two important symmetries: with respect to the exchange of states 1 and 3 (let us denote this symmetry operation by R) and with respect to the exchange of subblattice A and B (operation denoted by P). Then the following relations hold:

$$\begin{aligned} R\psi_1 &= \psi_1 \\ R\psi_2 &= -\psi_2 \\ R\phi_1 &= \phi_1 \\ R\phi_2 &= -\phi_2 \\ P\psi_1 &= \psi_1 \\ P\psi_2 &= \psi_2 \\ P\phi_1 &= -\phi_1 \\ P\phi_2 &= -\phi_2. \end{aligned}$$

Therefore, ψ_1 may appear in all terms, while the rest of the order parameters are allowed to appear only in even terms. The most generic expansion can be given as:

$$\begin{aligned} A = A_0 + a_1\psi_1 &+ a_2\psi_1^2 + a_3\psi_1^3 + a_4\psi_1^4 + \\ &+ b_2\psi_2^2 + b_4\psi_2^4 + c_3\psi_1\psi_2^2 + \\ &+ d_2\phi_1^2 + e_2\phi_2^2 + f_4(\phi_1^2 + \phi_2^2)^2. \end{aligned}$$

As it can be seen, the Landau expansion predicts a first order transition between the disordered and ferromagnetic state and a second order transition between the disordered and the antiferromagnetic state, depending of course on the model parameters.

Theoretically, it would be possible to express all the ψ_i and ϕ_i in terms of n_k^L. Then, we could identify the Bragg-Williams approximation in this form, extended with higher order corrections, and eventually could find a direct relation between the coefficients of the Landau expansion and the a and b model parameters. According to this argument, and as we already pointed it out when discussing the Bragg-Williams approximation, the coefficients of the Landau expansion may depend not only on the temperature, but also on the model parameters. They may even vanish for certain combinations, changing the nature of the phase transition.

Phase transitions in the system

In order to have a first glimpse in the behaviour of the system we measured two quantities, the order parameter defined by

$$r = \frac{(n_1 - n_2)^2 + (n_1 - n_3)^2 + (n_2 - n_3)^2}{2N^2}, \tag{4.33}$$

where, again, n_k is the number of spins found in the state α_k and N is the total number of spins ($N = L^2$). The second quantity measured is the energy per spin:

$$E = \frac{\sum_{<i,j>} J_{\sigma_i\sigma_j}}{N}. \tag{4.34}$$

We generated samples using a simple Metropolis Monte Carlo procedure. In our attempt to show that the model is capable of interesting behaviour, we scanned only a fraction of

the parameter-space, namely the $a \in [0, 2]$ and $b \in [0, 2]$ intervals. Figure 4.13 presents the values of the order parameter for the specified intervals of a and b with nearest neighbour interactions between the spins. It is interesting to see that the location of the phase

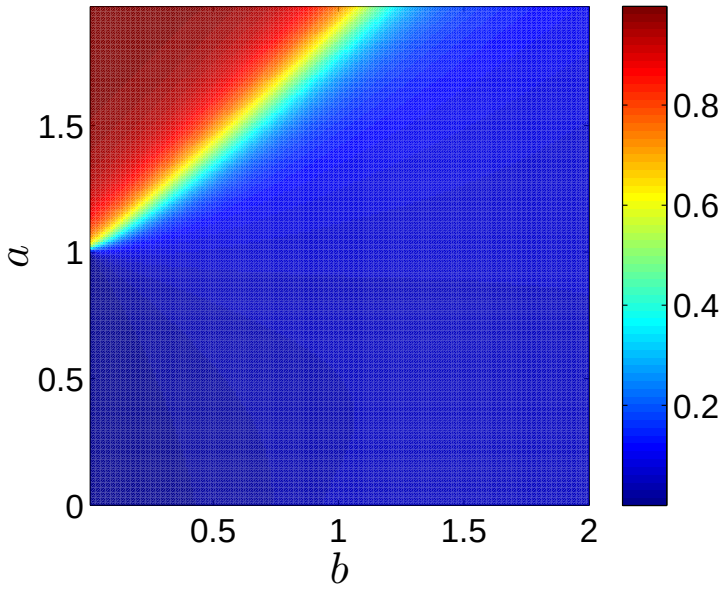

Figure 4.13: Order-disorder transition in the system. The order parameter (4.33) is represented by the color-code. We observe a shift in the phase transition also predicted by the Bragg-Williams approximation.

transition can be roughly defined by a straight line described by the $a \approx 1+0.96b$ empirical law, a shift already predicted by the Bragg-Williams approximation. For $b = 0$ (original Potts model), we get $a \approx 1$. The value is expected to be around $\ln(1 + \sqrt{q}) \approx 1.0051$, therefore, the agreement is relatively good.

The value of the energy per spin is presented in Figure 4.14. Analysing the plot we observe two low energy regions. One is for big a values and small b values. This regions corresponds to the ordered ferromagnetic phase. The other region is for small values of a and big values of b. This suggests that there is another type of ordering in this region, namely an antiferromagnetic ordering.

Figure 4.15 presents a particular configuration generated with parameters $a = 0.4$ and $b = 1.9$. As predicted in the Bragg-Williams approximation, if $b > a$, state α_2 (represented by the green coloured squares in the configuration) dominates the configuration, occupying roughly half of the spins. The other half is shared among state α_1 (orange) and state α_3. Dividing the configuration into two chessboard-like sub-lattices, α_2 occupies almost entirely one of the lattices. Furthermore, because we have only nearest neighbour interaction (spins do not interact diagonally), the other sub-lattice is simply a random configuration of states α_1 and α_3. Although it does not seem random, calculating the auto-correlation of a single spin, presented in Figure 4.16, we see that the autocorrelation decays with time exactly as it would with random spin-flip dynamics. Here we simulated systems at $a = 0$ and $b = 1$. As discussed earlier, the $\beta = 1$ case leaves a and b unchanged. The $\beta = \infty$ case (equivalent with $a = 0$ and $b = \infty$) corresponds to a system at zero

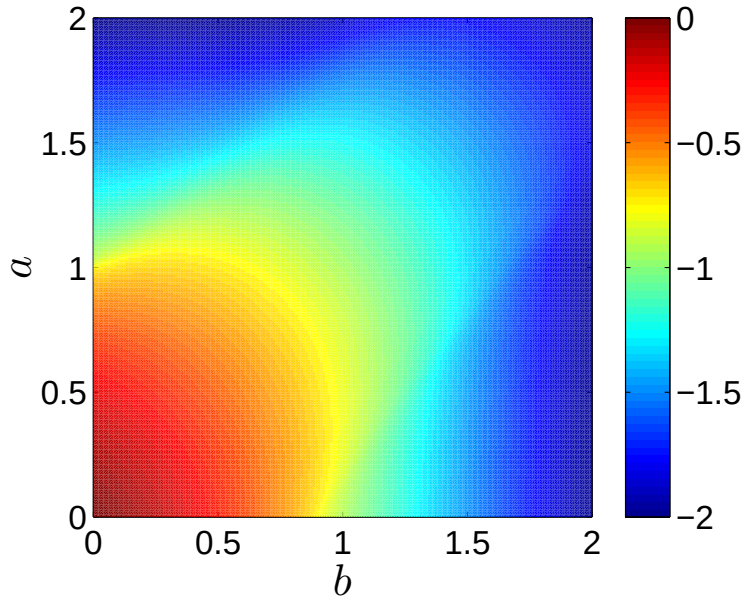

Figure 4.14: Value of the energy per spin. The values represented by the color-code are calculated with the formula (4.34). Note the low energy region in the lower right hand side of the figure. This region suggests the existence of an antiferromagnetic ordering as here $b > a$.

temperature ($T = 0$). In this case the system should remain forever in its ground state. Samples with random spin flips were achieved by randomly flipping spins and disregarding any interactions (as a matter of fact, this corresponds to the $\beta = 0$ case).

The energy of the system is also low when both a and b are large and approximately equal. However, there is no ordering at these combinations, since in this case neighbouring states α_1 and α_2 and neighbouring states α_2 and α_3 are "allowed" to mix as Figure 4.17 also illustrates this. This figure represents a configuration with parameters $a = 1.8$ and $b = 1.9$. Note how states α_1 and α_2 (orange and green) create interfaces, similarly do α_2 and α_3 (green and blue), while α_1 and α_3 are seldom next to each other.

This phase constitutes a relatively low energy transitional phase between the ferromagnetic ($b \ll a$) and antiferromagnetic ($b \gg a$) phases. Then, it is logical to think that the nature of the disorder-order phase transition changes as we increase the b parameter and from the $b \ll a$ phase we go through the $a \approx b$ phase with the mixed states and eventually reach $b \gg a$ as for large enough b parameters we reach the antiferromagnetic phase. Of course, this phase could be characterized with another order parameter which defines structures composed of a mixture of α_1 states with α_2 states and the mixing of α_2 states with α_3 as order. Studying this order parameter we may find another phase transition.

The chessboard-like pattern in Figure 4.15 was created as a result of nearest neighbour interactions (no diagonal interaction among spins). If we include interaction between next-nearest neighbours (diagonal spins), this pattern will disappear, as Figure 4.18 demonstrates this. The ground state in systems with nearest and next nearest neighbour interactions in the antiferromagnetic ($b \gg a$) case would be composed of rows or columns of spins in state α_1 and α_3 always separated by rows (or columns) of spins in state α_2.

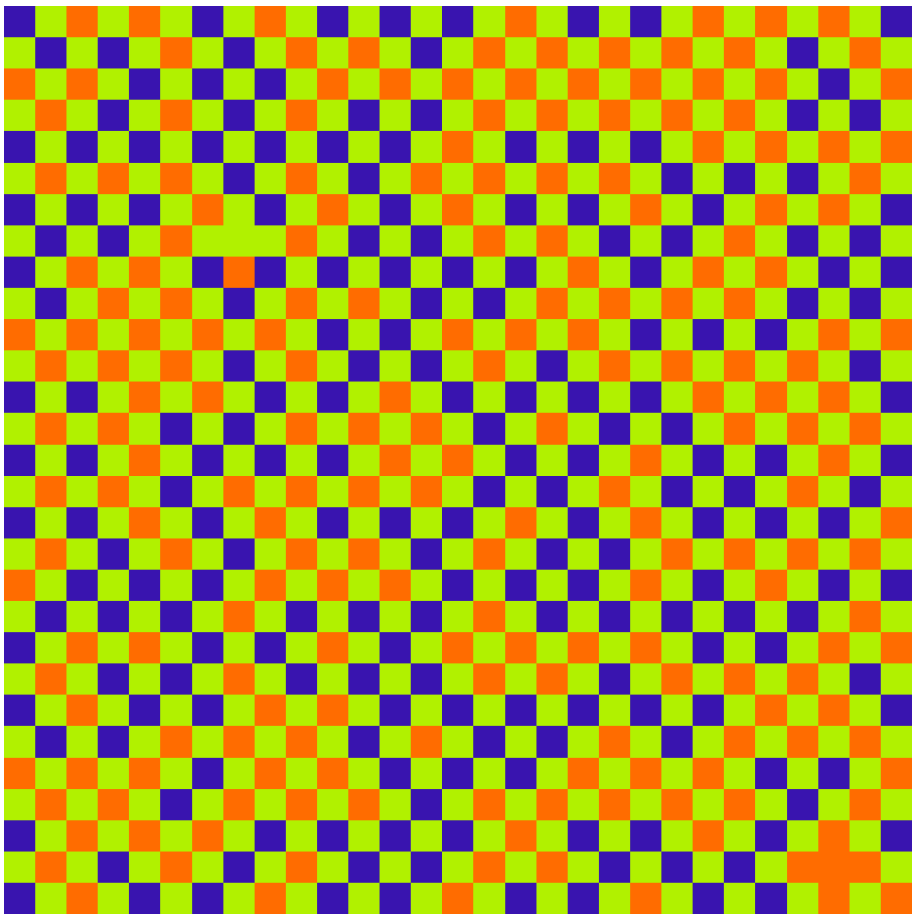

Figure 4.15: A configuration of the system for a parameter combination $a = 0.4$ and $b = 1.9$. State α_1 is coded by the colour orange, state α_2 is represented by the green colour squares, while state α_3 is indicated by the blue squares. Note that the neighbouring spins prefer to be in different states. If the system is divided into two chessboard-like sub-lattices, almost all of the spins in one of the sub-lattices are in state two denoted by the green colour.

Thus state α_2 would also dominate this phase. However, energies are not that reduced in this phase, as the energy per spin plot prepared for a system with nearest and next nearest neighbour interaction shows it in Figure 4.19. However, the disorder-ferromagnetic state phase transition still occurs (Figure 4.20).

4.2.3 The $q = 10$ case

In order to gain further insight into the behaviour of the system, we study the $q = 10$ case with a Wang-Landau type calculation [208]. Instead of sampling the system according to the Boltzmann distribution, the Wang-Landau algorithm aims to estimate the $g(E)$ density of states (or number of microstates for the energy E). Since the probability of hitting an energy E can be given as

$$p(E) = \frac{1}{Z} g(E) e^{-E/k_B T}, \tag{4.35}$$

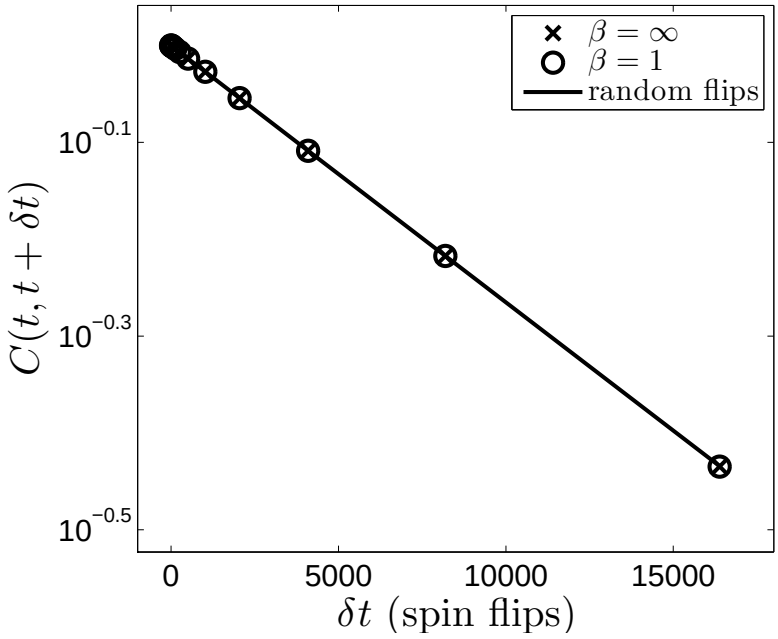

Figure 4.16: The autocorrelation of the spin states as the function of the number of flips in a 128x128 lattice of a system parametrized with $a = 0$ and $b = 1$. The "random flip" case corresponds to $\beta = 0$ ($T = \infty$).

thus we don't have to enumerate all the possible configurations to calculate Z, we can give it as a sum over the energy levels

$$Z = \sum_E g(E) e^{-E/k_B T}. \tag{4.36}$$

Then, we can calculate the free energy as

$$F = -k_B T \ln Z. \tag{4.37}$$

While here we will only look at the aspects of the phase transition, other quantities, of course, can be further derived from the free energy.

The idea of Wang and Landau is very simple and the algorithm approximates the density of states by visiting different random configurations and keeping record of the energy-histogram gathered over the visits by counting how many times was a given energy level E visited. Let us denote this quantity by $v(E)$. The configurations are visited according to a probability distribution dictated by the inverse of the current estimation of the density of states $g(E)$. Since, if randomly picking a configuration, the probability of hitting an energy E is proportional with $g(E)$, visiting a configuration with a probability $1/g(E)$ results in a uniform distribution over E. The aim is thus to construct a $g(E)$ so that the histogram of the visited states $v(E)$ is acceptably flat.

The algorithm initializes the densities by a constant value (e.g. $g(E) = 1, \forall E$). Also, the histogram is filled with zeros: $v(E) = 0, \forall E$. For the random walk in the energy space, we can use the Metropolis algorithm with modified transition probabilities. Since the probability of hitting a state is $1/g(E)$, the transition probability can be given as

$$\pi(E_X \to E_Y) = \min\left[1, \frac{g(E_X)}{g(E_Y)}\right], \tag{4.38}$$

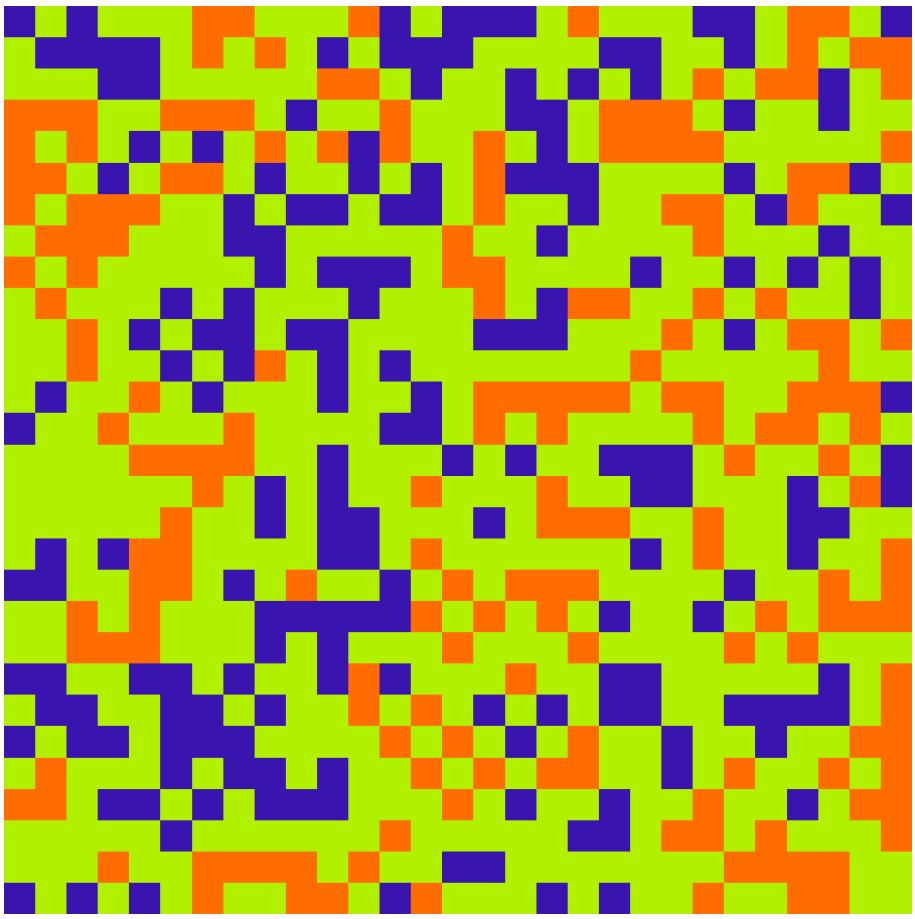

Figure 4.17: A configuration of the system for a parameter combination $a = 1.8$ and $b = 1.9$. State α_1 is coded by the colour orange, state α_2 is represented by the green colour squares, while state α_3 is indicated by the blue squares. Note the mixing of states α_1 with states α_2 and the similar behaviour of states α_2 and α_3. States α_1 and α_3 seldom create interfaces.

where E_X is the energy of the state (configuration) X (for details see the description of the Metropolis algorithm in Section 3.1.4). At each visit of an energy level E, the algorithm updates the histogram as $v(E) = v(E) + 1$ and the density by a multiplicative factor η: $g(E) = \eta g(E)$. While in the first steps v might be considered flat, it quickly develops a peak around the most probable energy values. After v becomes flat enough, η is decreased ($\eta = \sqrt{\eta}$) and the histogram is reset ($v(E) = 0, \forall E$). Then the whole the procedure of equalizing the histogram is repeated, however, this time the changes in $g(E)$ will be smaller because of the smaller factor η. These steps are repeated until a desired precision is achieved.

Note that the algorithm does not rely on the Boltzmann distribution an it is independent of the temperature T and thus from β. Knowing the density of states, we know the behaviour of the system for the whole temperature range.

We implemented the algorithm for the $q = 10$ generalized model defined by the coupling

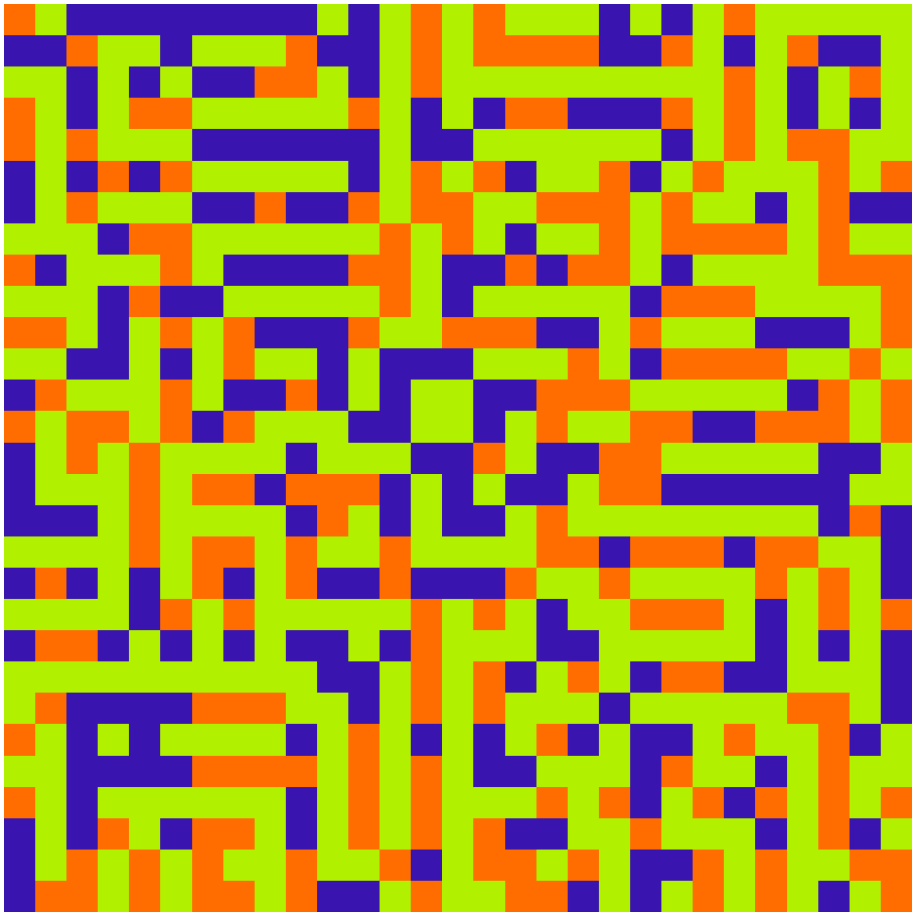

Figure 4.18: A configuration of the system with nearest and next nearest neighbour interactions for a parameter combination $a = 0.4$ and $b = 1.9$. State α_1 is coded by the colour orange, state α_2 is represented by the green colour squares, while state α_3 is indicated by the blue squares. Because of the interactions in the diagonal directions, the lowest energy configurations are rows (or columns) of states α_1 and α_3 always separated by a row (or column) with spins in state α_2. Thus state α_2 will dominate this configurations too. However, since here $a \neq 0$, we observe a mixture of horizontal and vertical bars.

matrix

$$J = \begin{pmatrix} a & b & 0 & 0 & \dots & 0 \\ b & a & b & 0 & \dots & 0 \\ 0 & b & a & b & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 0 & b & a \end{pmatrix}. \tag{4.39}$$

Fixing $a = a_0$ and $b = b_0$ parameters and calculating $g(E)$, we in fact scan a line passing through the origin of the a, b parameter space defined by $b/a = const = b_0/a_0$. For any a_1 value there is a $b_1 = b_0 a_1/a_0$ parameter. As we discussed this earlier, for any b_1/a_1 ratio we can find a temperature T_1 so that the behaviour of the system can be characterized by the Boltzmann distribution of the system defined by $J(a = a_0, b = b_0)$ at temperature T_1.

We implemented a Wang-Landau algorithm (reference) as this enables the investigation of the system for a whole set of parameters with a single run. The reasoning behind this

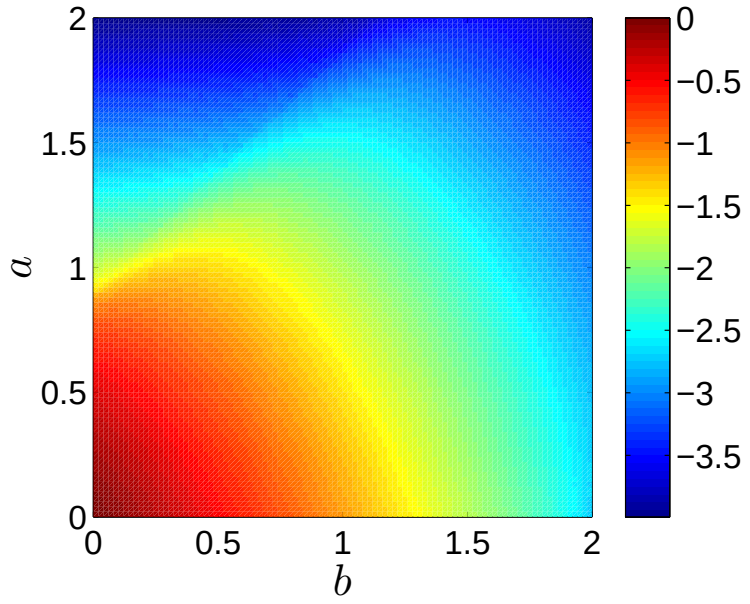

Figure 4.19: Value of the energy per spin. The values represented by the color-code are calculated with the formula (4.34). Note the low energy region in the lower right hand side of the figure. This region suggests the existence of an antiferromagnetic ordering as here $b > a$.

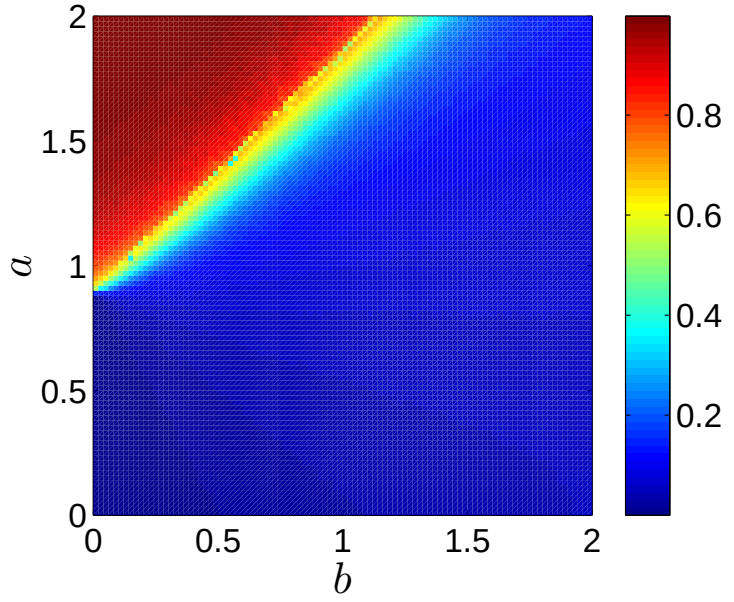

Figure 4.20: Value of the energy per spin. The values represented by the color-code are calculated with the formula (4.34). Note the low energy region in the lower right hand side of the figure. This region suggests the existence of an antiferromagnetic ordering as here $b > a$.

statement is the following: The Wang-Landau algorithm calculates the density of states $g(E)$. From the density of states we can calculate the probability of a given state E at a given temperature T by the relation

We calculated the density of states $g(E)$ for three b/a ratios: 0, 0.36 and 0.53. The corresponding curves are plotted in Figure 4.21. As the figure illustrates it, in the case of

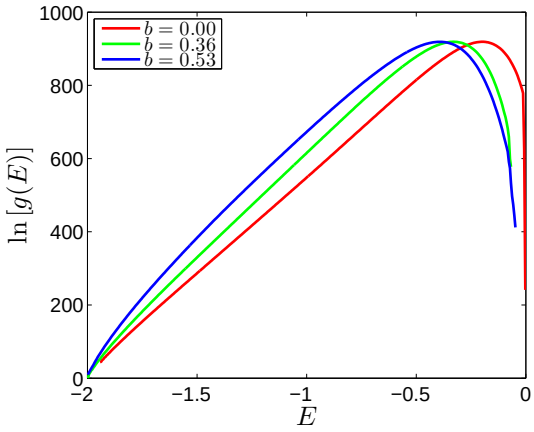

Figure 4.21: Density of states of the $q = 10$ model.

$b/a = 0.53$ the first order phase transition disappears. This is indicated by the concave shape of the function. A first order phase transition would be indicated by a "convex intruder" [209]. The middle point of the $b/a = 0$ case is convex, although this is hardly visible on the figure. Nevertheless, plotting the distribution of the energies at the transition temperature $T_C = 0.7574$, we see the double-peak structure characteristic for first order transitions (Figure 4.22). We found no clear double peak structure for the $b/a = 0.36$

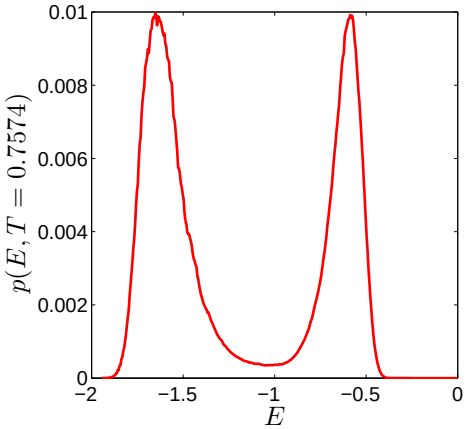

Figure 4.22: Distribution of the energies for the $b = 0$ case at the transition temperature $T_C = 0.7574$.

case, however, looking to the "dynamics" of the curve as we continuously increase T from a low temperature suggests that the phase transition is still present.

To get an insight what is happening in the system as we increase the value of the b/a ratio, we performed classical Metropolis Monte-Carlo simulations for the $q = 10$

case and estimated the transition temperature based on the samples generated by these simulations. We represented the b/a ratio by a single parameter, the angle θ corresponding to the arctangent of the ratio. We plotted the estimated critical temperature as the function of θ as presented in Figure 4.23. Critical temperatures were estimated by the location of the maxima in the specific heat curves. Although the estimation method might nor be the most precise, nevertheless, we can draw useful conclusions from the arrangements we observe. $b/a = 0.36$ corresponds roughly to $\theta = 0.34$ while the arctangent of $b/a = 53$ is around 0.49. While the first case is still in the "safe" range of the plot, where estimations were acceptable. However the transition temperatures estimated for $\theta > 0.45$ were unreliable. The plot indicates that there is no order-disorder transition roughly in the range $\theta \in [0.51]$. Considering numerical errors, this observation is in acceptable agreement with our conclusions made based on the Wang-Landau calculations, however, further calculations are needed to have a more precise comparison.

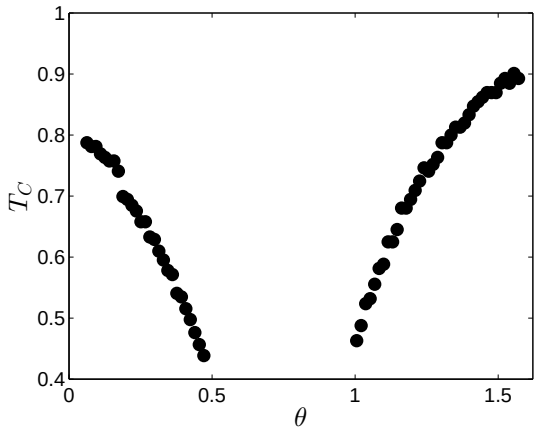

Figure 4.23: The estimated transition temperatures of the $q = 10$ generalized model for the different angles θ corresponding to different $a/b = \tan\theta$ parameter ratios. Estimations for the values corresponding to the interval $\theta \in [0.481]$ were either out of the simulated temperature range or very imprecise, therefore we did not include them.

4.2.4 Discussions and Conclusions

We presented a theoretical and computational study of a generalized Potts model. The generalization consisted of introducing extra couplings among neighbouring spin states. More precisely, if the states of the spins are denoted by $\alpha_1, \alpha_2, \ldots, \alpha_q$, then neighbouring spin configurations of $\alpha_k\alpha_k$ decrease the internal energy by a factor of a while neighbours configured as $\alpha_k\alpha_{k+1}$ decrease the energy by a factor of b ($k = \{1, 2, \ldots, q\}$, $q + 1 = 1$). We investigated the $q = 2$, $q = 3$ and $q = 10$ cases in two dimensions.

We showed by a simple analytical calculation, that the $q = 2$ case behaves like the classical Ising model on a shifted temperature. We elaborated a mean field theory using the Bragg-Williams approximation for the $q = 3$ case. Based on the obtained expression, we concluded that the transition point will shift to lower temperatures as we increase the b parameter of the model. Furthermore, we observed, that state α_2 is favoured by the energy function if the value of b is relevant compared to that of a. We could also predict form the topology of interactions that when $b > a$, the model will have a ground state resembling a chessboard-like arrangement. Because α_2 is preferred against α_1 and α_3, at

low temperatures and $b > a$, one of the sublattices (e.g. the white sites of the chessboard) is occupied by α_2 while the other sublattice corresponds to a random configuration composed of states α_1 and α_3.

Markov chain Monte-Carlo simulations of the $q = 3$ case supported this prediction. Furthermore, we simulated lattices with next-nearest neighbour interactions (diagonal interactions). In this case, the ground state consists of horizontal, or vertical lines of spins of states α_1 or α_3 always separated by a line of spins which are in state α_2. Thus state α_2 dominates this case also. On low finite (non-zero) temperatures, however, the system is in a state composed of a mixture of horizontal and vertical lines.

A Wang-Landau type of algorithm was implemented for the $q = 10$ case. We observed that the phase transition disappeared for certain combination of the parameters. To understand what happened in that combination, we also performed Markov chain Monte-Carlo simulations for the $q = 10$ case and estimated the transition temperature of the model for different combinations of the a and b parameters. We observed no phase transition for a range of parameter combinations. These simulations are thus in agreement with the results obtained by the Wang-Landau calculations.

Chapter 5

Intensitites, Distributions and Graphs

A Direct Insight in the Organization of the Chromatin

References

- G. Máté, M. Tark-Dame and D.W. Heermann, *Quantifying Effects of Light Stress in Arabidopsis Thaliana*, in preparation (2013).
- Y. Zhang*, G. Máté*, P. Müller, M. Hausmann and D.W. Heermann, *Measuring Structural Changes in Chromatin Induced by Ionizing Radiation*, in preparation (2013).
- G. Máté, L. Shopland and D.W. Heermann, *Spatial Relation of Lamin B Receptors and Nuclear Pore Complexes* , in preparation (2013).

*equal contributions

Chapter Summary

This chapter will describe our approaches based partly on the calculation of intensity distributions and their different momenta and partly on analysis of spatial graphs and their properties like the degree distribution or edge-length distribution.

Different types of microscopy images will serve as the basis our analysis. In some cases we will work directly on the images, using a more traditional image processing approach. In other cases, after some preprocessing steps, we derive a spatial graph characterizing the data based on the position of fluorescent markers. However, the relative intensity of these markers can be influenced by different factors (e.g., microscope gain, different densities of the dye molecules during staining, or structural changes). Assuming that images can be described by a probabilistic model, we also assume that the intensities are governed by such a law, too. This, in turn, allows us to correct for differences in microscope gain or to detect uniform/non-uniform structural alterations (e.g. uniform/non-uniform dilation). This is usually done by handling intensities, graph edge lengths, etc. as probabilistic variables, and apply a probability transformation on the distribution characterizing them. With the proper transformation, distributions can be scaled to each other and assessed whether they belong to the same family or not.

5.1 Reorganization of the Chromatin Fiber Under Light Stress

References

This section is adapted from our mansucript, which we intend to submit for publication,

- G. Máté, M. Tark-Dame and D.W. Heermann, *Quantifying Effects of Light Stress in Arabidopsis Thaliana*, in preparation (2013).

Different experimental studies reported an alteration in the chromosomal organization of the model plant *Arabidopsis thaliana* [210–212]. This is attributed to the acclimatization capabilities of the plant. Here we attempt to develop a framework for quantitative analysis of such differences between cells of plants grown under different light conditions.

Our experimental data consists of confocal microscopy images of the nuclei of Arabidopsis thaliana plants. The plants were divided into two groups, the first group was grown under normal light conditions, while the other group was covered with a mesh which block non-selectively 80 % of the light. While naming the first group as control group and the second group as "dark" group would be more appropriate, we will refer to the first group as *light-stressed* and to the second group as *non-light-stressed*. All plants are transinfected to express yellow fluorescence proteins (YFP) attached to H2B histones. Furthermore, A-T rich regions are marked with 4',6-diamidino-2-phenylindole (DAPI) stains.

We will characterize the samples by the density distribution of the chromatin in the nucleus. Assuming stoichiometry, the density should be proportional to the intensity recorded by imaging some uniform labelling along the chromatin. We will investigate whether different reorganizations of the chromatin can be modelled with a density distribution from the same family or not. Furthermore, we calculate different features of the distributions in an attempt to discriminate the two groups. Moreover, the number of chromocenters (regions of highly compactified chromatin) will also determined and compared between the two groups

5.1.1 Analysis

First, the images are subject to standard image processing techniques. A deconvolution with the point spread function of the microscope is performed, then the images are segmented so that regions outside of the nuclei are excluded.

Segmentation

The segmentation is based on a multiple thresholding algorithm applied on the channel defined by the YFP markers. First the intensity values in each image I are projected to the $[0, 1]$ interval. Then the images are thresholded at an intensity $i_t^{(1)}$. Objects defined by neighbouring voxels with intensities above the $i_t^{(1)}$ threshold are detected. The size of each object is determined and smaller objects are remove keeping only the largest one. The resulted binary image I_b thus contains only a single object. Each slice in the I_b image is eroded by 5 pixels.

In the next step, the original image is again thresholded at an intensity $i_t^{(2)} < i_t^{(1)}$, this resulting in an I_m binary image. Objects, that is, groups of connected voxels with intensities larger ore equal to $i_t^{(2)}$ are detected and the object in I_m overlapping with the object in I_b is kept, the rest of the objects are discarded. Since $i_t^{(2)} < i_t^{(1)}$, the volume of the object in I_m is larger or equal than the volume of the object in I_b and possibly contains more details.

A morphological closing [213] is performed on I_m, then holes are filled. Finally I_m is morphologically opened. The object in I_m obtained by this process will be used as a mask for I and will define the region/volume subject to the performed analysis.

The thresholds $i_t^{(1)}$ and $i_t^{(2)}$ were manually selected for the light-stressed and the non-light-stressed set to ensure a best performance of the masking procedure and their value is summarized in Table 5.1.

	non-light-stressed	light-stressed
$i_t^{(1)}$	0.15	0.15
$i_t^{(2)}$	0.12	0.13

Table 5.1: Table summarizing the threshold intensities used in the masking process.

Intensity Distributions

In order to analyse the images, we calculate the $f(i)$ intensity distributions. As Figure 5.1 presents it, the distributions are consistent within the different experiments and staining procedures.

The average distribution over the ensemble of images can be calculate for each particular experiment (Figure 5.2). The averages indicate that that chromatin structure differs between non-light-stressed and light-stress exposed cells. We attempted to fit the curves with known distributions, but we did not find a well-fitting distribution family.

In order to analyse whether differences in the average distributions indeed come from different structure of the chromatin or it is simply a result of different marker-densities (which might be the case especially with the DAPI markers) or different microscope gain, we handle the intensities as random variables. In fact, we already treated them as such when calculating the intensity distributions, as distributions are used to characterize random variables.

Assuming a linear change in the intensity, described by a $t(i) = \alpha i$ function, it can be shown that the distribution $f(i)$ of the rescaled intensities can be given as

$$f(i) = \frac{1}{\alpha} f_0(\frac{1}{\alpha} i), \tag{5.1}$$

where f_0 is the originally measured distribution.

We calculate the rescaled distribution for each of the curves from Figure 5.1, thus fitting them to each other, using α as fitting parameter. This process assures a minimal deviation between the different curves on each plot. After this step the scaled distributions are averaged.

The averaged distributions can be fitted to each other in a similar manner. There exists an α parameter which scales the average of the distributions calculated for the light-stressed cells so that the rescaled distribution has a shape which is the most similar to the

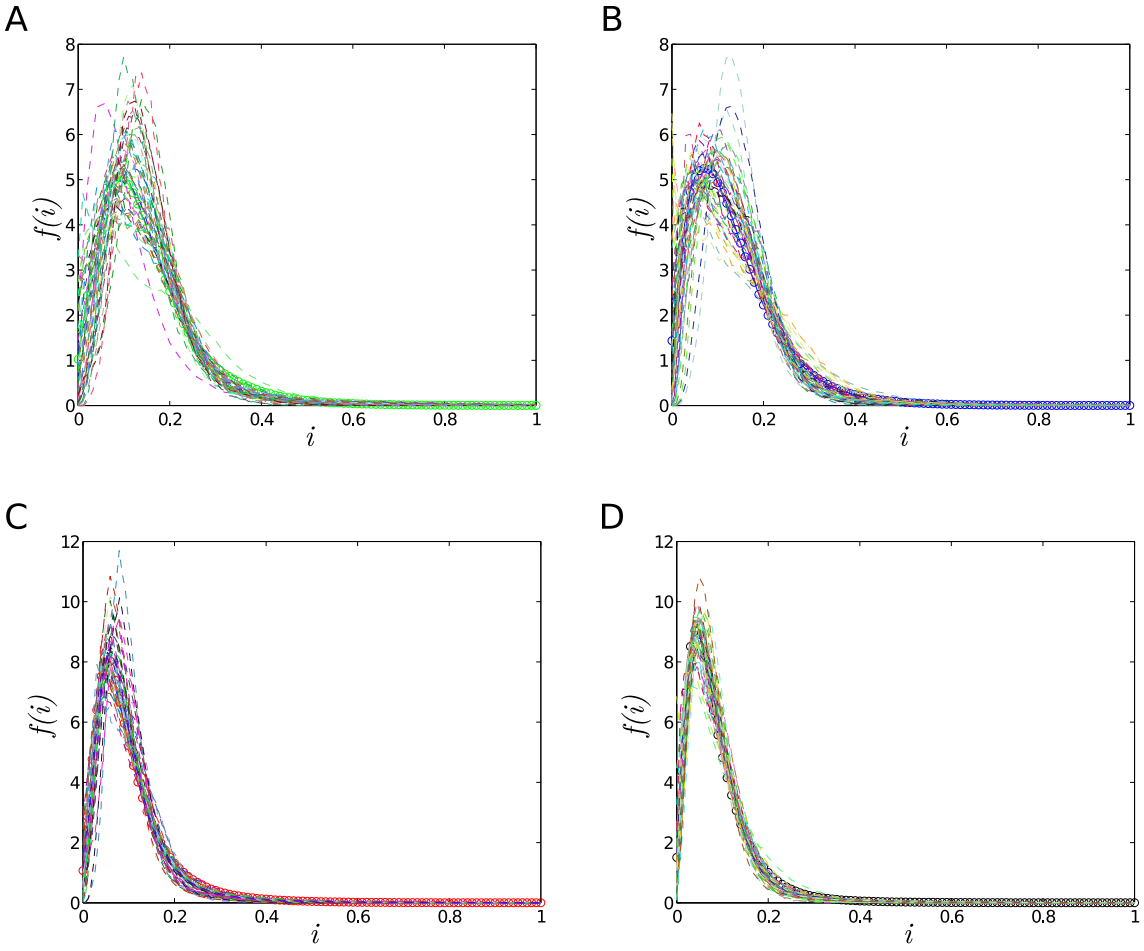

Figure 5.1: Distribution of the normalized intensities for each experimental image separately. Panel **A** presents the distributions for the non-light-stressed set stained with DAPI markers, Panel **B** contains the distributions for the light-stressed cells stained with DAPI. Panel **C** illustrates the distributions exhibited by the YFPs in the non-light-stressed set while Panel **D** presents the distributions calculated on the set of the light-stressed cells for the YFP markers. Dashed lines represent the distributions while lines tracing the circles represent the average of the distributions.

shape of the distribution obtained through the averaging of the distributions corresponding to the non-light-stressed cells.

These fits are presented in Figure 5.3 for the YFP markers and in Figure 5.4 for the DAPI markers. As Figure 5.3 illustrates this, intensity distributions of the YFP markers for the non-light-stressed and light-stressed samples seem to belong to the same family. However, the distributions for the DAPI markers are obviously different. Compared to the distributions calculated for the light-stressed samples, the distribution calculated over the non-light-stressed set grows slower for low intensities and it drops faster for bigger values.

Features of the Images

To further investigate the differences, image features proposed by the authors of [214] are calculated. The features can be calculated directly on the images, or on intensity values scaled by the α parameters obtained in the fitting processes.

Figure 5.5 presents the amount of chromatin in the medium density regions relative

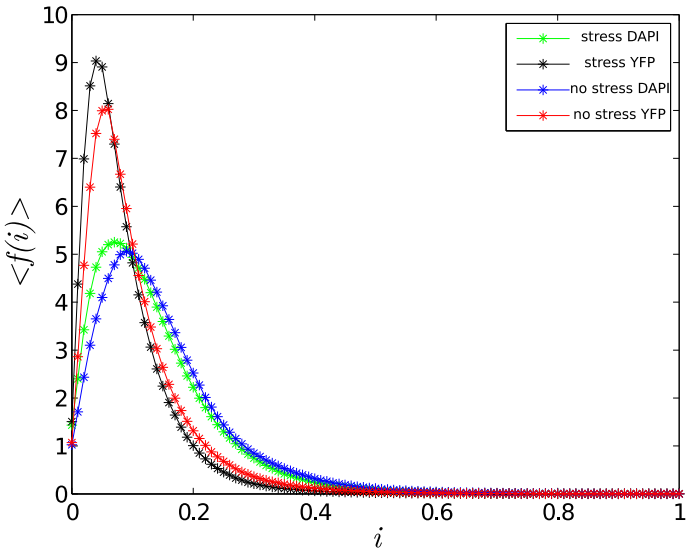

Figure 5.2: Averages of the intensity distribution for the different experimental sets.

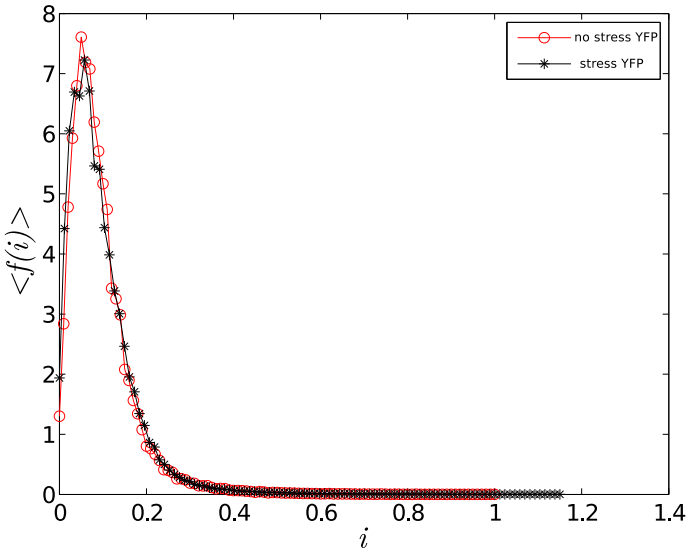

Figure 5.3: Rescaled averaged intensity distribution calculated for the YFP markers.

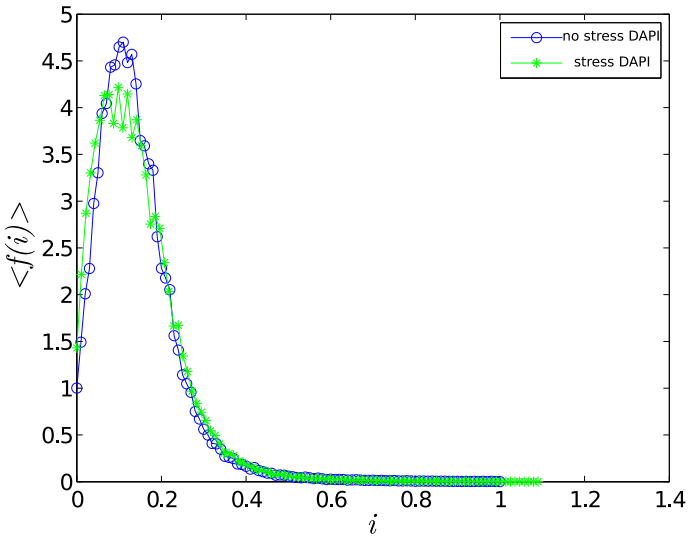

Figure 5.4: Rescaled averaged intensity distribution calculated for the DAPI markers.

to the total amount of chromatin, plotted as a function of the mean intensity. Medium density regions are defined as regions of the nucleus having an intensity ranging from $\langle i \rangle -$ stdev(i) to $\langle i \rangle +$ stdev(i), where stdev(i) is the standard deviation of the distribution and it can be given as $\sqrt{\langle i^2 \rangle - \langle i \rangle^2}$. Although the exact amount of chromatin can not be directly determined, because of stoichiometry the measured fluorescence intensity should be proportional to the density of the chromatin. Therefore, the amount of chromatin in medium dense regions is proportional to the integrated intensity of this region (the sum of the intensity values of all the voxels in this region). Similarly, the total amount of chromatin is proportional with the total integrated fluorescence intensity. Since here we calculate a ratio of amounts, the proportionality constant is canceled and the relative amount of the chromatin can be given as the ratio of the integrated intensity in the medium dense regions and the total integrated intensity.

Figure 5.6 illustrates the same quantity as Figure 5.5 but calculated on the rescaled intensities. The relative amount of chromatin in the mediumly condensed regions and the mean intensity seem to be comparatively good discriminators as domains for non-light-stressed and light-stressed samples can be separated in Figure 5.6. Inspecting the figure, we observe that most light-stressed cells show more chromatin in the medium condensed states than the non-light-stressed cells. The observation holds both for the YFP and DAPI markers.

Chromocenters

A further point to analyse is the difference in the number of the chromocenters in non-light-stressed and light-stressed cells. Chromocenters are detected through a simple thresholding with an i_c threshold, after which objects smaller than a predefined V_c volume are removed. The i_c threshold and the V_c volume were determined based on visual checks of the result. For YFP images i_c was set to 0.3 while $i_c = 0.35$ was used for DAPI markers. In both cases V_c was set to 30 in units of voxel-volume.

Figure 5.7 presents the histogram of the number of chromocenters counted on the YFP markers. The mean number of chromocenters for the non-light-stressed cells is 6.52 with

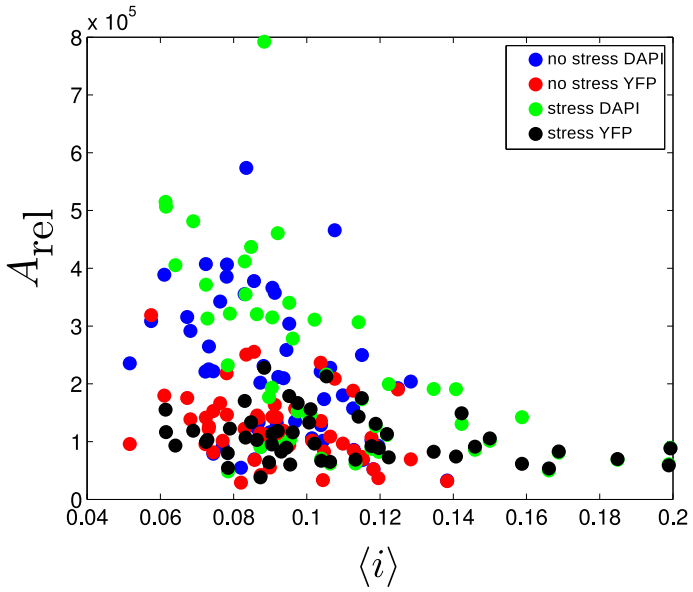

Figure 5.5: Relative amount of chromatin in medium density regions as a function of the average of the intensity distribution. Each point corresponds to an image.

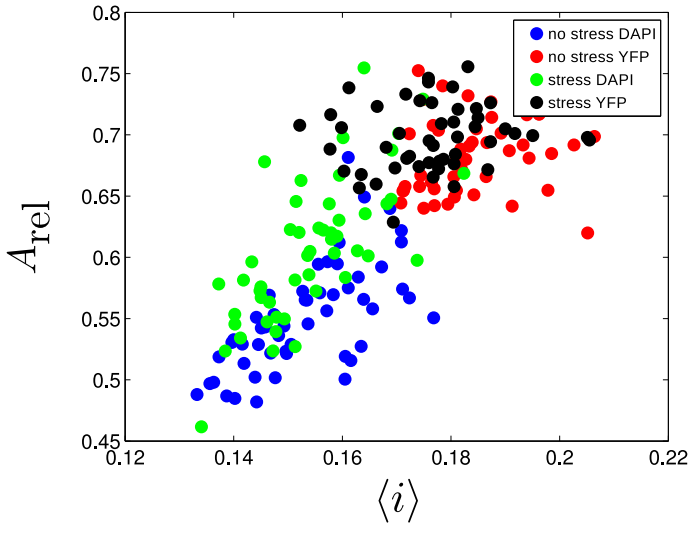

Figure 5.6: Relative amount of chromatin in medium density regions as a function of the average of the intensity distribution. The intensities are scaled so that the overlap of the distributions is maximal. Each point corresponds to an image.

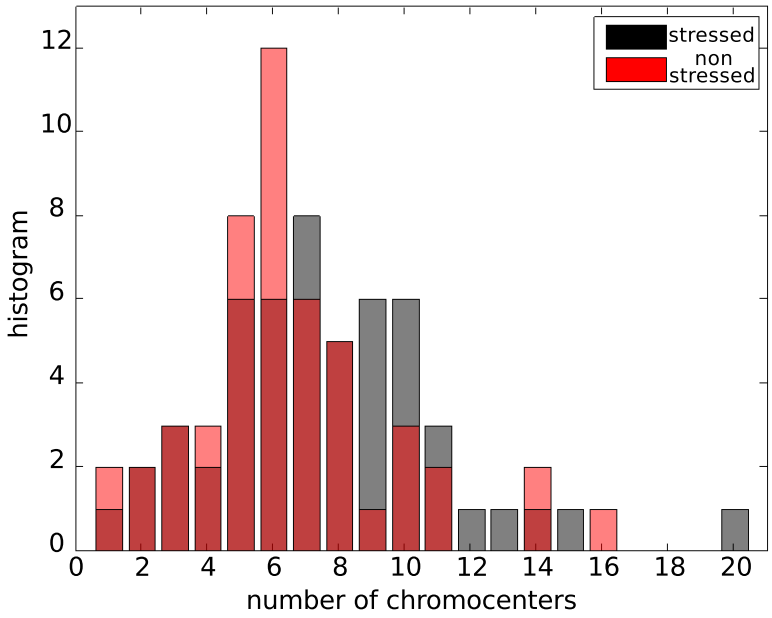

Figure 5.7: Histogram of the number of chromocenters for the YFPs. The average number of chromocenters is 6.52 $\pm$ 0.441 for the non-light-stressed cells, while it is 7.6415 $\pm$ 0.4797 for the light stressed ones

a standard error of 0.441. We find that light-stressed cells contain 7.6415 chromocenters, on avreage, the standard error being 0.4797.

Figure 5.8 illustrates the histograms for the number of chromocenters detected based on the DAPI markers. The average number of chromocenters for the non-light-stressed cells is 6.6 with a standard error of 0.358, while for the light stressed samples we detect 6.8302 chromocenters on average, with an error of 0.4797.

We observ a consistent increas of chromocenters for light-stressed cells, the measured difference of the averages being larger than the standard error for the YFP markers.

5.1.2 Discussion and Conclusions

We analysed confocal microscopy images of Arabidopsis thaliana. Plants were separated into two groups, one of them being grown in normal light condition, while plants in the other group received 80 % less light. The nuclei of plants were imaged with conventional confocal microscopy.

After segmenting the images, fluorescence intensity distribution were calculated. We rescaled the individual distributions in order to aceahve a best fit among the probability densities, thus eliminating effects of different microscope gain and possibly different densities of dye molecules which might appear especially in the case of DAPI markers.

Differences in the distribution between non-light-stressed and light-stressed cells were persistent even after rescaling of the intensities.

We illustrated that measures proposed by the authors of [214] also have a discriminative potential, however, certain measures present obvious correlation. Nevertheless, compared to the non-light-stressed cells, there is an increase in the relative volume of the medium density areas in the light-stress exposed cells.

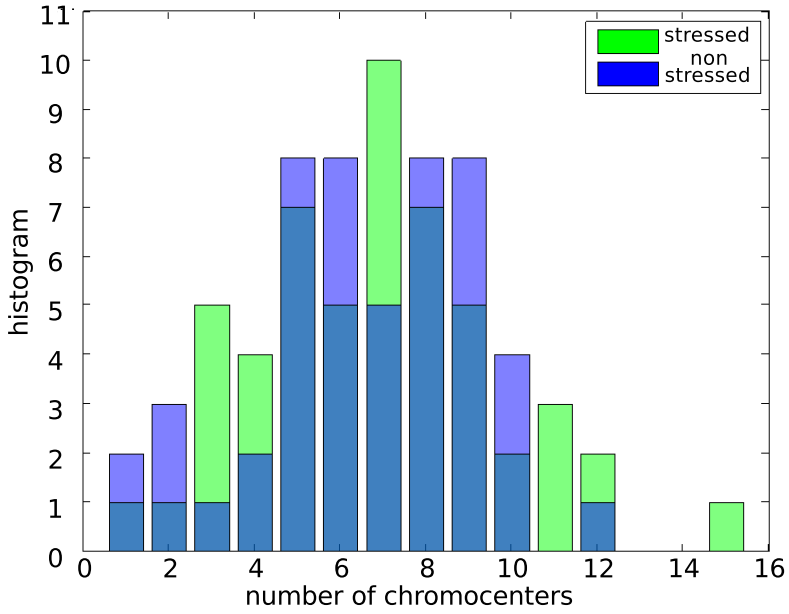

Figure 5.8: Histogram of the number of chromocenters for the DAPI markers. The average number of chromocenters is 6.6 ± 0.358 for the non-light-stressed cells, while it is 6.8302 ± 0.3883 for the light stressed ones

A significant change in the number of chromocenters was also assessed. On average, we counted roughly 1.1 more chromocenters for the light-stress exposed cells.

5.2 Structural Changes and Healing of Irradiated Cells

References

This section is adapted from our mansucript, which we intend to submit for publication,

- Y. Zhang*, G. Máté*, P. Müller, M. Hausmann and D.W. Heermann, *Measuring Structural Changes in Chromatin Induced by Ionizing Radiation*, in preparation (2013).

Masks and the radial distributions were calculated by YZ. Graphs and the measures related to it were computed by GM.

*equal contributions

Double strand breaks (DSBs) are probably the most harmful damages the DNA fibre may suffer as they can lead to the rearrangement of the genome [80]. DSBs are caused by ionizing radiation, oxidizing agents, replication errors and certain metabolic products of the cell [215]. As a response to the DSB, the repair process can take different pathways, the chosen one mostly depending on the phase of the cell cycle. It has been observed experimentally, that as a reaction to the presence of a DSB, the chromatin opens up in the region affected by a DSB [216]. This might be viewed as a mechanism which facilitates the repair.

Since DSBs are the causes of many pathologies, it is crucial to understand the repair mechanisms and the occuring structural reorganization. In this project, we will investigate the effects of ionizing radiation on the structure of the chromatin. Although biological implications are very important, in the following we emphasise the analysis approaches we used to gain an insight in the occurring structural changes.

Experimental Data

Experiments were performed on HeLa cell cultures. Cells were transfected to stably express green fluorescence proteins (GFPs) on histone H2B. Furthermore heterochromatic regions were marked with antibodies tagging the histone H4 Lysine 20 methylation. Images were captured by Spectral Position Determination Microscopy (SPDM), a relatively recently developed super resolution technique, already described in Section 2.2. As a result of the experiment, the most probable location of the fluorophores is recorded as pairs of x, y coordinates in a two dimensional plane. The two dimensional points represent the projection of a roughly 600 nm thick slice from the nucleus.

Cells were irradiated with doses of 0.5, 2 and 4 Gy doses. After irradiations cells were either allowed a 48 hour repair time or fixed after 30 minutes. Furthermore, a control group was also imaged.

5.2.1 Segmentation and Masking of Images

In the first step of the analysis we perform a segmentation of the image in order to detect regions of interest. In this process we exclude the areas that have a considerably lower than average number of localized fluorophores. This includes the area that is on the outside of

the cell nucleus but also areas inside the nucleus that only have a small chromatin content, e.g. at the site of nucleoli.

In the image, the fluorophores or antibodies represent a set of points in the two-dimensional plane. To perform the segmentation we calculate a density distribution for the fluorophores on the entire image by applying a gaussian kernel density estimation [217]. This means that for each point on the image we set a two-dimensional gaussian probability distribution with the coordinates of the point as its mean value. We choose radially symmetric gaussian distribution by setting the bandwidth matrix to a multiple of the unit matrix. The multiplier is a parameter in the model which defines a correspondence between the standard normal distribution and the physical scales in the experiment. We set it to $40\,nm$. The sum of all the gaussian probability distributions is then normalized yielding a probability distribution of the position of the points from the image. Since the values of the probability distribution are proportional to the density of points, the obtained probability distribution in fact represents the spatial density distribution in the nucleus.

Using the density distribution we calculate a mask that accepts all areas with a density that is above 25% relative to the lowest value and blanks out all other areas. In order to make sure that also the low density regions at the border of the cell nucleus are not included, we further erode the masked area by another $150\,nm$. Thus we efficiently exclude areas that are outside of the cell nucleus and areas which have a low marker density indicating sites of nucleoli.

5.2.2 Calculated Measures

In order to analyse the spatial arrangement of the points, we calculate different measures which may characterize the structure. Besides standard measures like the radial distribution function [218], we construct a spatial graph describing the neighbouring relations and analyse the structure of this graph.

Radial Pair Correlation Function

The radial pair correlation function $g(\mathbf{r_1}, \mathbf{r_2})$ is a measure of structure for many particle systems such as liquids. It is the probability distribution function that two particles are at the positions $\mathbf{r_1}$ and $\mathbf{r_2}$. Since absolute positions are not important, the correlation function is in fact only dependent on the directed distance between the two positions $\mathbf{r} = \mathbf{r_{12}} = \mathbf{r_2} - \mathbf{r_1}$. Thus the radial pair correlation function can be simply written as $g(\mathbf{r})$ and is given by

$$g(\mathbf{r}) = \frac{1}{N} \left\langle \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(\mathbf{r} - \mathbf{r_{ij}}) \right\rangle \tag{5.2}$$

The probability to find a second particle in the directed distance $\mathbf{r}$ of the first particle is $g(\mathbf{r})d\mathbf{r}$. Assuming an isotropic system, the pair correlation function only depends on the undirected distance $r = |\mathbf{r}|$ and the probability to find a second particle in the distance r of the first particle is given by $4\pi r g(r)dr$. The expression $\rho(r) = 4\pi r g(r)$ denotes the radial distribution (RDF).

For each image the pair correlation function is determined by calculating all pair distances between points inside of the masked regions of interest. Due to the finite system size, we cannot use all points for the calculation of $g(r)$ since points that are close to the border of the masked image do not have the correct surrounding. We therefore first

reduce the masked image isotropically from the border by a distance d_{shrink} thus yielding a second masked image that is considerably smaller than the first one. All points in the second mask have then equal environments for distances up to d_{shrink}. For each of the points in the second mask we then calculate the distances to all other points that lie in the original mask. The normalized distribution of these distances then give the radial pair distribution function $g(r)$ for the image.

We then average over the complete set of images to obtain the average $g(r)$ for each experimental setup. The standard deviation of the mean value is then used as a measure for the uncertainty of the pair correlation function $g(r)$ at each distance r.

The Spatial Graph of the Neighbourhoods

Another possibility to gain insight in the structure presented by the localization data is to build a graph on the skeleton defined by the localized points. We then investigate this graph by means of graph theoretical methods and conclude the topological or relational properties of the underlying structure.

Graphs [219] are abstract mathematical objects designed to capture interdependencies of certain entities. Entities are represented by nodes (or vertices) while the connections are encoded by edges between nodes. A graph is usually denoted by $G(V, E)$, where V is the set of nodes and $E = \{(a, b)|a \in V, b \in V\}$ is the set of edges. Graphs have been used to study a variety of structures and phenomena [220–224], most of these studies relying on the mathematical field called graph-theory.

We are mainly interested in observing local properties of the geometric arrangements formed by the points. Therefore, we choose the graph building procedure so that the resulting graphs emphasize the local relations among the fluorophores. We build the nearest-neighbor graph (NNG) in which each node is connected with its first-order neighbours.

There are many possible ways to build the NNG for a given set of points. The most widely used approach is the construction of the Delaunay triangulation [225] introduced in Section 3.4. However, for a better understanding, let us give a more intuitive description regarding the construction of the Delaunay triangulation. We start again by introduce the dual of this triangulation, the Voronoi tessellation [226]. By definition, the Voronoi tessellation of a set of points S is a tessellation in which each Voronoi cell V_i corresponding to the site S_i consists of all the points of the space closer to S_i than any other site. The faces of the Voronoi diagram consist of all the points in the space that are equidistant to sites corresponding to touching Voronoi cells. Thus the Voronoi cell V_i defines the space dominated by the site S_i. The Voronoi tessellation can be transformed into the Delaunay triangulation by connecting the sites of the neighbouring Voronoi cells. Two points will be considered neighbours if they are directly connected by an edge of the Delaunay triangulation or equivalently, if their corresponding Voronoi cells are touching. In Figure 5.9 we illustrate the Voronoi tessellation and the Delaunay triangulation for a given set of points.

We construct the NNG defined by the Delaunay triangulation for each experiment and calculate three properties of the obtained graphs: the degree distribution $h(d)$, the rescaled probability density $f(r)$ of the edge lengths and the conditional probability $p(r|r')$ of the edge lengths.

The degree of a node is defined as the number of connections of the node. The degree distribution is the probability distribution of the degrees over the whole graph. Since the degrees are integer numbers, $h(d)$ is a discrete distribution. It is in fact the histogram of

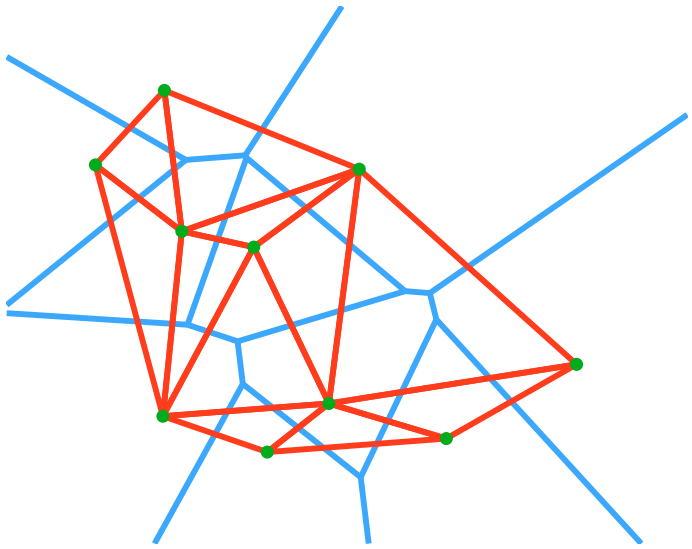

Figure 5.9: The Delaunay triangulation and its dual, the Voronoi tessellation for a random set of points. The blue lines are the segments of the Voronoi tessellation, the red ones are the edges of the Delaunay graph (triangulation)

the degrees of the nodes normalized such that the values in the histogram bins sum up to one. That is, if $\bar{h}(d)$ is the frequency of the degree d then the degree distribution is given as

$$h(d) = \frac{\bar{h}(d)}{\sum_{d'=0}^{d_{max}} \bar{h}(d')}, \tag{5.3}$$

where d_{max} is the maximal degree in the graph. This definition is equivalent with the one in Equation (3.29).

Let us denote the probability distribution of the length of the edges in a given graph $G(E, V)$ obtained by the Delaunay triangulation of a point-set corresponding to one of the localization images by $\bar{f}(r)$. For different experiments, we expect to get different $\bar{f}(r)$ distributions. These differences may stem either from different underlying structures or from different experimental conditions as these can vary from experiment to experiment (slightly different concentration of stains or different microscope gain). Provided that different experimental conditions have a liner effect on the density ρ of marked sites, variations in $\bar{f}(r)$ can be eliminated. This is achieved by rescaling the distributions to a reference density ρ_0. For this, we apply the following procedure: Since we are calculating the $\bar{f}(r)$ distributions, we are, in fact, treating the edge lengths in a probabilistic manner, that is, r is considered a random variable. Thus, the rescaling of the distributions to a reference density ρ_0 is in fact a probability transformation over r. To see this, let $d_{ij} = d[(x_i, y_i), (x_j, y_j)]$ denote the length of the edge between vertexes i and j. If we multiply the coordinates (x, y) of all the points with an positive real number $\alpha = \rho_0/\rho$, the following relation holds:

$$d[(\alpha x_i, \alpha y_i), (\alpha x_j, \alpha y_j)] = \alpha d[(x_i, y_i), (x_j, y_j)], \tag{5.4}$$

On the other hand, this multiplication corresponds to a uniform dilation (or contraction) of the system, that is, a uniform scaling of the density. Therefore, to calculate the scaled probability density, we have to apply the probability transformation defined by

$$t(r) = \alpha r. \tag{5.5}$$

Applying basic probability theory, we obtain the rescaled probability density

$$f(r) = \frac{1}{\alpha}\bar{f}(\frac{1}{\alpha}r). \tag{5.6}$$

We can either define a reference density ρ_0 and calculate the corresponding transformations for the different experiments, or, fit the probability densities to each other, using α as a fitting parameter.

Certain local properties are averaged out both by the $g(r)$ and the $f(r)$. For instance, non-regular density variations are not captured by these measures. Another example is a situations when certain regular structures appears multiple times but the size of the structures varies.

In order to detect these situations, we calculate the conditional probability $p(r|r')$ of the edge lengths of the nearest neighbour graph. This conditional probability is the probability of finding an edge with length r attached to a node which for sure has an edge with length r'. The conditional probability can numerically be represented in a matrix structure P where each row corresponds to a condition for a given interval over r' and each column represents an interval over r. The matrix entry P_{ij} corresponding to a given row i and a column j will give the probability of finding an edge with a length between r_j and $r_j + dr$ attached to a node which has at least one edge with a length between r'_i and $r'_i + dr$. In case the localization points are arranged according to a specific structure, the P matrix will indicate a preferential attachment. For instance, in a long chain which has constant link-lengths over larger domains, but different domains have different link lengths, the P matrix will be almost diagonal. Figure 5.10 shows the P matrix for two randomly generated data-set, one set containing points with coordinates distributed uniformly while the other set has additional Gaussian clusters. As the figure illustrates, uniformly distributed points will produce a rather uniform matrix, while point-sets with clusters will have a more emphasised diagonal.

5.2.3 Exposure to Ionizing Radiation Cause Local Changes

We first perform a segmentation of each of the images to cut out those areas in the image that do not belong to the cell nucleus. We also neglect areas inside the nucleus that have a very low density of markers since these areas are indicative for nucleoli. The segmentation is based on the calculation of the density distribution of markers in the image. We only take the areas with a density of more than 25% of the largest density into consideration. Figure 5.11 shows the image with the localized H2B histones and their density distribution and the final segmented image.

To assess how the overall organization of chromatin changes after cells are exposed to ionizing radiation compared to untreated cells we calculate the pair correlation function $g(r)$ for localized H2B histones in untreated cells and cells that were exposed to ionizing radiation. The pair correlation function is a measure for the positional correlation of points in a many-particle system. At this, $g(r)dr$ denotes the probability of finding two points separated by a distance between r and $r + dr$ for infinitesimal dr.

Results for the radial pair correlation function of H2B markers in untreated and irradiated cells are shown in Figure 5.12. We analyzed images of untreated cells and cells fixed $30\,min$ after they were exposed to $0.5\,Gy$, $2\,Gy$ and $4\,Gy$ of γ-irradiation. For all setups we can see deviations of $g(r)$ from unity only for distances up to $300\,nm$. It thus shows that structured organisation of chromatin in the cell nucleus is only apparent up to distances of roughly $300\,nm$. Below this critical distance, locations of labeled histones H2B are visibly

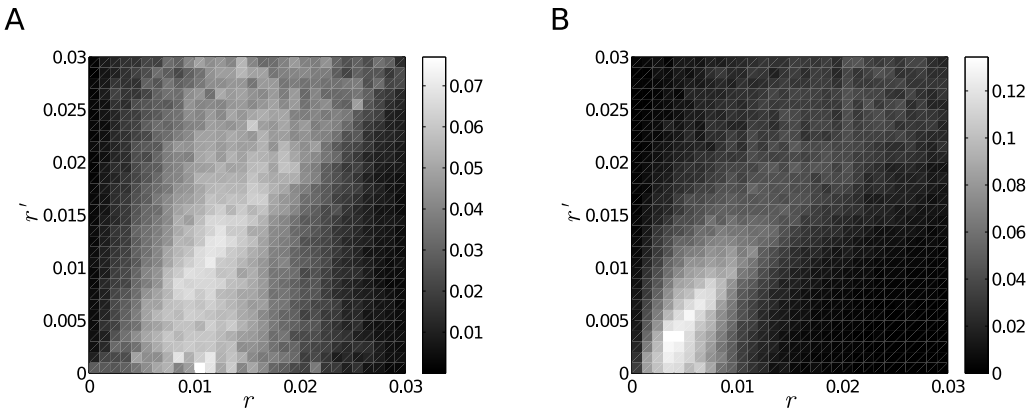

Figure 5.10: Conditional Probabilities of Edge Lengths for Random Data. The figures illustrate how the $p(r|r')$ conditional probability looks for randomly generated point positions. **A** Conditional probability of points with coordinates generated according to a uniform distribution. **B** Conditional probability of points with coordinates generated according to a mixture of uniform distribution and clusters of Gaussian distributions. In the latter example we observe an emphasised diagonal which is the result of the Gaussian clusters. Tightly packed points tend to produce short edges, while points from the edges of the clusters mostly have longer edges.

correlated. However, on distances larger than $300\,nm$ the radial correlation function drops to unity and the locations of histones can be viewed as being randomly distributed relative to each other. The radial distribution functions for the H2B histones are lying on top of each other and no deviations can be seen for cells that were exposed to ionizing radiation.

In order to assess the local positional correlations that are apparent up to a length of $300\,nm$ we pursued a graph theoretical approach. We first calculated the nearest neighbour distance distribution of the localized fluorophores. For this we performed a Delaunay triangulation to obtain graphs for all points in the segmented images and then calculated the length distribution of the edges of the graph. Results for H2B markers for untreated cells and cells fixed $30\,min$ after exposure to ionizing radiation are shown in Figure 5.13. We observe that there is a significant difference between them, which can be seen in the inset. However, the distributions belong to the same family since the rescaled distributions $f(r)$ are exactly the same. Therefore we can state that the observed differences may stem either from different experimental conditions such as different overall marker density or a uniform dilation/contraction of the system.

We also calculate the $p(r|r')$ conditional probability for the localized H2B markers. The matrix representations P of the conditional probabilities are plotted in Figure 5.14, the three panels corresponding to pre-irradiation configuration and to configurations fixed 30 minutes and 48 hours post-irradiation respectively. While the plots resemble a mixture of the plots from figure 5.10, they do not exhibit big discrepancies amongst each other. Figure 5.15 presents the differences between the three panels from 5.14. This plot shows that for samples imaged shortly after irradiation the diagonal of the conditional probability is more prominent but for samples recorded after a longer healing time the diagonal recedes. This means that, while upon irradiation the system changes towards a less uniform structure, that is, towards a less homogeneous P matrix, with longer healing time the changes are

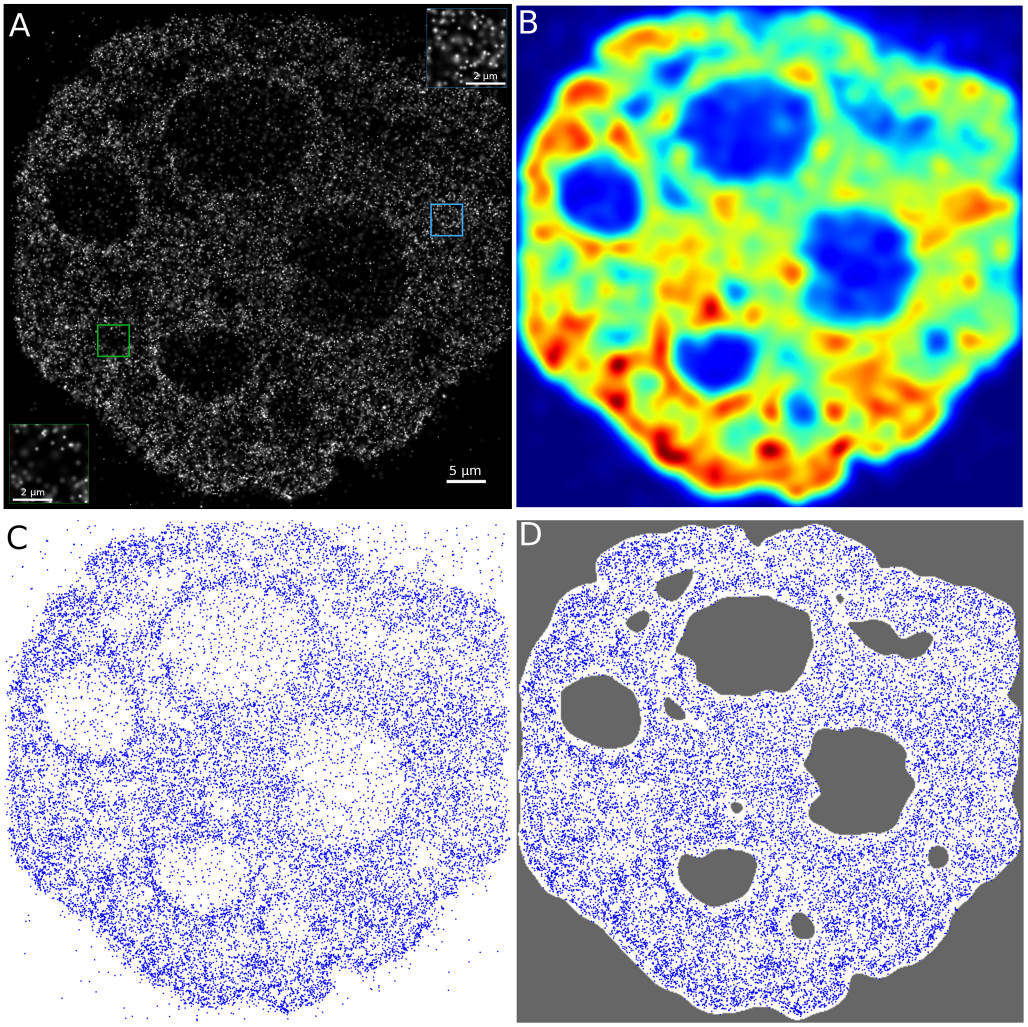

Figure 5.11: Localization Microscopy Images. **A.** The green inset shows an area with a low average density, which could correspond to euchromatin and the blue area corresponds to an area with high point density, possibly belonging to heterochromatin. **B.** This panel shows the calculated density distribution of the localized markers using a gaussian kernel density estimation with a uniform gaussian kernels. **C.** The figure shows the localized points of the image. From the points, areas with very low point density possibly corresponding to nucleoli can clearly be made out visually. **D.** Shown is the segmented image where only the area of interest is kept for the analysis. The segmentation was based on the intensity distribution with areas below an intensity threshold were discarded for analysis.

reverted.

Our results demonstrate that positional correlations of H2B histones are not altered by DNA damage caused by γ-irradiation. As H2B histones are distributed homogeneously along the genome we can conclude that ionizing radiation does not alter the overall organization of the chromatin in the cell nucleus.

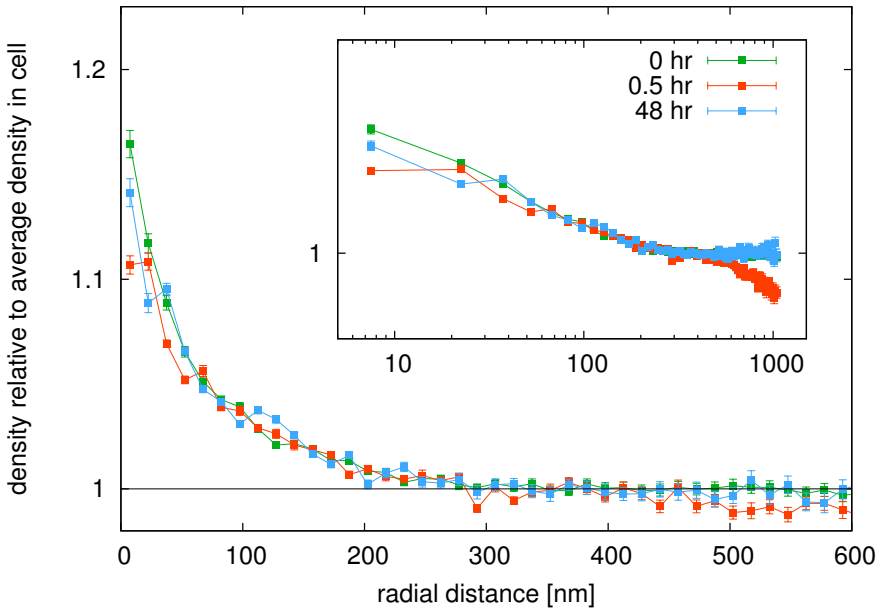

Figure 5.12: Results at different times after 0.5 Gy radiation dose. *left:* The radial pair correlation function $g(r)$ shows that correlations between the positions of labeled H2B histones exist up to a distance of roughly $300\,nm$. Marker and thus chromatin densities in the surroundings of each marker is elevated compared to the average density of markers in the cell nucleus. Above $300\,nm$ however, histone positions are uncorrelated and the marked histones can be viewed as being positioned randomly relative to each other. Furthermore, the radial pair correlation function for cells exposed to $0.5\,Gy$ γ-irradiation is the same as for untreaded cells, regardless of the time passed after irradiation.

5.2.4 Heterochromatic Regions Show a Decondensation

Antibodies for H4K20 made visible with the same fluorescence technique were used as markers for heterochromatic regions. Microscopy for antibodies were performed at the same time with the cells on the stably expressing histone H2B variant. Images were processed and segmented with the same methods described above and in the Methods subsection. In Figure 5.16 we show the density distribution of the heterochromatin antibody markers on panel A and the segmented image on panel B. The density distribution shows clearly how heterochromatic regions can be seen as coarse clumps located within the nucleus. These clumps are visible as bright spots in the density distribution.

The calculated radial pair correlation function for localized H4K20 antibodies in untreated cells and cells exposed to $0.5\,Gy$ irradiation after different times are shown in Figure 5.17. The correlation function for the antibodies show apparent differences to the correlation function for H2B histones. The high value of up to 20 times the average density that can be found at small distances r reflects the fact that antibodies are clustered together in small spots. The quick decay of $g(r)$ to the average density in the nucleus indicates that the clumps are relatively small confirming the visual impression of the images.

Upon irradiation, the value of the correlation function drops for small distances r. This means that the average density of antibodies in the surrounding of each antibody is lower after exposure to ionizing radiation. It indicates that the overall density becomes smaller in the heterochromatic clumps. We can conclude that heterochromatic regions on average

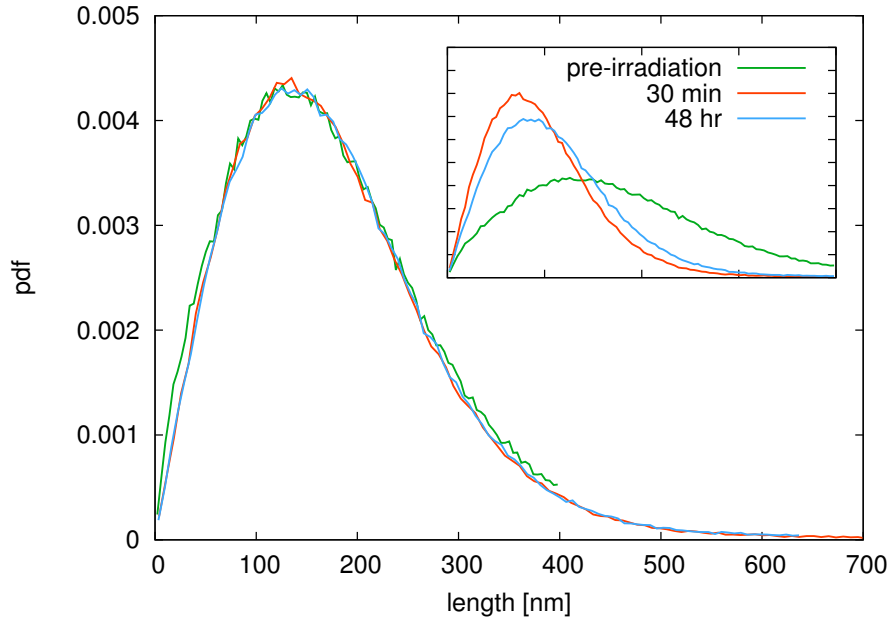

Figure 5.13: Results at different times after 0.5 Gy radiation dose. Shown are the rescaled distribution of the length of edges in a Delaunay triangulation of the H2B marker positions. The inset shows the original distributions for the different images. The differences in the edge length distributions are due to the different marker densities in the images. We therefore performed a rescaling of the distributions with respect to the point density to clear out this effect. The rescaled distributions can be seen in the main panel and show that the distributions belong to the same family.

become less compact and the strongly compacted organisation opens up and adopts a more loose structure upon exposure to ionizing radiation. We observe an average drop of 70 % of the mean antibody density in a sphere with a radius of 30 nm around an antibody for cells irradiated with 0.5 Gy γ-irradiation after 30min. Therefore, initial repair of DNA double-strand breaks caused by irradiation in heterochromatin seems to require a drastic decrease of the chromatin density and a strong relaxation of the compact organisation of the chromatin fiber here. 48 h after irradiation, this value is only at around 30 % which indicates that structures seem to have recovered after successful repair of the broken DNA strands.

The conclusions drawn from the radial pair correlation function are verified by the graph theoretical analysis. In Figure 5.18 the length distribution of the edges in the Delaunay triangulation of the marked H4K20 antibodies is shown. The distribution for unirradiated cells shows a very characteristical peak at small distances centered at around 30 nm. This emphasizes that there is a characteristical nearest-neighbour distance for the H4K20 markers thus meaning a preference of them to form clumps.

In cells at 30 min after exposure to ionizing radiation, the sharp peak in the edge length distributions vanishes. Instead, the distribution becomes a uniform distribution up to large lengths. The disappearance of the peak is a clear indicator that the heterochromatic regions are no longer organized as compact clumps in the nucleus. This difference can be seen visually in Figure 5.16C. Compared to the untreated cell nucleus shown in Figure 5.16A we clearly see that there the small bright spots have mostly vanished and the heterochromatic regions in the irradiated cell are much more smeared out. Our graph

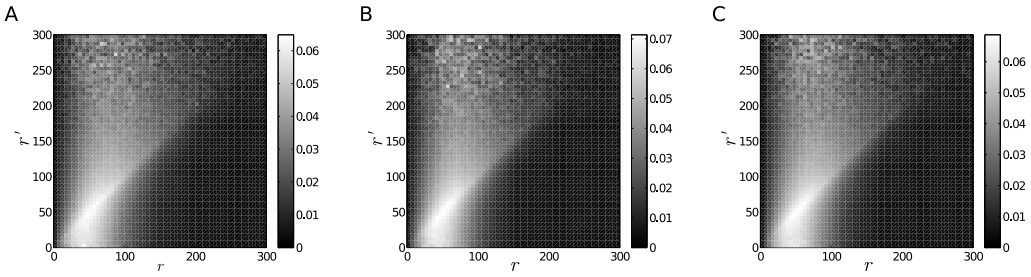

Figure 5.14: Conditional Probability Distribution of the Edge Lengts. Panel **A**: The panel shows the conditional probability distribution $p(r|r')$ for the H2B markers before irradiation. The relatively prominent diagonal indicates locally a varying density. Panel **B**: The panel shows the conditional probability for the H2B markers 30 minutes after irradiation. Panel **C**: The panel shows the conditional probability for the H2B markers 48 hours after irradiation.

theoretical analysis therefore verify our observations of the behaviour of the pair correlation function. The heterochromatic regions undergo a process of opening up after exposure to ionizing radiation.

At $48\,h$ after irradiation, the peak in the edge length distribution emerges again. Just as the pair correlation function is again very similar to untreated cells, the edge length distribution has now also the same shape as in the case of untreated cells.

Calculating the conditional probability $p(r|r')$ we see that the corresponding matrix-representations are strongly diagonal (figure 5.21). This is due to the tight clusters of the antibodies marking heterochromatic regions. This structure of the conditional probability matrices support our previous conclusions. Upon irradiation the conditional probability (panel **B.** in figure 5.21) becomes more homogeneous, indicating a more uniform structure. 48 hours after irradiation the conditional probability is again diagonal (panel **C.** in the same figure). Note that in this case the colour-map has a wider range and values on the diagonal are in fact very close to values from panel **A.**, except for very small radii.

The differences of the conditional probabilities are plotted in figure 5.22. While the changes upon irradiation barely depend on the condition (the value of r'), we observe that before irradiation smaller radii were more abundant (red colour in panel **A.**). The difference between the structure 30 minutes after irradiation and the structure we see 48 hours after hours after irradiation (in panel **C.**) indicates an almost uniform and unconditioned increase in the probabilities of the short edges, just as we saw it in the case of the unconditioned edge length distribution.

Our results here show that heterochromatic domains undergo structural reorganizations after exposure to ionizing radiation. At $30\,min$ after irradiation the previously very compact and densely organized domains open up and adopt a more loose organization. After $48\,h$ the structures heal again and the organization approaches again the initial configuration of the untreated cells.

5.2.5 Discussions and Conclusions

We analysed the effects of the reorganization of chromatin upon exposure to ionizing irradiation. Samples were imaged by Spectral Position Determination Microscopy (SPDM). The samples were subject to a preprocessing step in which regions of interest were detected.

A nearest neighbour graph was built by calculating the Delaunay triangulation of the localization points. The spatial organization of the fluorophores was characterized by

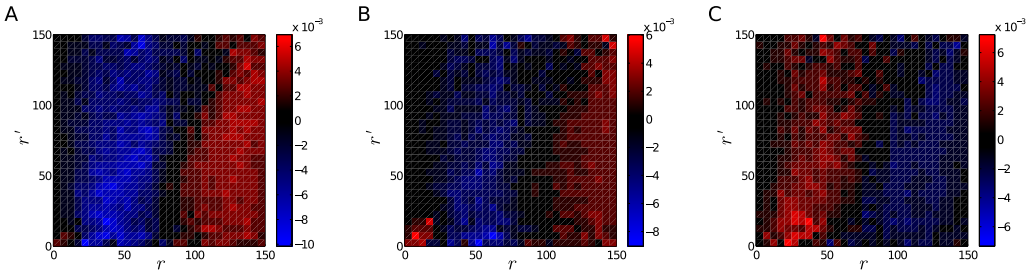

Figure 5.15: Difference in the Conditional Probability Distribution of the Edge Lengts. Panel **A**: The panel shows the difference of the conditional probabilities $p(r|r')$ measured for structures recorded before irradiation and for structures registered 30 minutes after irradiation. Shades towards red indicate values which are larger in the samples before irradiation, while values towards the shades of blue indicate probabilities which are larger in images registered 30 minutes after irradiation. The plot indicates slightly increased values along the diagonal after irradiation. This might mean a slightly increased clustering of the points. Panel **B**: The panel shows the differences of the conditional probability before irradiation and 48 hours after irradiation. Entries in shades of red are larger in samples recorded before irradiation, while entries in shades of blue are larger in the samples recorded 48 hours after irradiation. The trend is similar to that observed in Panel **A**, however differences are less prominent. Panel **C**: The panel shows the differences if the conditional probability measured 30 minutes and 48 hours after irradiation respectively. The red shades indicate larger probabilities in samples recorded 30 minutes after irradiation while blue shades indicate larger probabilities in samples registered 48 hours after irradiation. Here we observe a reversed trend compared to Panel **A**.

different graph theoretic measures. Moreover, the radial distribution function was also calculated.

Comparing the quantities calculated for non-irradiated and irradiated samples, we found that although overall the chromatin may compactify to a certain degree, the local neighbouring properties of the localized points do not change. For instance, the distribution of the edge-lengths belong to the same family for irradiated and non irradiated samples. At the same time, we found that heterochromatic regions, marked separately, open up open irradiation. Furthermore, assuring a long enough healing time after irradiation, we observe a recover of the heterochromatin from the mentioned opened state. Our findings are in agreement with other experiments observing structural changes caused by the presence of the double strand breaks.

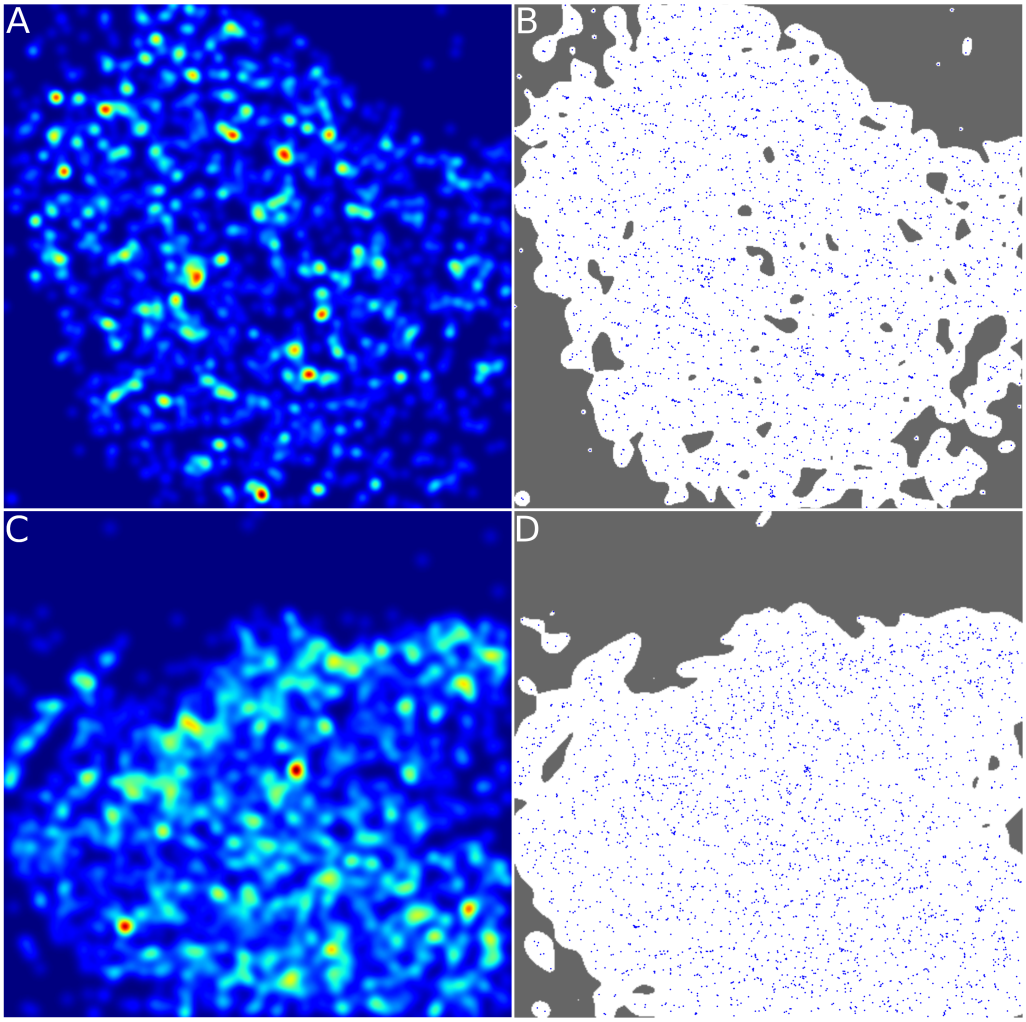

Figure 5.16: Localization Microscopy Images of Heterochromatin Markers. **A.** Shown is the density distribution of the localized markers in a cell prior to irradiation. Small, bright spots where markers are agglomerated can be seen. This means that heterochromatin is mainly organized in coarse clumps. **B.** Shown is the segmentated image of the not irradiated cell that is used for subsequent analysis of the marker distribution. **C.** The density distribution of a cell at $30\,min$ after irradiation with $0.5\,Gy$ is shown here. Differences between this cell and the not irradiated cell can be made out by visual inspection. We observe that the density has much less agglomerated and bright spots and is instead much more homogeneous. **D.** This effect can also be seen by visual inspection of the heterochromatin markers directly. Marker positions are visibly more spread out and less strongly clumped together. Heterochromatin clearly undergoes structural changes upon irradiation.

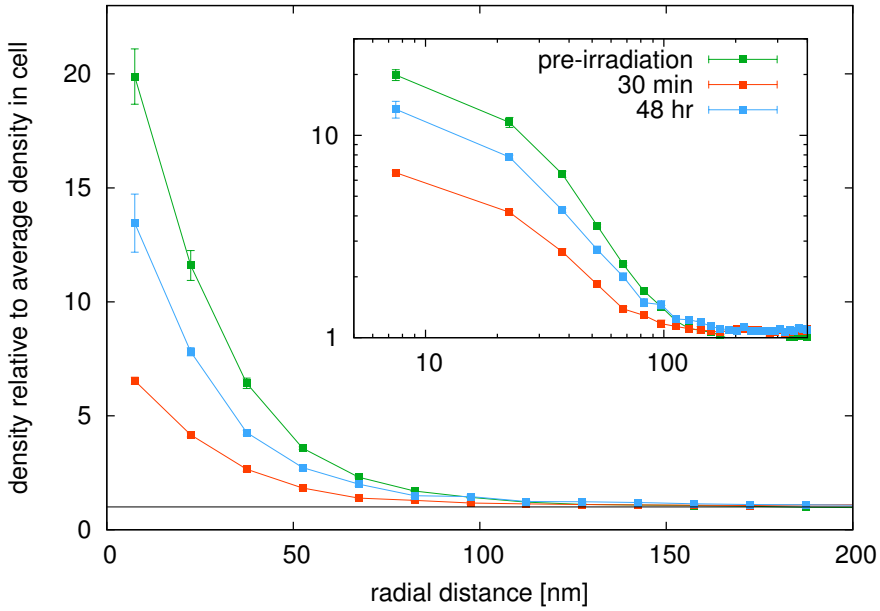

Figure 5.17: Results at different times after 0.5 Gy radiation dose. The figure shows the radial pair correlation function for methylated histone variants H4K20 antibodies representing heterochromatin for unirradiated and irradiated cells. Error bars represent the standard deviation of the mean value after averaging over the sample of cells. The value for g(r) at small distances goes up to around 20, indicating the high marker densities in regions where heterochromatin is located. The rapid drop off of the pair correlation function within a distance of less than $100\,nm$ shows that heterochromatin forms small clumps that are spread throughout the cell nucelus. Upon exposure to $0.5\,Gy$ γ-irradiation, a dramatic change in the correlation function can be observed in cells that were microscoped after $30\,min$. The value at small radial distances drops to around 6, or around 70 % smaller than in unirradiated cells. This indicates that the density in the heterochromatic regions is on average much lower in irradiated cells, requiring that the organisation of the chromatin fiber in these regions has to have loosened compared to before due to DNA damage such as double-strand breaks. In cells measured $48\,h$ after irradiation, the correlation function have recovered again and the value at small r is at around 14, only 30 % less than in unirradiated cells.

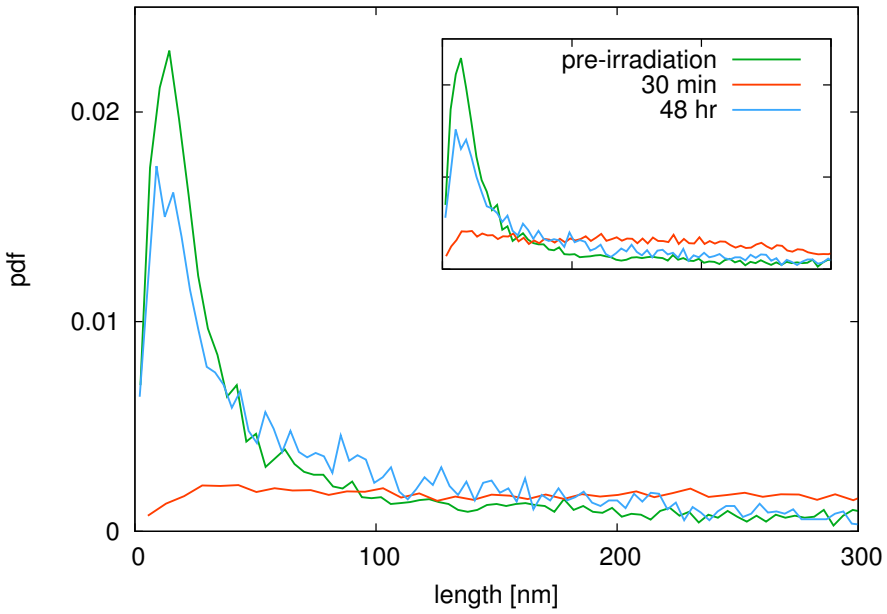

Figure 5.18: Results at different times after 0.5 Gy radiation dose. The distribution of edge lengths in the Delaunay triangulation of the markers confirms these observations. A sharp peak in the distribution at around 30 nm can be seen in untreated cells. In 30 min post-irradiation cells the peak vanishes and a spread distribution can be seen. In 48 h post-irradiation cells however, the peak reappears again but less pronounced than in untreated cells.

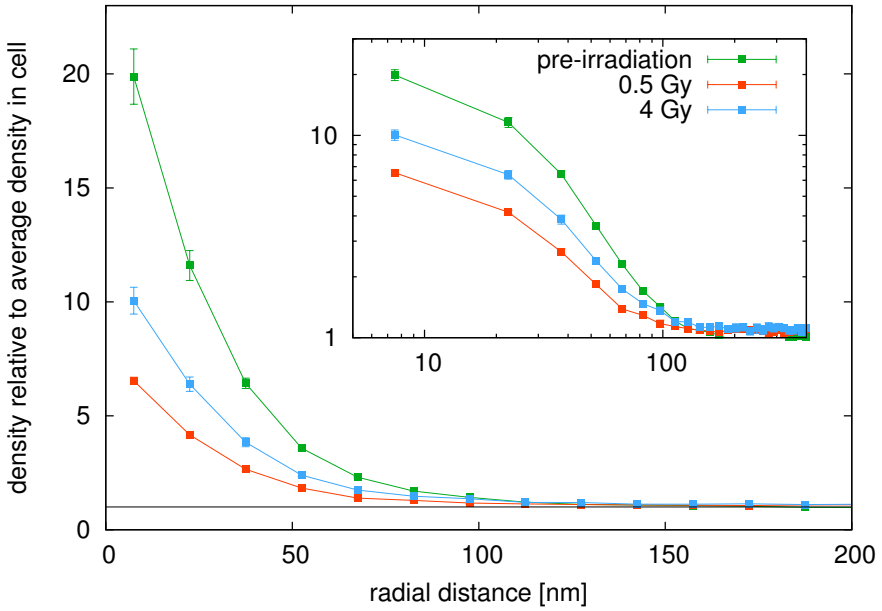

Figure 5.19: Results for different doses 30 min post irradiation.

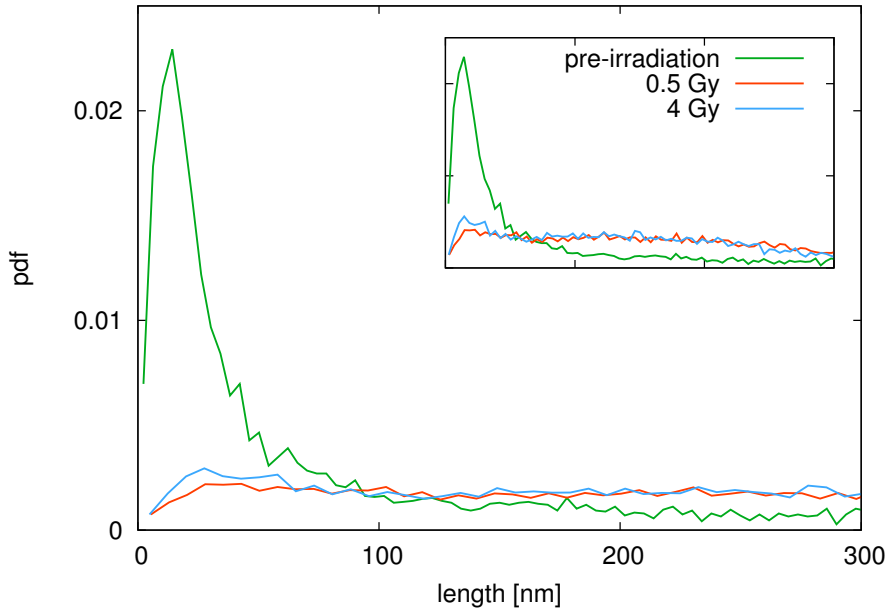

Figure 5.20: Results for different doses 30 min post irradiation.

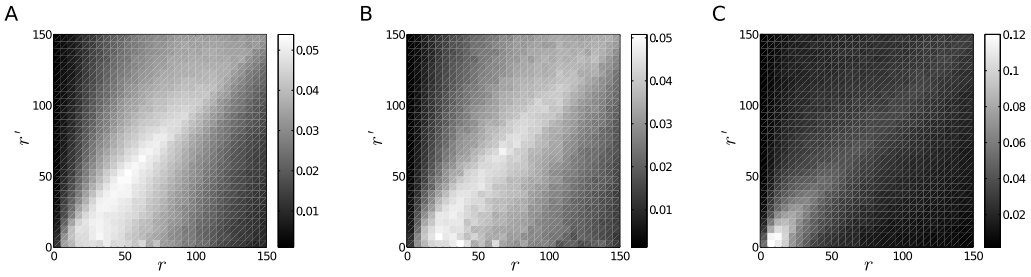

Figure 5.21: Conditional Probability Distribution of the Edge Lengts for Heterochromatin Markers. Panel **A**: The panel shows the conditional probability distribution $p(r|r')$ calculated for the positions of the antibodies marking heterochromatic regions before irradiation. Panel **B**: The panel shows the conditional probability for the heterochromatin markers 30 minutes after irradiation. Panel **C**: The panel shows the conditional probability for the heterochromatin markers 48 hours after irradiation. In all three cases the diagonal is very emphasised indicating preferential spatial distribution of the edges. This may stem from the clustering of the heterochromatin markers. Note that although shades along the diagonal are darker in Panel **C**, the value of the corresponding probabilities are very close to the probabilities along the diagonal of Panel **A** except for small radii.

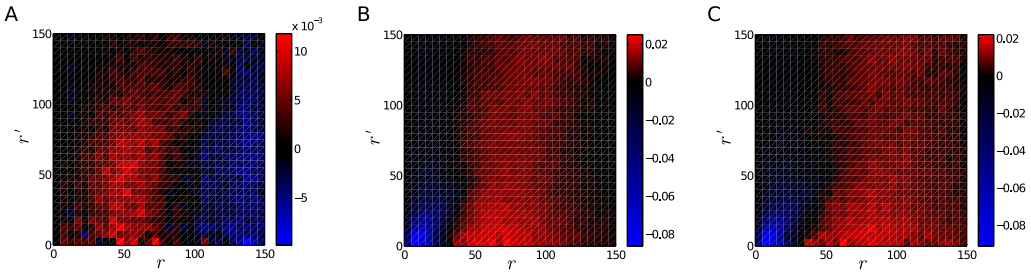

Figure 5.22: Difference in the Conditional Probability Distribution of the Edge Lengts for Heterochromatin Markers. Panel **A**: The panel shows the difference in the conditional probability distribution $p(r|r')$ before irradiation and 30 minutes after irradiation. A red shade of the colour-map means that the probability is higher before irradiation while a blue shade means that it is higher after irradiation. The panel indicates a slightly stronger change along the diagonal indicating a more homogeneous system after irradiation. However the change is almost independent of the value of the condition r'. Panel **B**: The panel shows the difference of the conditional probability distribution calculated for samples before and 48 hours after irradiation. Shades of red indicate higher probabilities for the samples recorded before irradiation while shades of blue indicate higher probabilities in samples recorded after irradiation. Panel **C**: The panel illustrates the difference of the conditional probability distribution 30 minutes and 48 hours after irradiation. Red shades correspond to higher probabilities 30 minutes post-irradiation while shades of blue indicate larger values 48 hours after irradiation. The difference between structures observed 30 minutes after irradiation and 48 hours after irradiation indicate a reversed trend compared to Panel **A**.

5.3 Relation of Nuclear Pore Complexes and Lamin B Receptors

References

This section is adapted from the mansucript, which we intend to submit for publication,

- G. Máté, L. Shopland and D.W. Heermann, *Spatial Relation of Lamin B Receptors and Nuclear Pore Complexes* , in preparation (2013).

Nuclear pores are large protein complexes that perforate the nuclear envelope and allow the transport of molecules through the membrane [227]. Nuclear Pore Complexes (NPCs) are thought to contain more then 50 proteins called nucleoporins. On the other hand Lamin B Receptors (LBRs) are receptor proteins that bind to the inner side of the nucleus and serve as an anchor for the nuclear lamina [59]. In this project, we will explore the spacial relation of the NPCs and LBRs by studying graphs constructed on positions defined by fluorophores attached to respective proteins (or complexes).

Nuclear pore complexes were marked with Cyanine (Cy5) molecules while Lamin B receptors were stained with antibodies raised against them (FITC). Images were recorded by Three Dimensional Structured Illumination Microscopy (3D-SIM). This technique is based on projecting a patterned (structured) light onto the sample. The light-pattern interacts with the fluorescent probes in the samples and generates interference patterns, so-called *Moiré fringes*. Patterns are analysed by extensive computer algorithms and finally the image is reconstructed reaching a lateral and axial resolution double the resolution of conventional wide-field microscopy [228].

5.3.1 Masking and Detecting Pores and Receptors

First, images are thresholded at an i_t threshold intensity, calculated with Otsu's method [229] which assumes that pixels in the image belong to two classes: the region of interest (ROI) and the background. Otsu's method tries to minimize the intra class variance of the intensity distributions. As a result of the thresholding, we have an I_b binary image, where each voxel has a value of 1 (or $true$) if the corresponding voxel in the original image I had an intensity above the i_t threshold. Otherwise the voxel in I_b is set to 0 (or $false$).

The achieved binary image is improved in a quasi three dimensional procedure: Each slice from the z-stack is processed together with its two neighboring slices. After a morphological closing of the three slices, holes are filled. This is followed by the extraction and a morphological opening of the middle slice. The resultant binary image is stored in the corresponding z index of the mask.

After scanning the z-stack with the described procedure, connected components are detected, their size is determined and smaller objects are removed from the mask, keeping only the largest connected block of voxels. Two voxels are connected if they share a common face. Let us denote the obtained mask by an M three dimensional matrix, so that, M_{ijk} represents a voxel in the intensity image and it is either $true$ or $false$. If it is $true$ than the corresponding voxel belongs to the ROI, if it is $false$, then the voxel belongs to the background.

In the next phase, we extract the positions of the LBRs and the NPCs. This is done

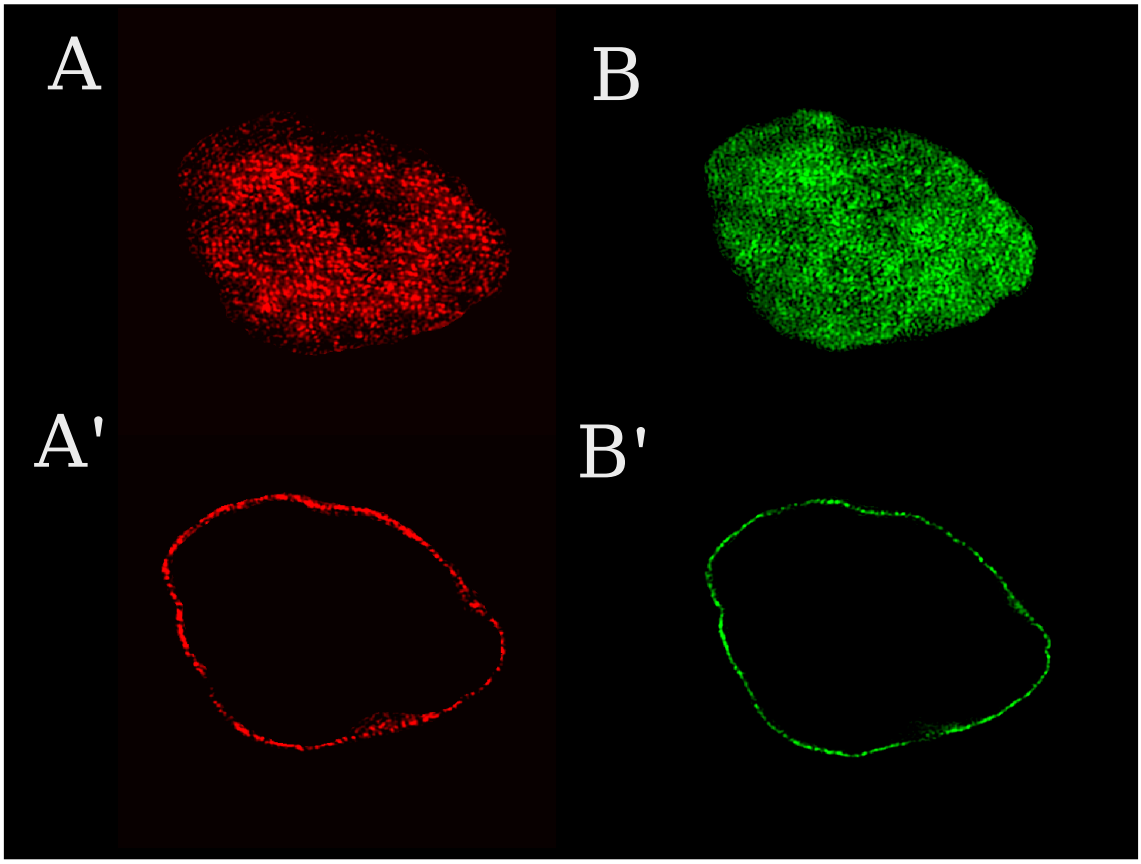

Figure 5.23: Intensity images presenting the (A, A') nuclear pore complexes and the (B, B') marked Lamin B receptors. Panels A and B contain slices from the bottom part of the cell while panels A' and B' illustrate slices from the middle of the same cell.

by a local thresholding of the original intensity images in the following way: First, the mask constructed in the previous step is eroded six voxels in the $x - y$ plane and two voxels in the z direction. These erosions roughly correspond to $240nm$s, voxels in the $x - y$ direction having a size of $39.5nm$s, while their dimension along the z axis is $125nm$s. The eroded mask is then subtracted from the original mask, creating an M' shell-mask covering a roughly $240nm$s thick shell of the nucleus. We exclude all the voxels from I which are not marked as *true* by the M' shell-mask, obtaining this way an I' intensity image. In the rest of the procedure we work with the I' image.

Starting with an initial intensity threshold of 15 % of the maximal intensity in I', we threshold the image and detect all the connected groups of voxels in the thresholded image. We count the numbers of voxels in each connected group and if there are groups with more then 125 voxels (this corresponding to a box with a side of ca. $280nm$s), we increase the applied threshold by 5 % and re-threshold the regions in I' corresponding to these large voxel groups. We repeatedly re-threshold the large groups, each time also increasing the threshold intensity by 5 %, until all the connected groups have a voxel-count below 125. We scale the x, y and z coordinates according to the resolution of the image, so that the positions, and consequently all the calculated distances, will be expressed in units of nanometers. This means that x and y coordinates were multiplied by 39.5 while z coordinates were scaled by 125.

As a last step, we define the positions of the NPCs and LBRs as the centers of gravity of the different groups of connected voxels. As the intensity images for the NPCs and LBRs were recorded on separate channels, we can detect the NPCs and LBRs separately.

5.3.2 Constructing a Network

Analyzing the spatial arrangement and positioning of the NPCs and the LBRs is a complex task. There are different approaches one might explore. One may count the NPCs and the LBRs, measuring the densities and comparing the resulted values for the two different sets or calculate the deviation of these values among the different cells.

Here, however, we apply a different approach. Since both the NPCs and the LBRs lie along the surface defined by the nuclear envelope, instead of taking a three dimensional approach, we perform our analysis along this surface. Also because of the confining effect, the points defined by the NPCs and LBRs are embedded in a (quasi) two dimensional space, it is very convenient to study the properties of the organization of the NPCs and LBRs through the analysis of the network defined by the neighboring relations of the objects.

In general, networks consist of a set of entities and certain relations among these. Systems from different fields has been studied in the framework of networks [220–224]. Most of the studies rely on the mathematical representation of networks, called graphs [219]. Graphs are abstract mathematical objects, composed of nodes or vertexes (representing the entities of the network) and edges or links (representing the relations among entities). Graph theory provides reliable and useful tools for the study of networks. It enables us to analyze the properties of our networks by calculating measures which characterize the relations among the nodes (that is, the NPCs and LBRs, in the present case).

The two dimensional space allows an easy way to define neighbors by means of a triangulation of the nodes. We will describe this procedure for the positions of the NPCs, but we can follow the very same process for the position of the LBRs.

Let us assume that the positions of the NPCs define a point set A in the three dimensional space. First, we calculate the Delaunay triangulation [230] of A. In general, the $Dt(P)$ Delaunay triangulation of an arbitrary set of three dimensional points P defines a set of tetrahedrons on P, such that, no point is inside the sphere defined by the corners of any of the tetrahedrons defined by the triangulation. Similarly, if P is two dimensional, then the $Dt(P)$ triangulation defines a set of triangles on P, such that, no point from P is inside the circles circumscribing the triangles in $Dt(P)$.

The Delaunay triangulation is often used to define the nearest neighbor relations [231–233] among members of a given set of points. After calculating the $Dt(A)$ triangulation of A, we discard all the triangles containing edges which cross regions outside the M' shell-mask. Let us denote the set of remaining edges by E_A ($E_A = \{(\alpha, \omega) | \alpha, \omega \in A\}$). Then, by following the described procedure we obtain a $G(A, E_A)$ graph structure which describes the local geometric relations of the points. This graph structure is visualized in figure 5.24 for one of the experimental images.

The $G(B, E_B)$ can be calculated in a similar way for the B set of points defined by the centers of the LBRs. This is illustrated in figure 5.25.

5.3.3 Calculating Properties of the Network

In the following we detail the analysis of the constructed networks. There are a few main question we seek to answer: How similar the graphs are? Is there any ordering with respect

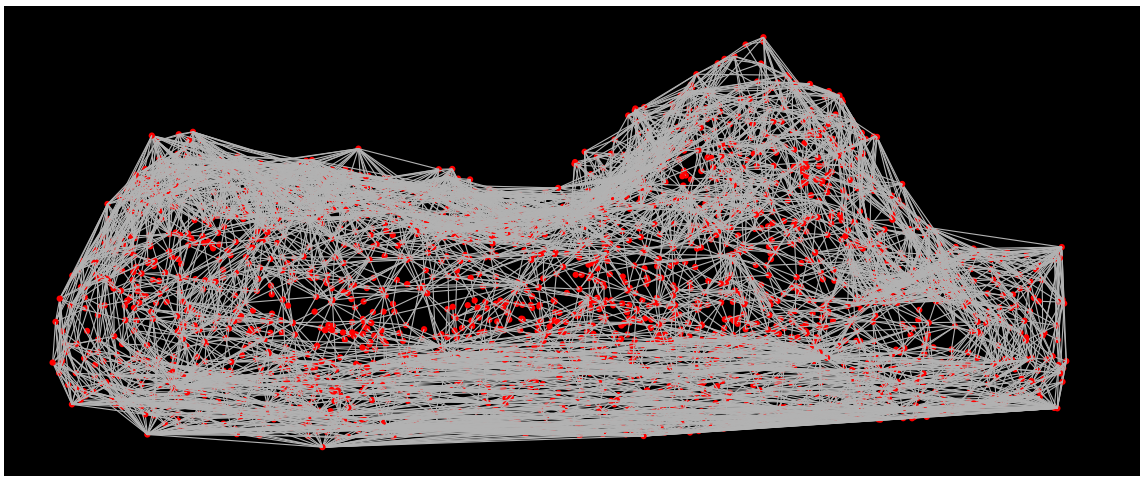

Figure 5.24: Network defined by the position of the NPCs.

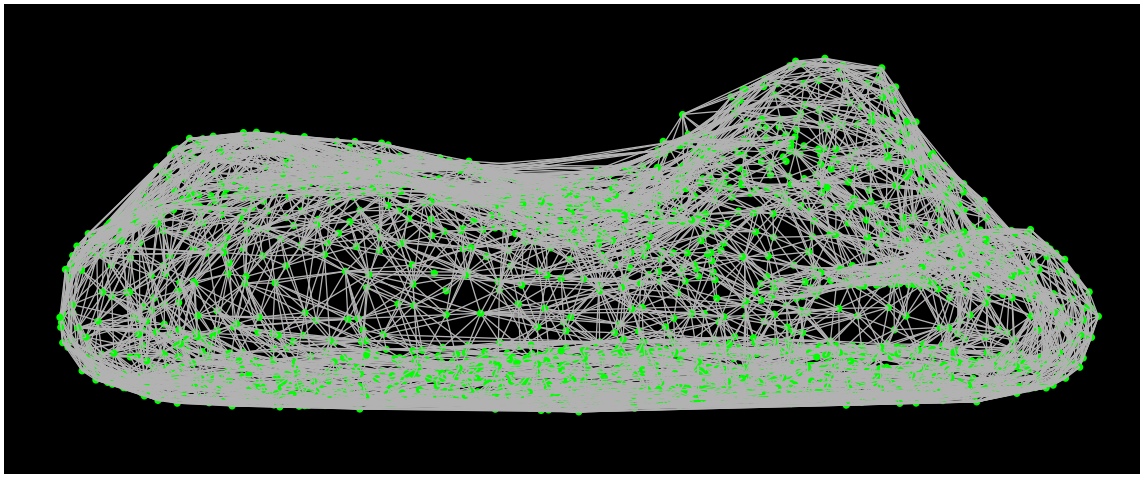

Figure 5.25: Network defined by the position of the LBRs.

to the position of the points? How similar the neighboring relations in the two graphs are? And finally, we are also interested whether the positions of the LBRs depend on the location of the pore complexes or not.

In order to characterize the networks and analyze the relation of the nodes, first, we calculate three different quantities: the $h(d)$ degree distribution [234], the $g(r)$ radial distribution function [218] calculated along the surface and the probability density function of the length of the edges in E_A (respectively E_B), denoted by $f(r)$.

The degree distribution is a discrete probability distribution and it simply measures the probability of finding a node with a given number of edges, that is, $h(d)$ is the probability of finding a node in the graph with exactly d edges.

The $g(r)$ radial distribution function is a function which is often applied in solid state and statistical physics and is calculated to characterize the geometric relations among points from a given set. It describes how the density varies as a function of the distance from an arbitrary reference particle (it is averaged over all particle-positions). In the case when points are positioned in a regular geometric structure (called a crystalline structure

in physics), the $g(r)$ would show that the density changes periodically. If ρ is the average density of the considered points, then $g(r)\rho$ is the expected density at a distance r from an arbitrary particle.

The $g(r)$ radial distribution function was devised in physics for studying fluids and crystals. Its definition implies that $g(r)dr$ is proportional to the probability of finding an atom in a dr thick shell of radius r centered around another atom. In three dimensions, for instance, $4\pi\rho g(r)r^2dr$ gives the mean number of atoms in this volume-shell.

Given an n_i histogram of pair separations which counts the number of pairs of atoms which are separated by a distance of ca. r_i, that is,

$$(i-1)dr \leq r_i < idr, \tag{5.7}$$

we can numerically estimate $g(r)$ for a two dimensional system by the following relation [235]:

$$g(r_i) = \frac{An_i}{\pi N^2 r_i dr}, \tag{5.8}$$

where A is the area of the system under investigation and N is the total number of particles.

In the present case, however, calculating the $g(r)$ is not straightforward, mainly because our points are constrained to a surface. Therefore, we will have to measure shortest distances on the surface, that is, along paths allowed by M'. For this purpose we will use the mesh defined by the triangulation and will approximate shortest distances by calculating shortest paths on the graph G, using Dijkstra's algorithm [236].

Dijkstra's algorithm finds the shortest path from an α start node to every other node in a given undirected graph $G(V, E)$ with node set V and edges E. It can also be used to find the distance from node α to a given end node ω by stopping the algorithm once ω is reached. A pseudo-code is detailed below in Algorithm 4.

When estimating the $f(r)$ probability distribution, we think about the edge lengths as random variables. Imagine that we write the length of each edge on indistinguishable pieces of paper and put these papers in an urn. Then, $f(r)dr$ gives the probability of drawing a paper with a number between r and $r + dr$, that is, it is the probability of finding an edge with a length between r and $r + dr$.

$f(r)$ has an advantage over $g(r)$. Since f is a probability density function, it can be rescaled to eliminate different experimental conditions like deviations in marker-densities and microscope settings.

As already said, when calculating $f(r)$, we assume that the lengths are random numbers. Obviously, the lengths depend on the position of the markers recorded through the imaging process. Assuming a linear dependence of the recorded intensity on the microscope gain and dye concentration, the density of the recorded points also depends linearly on these experimental conditions, just as it would depend on the "stretching" of the whole system. In this sense, the observed distances between the recorded points in a low density system should look the same as in a system with high density after undergoing a linear transformation (stretching of the coordinate axis). Therefor, the probability density of the distances can be rescaled to different marker-densities.

Given a probability density function f and a monotonic function g, the rescaled f_r probability density is given as:

$$f_r(y) = f(g^{-1}(y))\frac{d}{dy}g^{-1}(y). \tag{5.9}$$

Algorithm 4 Calculating the length of the shortest path between node α and node ω on graph $G(V, E)$.

```
1: procedure DIJKSTRA(α, V, E)
2:     distance[α] = 0;                          ▷ the distance to the starting node is 0
3:     for γ ∈ V \ α do          ▷ initially, set the distance to every other node to ∞
4:         distance[γ] = ∞;
5:     end for
6:     U = V                                                    ▷ unvisited nodes
7:     while U ≠ ∅ do
8:         select ψ ∈ U, where distance[ψ]==min(distance)   ▷ visit the node with the
   smallest distance, denote it by ψ
9:         W = {γ|(ψ, γ) ∈ E}              ▷ W is the set containing the neighbors of ψ
10:        for γ ∈ W do                                  ▷ visiting all the neighbors of ψ
11:            tempdistance = distance[ψ] + d(ψ, γ)  ▷ d(ψ, γ) is the distance between ψ
   and γ
12:            if distance[γ] > tempdistance then
13:                distance[γ] = tempdistance
14:            end if
15:        end for
16:        U = U \ ψ                                          ▷ marking ψ as visited
17:    end while
18: end procedure
```

Results for the $h(d)$, the $g(r)$ and the $f(r)$

The $h(d)$ degree distributions for the NPCs and LBRs were calculated and averaged over the different images. The results are illustrated in figure 5.26. The distributions look quite similar, they also decay in a similar way. The mean degree for the NPC network is 10.008553 while it is 11.103745 for the network of the LBRs. These two numbers are relatively close to each other. Thus, with respect to connectivity, these networks are very similar to each other.

The radial distribution function, averaged over the different experimental data is presented in figure 5.27. The $g(r)$ is low for small values of r and it has a first peak around $350nm$s for the NPCs and ca. $300nm$s for the LBRs. Afterwards the $g(r)$ drops and stays constantly around 1. First of all, this means that the LBRs are more tightly packed than the NPCs. It also means that the NPCs prefer to be separated by a minimal distance and perceive each other as hard spheres with a radius of around $175nm$s. This is also true for the LBRs, when binding to the wall of the nuclear envelope. They perceive each other as spheres with a radius of ca. $150nm$s. Of course, this rather is a consequence of complex binding mechanisms and the packing of the heterochromatin and it does not stem from the physical interactions between the LBRs.

Figure 5.28 depicts the averaged distribution of edge-lengths for the NPCs and LBRs. As the figure shows, edges among LBRs are on average shorter then among NPCs, thus the system of the LBRs is denser than that of the NPCs as this already was predicted by the $g(r)$. On the other hand, we cannot be sure that the difference in the density is caused by differences in certain experimental conditions. Therefore we rescale the densities

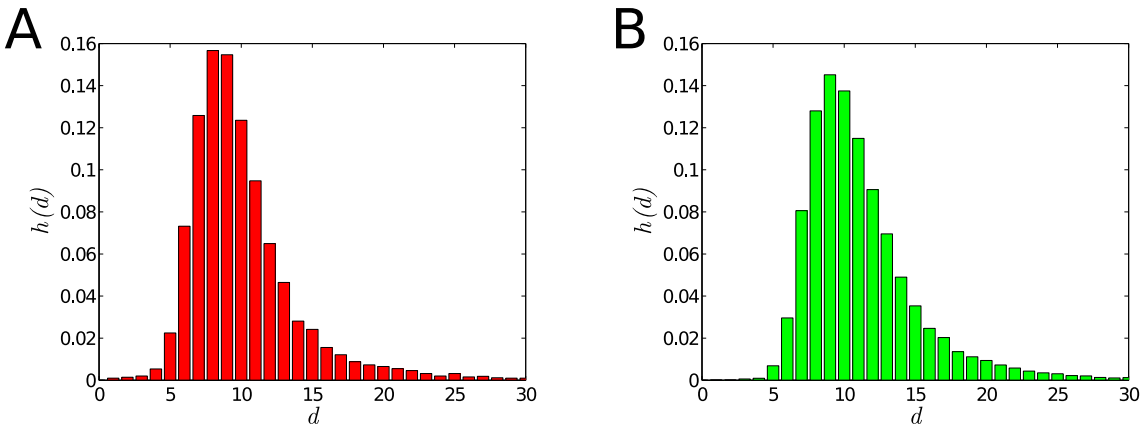

Figure 5.26: The $h(d)$ degree distribution calculated for the networks of the NPCs and the LBRs, respectively. The mean degree for the NPCs is 10.008553 and it is 11.103745 for the LBRs, the two values being very close to each other.

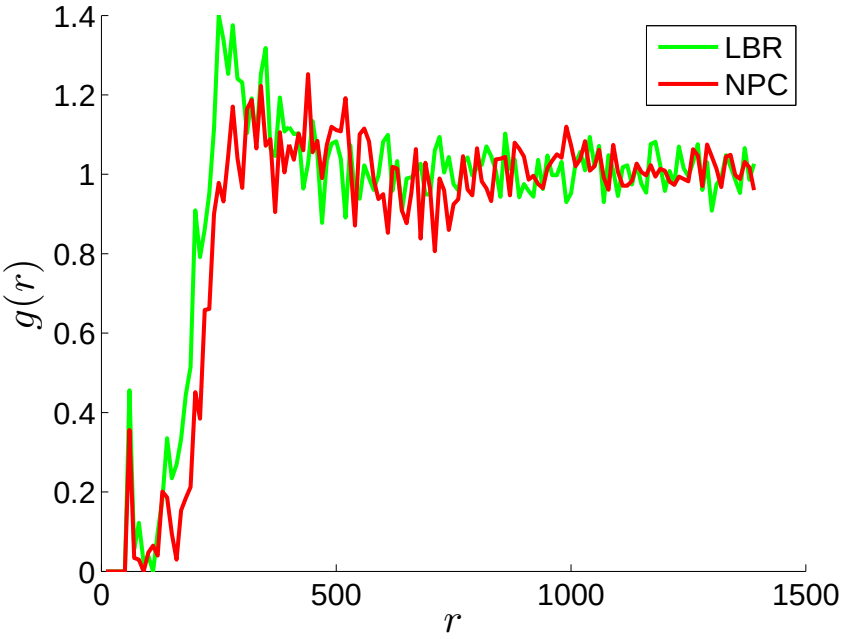

Figure 5.27: The $g(r)$ radial distribution, calculated for the position of the NPCs and the LBRs, respectively. The $g(r)$ is low for small values of r, meaning that there is a certain excluded volume interaction, that is, pores and receptor proteins behave like hard spheres when positioned on the membrane and they like to be placed at a certain minimal distance from each other. This distance is indicated by the first large peak of the $g(r)$. The approximate radius of these spheres is around $200nm$s for the NPCs and is slightly smaller for the LBRs (the $g(r)$ for the LBRs peaks at a smaller value).

and fit them on each other. The fit is presented on figure 5.29. The figure shows that while the shorter and longer edges are expected with the same probabilities for the density normalized NPC and LBR networks, the preference regarding edges with average lengths deviates in the two graphs. This clearly indicates that the two networks are different and

there definitely is no one-to-one correspondence. We can safely state that the neighboring relations are different for the two networks. However, based on the $f(r)$s, we cannot state anything certain regarding this difference.

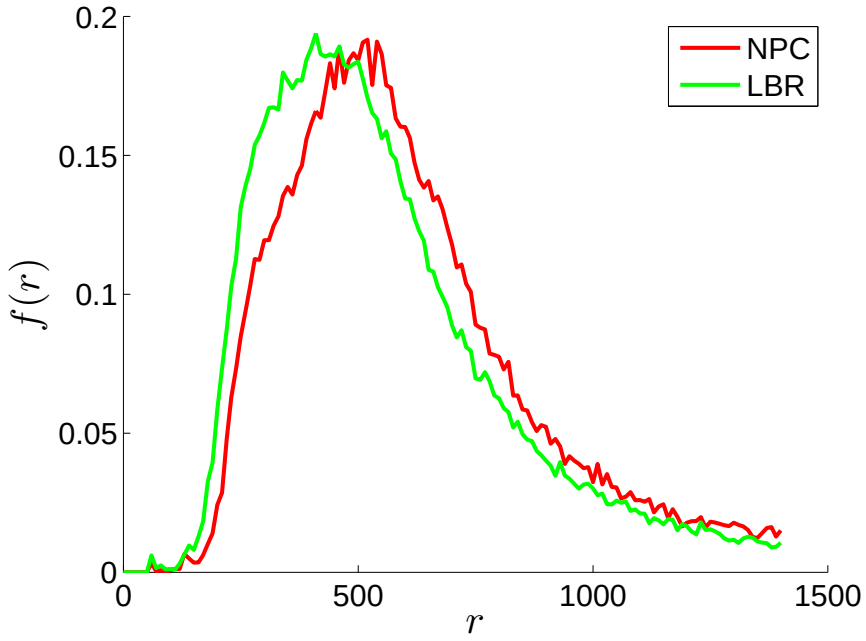

Figure 5.28: The $f(r)$ distance distribution, calculated for the position of the NPCs and the LBRs, respectively.

5.3.4 Connecting the Networks

To analyze whether the positioning of the LBRs is indeed influenced by the locations of NPCs, we construct a third graph by unifying the A and B sets. This way, a $G(C, E_C)$ graph can be constructed on the $C = A \cup B$ united set of points, as described in subsection 5.3.2. In order to represent solely the relation of the two different sets, we remove edges connecting nodes only from A or only from B, keeping only edges defined by the following relation:

$$E'_C = \{(\alpha, \omega) | (\alpha, \omega) \in E_C, (\alpha \in A \wedge \omega \in B) \vee (\alpha \in B \wedge \omega \in A)\}. \tag{5.10}$$

In addition, we add an extra edge for each point in C, so that, each point from A will be connected to the closest point from B – let us denote these edges by $E_{A \to B}$ – and each point form B will be connected to the closest point in A – we denote these edges by $E_{B \to A}$. This is required as in situations when points either only from A or only from B cluster, certain nodes might be "shaded" from the other set and get disconnected as a result from the removal of the edges which connect nodes from the same set. Eventually, we can define the edge set as $E_C = E'_C \cup E_{A \to B} \cup E_{B \to A}$. A concrete $G(C, E_C)$ graph is illustrated in figure 5.30.

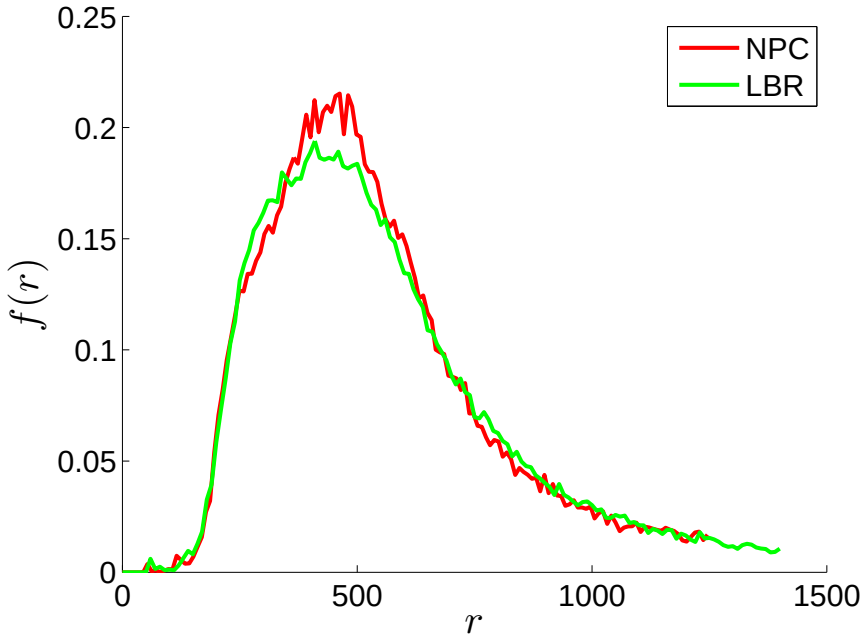

Figure 5.29: The rescaled $f(r)$, calculated for the position of the NPCs and the LBRs, respectively. Although the two distributions for well for short and long edge lengths, they deviate around the average value, thus the distributions are in fact different and in most cases the NPCs and LBRs have a different arrangement except when tightly or very loosely packed.

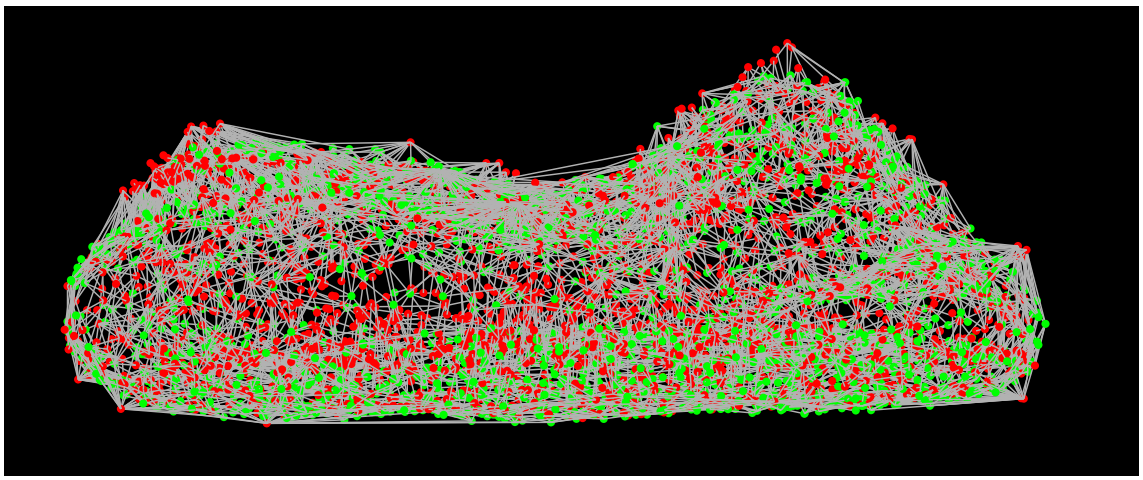

Figure 5.30: Network defined by the position of the NPCs and LBRs.

Results for the $g(r)$ and the $f(r)$

The $g(r)$ for the network of the union of the NPCs and LBRs is plotted in figure 5.31. According to the radial distribution, there is a certain degree of excluded volume interaction among members of the union, however, there is no preference regarding the separation distance.

We can calculate a special $g_{AB}(r)$ radial distribution by measuring the probability of

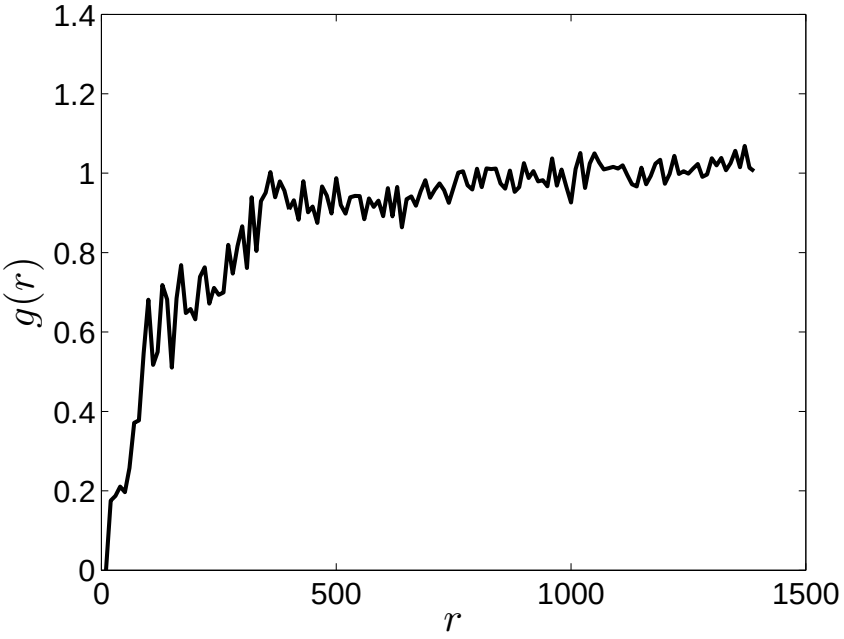

Figure 5.31: $g(r)$ calculated for the unified set of the position of the NPCs and the LBRs. According to this measurement, there is some excluded volume interaction between the pore complexes and the receptor proteins, however, there is no preference regarding the separation distance.

finding a LBR at a distance r from a NPC. Similarly, we can also measure the probability of finding a NPC at a distance r from a LBR. By doing so, we get the radial distributions presented in figure 5.32. This figure indicates that the LBRs are closer, on average, to the NPCs than the pore complexes are to the LBRs. This suggests that certain pore complexes are further away from most of the Lamin B receptors.

The $f(r)$ distribution of edge-lengths in the union network deviates from that of the NPC or LBR network on all scales, as figure 5.33 shows this. This means that the NPCs and LBRs "mix", and because of this mixing, the nature of the neighbor-network of the union is qualitatively different.

Distribution of Shortest Distances

Neither the $g(r)$, nor the $f(r)$ signals the presence of any special relation between the position of the NPCs and the LBRs. However, the $g_{AB}(r)$ radial distribution suggested that there is certain relation among the positioning of the LBRs and the positions of the NPCs. To investigate this relation in more details, we calculate a further quantity, the $s(r)$ distribution of the length of the shortest edges for each of the pores and receptors, respectively.

For any particular pore complex, there is a closest receptor, similarly, for any particular receptor there is a closest pore complex. Note that we already know the edges corresponding to these shortest distances. They are stored in the $E_{A\to B}$ and in the $E_{B\to A}$ edge sets. In fact, $s(r)$ is the distribution of the lengths of the edges in $E_{A\to B}$ and in $E_{B\to A}$, respectively. We calculate the distributions for the two sets separately.

Figure 5.34 illustrates these distributions. According to the $s(r)$s, most of the pores are

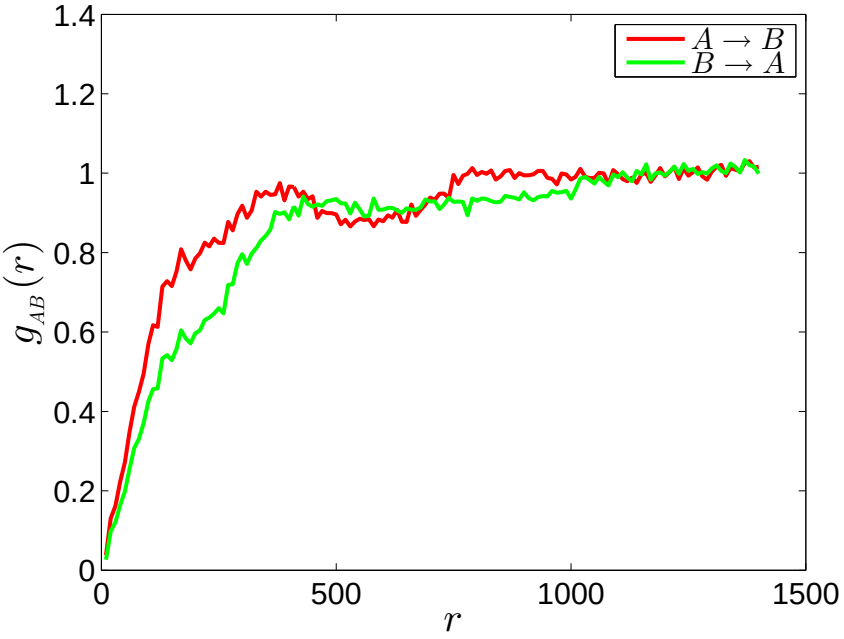

Figure 5.32: The $g_{AB}(r)$ radial distribution gives the radial distribution of the positions of the LBRs – in red – (NPCs – in green) measured from centers defined by the positions of the NPCs (LBRs, correspondingly). The plot suggests that the LBRs on average are closer to the NPCs than the NPCs are to the LBRs. This means that there are certain pore complexes which are further away from most of the LBRs.

placed to a preferred minimal distance of around 200nms from the receptors. However, larger distances, even above 1000nms also have significant probabilities. On the other hand, most receptors are positioned at a distance of about 250nms from some of the pores and there are no receptors further away than 500nms from a pore complex. This means that the position of the Lamin B receptors are somewhat influenced by the position of the pore complexes. While not all pore complexes will be surrounded by a receptor, receptors will always bind close to some of the pore complexes. In a more biological interpretation, we could say that since NPCs are embedded in the envelope, they are less mobile. In turn, LBRs can diffuse more easily and prefer to bind close to a pore complex.

Conditional Probabilities

Another way to investigate local structural properties of the constructed networks is by calculating the conditional probability distribution describing the length of edges connected to a given node, conditioned on finding an edge with a given length connected to the respective node. In this case, the question we ask is the following: What is the probability of finding an edge of length r connected to a given node if we know for sure that the node has an edge which has a length of r'. This probability can be shortly denoted by $p(r|r')$ and can be numerically represented as a matrix. A structure in the graph, in which certain nodes have only long edges while others only short ones and yet others have edges with intermediate lengths, would be averaged out in the calculation of $g(r)$ as the radial distribution is averaged over all the nodes. It obviously would not count in the calculation of $f(r)$, as in this case we look only at the length of the edges and do not care

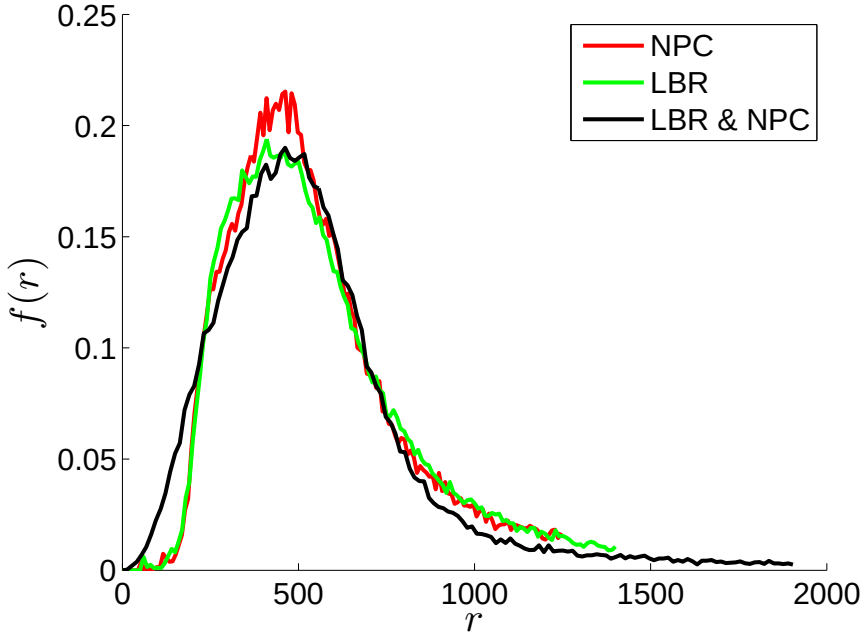

Figure 5.33: The rescaled $f(r)$, calculated for the union of the sets of the position of the NPCs and the set of positions of th LBRs (in black), compared to the $f(r)$ for the NPCs (in red) and for the LBRs (in green) . The $f(r)$ for the union deviates completely from the other two distributions. Thus the local structure of the union is completely different from that of the NCSs and LBRs.

to which nodes are the different edges connected to. The $p(r|r')$ conditional probability reveals this kind of information. If the graph is biased towards a structure in which edges are organized in the mentioned way, $p(r|r')$ would have large values when r is close to r' and smaller values otherwise. This effect is demonstrated in figure 5.35. Here we show a set of points which have some visually obvious features. However, both the $g(r)$ and the $f(r)$ fail to indicate the presence of a structure in the set. On the other hand, the uneven spacial distribution of the points is clearly indicated by the $p(r|r')$ conditional probability. The brighter diagonal in the matrix-plot indicates the presence of the non-uniform spatial arrangement.

Calculating these conditional probability distributions for the graphs defined by the positions of the NPCs, LBRs and the unified set (figures 5.36, 5.37 and 5.38, respectively), we observe only a slight preference driven by the r' condition. This preference is slightly more prominent in the case of the unified set. Nevertheless, the structures seem to be relatively uniform.

5.3.5 Discussion and Conclusions

We presented an analysis of a set of 3D-SIM microscopy images of the nuclear envelope in which nuclear pore complexes and Lamin B proteins were separately stained. We described a procedure to detect the position of the pore complexes and the receptors. We analyzed the spatial organization of the complexes and receptors within a network-theory framework. We built the neighbor-network of the objects by means of a Delaunay triangulation. Instead of applying a three dimensional approach, we removed all the edges

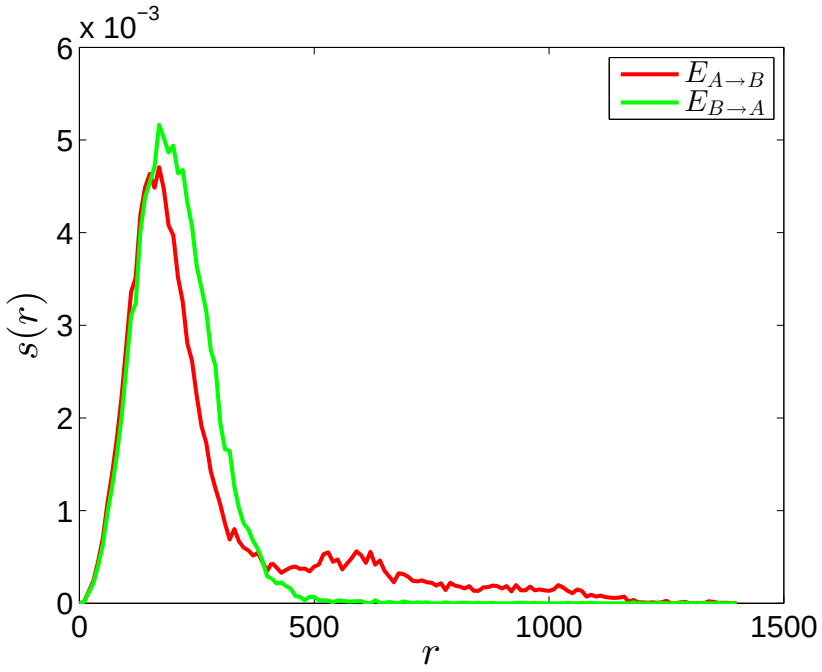

Figure 5.34: The $s(r)$ distributions of the shortest distances measured from each of the pore complexes to any of the receptors (in red) and from each of the receptors to any of the pore complexes (in green). While most of the pores are placed to a preferred minimal distance of around $200nm$s from the receptor proteins, larger distances, even above $1000nm$s, also have significant probabilities. On the other hand, most receptors have a minimal distance of about $250nm$s from the pores and no receptor is placed at distances larger than $500nm$s

from the network which do not run along the nuclear envelope, therefore, we studied the system in a (quasi) two dimensional space. To gain information regarding the organization of the pore complexes and the receptor proteins we calculated quantities like the degree distribution, the radial distribution function, the distribution of the edge lengths (also conditioned on the presence of a given length), the distribution of the shortest edges.

We found that the neighbor-networks of the pore complexes and the Lamin B receptor proteins have a similar nature in terms of graph-topology, they have a similar degree distribution, however, they are different in a geometrical sense: the arrangement of the pore complexes is different from that of the Lamin B receptors. Both the pore complexes and Lamin B receptors can be imagined as hard spheres when considering only the NPC–NPC and LBR–LBR relations. In turn, an NPC versus an LBR is like a system composed of two spheres which penetrate each other to a varying extend.

We showed that while the spreading of the pore complexes and the receptors along the surface is rather uniform, the Lamin B receptors prefer to be located close to a pore complex.

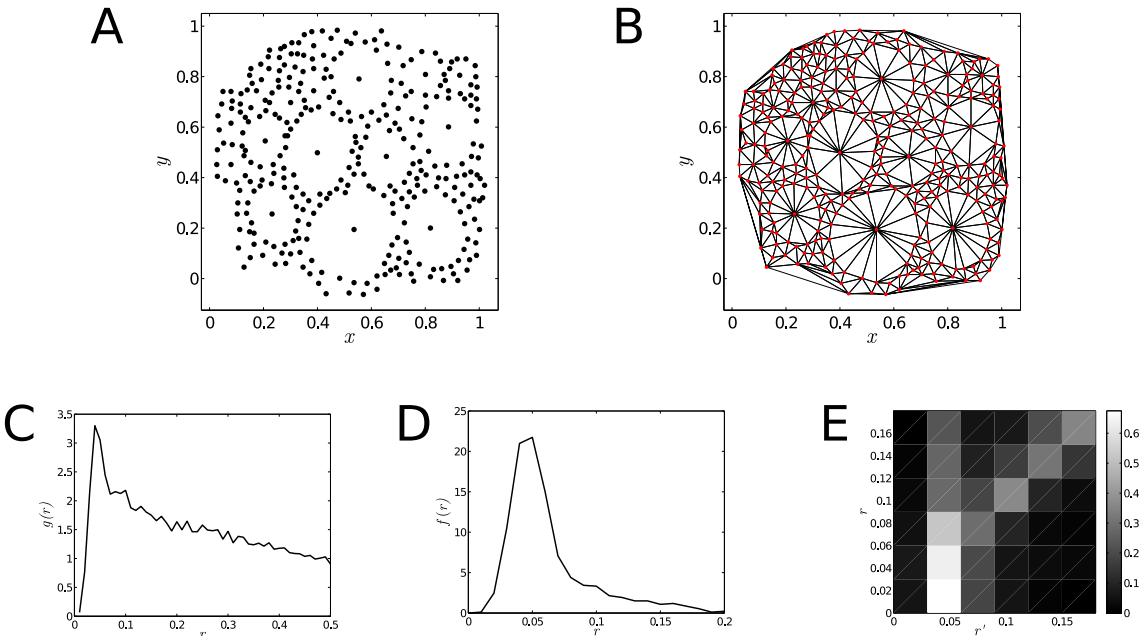

Figure 5.35: Schematic figure illustrating the usefulness of $p(r|r')$ conditional probability. Panel A: a set of points in the two dimensional space, exhibiting a spatial arrangement where certain points "occupy" relatively large areas compared to other points. Panel B: Delaunay triangulation of the point-set. Panel C: the $g(r)$ of the points indicates that the points do not overlap (its value is close to zero for very small distances) but reveals no other structure. Panel D: the $f(r)$ qualitatively is very similar to what we observe in the case of LBRs or NPCs. Panel E: the $p(r|r')$ indicates the special arrangement by the higher probabilities along the diagonal. Shades correspond to probability values indicated by the color-bar. The matrix is a right stochastic matrix, meaning that entries in a row sum up to 1.

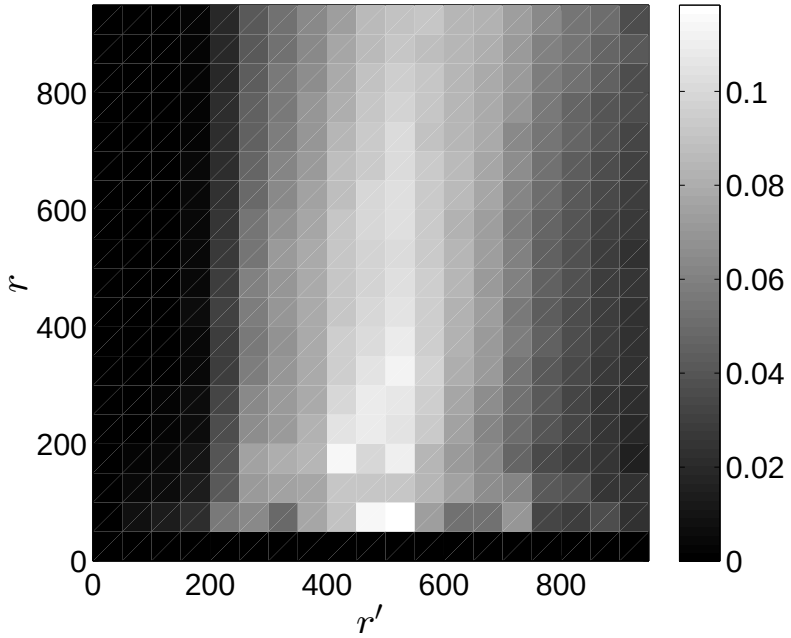

Figure 5.36: Conditional probability distribution defined by the NPC graph.

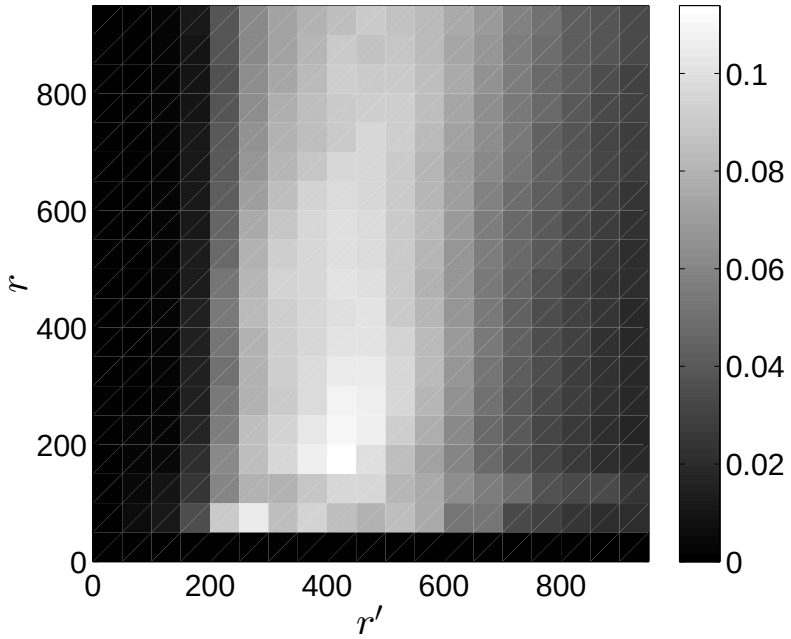

Figure 5.37: Conditional probability distribution defined by the LBR graph.

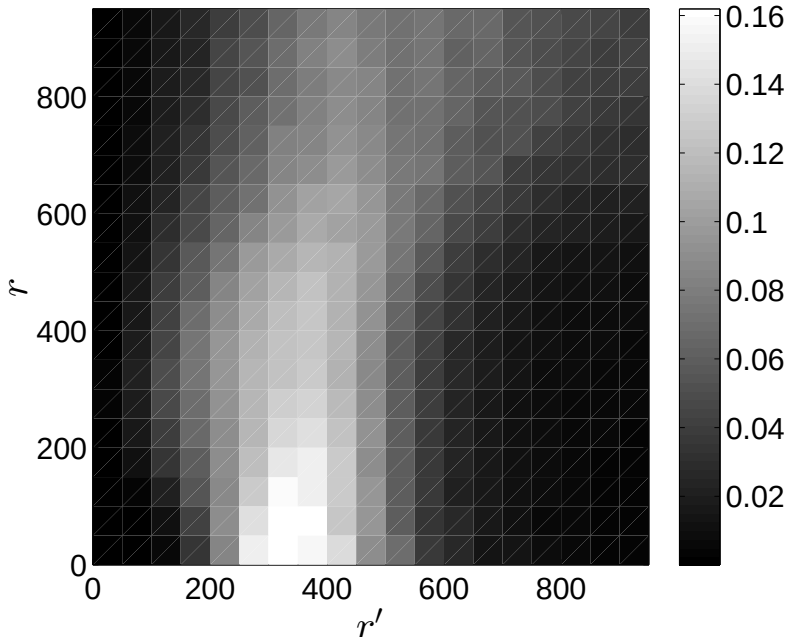

Figure 5.38: Conditional probability distribution defined by the NPC+LBR graph.

Chapter 6

Mixing Topology and Geometry: Barcodes

References

- C.J. Feinauer, A. Hofmann, S. Goldt, L. Liu, G. Máté and D.W. Heermann, *Zinc finger proteins and the 3D organization of chromosomes.*. Adv Protein Chem Struct Biol. 2013;90:67-117. doi: 10.1016/B978-0-12-410523-2.00003-1.
- G Máté, A Hofmann, N Wenzel and D.W. Heermann, *A topological similarity measure for proteins*, Biochimica et Biophysica Acta (BBA) - Biomembranes, Available online 10 September 2013, ISSN 0005-2736.
- G. Máté and D.W. Heermann, *Statistical Comparison of Topological Features of Proteins* , PLoS One, under peer review (2013).
- G. Máté and D.W. Heermann, *Persistence Intervals of Fractals* , Physica A, under peer review (2013).

Chapter Summary

This chapter is organized around the concept of persistence diagrams introduced in Section 3.4. While most of the sections of this chapter deal with assessing a similarity between chemical structures, we will deviate from this topic in the last section, presenting a more generic application of the persistence diagrams. Although the application is presented from a theoretical point of view, it indicates that an analysis based on the mentioned topological invariants can offer deeper insight in a particular folding of a protein.

We showed in our publication [57] that flexibility is a very important feature of proteins. The theoretical model presented in the paper indicates that increasing flexibility, the binding affinity may increase by orders of magnitude. We argued, that the *CCCTC-binding factor* (CTCF) or 11-*zinc finger* protein, often called the master weaver of the genome, is able to bind to many different sites partly because it is very flexible. In fact, it behaves like a self avoiding random walk, with monomer size roughly corresponding to the diameter of a zinc finger. Therefore, it is very important how we treat proteins when comparing them. Flexibility is obviously a feature which must be considered properly.

In this chapter, we present the generic ideas of a framework which enables a comparison considering geometry and flexibility in a natural way. We describe three different variations of the approach. While they are all based on the calculation of persistent homologies, each of the approaches has its own particularities. In order to ensure the understandability of the different approaches, we will introduce the concept of Vietoris-Rips and alpha complexes and that of the persistent diagrams in a more intuitive manner at the appropriate sections. Since the sections are adapted versions of already published papers or manuscripts which are currently under peer review, each section can be read independently from the others. The mentioned approaches are described in Sections 6.1, 6.2 and 6.3.

We recognize that if the persistent diagrams indeed represent the topology of objects, they also must encode properties like fractality, for instance. We show in Section 6.4 that there is a simple and intuitive relation between the fractal dimension and the structure of the diagrams and that this relation can be used to calculate the fractal dimension of certain objects.

6.1 The Hausdorff Distance Based Topological Similarity

References

This section is adapted from its published version, the second part of our paper,

- C.J. Feinauer, A. Hofmann, S. Goldt, L. Liu, G. Máté and D.W. Heermann, *Zinc finger proteins and the 3D organization of chromosomes.*. Adv Protein Chem Struct Biol. 2013;90:67-117. doi: 10.1016/B978-0-12-410523-2.00003-1.

The results presented in the section are the works of AH and GM.

As we showed it in our publication [57], proteins are flexible structures. Then, it is logical to ask how do we threat these flexible proteins? How do we compare them to other chemicals? How do they bind to different structures? Can we predict potential binding sites? It is obvious, that predicting these sites is not a lock and key problem. The following subsection aims to introduce a very engaging method which promises to answer all these questions.

To have a reference base, first we will summarize the presently still generic geometric treatment of proteins which presumes that they are crystalline structures. Then we introduce a new approach to the problem and illustrate its potentials with a few examples.

6.1.1 Geometric Similarity

Let us now shortly describe how geometric similarity can be determined. Our aim here is not to exhaust the meaning of geometric similarity, instead we would like to provide a simple basis for the comparison with our approach.

In order to determine the similarity of chemical structures first we need to find an adequate shape representation of the molecules. Calculating geometric similarity practically means calculating the similarity of shapes of the molecules. The shape of chemical structures can be estimated by placing a sphere to the center of the structure's each atom with a radius corresponding to the atom's Van der Waals radius. These spheres can be Gaussian surfaces but in the most unsophisticated case we would use simple hard spheres.

A very important step when estimating geometric similarity is the calculation of the best alignment of the structures as geometric similarity measures heavily depend on the orientation of the shapes. In other words this means that no matter what the definition of the geometric similarity is, we need to maximize it over all rotations and translations of one of the structure while keeping the other one fixed. This is a computationally very expensive process.

After we achieved the best alignment we can proceed with the last step of calculating geometric similarity. For this reason we overlay the shapes and calculate the subsection and union of volumes as it is illustrated in Figure 6.1 with the help of adenine (A) and guanine (B).

The Hodgkin measure of geometric similarity [237] is defined as the Tanimoto measure [238] of the volumes, that is the volume of the subsection over the volume of the union. In more general case the volume can be replaced with other descriptors however, we will constrain ourselves to calculating only volumes as it already provides the necessary base for comparison.

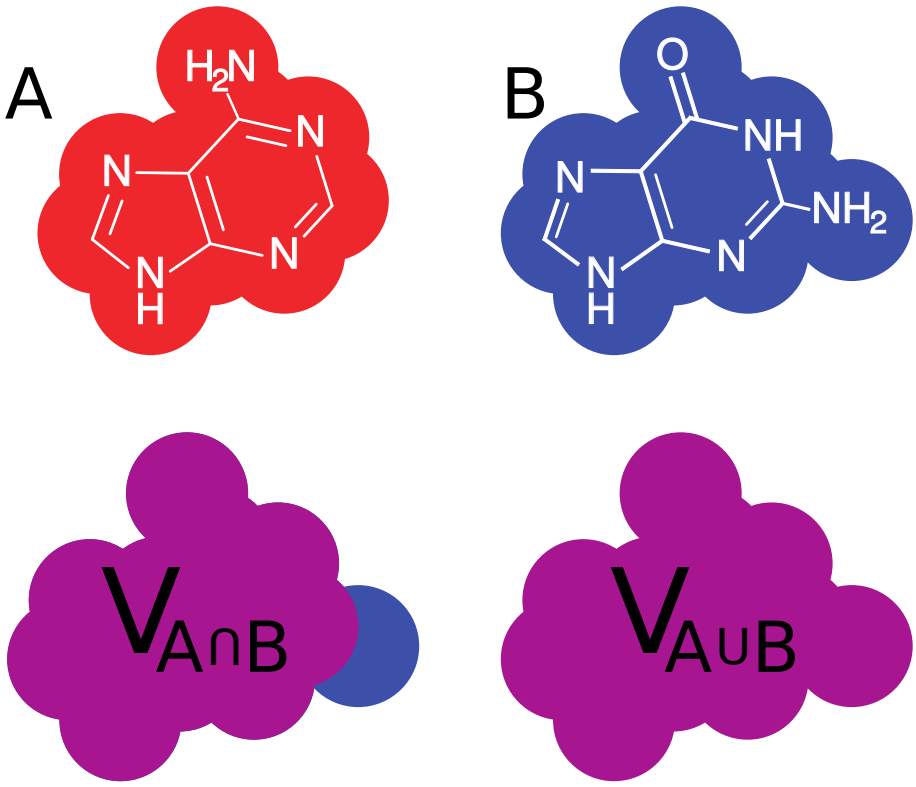

Figure 6.1: Geometric similarity is calculated as the ratio of the volumes of the subsection ($V_{A\cap B}$) and the union ($V_{A\cup B}$) of the shapes representing the chemical structures.

A very good and detailed review of geometric similarity measures can be found in [237].

6.1.2 Geometry vs Topology

As it was pointed out, geometric similarity measures focus on how the actual shape of the molecules compare. However, a vast amount of proteins we encounter in living organisms are very flexible structures. This flexibility allows them to frame to different constraints or to other shapes. On the other hand, the capability of such adaptability automatically triggers the fact that there must be an almost unlimited number of shapes, an ensemble of shapes, a protein could don. Some of these shapes are of course similar to each other, but some of them might be totally different. If we randomly pick two shapes from the ensemble odds are high that the geometric similarity measure indicates a reduced similarity between the two. This is of course correct if we are talking strictly about geometry. However, proteins are flexible for a good reason: to be able to adapt and bind to different shapes. Thus, when talking about similarity measures as means of identifying potential bindings we must not neglect the capability of a protein of changing its conformation. We should be able to tell that the two conformations came from the same ensemble that is, they represent the same protein or, at least, we should be able to indicate that small scale (local) structures are built up in the same way. This means that structures like bonds, rings, loops, etc. must be preserved. In a more mathematical language we would say that the *topology* of the two systems must be the same.

We introduced here the key concept of our method: topology. To briefly summarize, topology is the field of mathematics which studies properties of objects which are preserved

under certain deformations like stretching and bending. Note that these deformations are exactly the ones which allow a protein to change its shape. Thus, if we look at the topology of proteins instead of their geometry, we should be able to decide whether two instances are chosen from the same ensemble or not.

6.1.3 Using Topology to Compute Similarity

Looking at the topology of a given molecule carries a huge advantage. The topology of a structure is invariant with respect to similarity transformations. This means that for a given object no matter what the point of reference is, the topology is the same. Thus, by looking at the topology of the objects one avoids the expensive calculation of the best alignment.

When it comes to investigating the topology of a protein, we can think about a few different approaches. One is to treat the proteins as graphs – abstract mathematical objects consisting of nodes interconnected by edges or links – in which the nodes would represent the atoms and edges the chemical bonds. Afterwards we could forget about the protein itself and could carry on looking only at the graph. We can pick from a vast amount of well-established and studied measures introduced by the experts of the field which would characterize the topology of the graph. However, in this case we have to be careful as these measures are usually independent of the physical size of the investigated system that is, these measures are invariant with respect to the rescaling of the system. In the present situation this behavior is not desired.

Another approach is to take advantage of the recent developments of computational geometry as it promises suitable techniques crafted specifically to analyze topological properties of a given object by calculating so-called *topological invariant*s. The topological invariants we will focus on are the number of connected components, the number of holes and the number of voids in the investigated object. Connected component in this context means parts which are connected to each other. For example, a regular ball has a single connected component (since it is a single piece), no holes and a single void (inside). A piece of paper also has a single connected component, no holes and no voids. If we would tear the paper in two, the system composed from the two (now separated) pieces of paper would have two connected components (since the two pieces are not connected to each other anymore), no holes and no voids. If we would take a pencil and would poke a hole in one of the papers we would end up with a system with two connected components, one hole and no void. In the field of topology the mentioned topological invariants correspond to the so-called *Betti numbers* [239]. The number of connected components is the 0th Betti number, the number of holes is the 1st Betti number while the number of voids is the 2nd Betti number.

If we would try to compare two objects relying only on the number of connected components, holes and voids, we would quickly run into trouble. For instance, we would find that a regular coffee cup (one component, one hole in the handler and no void) is similar to a donut (one component, one hole in the middle and no voids).

A Barcode Representation of Topology

In order to avoid such blunders when comparing proteins we resort to the following abstraction. We remove all the bonds from the proteins keeping only the atoms. From now on we are not interested in the physical meaning of atoms, thus we will refer to them simply as points. Then we adopt the following procedure:

- we define a distance scale d
- we connect all point pairs that are closer then d with a line
- we calculate the topological invariants for the obtained structure

We repeat these steps for different d values and follow how the topological invariants behave as we vary d. To do this consistently, we define a minimal and a maximal value for d and fix the number of values d will have between its extremities, that is we divide the range defined by the minimum and the maximum to equal intervals. First, we set d to its minimal value, register the values of the topological invariants at this value then increase d to its next allowed value. We will repeat these steps until we exhausted the allowed values for d.

At this point it is a valid question what is counted as a hole and how can voids form when all we do is connecting points with lines. The answer lies in the definition of topological building blocks which are points, lines, triangles, tetrahedrons and their higher dimensional analogs. According to this, triangles do not count as holes, instead they constitute faces. Similarly, the space enclosed by a tetrahedron is not counted as void. Thus, any polygon which is not a triangle constitutes a hole, similarly, any polyhedron which is not a tetrahedron contains a void. Note that the faces of the polyhedrons can only be triangles – otherwise there would be a hole in the wall of the polyhedron, thus it would not be a polyhedron anymore.

After we scan the system with the procedure described above we know the numbers of connected components, holes and voids for each value of d. The acquired information can be summarized in a diagram in the following way:

- each instance of connected component, hole and void will be represented by a bar
- the start point of the bar will correspond to the value of d when the instance came into existence
- the end point of the bar will correspond to the value of d at which the instance ceased to exist

The bars for connected components are somewhat special as connected components unite as d increases. This process can be viewed as one of the connected components embeds the other one. Accordingly, the bar of the embedded component will end at the point where the component was embedded while the bar of the embedder component will continue until the letter will be embedded in another component. The role of embedded and embedder is arbitrary. It is easy to see that one of the bars for connected components will persist even at the highest values of d as there will always be at least one connected component, thus this bar can be neglected as it does not carry any information.

The diagram compiled in the previously described way will be a barcode-representation of the topology of the system in which each bar represents the interval of d over which the corresponding topological feature persists (persistence interval). An example for such a barcode can be seen on Figure 6.10. This representation was developed by Carlsson and his collaborators and a very good review of their work can be found in [240]. Betti numbers and calculating persistence intervals are discussed with more details in the theoretical introduction of this thesis, in Section 3.4.

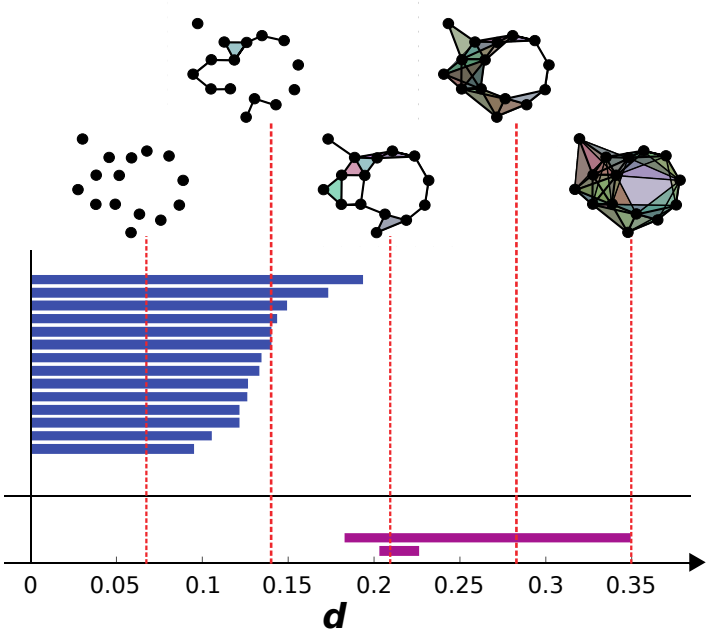

Figure 6.2: Barcodes for a given set of points in 2D. The horizontal axis represents the distance parameter. Persisting "features" are arranged on the vertical axis in an arbitrary order. On the top of the figure the procedure of connecting points is illustrated for a few values of d. Shaded faces in these illustrations signal formed triangles. Each triangle has a different color.

A Barcode Based Topological Similarity Measure

So far we described how to represent the topology of proteins with barcodes. Since these barcodes encode the topology, by comparing them we actually compare topologies. A very important extra feature of the barcodes is that while they encode topological invariants, they do not neglect the physical configuration of the proteins as they carry information about distances. In principle it would be possible to reconstruct a configuration from its barcode representation, however this would require solving an optimization problem. Thus, comparing objects by looking at their barcode representation should correlate with measuring their geometric similarity. In addition, it should be able to match geometrically different but topologically similar structures.

Comparing barcodes to each other requires a definition of a distance measure. For this purpose, we need to remember that each bar can be characterized by two numbers: its start and end point. This means that we can represent the barcodes for a given topological invariant as points in a two-dimensional plot in which one of the axes represents the start points while the other axis represents the end points. If we plot for instance two sets of barcodes for holes, we would get a diagram similar to the one in Figure 6.3 panel **a**. Then for each point from one of the sets we could find the closest point from the other set (figure 6.3 panel **b**).

We measure the distances between the pairs of closest points. Choosing the largest distance as a similarity distance assumes a worst-case similarity of the sets. In set theory this is called the Haussdorff metric [241].

The concept of Hausdorff distance (or Pompeiu-Hausdorff distance) was developed

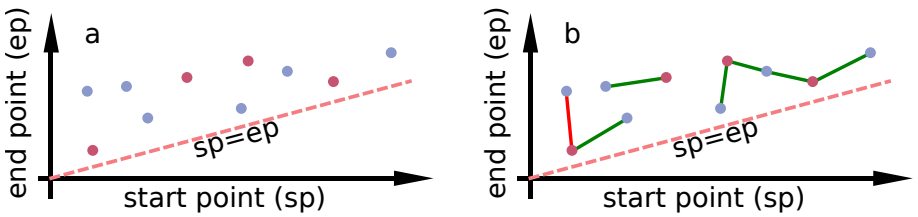

Figure 6.3: A different representation of barcodes. The start and end values of the bars are represented on different axes. Each bar corresponds to a point. Note that all points should be above the line which is defined by the start values being equal to the end values since each bar ends later than it starts. Blue and magenta points correspond to different shapes. The red and the green lines connect closest points of the different sets. The red line indicates the largest distance between the closest points.

in the field of set theory. Generally, the definition of the Hausdorff distance of the sets $C, D \subset \mathbb{R}^n$ is given by

$$d_H(C, D) = \sup_{x \in \mathbb{R}^n} |d_C(x) - d_D(x)|,$$

where $d_C(x)$ is the distance of the point x from the set C [241]. A more intuitive definition can be given as follows:

$$d_H(C, D) = \max\{\sup_{x \in C} \inf_{y \in D} d(x, y), \sup_{y \in D} \inf_{x \in C} d(y, x)\}.$$

In our case we are dealing with subsets of $\mathbb{R}^2$ as each element –each bar– is characterized by two numbers. Thus, the Hausdorff distance can be applied straightforwardly.

We represent each of our objects with three different sets: persistence intervals for number of connected components, number of holes and number of voids respectively. It is not correct to compare the number of holes to the number of connected components. Therefore, we have to calculate the Hausdorff distance for the corresponding pair of sets. In a more formal description, let A and B be two objects for which the set representing the barcodes for connected components is denoted by CC_A and CC_B, the set representing the barcodes for holes is denoted by HL_A and HL_B while the set representing the barcodes for voids is denoted by VD_A and VD_B respectively. We chose to define our topological similarity of A and B as

$$TSM = 1 - \left[d_H(CC_A, CC_B) + d_H(HL_A, HL_B) + d_H(VD_A, VD_B)\right].$$

This way, values of TSM close to one mean very similar topologies while smaller values indicate less similar topologies.

6.1.4 A Comparison between Geometric and Topological Similarity

In order to verify the previously defined topological similarity measure, we test it on a benchmark database and compare it against the Hodgkin geometric similarity measure. For this purpose we chose the DUD [242] database. We present the results for two target proteins: Acetylcholinesterase (AChe) and Thymidinkinase (TK). The DUD provides 105 ligands and 3732 decoys for AChe while for TK we have 22 ligands and 785 decoys. We divide each ligand and decoy in 2-3 subsystems depending on their size and

homogeneously sample neighborhoods of similar sizes from AChe and TK. We compare each ligand and decoy to its corresponding target by comparing all possible combinations of pairs of neighborhoods between them. Topological invariants are calculated with the freely available *perseus* software [243], while geometric similarities are calculated with an in-house developed library. We mark the results of the comparison on a plot with the x axis representing the topological similarity measure while the y axis represents the geometric similarity. Since on these plots many of the points overlap we calculate the density of the points and present them as a density-plot.

It is already clear by inspecting Figure 6.4 – the plot for the AChe target – that the topological and geometric similarity measures do correlate, as it is naturally expected. We can also see that while the geometric similarity ranks relatively low for some of the neighborhood pairs they might still have very similar topologies. This suggests that the introduced topological similarity measure is able to pick up additional potential binding sites which is a very important achievement over the geometric similarity measure (lower right side of the figure). Similar observations can be made inspecting figure 6.5 which presents the results for the TK target.

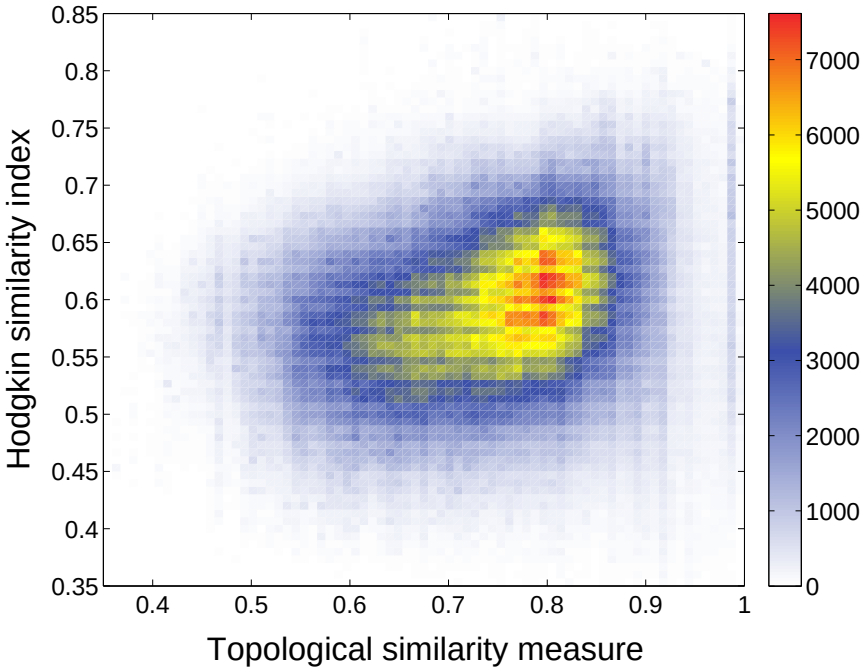

Figure 6.4: Topological vs geometric similarity for AChe.

In order to prove that our similarity measure is consistent and well-behaving, we calculate the distribution of the highest similarity measure values for the decoys and the ligands separately and expect a similar distribution for the two.

Examining Figures 6.6 and 6.7 we can deduce that our expectations are satisfied. Thus, we can conclude that the presented method works well on real-life scenarios and correlates with the geometric similarity which was an expected behavior. Moreover, our method has a huge advantage over the geometric similarity measure: it can indicate additional potential binding sites which were totally neglected by the geometric similarity measure.

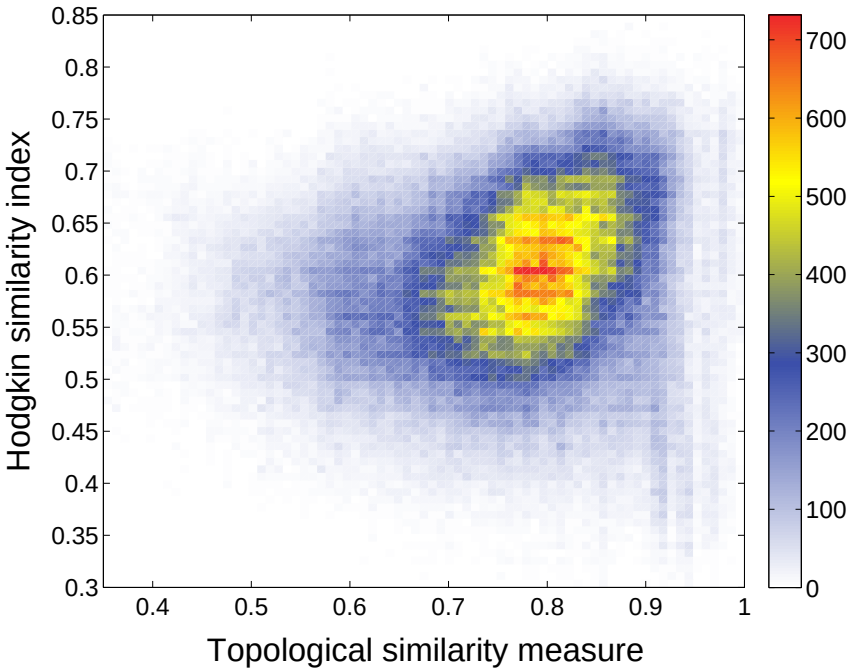

Figure 6.5: Results for the topological and geometric similarity measures for TK.

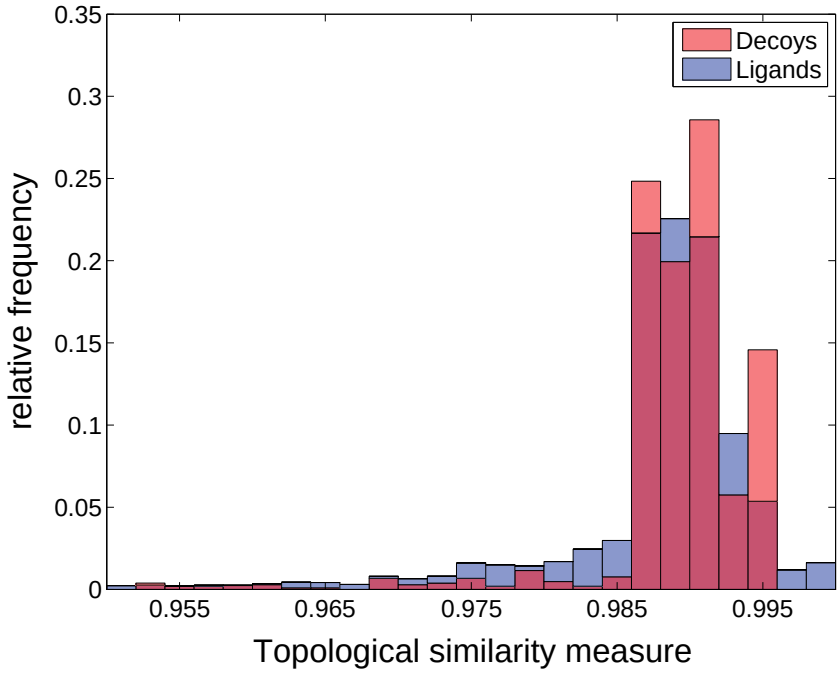

Figure 6.6: Distribution of the highest values of the similarity measure in case of AChe.

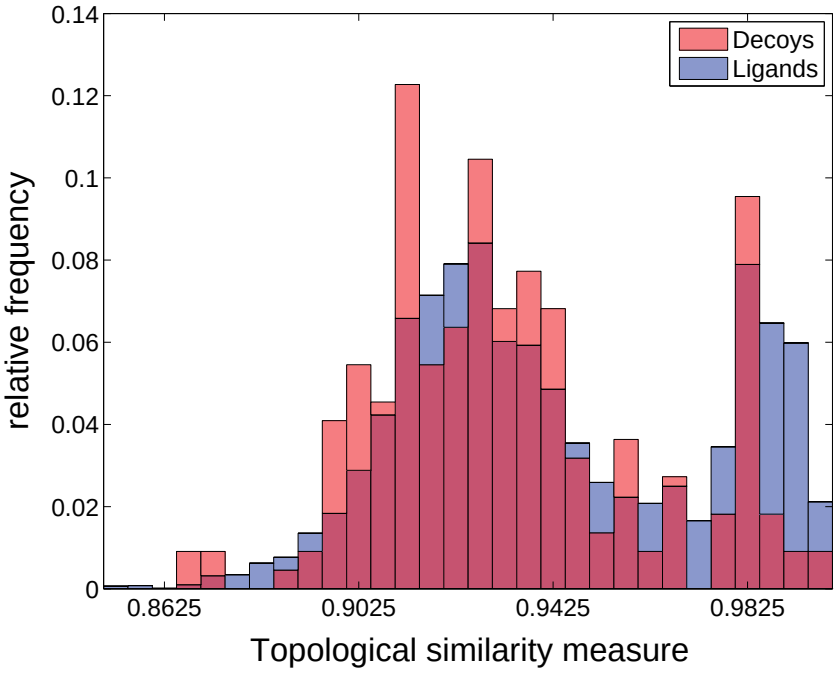

Figure 6.7: Distribution of the highest values of the similarity measure in case of TK.

6.1.5 An Application to CTCF

As a first test-application on CTCF we considered several configuration of the protein achieved by bending some of the linkers and then relaxing the system. In our test we chose the subunits composed by the bent linkers and the two zinc finger domains at the end of the former and attempt to demonstrate that although these units changed their geometry, we are still able to show that they are similar. The units are illustrated in Figure 6.12.

To do this, first we coarse-grain the selected units as proper coarse-graining must not change the topology and it improves computation. Then we calculate the topological and geometric similarity measures. Results are summarized in Table 6.2.

Tests	Hodgkin similarity	Topological similarity
subunit 1 vs subunit 2	0.532	0.98847
subunit 1 vs subunit 3	0.641	0.99012
subunit 1 vs subunit 4	0.602	0.99012
subunit 2 vs subunit 3	0.669	0.98378
subunit 2 vs subunit 4	0.667	0.98413
subunit 3 vs subunit 4	0.955	0.99997

Table 6.1: Table presenting results for the geometric and topological similarity comparison of CTCF subunits.

Although geometric similarity indicates differences, it can be seen that topological similarity clearly shows that the domains have similar topologies.

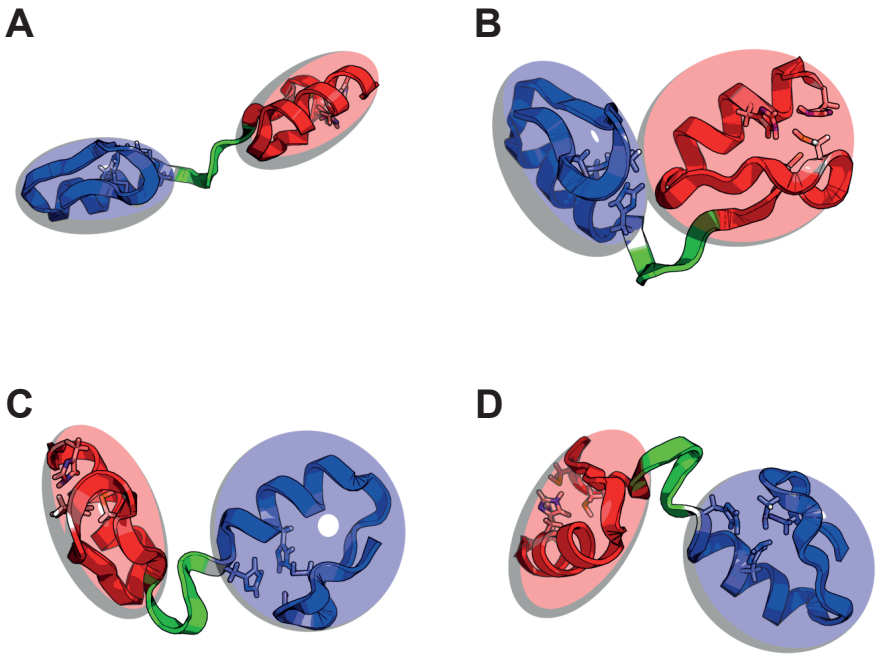

Figure 6.8: CTCF subunits used in testing geometric versus topological similarity. They are composed of two zinc finger domains and a flexible linker. Panel A subunit 1, Panel B subunit 2, Panel C subunit 3 and Panel D subunit 4 (see table 6.2).

6.1.6 Discussions and Conclusions

In this section, we presented our first attempt for constructing a framework to measure the similarity of flexible structures.

We introduced the concept of barcodes, that is, persistence diagrams, in an intuitive manner. We defined the similarity measure as one minus the Haussdorff distance. We tested the concept on ligands known to bind to two particular proteins, comparing them against decoys, which do not bind in turn, although they are geometrically similar to ligands. We found that the defined topological similarity correlates with the geometric similarity. However, this is an expected behavior. On the other hand, the topological measure assigns high similarity to extra pairs which are geometrically less similar. We also tested our measure on different foldings of two zinc fingers connected by a flexible linker protein. While geometrically the conformations were obviously different, the topological similarity correctly assessed close resemblance.

6.2 Average Jaccard Measure of Best Overlaps

References

This section is adapted from its published version,

- G Máté, A Hofmann, N Wenzel and D.W. Heermann, *A topological similarity measure for proteins*, Biochimica et Biophysica Acta (BBA) - Biomembranes, Available online 10 September 2013, ISSN 0005-2736.

AH and NW contributed in the development of the used software, in addition AH calculated many of the persistence diagrams. We would like to thank Lei Liu for providing the CTCF configurations and to Yang Zhang for the useful discussions.

In the previous section we introduced a first attempt to apply computational topology algorithms for assessing similarity among chemical structures. We saw that a persistence-diagram based method considers both topology and geometry, furthermore, our experiments indicated that the approach is more then a mere geometric comparison and is able to predict extra similarities.

Now we want to establish an alternative way to perform the same task, that is, to determine molecular similarity among chemical structures based exclusively on the physical configuration of these. We aim to develop a method which takes into account the topological features and the geometry of the investigated structures.

We validate the method by calculating the similarities of different configurations of two zinc fingers connected by a flexible linker protein. Then we apply the method on the Directory of Useful Decoys: Enhanced (DUD-E) database [244], a docking database which contains many membrane-proteins, ligands binding to these and decoys specifically selected so that they do not bind.

For the sake of completeness, we will review all the previously presented concepts.

6.2.1 Comparing Proteins

The two generic approaches (comparing geometry and comparing the topology) may be the easiest way to determine similarity among molecules, however, when purely applying one or the other we discard important information. In order to demonstrate the flaws of these methods, we briefly introduce them.

Topological Approach

Topology is the field of mathematics which investigates properties of objects which are invariant under certain deformations, i.e., stretching, bending – excluding breaking and tearing. Topological approaches (in fact all mathematical approaches) always require a good representation of the investigated objects. For instance, it would be really hard and thus unfeasible to represent a molecule with an abstract function.

Since, from the chemical point of view, the connectivity of the atoms is of crucial importance, these approaches intend to capture this aspect when representing a chemical structure. This information is easily stored by the chemical formula, but also by a more complex mathematical object in form of a graph. Graphs are specially designed to capture

connectivity information among different entities – atoms in the case of proteins, but they are suitable to represent any kind of structures composed of separable but interacting parts, commonly called *networks*. Usually, the interacting entities (e.g. representations of atoms) are referred to as vertices or nodes while connections between nodes (encoding chemical bonds, for instance) are represented by edges or links. Graphs can also be used to represent, for instance, computer networks [245], where the connected entities are the computers, online social networks [246], where nodes represent persons and edges represent friendships but also the complex connectivity characterizing the human brain [247].

In a purely topological, graph-theory based approach each atom of the investigated molecules is represented by a node and each bond is represented by an edge. The set of nodes and edges corresponding to a given molecule is a well-defined mathematical object, and there is a whole mathematical field built around these objects, called graph-theory.

Besides laying the foundations and defining the framework for handling graphs, graph-theory also provides the necessary measures and algorithms to compare graph-objects [248]. Without detailing these measures and methods, it is easy to understand now, that such an approach completely neglects any geometric or physical constraint since the representation of the data deals only with the connectivity information. Therefore, a purely topological approach could assess high similarity between a molecule and a physically and chemically incorrect copy of itself.

Geometric Approach

Comparing molecules from a geometric point of view in turn supposes representing molecules as a form of volume. The easiest and perhaps the most realistic way to do this is by modeling each atom by a hard sphere with a radius corresponding to van der Waals radius of the atom. In this case, one can define geometric similarity as the Tanimoto or Jaccard measure of the volumes [237, 249] calculated for the best alignment. This measure is defined as:

$$S_G(O_A, O_B) = \frac{V_A \cap V_B}{V_A \cup V_B}, \tag{6.1}$$

where O_A and O_B denote two different molecules, while V_A and V_B denote the volumes of O_A and O_B, respectively. The operation $V_A \cap V_B$ yields the subsection of the volumes while the operation $V_A \cup V_B$ yields the union of the volumes. Calculating the geometric similarity supposes that we previously calculated the best alignments, i.e., we tried to maximize this measure as a function of all possible rotations and translations. This is a computationally very costly procedure.

Although this measure performs very well when one is strictly interested in geometric similarity, proteins are flexible structures and flexibility turns out to be a very important property as it influences binding affinity [250] and function [251]. By calculating only geometric similarity we assess very reduced similarity between two different foldings of the same protein, which is obviously a bad result. On the other hand, geometric similarity is also sensitive to the difference in number of atoms.

6.2.2 Similarity and Topological Invariants

Based on the previous descriptions, it is clear that considering only topology or only geometry may lead to an incorrect conclusions. There is a need for a method which is able to handle flexible structures and assess the correct similarity value even in complicated cases, for instance, when one compares two different distortions of the same object.

We build our similarity measure around two concepts: topology and physical constraints. Considering only topology would result in high similarity between a structure and its stretched version, which is an unwanted behavior. Note that considering physical constraints means that to some extent we are also interested in the geometry of the structures we want to compare.

A possible way to characterize topology is to record properties of the structures which are invariant under certain deformations of the object. Deformations which might fragment the structures (breaking, tearing, gluing, etc.) are excluded. In a more mathematical language, these deformations must correspond to a continuous transformations of the topological space defined by the structures.

Just as in the Section 6.1, we will focus our attention to three quantifiable properties: the number of components which are independent from each other and connections only exist within components, the number of holes on the surfaces and the number of voids inside the structures. Remember, the field of algebraic topology has special names for these properties, they are called the Betti numbers of dimension zero, one and two, respectively, and they turn out to be very important topological invariants which help distinguishing between different topological spaces [239, 252].

By comparing these quantities of two solid objects we can decide whether they have the same topology or not. But molecules are not solid objects. They are better described by the point-set defined by the coordinates of the atoms. Thus we need a method through which we can actually define what we mean by components, holes and voids.

To accomplish this, we, in fact, need to convert the point-set into a solid object in a similar fashion we did in the previous approach: First, we take the point-set defined by the coordinates of the atoms and discard all the bond-information. From now on we will work only with these points. Next, we want to define a geometric relationship among the points. For this, we start growing spheres around each of them. Whenever two spheres mutually embed each-other's center we connect the centers of the spheres by a line/edge. Points connected by an edge are considered to belong to the same component. Any two points which are connected by a path through the existing edges are in the same component. As we increase the radii of the spheres we can record each event of connecting two previously disjoint components. By this we actually can follow how the number of components changes as a function of the radius. First each point is a separate component, while for a radius large enough each point is connected, and we end up with a single component.

The definition of holes and voids also stems from this process. In order to build a solid, beside points and lines, we need face and volume building blocks. For this, we will use the simplest polygon and polyhedron, namely the triangle and the tetrahedron. Whenever three edges form a triangle we consider not only the edges but also the face of the triangle. Similarly, whenever four triangles form a tetrahedron, we consider the volume of the tetrahedron as solid volume. The described procedure is presented in Figure 6.9 for a particular set of points.

Once the surfaces and volumes are defined we can proceed and count the holes and the voids. In fact, it is possible to register their number and also the number of components for every separate value of the radius of the spheres. This will be important in the next stage.

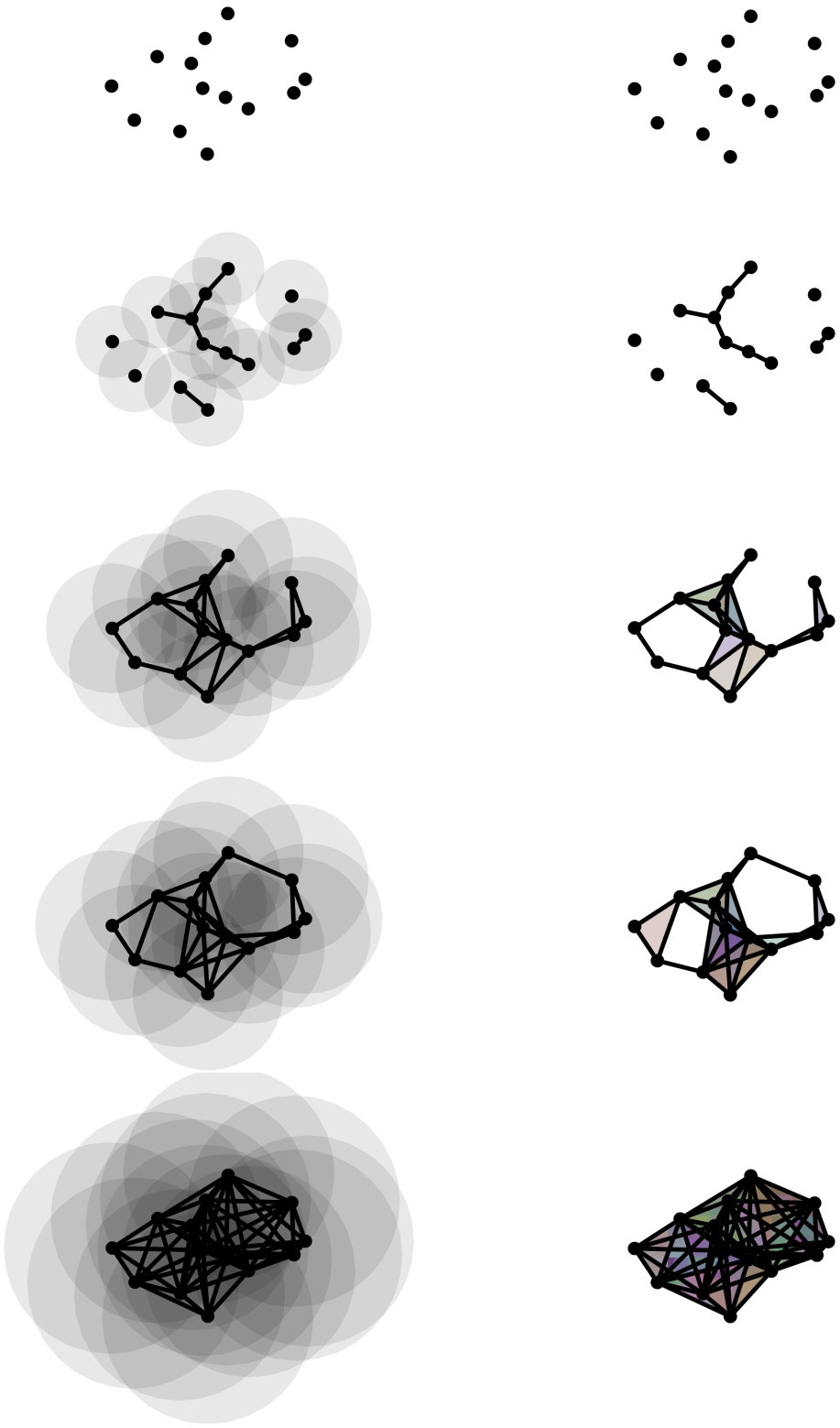

Figure 6.9: Converting a point-set into a solid object. As the growing spheres mutually embed the center of each-other the corresponding centers are connected by an edge (as shown in the left column). Whenever a triangle/tetrahedron is formed, it is included in the solid as a face/volume element(illustrated in the right column).

A Barcode Representation of the Structure

We construct the barcodes in the similar way we did in the approach described in the Section 6.1:

- each instance of component, hole and void will be represented by a bar
- the position and length of a bar represents the "lifetime" of the corresponding component/hole/void
- the start point of the bar will correspond to the value of the radius at which the instance came into existence
- the end point of the bar will correspond to the value of the radius at which the instance ceased to exist

The bars, in fact, are graphical representations of the intervals of the radii over which certain topological features (components, holes, voids) persist and they are called persistence intervals. The set of these bars characterizes how the topology of the object changes as we coarsen the representation of the structure and it can be viewed as a barcode of the topology on different scales. The barcode corresponding to point-set in Figure 6.9 can be seen in Figure 6.10.

Note that for a given object we will have three different barcodes: one for components, one for holes and one for voids. In a mathematical terminology they are often referred to as dimension 0, dimension 1 and dimension 2 intervals, respectively.

To have a more physical understanding of the concept of components, holes and voids, imagine a regular rubber ball. The ball obviously has a single component and a void (usually filled with air) enclosed by the shell. If we poke a hole on the shell of the ball, we practically destroy what we in the context of this paper would call a void, as through the hole the air can escape. Now, in theory at least, we can grab the shell from the sides of the hole and stretch the rubber ball to a flat surface. In a mathematical language, we say that the ball with the hole is homeomorphic to a plane. Therefore, a single hole on a closed surface is in fact not a hole. Perforating the shell again and stretching from one of the holes results in an object homeomorphic with a plane with a "real" (topological) hole on it. Note that all the so far created objects had just a single component. In order to have two components we would need to cut the ball into two separated parts.

The bars/intervals for connected components (green lines on Figure 6.10) are somewhat special as connected components unite as the radius increases. This process can be viewed as one of the connected components embeds the other one. Accordingly, the bar of the embedded component will end at the point where the component was embedded while the bar of the embedder component will continue until the latter will be embedded in another component. The role of embedded and embedder is arbitrary. It is easy to see that one of the bars for connected components will persist even at the highest values of the radius as there will always be at least one connected component, thus this bar can be neglected as it does not carry any information. For this reason, this bar may even be removed from the barcode.

Note that we are looking at the way the topology changes as we coarsen the representation of the structure we are investigating. By this we in fact implicitly consider geometric information without having to perform the expensive calculation of the best alignments. To understand how geometry is encoded in the barcodes let us return to the example with the ball. As already pointed out, this ball has a single connected component (its shell),

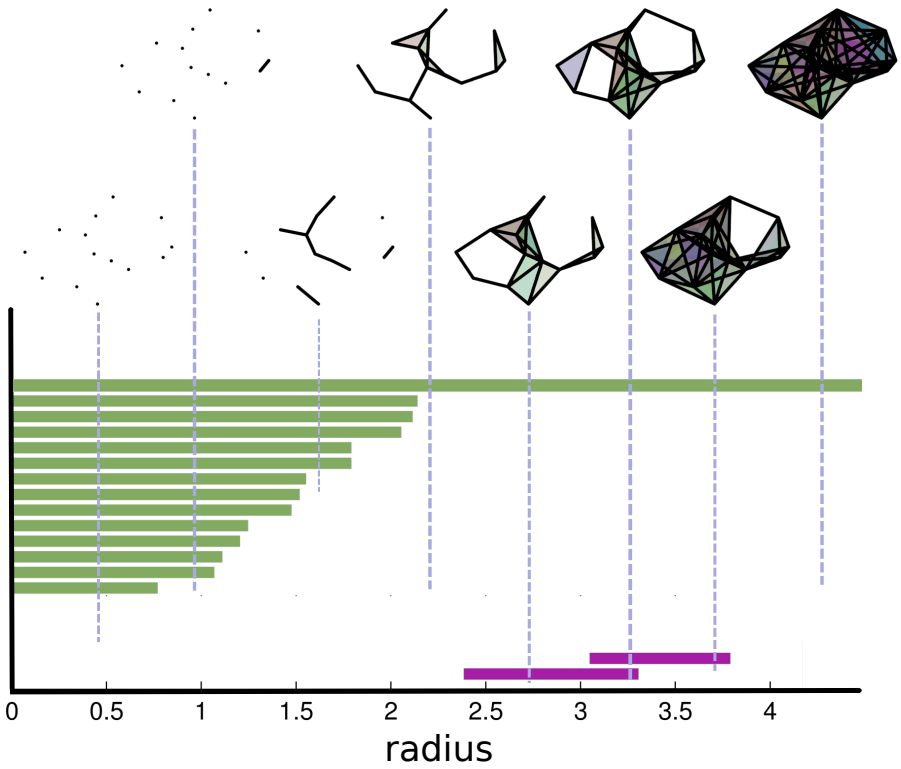

Figure 6.10: Barcodes for a particular set of points in 2D. The horizontal axis represents the radius of the growing spheres. The green bars correspond to components while the purple ones correspond to holes. Persisting "features" are arranged on the vertical axis in an arbitrary order. On the top of the figure the procedure of connecting points is illustrated for a few values of the radius. Here, shaded faces signal formed triangles. Each triangle has a different color. Note how first each point constitutes a component, then as the radius increases the points start to connect to each other, thus the number of separate components decreases. Also note that the first hole forms at a radius value around 2.4 while at a radius if 4.3 everything is connected, every hole is filled.

no holes (otherwise the air would escape) and a single void inside the shell. Thus, there would be a single bar of a length corresponding to the diameter of the ball in the barcode representing voids. It is clear that if we change the geometry of the ball by flattening it for instance, we immediately would see the result of the change in geometry by the shrinkage of the bar representing the void inside the ball.

Similarity Based on the Barcodes

At this point we are able to calculate a barcode-representation of certain important topological features for a given structure. As we argued above, these barcodes also encode geometry. It is natural then to assess the similarity of two structures which may be of high complexity through comparing their barcodes, the latter being rather simple mathematical representation of the structures.

Since a barcode is in fact a set of bars, the first thing that comes to mind is the so-called

Hausdorff distance [241] of the bar-sets. Although this approach would already provide an insight regarding similarity [250], the Hausdorff distance is a distance and not a similarity measure. It indicates the dissimilarity between two sets and its magnitude depends on the magnitude of the set-elements, that is, it is impossible to decide from the value of the Hausdorff distance of two sets whether the two sets are similar or not. We always have to provide a frame of reference. Although interpreting values of similarity measures defined on the interval $[0, 1]$ is not straightforward either, at least we know that values closer to one indicate high similarity, while values closer to zero mean reduced similarity.

Another classical way to compare sets is calculating their Jaccard or Tanimoto index (or measure) [249]. The Jaccard index is in fact the count of the elements present in both sets divided by the total number of elements, that is,

$$S_J(M, N) = \frac{|M \cap N|}{|M \cup N|}, \tag{6.2}$$

for any *nonempty* M and N sets. Unfortunately, in the case of the sets of the bars (the barcodes) it is not straightforward to apply the Jaccard similarity index since, for example, the coordinates are real valued numbers and bar-lengths may differ already because of experimental errors, thus deciding whether two bars from two different barcodes are equivalent or not is not a simple task. Also, we may consider two circles/rings similar even if their radius differs (different radius would mean different bar lengths). However, it is possible to define a measure based on the Jaccard index in the following way:

- we can calculate the Jaccard measure for every pair of intervals from two different barcodes
- for each bar from one barcode there exists a bar from the other barcode for which the Jaccard index is the highest
- we define our similarity measure as the average of these highest Jaccard measures

Within a more mathematical framework, we can define this barcode-overlap similarity measure as

$$S_{BO}(A, B) = \frac{1}{|A| + |B|} \left[\sum_{a \in A} sup_{b \in B} \frac{a \cap b}{a \cup b} + \sum_{b \in B} sup_{a \in A} \frac{a \cap b}{a \cup b} \right], \tag{6.3}$$

where A and B denote two different barcodes while a and b denote different bars from barcodes A and B, respectively. Figure 6.11 attempts to illustrate the calculation of this similarity.

For the definition given in (6.3) it is possible to show that S_{BO} is a similarity measure in the mathematical sense. In the following, we present the proof of this statement.

Definitions

Let A, B and C be three *nonempty* sets:

$$A = \{a | a = [a_s, a_e], a_s, a_e \in \mathbb{R}_+, a_s \leq a_e\} \tag{6.4}$$
$$B = \{b | b = [b_s, b_e], b_s, b_e \in \mathbb{R}_+, b_s \leq b_e\} \tag{6.5}$$
$$C = \{b | c = [c_s, c_e], c_s, c_e \in \mathbb{R}_+, c_s \leq c_e\}, \tag{6.6}$$

where $[x, y]$ denotes a closed interval with limits x and y.

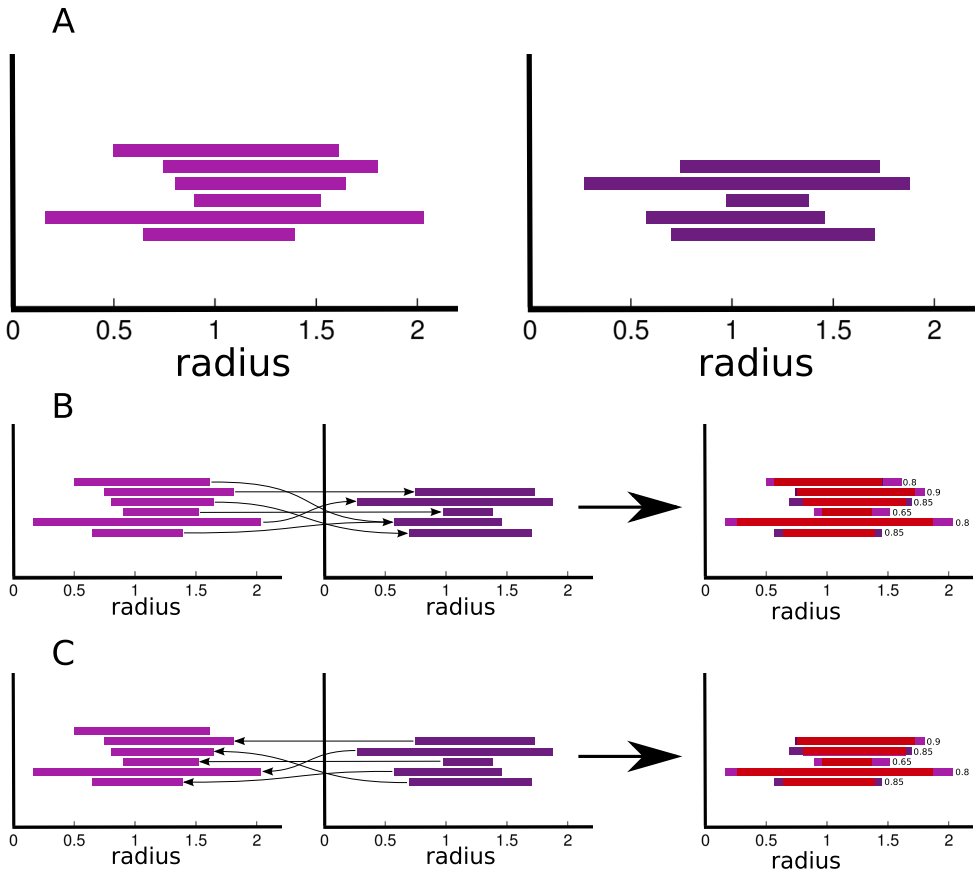

Figure 6.11: In this figure we illustrate the calculation of the proposed similarity measure. Panel A presents two barcodes from two different molecules. Panel B illustrates the process of selecting for each bar from the first barcode from Panel A, those bars from the second barcode from Panel A for which the Jaccard index is the highest. Panel C illustrates this process for each bar from the second barcode. Overlaps are illustrated in red in the rightmost plots of Panels B and C. The (approximate) Jaccard indexes are also printed next to the illustrated overlaps. Our similarity measure is, in fact, the average of these indexes, which, in the presented case, would give a similarity of 0.8091.

Let $S_{BO}(A,B)$ be a mapping defined as in Equation (6.3):

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right]. \tag{6.7}$$

Aim

We intend to prove that S_{BO} is a proper similarity measure. According to [253] S_{BO} is a similarity relation if it satisfies the following conditions:

$$0 \le S_{BO}(A,B) \le 1 \tag{6.2.C8}$$

$$A = B \Rightarrow S_{BO}(A,B) = 1 \tag{6.2.C9}$$

$$S_{BO}(A,B) = S_{BO}(B,A) \tag{6.2.C10}$$

$$A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(A,B) \tag{6.2.C11}$$

$$A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(B,C) \tag{6.2.C12}$$

Proofs

Since for any $a \in A$ and $b \in B$ $|a\cap b|/|a\cup b|$ is between 0 and 1 for any A and B, $S_{BO}(A,B)$ will also be bounded by 0 and 1, thus (6.2.C8) is true.

For $A = B$ $sup_{a\ inA}|a\cap b|/|a\cup b| = 1$ for any $b \in B$ and also $sup_{b\ inB}|a\cap b|/|a\cup b| = 1$ for any $a \in A$. Therefore, $S_{BO}(A,B) = (|A|+|B|)/(|A|+|B|) = 1$, that is (6.2.C9) is true.

Condition (6.2.C10) is true by definition.

Condition (6.2.C11) Proving $A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(A,B)$.

Because of the relation $A \subseteq B \subseteq C$, the definition (6.7) for $S_{BO}(A,B)$ and $S_{BO}(A,C)$ can be rewritten in the following forms:

$$\begin{aligned} S_{BO}(A,B) &= \frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right] \\ &= \frac{1}{|A|+|B|}\left[|A| + \sum_{b\in A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right] \\ &= \frac{1}{|A|+|B|}\left[2|A| + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right], \end{aligned}$$

that is,

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[2|A| + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right], \tag{6.2.13}$$

similarly,

$$S_{BO}(A,C) = \frac{1}{|A|+|C|}\left[2|A| + \sum_{c\in C\setminus A}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{6.2.14}$$

Equation (6.2.14) can be further rewritten:

$$S_{BO}(A,C) = \frac{1}{|A|+|C|}\left[2|A| + \sum_{c\in B\setminus A}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|} + \sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{6.2.15}$$

Denoting

$$\sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|} =: x, \tag{6.2.16}$$

we finally have

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[2|A| + x\right], \tag{6.2.17}$$

and

$$S_{BO}(A,C) = \frac{1}{|A|+|C|}\left[2|A| + x + \sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{6.2.18}$$

Then we can proceed as follows:

$$S_{BO}(A,C) \le S_{BO}(A,B) \quad \Leftrightarrow \tag{6.2.19}$$

$$\frac{1}{|A|+|C|}\left[2|A| + x + \sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right] \quad \le \tag{6.2.20}$$

$$\frac{1}{|A|+|B|}\left[2|A| + x\right]. \tag{6.2.21}$$

But since

$$\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|} \le |C| - |B|, \tag{6.2.22}$$

proving that

$$\frac{1}{|A|+|C|}\left(2|A| + x + |C| - |B|\right) \le \frac{1}{|A|+|B|}\left(2|A| + x\right) \tag{6.2.23}$$

is a stronger condition. From equation (6.2.23) we can proceed in the following way:

$$\frac{2|A| + x + |C| - |B|}{|A|+|C|} \quad \le \quad \frac{2|A| + |B| - |B| + x}{|A|+|B|} \Leftrightarrow \tag{6.2.24}$$

$$1 + \frac{|A| + x - |B|}{|A|+|C|} \quad \le \quad 1 + \frac{|A| - |B| + x}{|A|+|B|} \Leftrightarrow \tag{6.2.25}$$

$$\frac{|A| - |B| + x}{|A|+|C|} \quad \le \quad \frac{|A| - |B| + x}{|A|+|B|}. \tag{6.2.26}$$

Inequality (6.2.26) is obviously true since $|A| + |C| \ge |A| + |B|$ as $A \subseteq B \subseteq C$. Thus (6.2.C11) is proved.

Condition (6.2.C12) Here we prove that $A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(B,C)$.

The formula for $S_{BO}(B,C)$ can be rewritten similarly to (6.2.14) form of $S_{BO}(A,C)$, that is,

$$S_{BO}(B,C) = \frac{1}{|B|+|C|}\left[2|B| + \sum_{c\in C\setminus B}\sup_{b\in B}\frac{|b\cap c|}{|b\cup c|}\right]. \tag{6.2.27}$$

Let

$$y := \sum_{c \in C \setminus B} \sup_{b \in B} \frac{|b \cap c|}{|b \cup c|}. \tag{6.2.28}$$

Therefore, (6.2.27) simplifies to

$$S_{BO}(B, C) = \frac{1}{|B| + |C|} \left(2|B| + y\right). \tag{6.2.29}$$

Then, the statement we want to prove is

$$\frac{1}{|A| + |C|} \left[2|A| + x + \sum_{c \in C \setminus B} \sup_{a \in A} \frac{|a \cap c|}{|a \cup c|}\right] \leq \frac{1}{|B| + |C|} \left(2|B| + y\right). \tag{6.2.30}$$

Note that since $A \subseteq B$ the fallowing inequality holds:

$$\sum_{c \in C \setminus B} \sup_{a \in A} \frac{|a \cap c|}{|a \cup c|} \leq \sum_{c \in C \setminus B} \sup_{a \in B} \frac{|a \cap c|}{|a \cup c|}, \tag{6.2.31}$$

that is,

$$\sum_{c \in C \setminus B} \sup_{a \in A} \frac{|a \cap c|}{|a \cup c|} \leq y. \tag{6.2.32}$$

Therefore, if we can show that

$$\frac{1}{|A| + |C|} \left(2|A| + x + y\right) \leq \frac{1}{|B| + |C|} \left(2|B| + y\right) \tag{6.2.33}$$

is true, then relation (6.2.30) will also hold.

From (6.2.16) we see that $x \leq |B| - |A|$ and from (6.2.28) it results that $y \leq |C| - |B|$. Since $|A| + |C| \leq |B| + |C|$, one being the denominator on the left hand side of equation (6.2.33) the other being the denominator on the right hand side of the same equation, replacing y on both sides of the equation with $|C| - |B|$, will have a larger contribution on the left hand side. Therefore, if the resulting inequality still holds, it means that (6.2.33) also holds and therefore (6.2.30) holds, too.

By carrying out the substitution we get the following:

$$\frac{2|A| + |C| - |B| + x}{|A| + |C|} \leq \frac{2|B| + |C| - |B|}{|B| + |C|} \Leftrightarrow \tag{6.2.34}$$

$$\frac{|A| + |C| + |A| - |B| + x}{|A| + |C|} \leq \frac{|B| + |C|}{|B| + |C|} \Leftrightarrow \tag{6.2.35}$$

$$1 + \frac{|A| - |B| + x}{|A| + |C|} \leq 1 \Leftrightarrow \tag{6.2.36}$$

$$\frac{|A| - |B| + x}{|A| + |C|} \leq 0. \tag{6.2.37}$$

Since $|A| + |C| > 0$, (6.2.37) is equivalent with $|A| - |B| + x \leq 0$. But from (6.2.16) we already saw that $x \leq |B| - |A|$, therefore, our last statement is true which means that (6.2.30) is true, that is, (6.2.C12) is true.

By this we proved that S_{BO} is a proper similarity measure. However, since we may encounter the case when there are no holes or voids in our structure, we need to extend the definition of our similarity measure so that we can handle these exceptions.

Extension to Empty Sets

The extension can be achieved by recognizing that an empty set is completely similar to another empty set. Therefore, we assign a value of 1 as the similarity between two empty barcodes. Also, note that the case when there are no bars in the barcode is quite different from the case when there are bars. Therefore, we assign a 0 similarity for this case. Compressing these in a mathematical formula, we get the following:

$$S_{BOE}(A,B) = \begin{cases} S_{BO}(A,B) & A \neq \emptyset \text{ and } B \neq \emptyset \\ 1 & A = \emptyset \text{ and } B = \emptyset \\ 0 & (A = \emptyset \text{ and } B \neq \emptyset) \text{ or } (A \neq \emptyset \text{ and } B = \emptyset) \end{cases} \tag{6.2.38}$$

Based on this definition, it is also possible to show that S_{BOE} is a proper similarity measure.

Definitions

As the Jaccard index is not defined for empty sets, here we extend the proof presented above to the case which allows comparing empty sets. Since the empty set is similar to itself, we define the similarity of two empty sets as total similarity, taking the value 1. Furthermore, since the empty set is totally different from any non-empty set, we assign the value 0 to the similarity between the empty set and any nonempty set. In mathematical terms, this means the we need to prove that the measure defined as

$$S_{BOE}(A,B) = \begin{cases} S_{BO}(A,B) & A \neq \emptyset \text{and} B \neq \emptyset \\ 1 & A = \emptyset \text{ and } B = \emptyset \\ 0 & (A = \emptyset \text{ and } B \neq \emptyset) \text{ or } (A \neq \emptyset \text{ and } B = \emptyset) \end{cases} \tag{6.2.39}$$

is a similarity measure.

Proof

The proofs for the conditions (6.2.C8), (6.2.C9) and (6.2.C10) are relatively simple:

- Since $S_{BO} \in [0,1]$, S_{BOE} is also constrained to the interval $[0,1]$, therefore, (6.2.C8) is true.
- If $A = B$, this means that both are either empty or not. If both are empty, then according to the (6.2.39) definition $S_{BOE}(\emptyset,\emptyset) = 1$. If they are not empty then $S_{BOE}(A,B) = S_{BO}(A,B)$. But we already saw that if $A = B$ then $S_{BO}(A,B) = 1$. Therefore, (6.2.C9) is true.
- S_{BOE} is symmetric by definition, that is (6.2.C10) is true.

Proving (6.2.C11) and (6.2.C12) In order to show that (6.2.C11) and (6.2.C12) both hold, we need to consider four different cases of the condition $A \subseteq B \subseteq C$:

$$A \neq \emptyset,\ B \neq \emptyset,\ C \neq \emptyset \tag{6.2.C40}$$
$$A = \emptyset,\ B \neq \emptyset,\ C \neq \emptyset \tag{6.2.C41}$$
$$A = \emptyset,\ B = \emptyset,\ C \neq \emptyset \tag{6.2.C42}$$
$$A = \emptyset,\ B = \emptyset,\ C = \emptyset \tag{6.2.C43}$$

We now go through these different cases.

- case (6.2.C40) is obviously the case when $S_{BOE} \equiv S_{BO}$, therefore, both (6.2.C11) and (6.2.C12) hold in this case.
- in case (6.2.C41) $S_{BOE}(A, B) = 0$, $S_{BOE}(A, C) = 0$, $S_{BOE}(B, C) = S_{BO}(B, C) \in [0, 1]$. Therefore, condition (6.2.C11) is equivalent with $0 \leq 0$, while condition (6.2.C12) can be written as $0 \leq S_{BO}(B, C)$. It is evident that both of these affirmations hold, therefore, both conditions are satisfied.
- in case (6.2.C42) $S_{BOE}(A, B) = 1$, $S_{BOE}(A, C) = 0$, $S_{BOE}(B, C) = 0$. Therefore, condition (6.2.C11) is equivalent with $0 \leq 1$, while condition (6.2.C12) can be written as $0 \leq 0$. These affirmations again hold, therefore, both conditions are satisfied.
- in case (6.2.C43) $S_{BOE}(A, B) = 1$, $S_{BOE}(A, C) = 1$, $S_{BOE}(B, C) = 1$. Therefore, condition (6.2.C11) is equivalent with $1 \leq 1$, while condition (6.2.C12) can be written as $1 \leq 1$. Since these are all true, the original conditions are again satisfied.

Based on the previous points, we see that if S_{BO} is a proper similarity, then S_{BOE} is also a similarity measure. The pseudocode describing the calculation of the S_{BOE} similarity measure is given in the Algorithm 5.

The next question we are facing is how to unify the three similarity values we get from comparing the barcodes of connected components, holes and voids. Unfortunately, there is no unique way to do this. For example, we could take the average of the three numbers but we could also take the normalized Euclidean sum of the three, that is, summing the square of the three numbers, divide the outcome by three and then take the square root of the result. In fact, we could construct any method of unifying the values keeping in mind a single constraint: the method should not change the ordering of classification, that is, if a pair of objects is more similar in all the different barcodes then another pair of objects, the resultant unified similarity should be higher for the first pair. Mathematically speaking, we could apply any monotonically increasing function f which for any combination of input arguments from the range between zero and one would yield a result constrained to the same range, that is,

$$f : [0, 1]^3 \longrightarrow [0, 1]$$
$$f(x_1, x_2, x_3) \leq f(y_1, y_2, y_3), \forall x_i \leq y_i, x_j = y_j, i \neq j.$$

Important to note is that we must be consistent in our choice. It is not possible to compare two similarity values produced by two different forms of f. It makes even less sense to directly compare numerical values of geometric similarity to the values produced by S_{BOE} or any function of the latter.

For the sake of simplicity, we will define f as an average over the three arguments, that is:

$$f(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{3}, \tag{6.2.44}$$

and thus, we define the unified similarity measure as

$$S(O_A, O_B) = \frac{S_{BOE}(A_{cc}, B_{cc}) + S_{BOE}(A_{hl}, B_{hl}) + S_{BOE}(A_{vd}, B_{vd})}{3}, \tag{6.2.45}$$

Algorithm 5 Calculating the S_{BOE} similarity

```
1: procedure S_BOE(A, B)
2:     if A = ∅ AND B = ∅ then
3:         return 1
4:     else if (A = ∅ AND B ≠ ∅) OR (A ≠ ∅ AND B = ∅) then
5:         return 0
6:     else
7:         pos ← 1
8:         for a ∈ A do                    ▷ calculating the first sum from equation (6.3)
9:             jac[pos] ← 0
10:            for b ∈ B do
11:                if Jaccard(a, b) > jac[pos] then
12:                    jac[pos] ← Jaccard(a, b)
13:                end if
14:            end for
15:            pos ← pos + 1
16:        end for
17:        for b ∈ B do                    ▷ calculating the second sum from equation (6.3)
18:            jac[pos] ← 0
19:            for a ∈ A do
20:                if Jaccard(a, b) > jac[pos] then
21:                    jac[pos] ← Jaccard(a, b)
22:                end if
23:            end for
24:            pos ← pos + 1
25:        end for
26:        sim ← 0
27:        for i ← 1, pos − 1 do                              ▷ averaging the results
28:            sim ← sim + jac[i]
29:        end for
30:        return sim/(pos − 1)
31:    end if
32: end procedure
33: procedure Jaccard(a, b)                ▷ Calculates the Jaccard index of two bars
34:    s ← a ∩ b
35:    u ← a ∪ b
36:    return |s|/|u|
37: end procedure
```

where O_A and O_B denote two different objects/structures, A_{cc} and B_{cc} are the barcodes corresponding to connected components of the structures O_A and O_B, A_{hl} and B_{hl} are the barcodes for holes of the structures O_A and O_B, A_{vd} and B_{vd} are the barcodes representing voids of the structures O_A and O_B, respectively.

Validation of the Method

As a validation of the method, here we calculate the S_{BOE} measures of the barcodes and the geometric similarity for four conformations of two zinc finger domains connected by flexible linker proteins, extracted from different configurations of CCCTC-binding factor (11-zinc finger protein) as presented in Figure 6.12. Best overlaps among the configurations are illustrated in Figure 6.13. We summarize the results of the comparison in table 6.2.

The first observation in these comparisons is that configuration A and B have the smallest geometric resemblance. Comparing A against C yields a slightly larger geometric similarity, while this comparison gives a large value for the S_{BOE} similarity. Configurations B and C show higher geometric similarity than A and C, while the S_{BOE} measure indicates a slightly reduced similarity compared to the A-C case.

Note that the C and D configurations are almost identical, they indeed have a very high geometric similarity, showing an increase of 0.3 compared to the A-C case, while the S_{BOE} similarity barely changes, ranking both pairs as very similar. Also note that comparing configuration B against any of the others consistently yields relatively reduced (but still high) S_{BOE} similarity, probably because of the particular features in the fold, while the geometric similarity of B to the other configurations is comparable to the values of similarity we get when comparing configuration A to the others, although A and B have the most reduced geometric similarity.

We remark that the S_{BOE} similarity of a value around 0.6 configuration B shows when compared to the other configurations is considered relatively high as, comparing any of these configurations against a completely random configuration of comparable size returns a value averaging around 0.3 both for the S_{BOE} and the geometric similarity.

Tests	geometric similarity	S_{BOE}
config. A vs config. B	0.532	0.63853
config. A vs config. C	0.641	0.97505
config. A vs config. D	0.602	0.97791
config. B vs config. C	0.669	0.63293
config. B vs config. D	0.667	0.63615
config. C vs config. D	0.955	0.98984

Table 6.2: Table presenting results for the geometric similarity and the introduced S_{BOE} similarity measure among the configurations of the zinc finger proteins presented in Figure 6.12.

6.2.3 An Application

As a first application, we chose to compare the structures found in the Database of Useful (Docking) Decoys: Enhanced (DUD-E) database [244]. This database contains active ligands known to bind to given target-molecules and decoys which have geometries similar to those of the ligands, but they are chemically different. Decoys were selected from a vast amount of candidates and included in DUD-E based on two criteria. First, molecules were

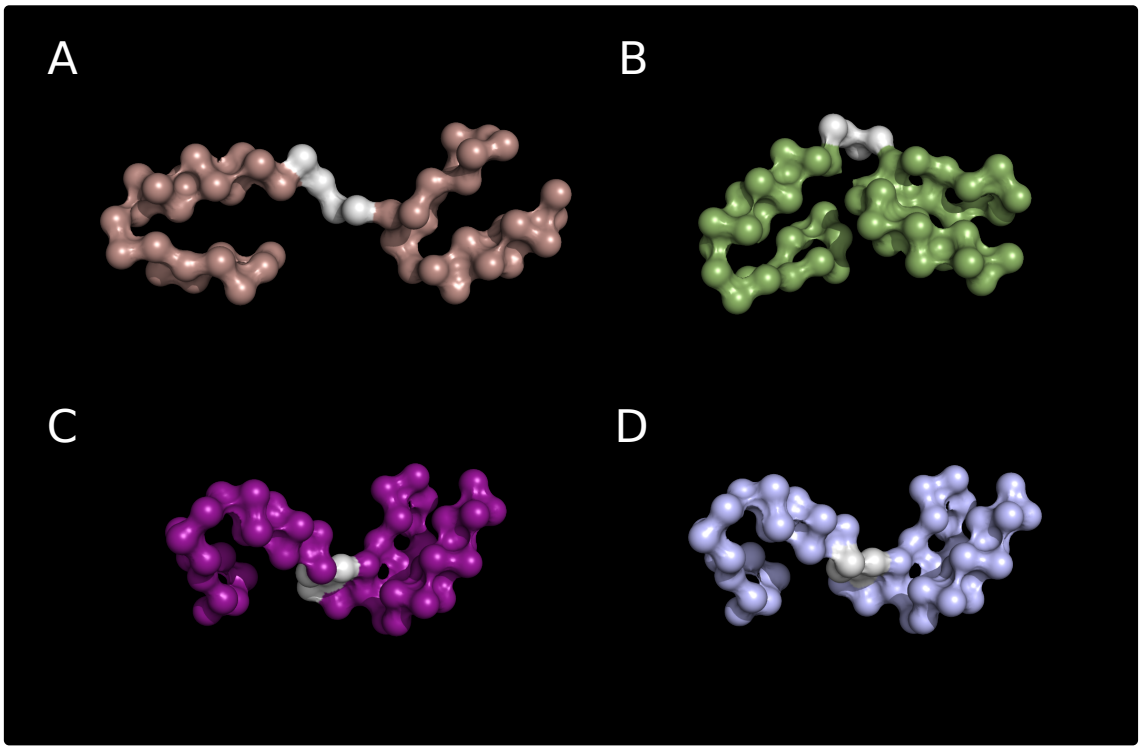

Figure 6.12: Different configurations of two zinc finger proteins connected by a flexible linker protein. We use these configurations to validate our similarity measure (see table 6.2) for the comparison.

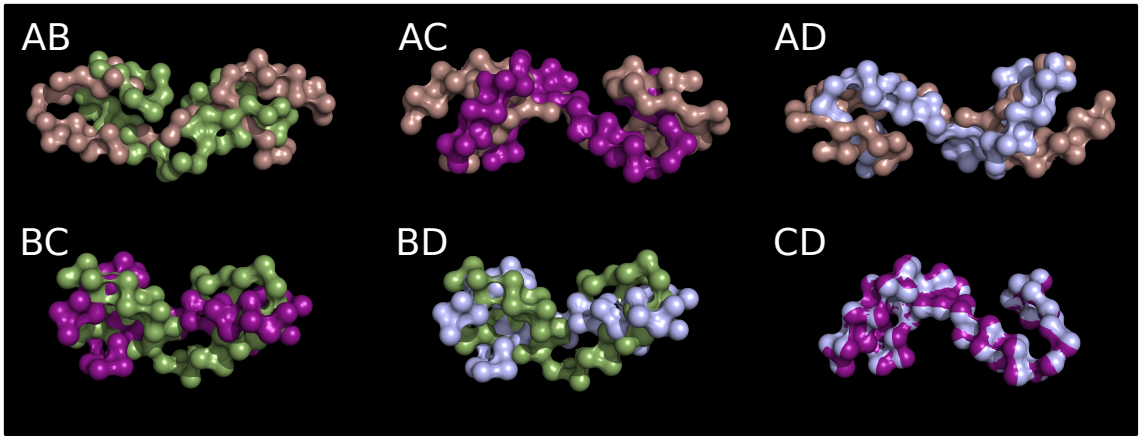

Figure 6.13: Best alignments of the pairs of configurations from Figure 6.12. For the values of the similarity measures for these pairs see table 6.2.

selected so that they have a high geometric similarity to one of the ligands, second, only those molecules were included in the database which were found to be inactive (molecules which do not bind to the target proteins – thus the name decoy).

We selected ligands grouped around fifteen proteins (AA2AR, ABL1, ACE, ADA, ADRB1, AKT1, ALDR, ANDR, AOFB, BRAF, CAH2, COMT, CP2C9, DEF, HIVPR), each of them known to bind at least to one of the targets. In this experiment we compare the ligands against the decoys from the same groups.

Although DUD-E was designed as a docking database, we use it for testing purposes. Since chemical differences must show up in the topology of the molecules, decoys and actives must present such differences. Therefore, it is a perfect sandbox for testing our similarity measure and to demonstrate that our measure picks up geometric similarity but it is not equivalent with it.

The calculations have two stages. First, there is a preprocessing step in which the barcodes are calculated. For this we used the Perseus software [243]. The calculated barcodes can be stored and there is no need to recalculate them at every comparison. A barcode, on average, can be calculated in roughly 12 seconds on a computer with a processor having a clock rate of 3.2 GHz. After barcodes for the present dataset of fifteen proteins were constructed, the similarities were calculated with a MatLab script. Using the mentioned hardware, the runtime of the similarity calculations was 2684.2 seconds. Thus a comparison is performed in 0.0020629 seconds which roughly corresponds to 484 comparisons per second.

By looking at the distribution of the values of the geometric similarity (Figure 6.14), we see that the values are centered around a well-defined mean value. It is possible to show, that these values actually follow a Gaussian distribution with a mean value around 0.6.

Looking now at the distribution of the values of the S_{BOE} similarity illustrated on Figure 6.15, we see that instead of having a single peak, a second peak may appear, which is caused by the unification of the different similarity values extracted from the barcodes of connected components, holes and voids as these different features may emphasize different aspects of the similarity. If we concentrate on the large peaks, we could say that the mean values are roughly around 0.75.

In Figure 6.16, we present the values of the geometric similarity versus the values of the S_{BOE} index. Pairs for which the values are presented were selected so that the geometric similarity is among the largest values, roughly ranging from 0.8 to 0.95, well beyond the 0.55 average value. Note that almost all the corresponding S_{BOE} similarity values are also above their 0.75 average, most withing the range between 0.75 and 0.92, that is, high geometric similarity implies high S_{BOE} values. Figure 6.17, on the other hand, is prepared so that the values of the S_{BOE} similarity index are among the highest ones. Note that though the average of the corresponding geometric similarity is higher than the global average, its values do not present such restriction as the values of the S_{BOE} similarity did in the previous case. This experiment clearly shows, that restricting geometric similarity to high values also restricts the S_{BOE} similarity index to higher values, while this is less true the other way around. This clearly indicates that the S_{BOE} measures more than the simple geometric similarity. In fact, it measures the similarity of the topological features on given geometric scales.

The same effect is also noticeable when looking at the ligands and the decoys themselves. In Figure 6.18 we plotted pairs of ligands and decoys with the highest geometric similarity, while in Figure 6.19 we show pairs of ligands and decoys for which both the

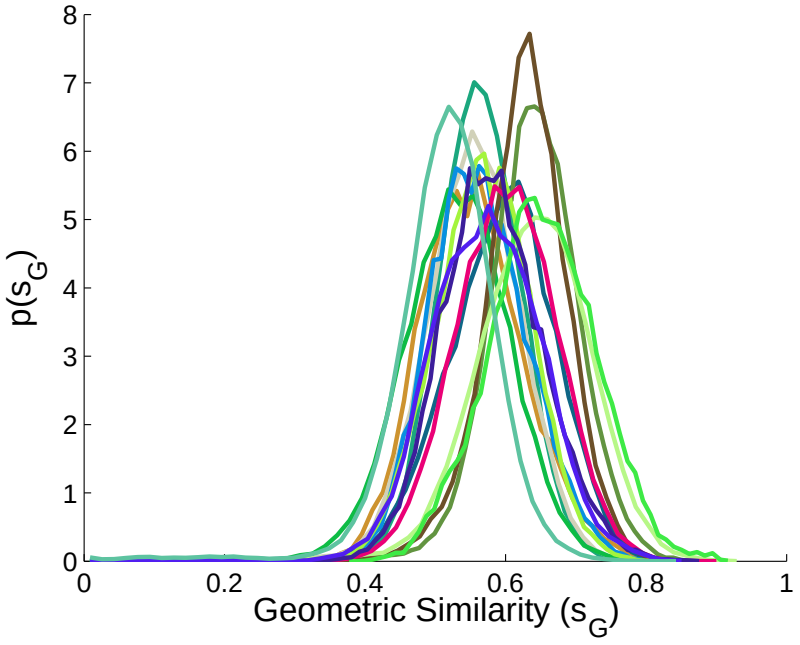

Figure 6.14: Distribution of all the geometric similarity values among all the decoys and ligands from the 15 target proteins. Colors correspond to the different target proteins.

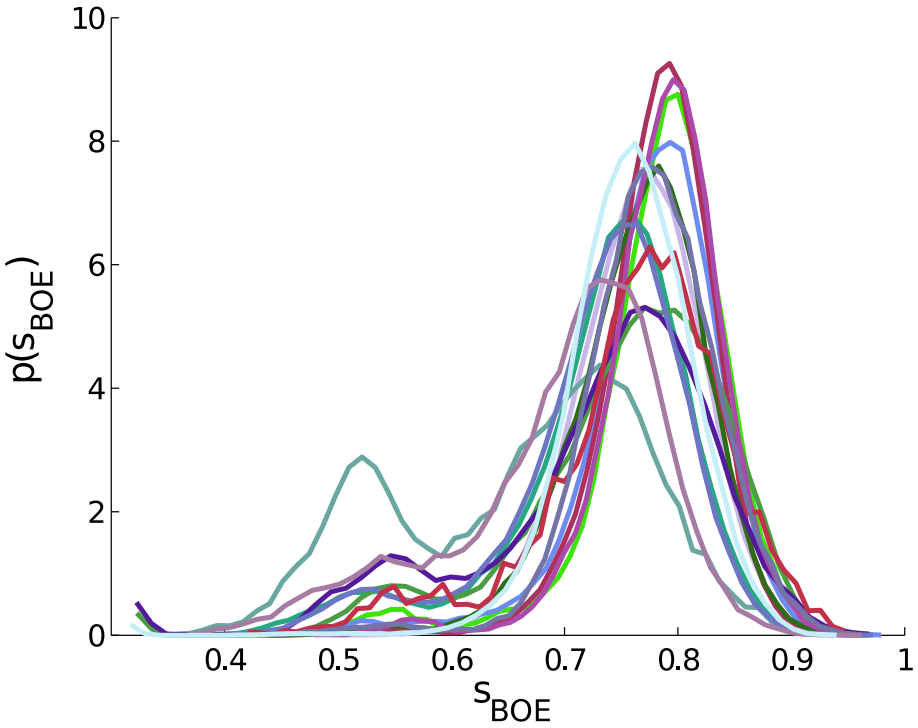

Figure 6.15: Distribution of all the S_{BOE} values among all the decoys and ligands from the 15 target proteins. Colors correspond to the different target proteins..

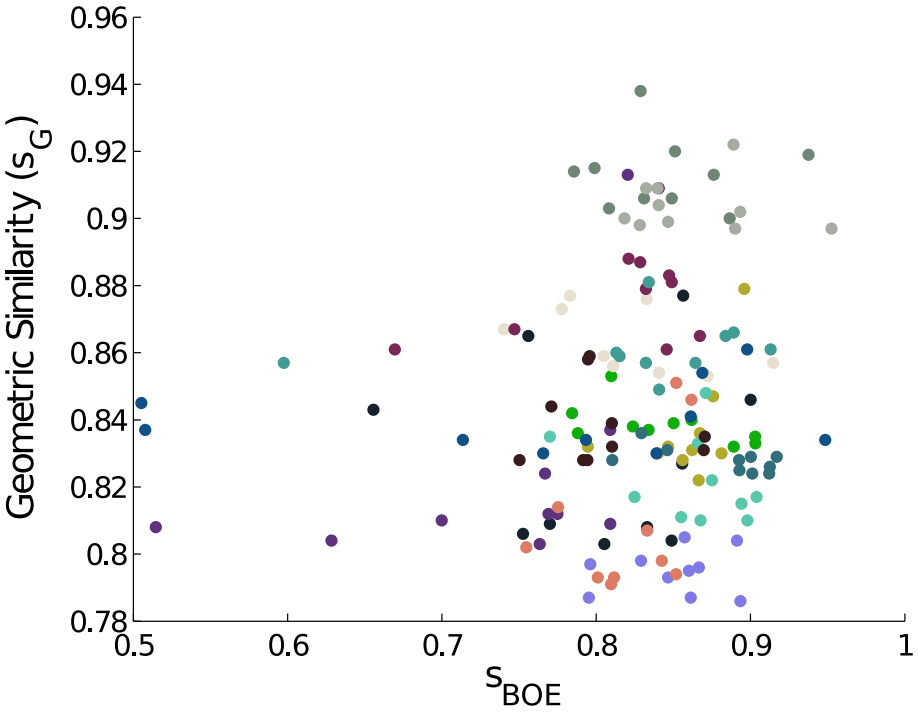

Figure 6.16: Geometric similarity index versus S_{BOE} for pairs of decoys and ligands. The pairs were selected so that their their geometric similarity is among the largest values. Colors stand for the different target proteins.

geometric and S_{BOE} similarities rank high. As it can be seen, pairs geometrically resemble each other even when comparing them between the two figures. In Figure 6.20, on the other hand, we show pairs with the highest S_{BOE} similarities. As it can bee seen, these configurations are very different from the configurations seen in Figures 6.18 and 6.19.

6.2.4 Discussion and Conclusions

In this section we introduced an alternative approach to assess similarities based on well-established computational topology algorithms. The measure was designed for assessing the similarity of different chemical structures but it may also be applicable in other fields. We proved that our definition is rigorous and it satisfies the mathematical requirements which are often neglected when new similarity measures are introduced.

Although the meaning of similarity is not clear-cut, being consistent in our choice is probably the most important principle to follow. It was easy to understand already based on our arguments that geometric similarity is not reliable and in certain cases it may fail. If we require consistency, mixing the values yielded by a given geometric similarity with other type of similarity measures is not viable. Therefore, we must construct similarity measures which, on the one hand, are proper measures, and, on the other hand, consider geometry, topology and other important factors at the same time. We believe that our method may be a good starting point for such an approach as we observed a logical path while welding geometry and topology and it is straightforwardly applicable when one is strictly interested in conformational similarities.

It is also important to form a good idea about the meaning of similarity. This is straightforward when it comes to geometry but it may not be so simple when one considers

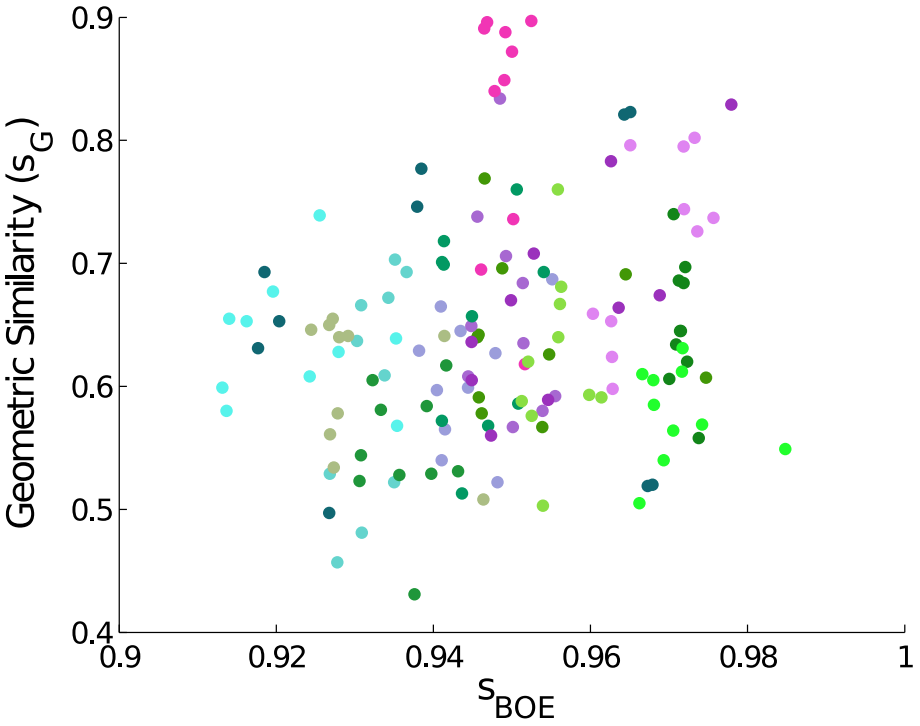

Figure 6.17: Geometric similarity versus S_{BOE} for pairs of decoys and ligands. The pairs were selected so that their their S_{BOE} similarity is among the largest values. Colors stand for the different target proteins.

other features. As for our method, we would like to emphasize again, that our aim was to elaborate a measure which considers similarity beyond geometric resemblance, looks at the number of rings and other topological features, takes into account all the scales, but it is not scale invariant, while sticking to a rigorous mathematical background. Of course, the method is easily extendible.

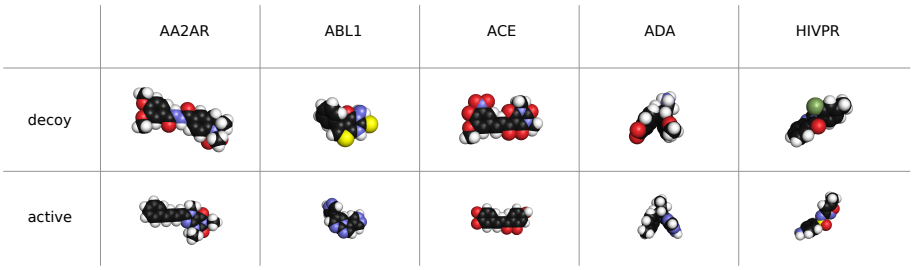

Figure 6.18: Decoys and actives with the highest geometric similarity values.

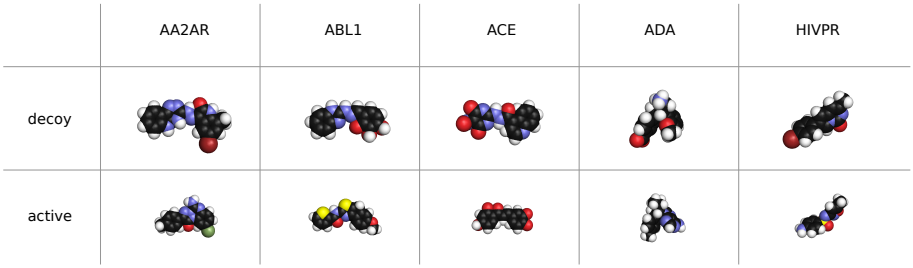

Figure 6.19: Pairs of decoys and ligands with high geometric similarity and high S_{BOE} similarity values.

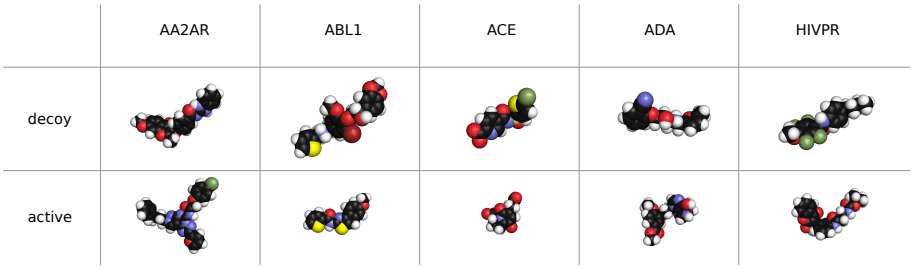

Figure 6.20: Decoys and actives with the highest S_{BOE} similarity values.

6.3 The Wasserstein Distance of Barcodes

References

This section is adapted from the manuscript submitted for publication,

- G. Máté and D.W. Heermann, *Statistical Comparison of Topological Features of Proteins* , PLoS One, under peer review (2013).

As a continuation of the previously presented two approaches for assessing similarity of proteins, we would like to introduce our last method. However, as this method is a bit different from the other two, being more a classification method, we would like to motivate our work and to place it in context by enumerating a few of the existing approaches.

First of all, the comparison methods can be divided into two broad classes: superposition and descriptor methods [254]. The former aim to calculate the best alignment of the molecules and quantify the similarities as some measure of the overlap while methods in the latter category describe the molecules with certain feature vectors and assess similarities by comparing the features, thus being independent from molecular orientation. Most methods in both categories treat molecules as rigid objects, however, in the last decade plenty of methods emerged which address flexibility, too.

Aligners first choose a scoring function which indicates the overlap of the compared molecules. Once the choice is made, the correspondence among molecules has to be found, which is an optimization problem. Thus, the crucial step here is the choice of the scoring function. Although there are few empirically parametrized models, the methods are based on "ad hoc" scoring functions [255]. Flexibility is usually treated in the mechanical sense, aligners define rigid substructures but allow movements at the joints of these [255].

Descriptor based methods seek to build a rotation and translation invariant signature to represent a molecule and use the signature to compare these molecules [256]. However, similarly to the scoring functions of the aligners, these signatures are mostly based on heuristic algorithms and they seldom have a rigorous mathematical motivation. Perhaps the methods with the most theoretical foundations are the ones based on graph theory [257]. Additionally, descriptors come short to deal with flexible molecules as addressing this issue in terms of signatures is still challenging [256].

Although, as noted above, many approaches exist, it is evident that they often lack mathematical rigor, especially when treating flexible molecules, despite the fact that a solid mathematical basis is required in order to ensure reliability. While methods, such as the ones relying on geometric comparison of molecules, may fail when handling flexible structures, these approaches possess a proper theoretical foundation. Geometric comparison methods, for instance, are usually based on volume overlaps, that is, set intersubsections in a mathematical terminology, and they perform extremely well on rigid bodies. To achieve a similar performance for flexible structures, it is indispensable to base the approaches on mathematics specifically developed for studying flexible manifolds, namely mathematical topology. Proper mathematical handling should be, of course, only the basis on which methods should build the knowledge from chemistry, physics and biology. This is especially the case since in recent years flexibility turned out to be a very important feature of many proteins [258] as it may influence binding affinity [250] and functionality [251]. Thus, it is crucial to minimize possible flaws and place methods addressing the comparison

of flexible molecules in a proper mathematical context.

We, again, approach the problem of comparing flexible structures from this perspective, applying computational topology algorithms and following a proper mathematical logic. Our approach is intended to demonstrate a basic comparison method relying on the calculation of certain topological properties of the molecules on different geometric scales. A given configuration of a protein is a representation of a topological space which is homeomorphic with all the possible foldings of the given protein. Based on this, we elaborate a method which enables a comparison which takes into account the possibly flexible nature of certain molecules. We do not intend to build in information regarding the chemical structure.

6.3.1 Topological Invariants

Our method relies on the calculation of the *persistence intervals of the Betti numbers* [239] for the investigated structures. Betti numbers [252] are the counts of different topological features like connected components (0th Betti number), holes (1st Betti number), voids (second Betti number) and their higher-dimensional generalizations. Of course, since real world structures are three-dimensional ones, we do not have to deal with these generalizations. The persistence intervals of these features denote the geometric scales on which the given features do not change. To have a better understanding of the concept, let us consider the following scenario. Let $S = \{(x, y, z)|x \in R, y \in R, z \in R\}$ be a point-set sample of an unknown O object embedded in the three-dimensional space, where R is the set of real numbers. Note that O could consist of multiple pieces (components). In order to calculate the Betti numbers, that is, count the components, the holes and the voids present in O based on the S sample, we have to reconstruct O from S. One could conceive different ways of reconstructing the object. Perhaps the most straightforward method is to connect each point with its nearest neighbors. We can define the nearest neighbors of the points by calculating the Delaunay triangulation [230] of S and discard the edges which are larger than an l_c cutoff length. This cutoff length can be defined as some fraction of the maximal edge-length in the triangulation, for instance. The remaining triangles are considered face elements and the tetrahedrons are treated as solid volume.

Components are relatively easy to count. Any two points from S which have a path between them along the edges of the triangulation (that is, they are connected) are in the same component. Two points connected by no path are in different components. Components thus can be counted by counting the subsets of S which are not connected to each other through the edges of the triangulation.

Counting holes and voids is a bit more difficult. In order to illustrate the problem, imagine a ball. It has a single component (everything is connected), no holes (otherwise the air would escape) and a single void (the space enclosed by the shell of the ball). A single perforation on the surface of the ball is not considered a hole. The reasoning behind this is the following: in theory, we could hold the ball membrane from the boundaries of the perforation and stretch it out until the membrane flattens out completely. Thus a ball with an opening on a membrane is homeomorphic to a plane without holes. If we puncture the shell again, we will have an object homeomorphic with a plane with a hole on it. Thus only the second hole on the surface of the ball is counted as a hole. Note also that as soon as we created the first perforation, the void disappears because of the homeomorphism with the plane. When counting holes and voids, one has to take into account these effects. For instance, only the triangle-faced polyhedrons create voids.

Considering these criteria, we can proceed with the calculation of the Betti numbers for

the object obtained through the reconstruction process based on S. If S is a good sample and is dense enough, that is, the distance between nearest neighbors in S is roughly uniform and is much smaller than the diameter of O, then S captures well the topology of O and the Betti numbers measured on S will be good descriptors of the topology of O.

However, if S is a sparse sample of O, the reconstruction procedure may yield an incorrect representation of O. To render the method more robust, instead of considering only one geometric scale defined by the fixed cutoff-length of the triangle edges, we consider more geometric scales by varying l_c from zero to infinity. We calculate the Betti numbers for each value of l_c and register it. Calculating the Betti numbers infinitely many times is not feasible of course, however, in practice, it is enough to consider the length of the longest edge in the triangulation as the upper bound for l_c.

In principle we could use any triangulation or any (even non-planar) graph defined on the S set. The Delaunay triangulation, however, is a good compromise between calculation complexity and memory efficiency when calculating the Betti numbers. Considering the complete graph on S is, in computational topology terms, equivalent to the construction of the Rips-complex [259], while the Delaunay construction is analogous to the calculation of the α-complex [260].

Given that a hole exists at a particular l_c cutoff, there is a largest $l'_c \leq l_c$ cutoff, for which the hole is not yet present in the reconstructed object and there is a smallest $l''_c \geq l_c$ for which the hole is filled in. The interval (l'_c, l''_c) is the persistence interval of the mentioned hole.

Betti numbers are topological invariants as their value is invariant under continuous deformation of the objects such as stretching or bending, for instance (tearing and gluing are not continuous deformations). Continuous deformations do not change the topology of the objects, thus Betti numbers are handy invariants when comparing different topologies.

6.3.2 A Graphical Representation of the Topology

There is a convenient way to represent the information gained through the scanning of S described above. Instead of simply counting the components, holes and voids, we construct a diagram for each of the Betti numbers. The horizontal axis of the diagram will correspond to the l_c cutoff. We represent each instance of components, holes and voids on the corresponding diagram with a horizontal bar. The starting-point of the bar corresponds to the cutoff value at which the instance was created while the end-point of the bar is the cutoff value at which the instance ceased to exist. The vertical ordering of the bars is arbitrary. This representation was developed by Carlsson and his collaborators (see, for instance, [239]. For a short review see [261]). In figure 6.21 we present such a plot for a particular point-set.

Carlsson's diagrams can be viewed as a fingerprint, a barcode of the structure. It encodes all the information regarding the Betti numbers on different scales. Betti numbers can be extracted by drawing a vertical line at any cutoff value and counting the numbers of the intersubsections with the bars of the diagram. Barcodes for components, holes and voids are also called dimension zero, dimension one and dimension two intervals/barcodes, respectively. These barcodes constitute the topological basis of our approach.

Dimension zero intervals are somewhat special. Since all the points exist for $l_c = 0$, and none of them are connected at this cutoff value, all the points are in different components. There will be as many zero dimension intervals as many points there are and all these intervals will have a starting point of zero. As we increase the l_c cutoff, points will start to be connected. Whenever two points from two different components are connected

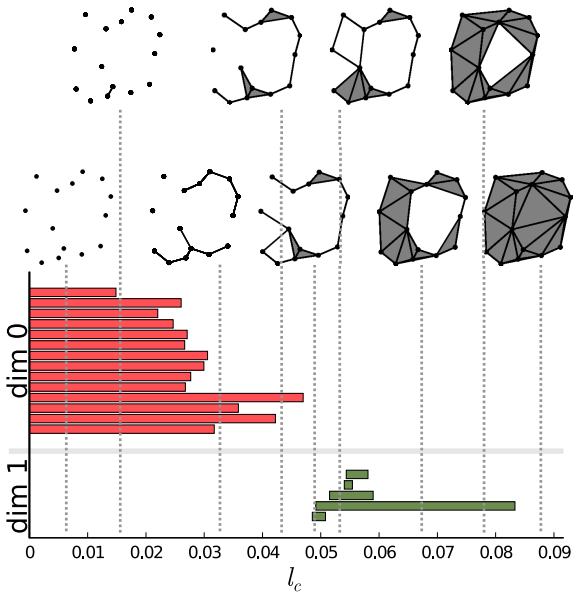

Figure 6.21: Persistence intervals for a particular set of points. Light gray bars represent the dimension 0 intervals, dark gray bars represent dimension 1 intervals. Note that for connected components there is an interval which closes at ∞. Since this bar is present for any nonempty point-set, it carries no information, therefore it can be removed from the representation. The process of connecting the points is also presented on the upper side of the figure for certain values of the l_c cutoff.

the components will be unified and the number of components is decreased, thus one of the intervals representing the just connected components is closed. Note that for any nonempty point-set there is always at least one component. Therefore, one of the intervals will always range from zero to infinity. Being always the same, it carries no information, thus it can be removed from the set of intervals.

The Distribution of Topological Features

Objects, in general, can be characterized by the size of their components, the way these are joined together and the size of the holes and voids that form during the building process. On the other hand, the end-points of the dimension zero bars have values statistically proportional to the spacing between sub-components of the system, while their number carry information regarding the size of the represented structure. The end-points of the dimension one bars have values statistically proportional to the diameters of the holes in the system. Similarly, end-points of the dimension two bars have values statistically proportional to the diameters of the voids. The dimension zero intervals always start at zero, thus it is only the end-point which matters in this case. The starting points of the dimension one and two intervals would mostly depend on the density of the points. In this sense, it is enough to describe the objects with the end-points of the intervals. Even more, we can replace the set of the end- points by the distribution of these, thus representing an object with a probability distribution. Then we can measure the similarity/dissimilarity between two objects as the similarity/dissimilarity between the representing distributions.

The Wasserstein Distance

There are a number of ways to compare two distributions. One can calculate any of the suitable f-divergence measures [262], for instance, the Kullback-Leibler divergence [263]. However, these measures are not necessary proper distances, in particular, they may not be symmetric or transitive. Another approach is to calculate the Wasserstein (or Vasershtein in the original spelling) distance (for a comprehensive review see [264] between the probability densities). The Wasserstein distance is a proper metric and can informally be introduced with a simple analogy: the distance is proportional to the physical work needed to transform a pile of earth shaped like one of the density functions to a pile shaped like the other density function. Based on this analogy, the Wasserstein distance is sometimes referred to as the earth movers distance (EMD). In fact, the Wasserstein distance is a class of distances parametrized with a $p \geq 1$ parameter in which the EMD corresponds to the 1st ($p = 1$) Wasserstein distance. As in the present work we only use the 1st Wasserstein distance, we may drop the notation regarding the parametrization or we will simply refer to it as EMD.

Given two probability density functions f and g, a more mathematical definition of the EMD can be given as

$$d_{EMD}(f, g) = \inf_{X \sim f, Y \sim g} E\left[d(X, Y)\right], \tag{6.3.46}$$

where $d(X, Y)$ is a distance function and in the simplest case is the absolute difference of the arguments, that is, $d(X, Y) = |X - Y|$.

6.3.3 Application of the Method

We measure dissimilarity between two chemical structures as indicated in the previous subsections. We treat a molecule as a point-set defined by the coordinates of its atoms. We calculate the persistence intervals and compute the distribution of the upper boundaries of the intervals. We proceed in this manner for each molecule we want to classify. Finally we calculate the Wasserstein distance among each pair of distributions, constructing thus a fully connected weighted graph of the molecule ensembles with the weights corresponding to the Wasserstein distances.

In order to classify the molecules, we simply need to cluster the obtained graph. For this purpose we apply the k-means algorithm [265].

We used different software to conduct our studies. We calculated the persistence intervals using the Dionysus software [266], we computed the Wasserstein distances with a code provided by the authors of [267], available for free online on their website. We carried out the clustering step with the built-in k-means algorithm of the MatLab's statistical toolbox, but, of course, any implementation of k-means is suitable.

The Ensemble Protein Database

For testing and demonstrative purposes, we apply our approach to a set of structures obtained from the Ensemble Protein Database (EPDB) [268]. We analyze five approximate ensembles constructed for the following proteins: Barstar (1A19), Calmodulin (1CFD), Ferredoxin-2 (1FXD), Alpha-Amylase inhibitor (1HOE) and Human CDC25B Catalytic Domain (1QB0). There are 191 configurations for Barstar, 196 for Calmodulin, 141 for Ferrodexin-2, 129 for the Alpha-Amylase inhibitor and 495 for the Human CDC25B Catalytic Domain.

Feeding the configurations to our method, without including any information about the origin of the conformations, we expect that the approach is able to distinguish between the different proteins. We will compare each protein configuration with every other configuration and calculate the Wasserstein distance for all of the pairs. It is convenient to display the results of the comparison in a color-coded matrix where each row and column corresponds to a protein. The ordering of the proteins in the rows and the columns are the same. Throughout the rest of the paper we apply the same ordering of the proteins in each figure, where the first 191 rows/columns represent the Barstar protein, the next 196 represent the Calmodulin, the next 141 contain results for the Ferrodexin-2, the following 129 represent the Alpha-Amylase inhibitor, while the last 495 rows/columns display results for the Human CDC25B Catalytic Domain.

Figure 6.22 presents the calculated Wasserstein distances for the dimension zero intervals. Looking at the figure, we see that the Wasserstein distances within certain groups are smaller than the inter-group distances and we actually can separate five groupings. In order to give an explicit grouping of the configurations by applying the k-means algorithm, we need to make sure that our guess of requesting $k = 5$ clusters based on the visual inspection of figure 6.22 is indeed a good choice. For this reason, we calculate clusterings for different cluster numbers, letting k run from 1 to 10. For each k value, we randomly select ten proteins from each class and we feed the fifty selected proteins to the k-means algorithm. In order to decide whether a clustering is good or not, we calculate the mean distance to the center for each cluster and characterize a clustering with the sum of these means. In mathematical terms, we define this sum as:

$$S(k) = \sum_{i=1}^{k} \langle ||Z_i - c_i|| \rangle_Z, \tag{6.3.47}$$

where Z_i represents the "coordinates" of a protein in the ith cluster with center c_i ($c_i = \langle Z_i \rangle_Z$) and $||.||$ is the euclidean norm. The coordinates are in fact the Wasserstein distances to all other proteins, that is, a row in the distance-matrix. We consider a clustering with k clusters good if the $S(k)$ sum is low.

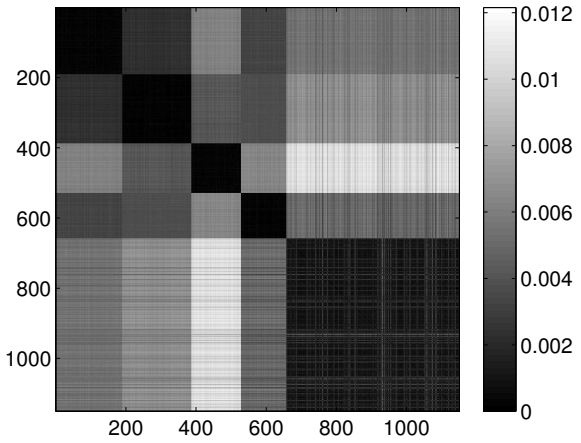

Figure 6.22: Wasserstein distances for the distributions representing the dimension 0 intervals. Each row and column corresponds to a protein-configuration, the ordering of the rows and the columns are the same. A darker shade means small distance while lighter shades imply larger distances. Note the dark blocks on the diagonal, they correspond to protein-groups which are close to each other.

To avoid problems caused by the probabilistic nature of k-means, we repeat the clustering many times for different samples, thus generating an ensemble of clusterings, and present the results averaged over this ensemble. In other words, the result of the clusterings are presented in a matrix form where each row and column corresponds to a protein and the matrix entry at the intersubsection of a given row and a given column is the probability of finding the two proteins corresponding to the row and the column in the same cluster, calculated based on the ensemble of the clusterings.

Figure 6.23 presents the $S(k)$ curve for the clustering of the dimension zero data. As it can be seen, the curve predicts that $k = 5$ or $k = 6$ gives us a relatively good clustering. If we look at the actual clusterings (shown for $k = 5$ in figure 6.24 and for $k = 6$ in figure 6.25) based on which $S(k)$ was calculated, we see that the clusterings for $k = 5$ and $k = 6$ are in fact equivalent. Setting k to 6 allows more flexibility for the clustering, than the $k = 5$ case, but it is clear that there are five groups, exactly corresponding to the different proteins. Thus, this clearly indicates, that the method is able to find the original groups.

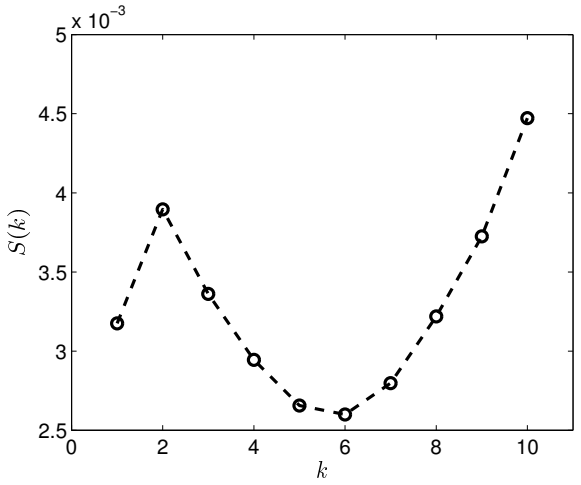

Figure 6.23: The $S(k)$ curve for the clustering of the dimension 0 data, calculated based on equation (6.3.47).

Clustering the entire dataset in five clusters gives the results presented in figure 6.26. We can clearly see the five groups of proteins, four smaller strongly coupled groups (corresponding to the proteins 1A19, 1CFD, 1FXD and 1HOE) and one larger group (corresponding to 1QB0). Members of the last group are not coupled as strongly as the members of the other groups but they always classify in the same way and do not mix with the other proteins. While there is some mixing in the first two and the fourth group, the core groups are clearly distinguishable.

Figure 6.27 presents the Wasserstein distances for the dimension one intervals. Looking at the figure it is obvious, that the dimension one intervals indicate two groups. Performing the same check as in the case of the dimension zero intervals, we see in figure 6.28 that the $S(k)$ measure also indicates that clustering the dataset into two clusters is a good choice in this case.

Performing the clustering for $k = 2$ we get the results shown in figure 6.29, which clearly gives two clusters, putting the first four groups of proteins (1A19, 1CFD, 1FXD, 1HOE) in the same class while the last group (1QB0) forms a different class. The explanation

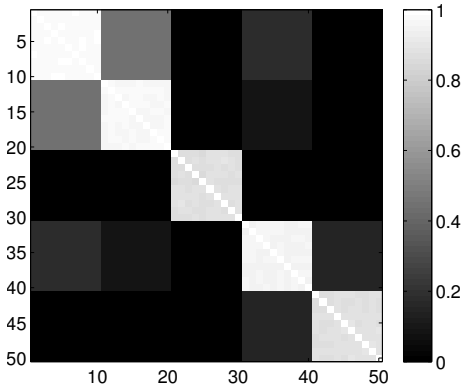

Figure 6.24: Probability of the event when two proteins are assigned to the same class when $k = 5$ classes are requested. The probability is calculated for each pair of proteins from different sub-samples of the ensemble. The proteins for a given position in a row/column were selected randomly from the ensemble with the constraint that they always belong to the same group. Probabilities are calculated by repeating the clustering multiple times and counting how many times the pairs were co-classified.

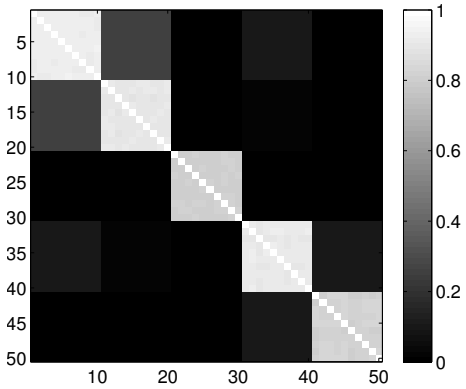

Figure 6.25: Probability of the event when two proteins are assigned to the same class when $k = 6$ classes are requested. The probability is calculated for each pair of proteins from different sub-samples of the ensemble. The proteins for a given position in a row/column were selected randomly from the ensemble with the constraint that they always belong to the same group. Probabilities are calculated by repeating the clustering multiple times and counting how many times the pairs were co-classified.

behind this result is that while the proteins corresponding to the first four groups are comparable in size (containing 89, 72, 58 and 74 residues, respectively), the last protein is much larger (177 residues). In fact the mixing of the first two and the fourth group we see in figure 6.26 is probably also a size-related effect as these groups are very close to each other in size, while the third group is a bit smaller. Nevertheless, it is now clear that by looking at the dimension one intervals, we in fact classify the proteins with respect to their sizes but we avoid calculating the geometric similarity which is a computationally very expensive procedure, as one needs to calculate the best overlaps among the structures.

For comparison, figure 6.30 gives the result for clustering the dimension one intervals into five clusters. As it can be seen, no additional clusters were found, just the probabilities for two proteins being in the same cluster decreased as the result of the non-optimal random

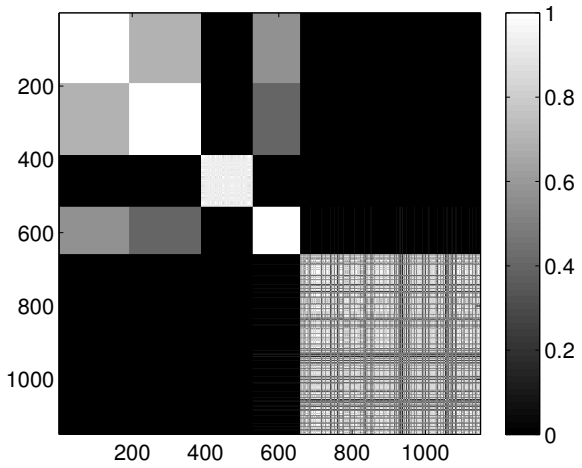

Figure 6.26: Probability of the event when two proteins are assigned to the same class when $k = 5$ classes are requested for the dimension 0 data. The probability is calculated for each pair of proteins by repeating the clustering multiple times and counting how many times the pairs were co-classified.

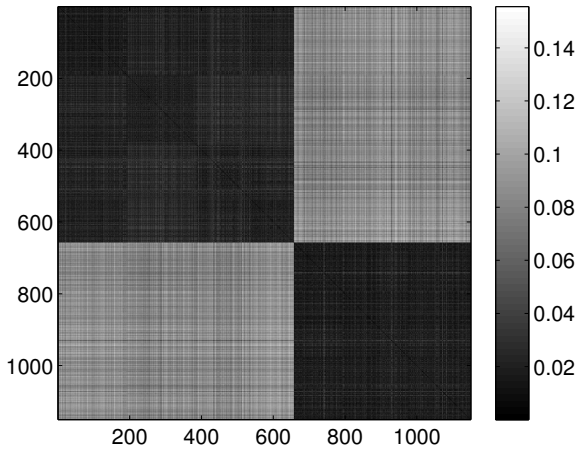

Figure 6.27: Wasserstein distances for the distributions representing the dimension 1 intervals. Each row and column corresponds to a protein-configuration, the ordering of the rows and the columns are the same. A darker shade means small distance while lighter shades impliy larger distances. Note the dark blocks on the diagonal, they correspond to protein-groups which are close to each other.

sub-grouping of the samples.

Last, figure 6.31 illustrates the Wasserstein distances for the dimension two intervals. Similarly to the distances for the dimension one intervals, we can distinguish two blocks, the first four groups in the first block and the last group of configurations in a separate block. However, groups two and three (1CFD, 1FXD) seem to have relatively reduced distances to group five (1QB0) perturbing a bit the block-structure. If we look at the corresponding $S(k)$ curve (figure 6.32), we see that it indicates a single cluster as the best solution, probably because of the coupling of the groups two and three to the fifth group. Still, the jump of $S(k)$ from $k = 1$ to $k = 2$ is less steep than the other increments, therefore we can consider a two-cluster structure. Performing the clusterings for $k = 2$

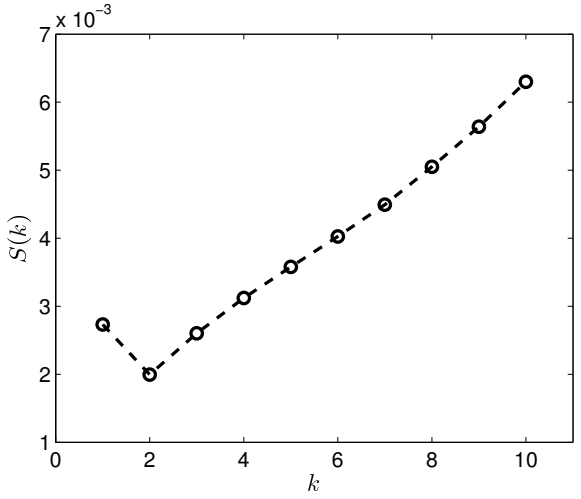

Figure 6.28: The $S(k)$ curve for the clustering of the dimension 1 data, calculated based on equation (6.3.47).

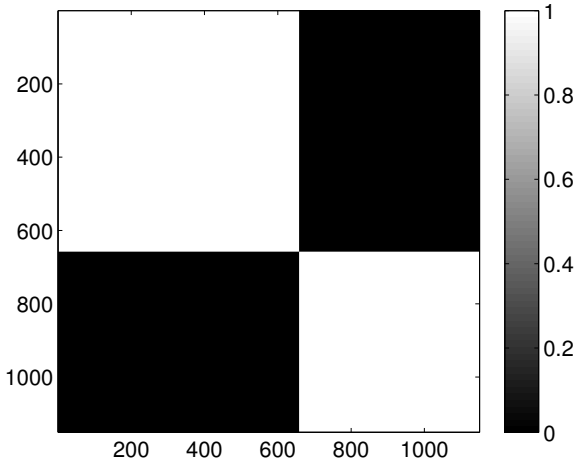

Figure 6.29: Probability of the event when two proteins are assigned to the same class when $k = 2$ classes are requested for the dimension 1 data. The probability is calculated for each pair of proteins by repeating the clustering multiple times and counting how many times the pairs were co-classified.

yields results presented in figure 6.33. Indeed, we find the two clusters which correspond to the clusters found in 6.29. However, if we try to find more clusters, as presented in figure 6.34, we see that there is an underlying structure of the clusters, which contains three clusters, groups one and four (1A19, 1HOE) corresponding to one cluster, groups two and three (1CFD, 1FXD) to a second one while the fifth group (1QB0) is again separated from the rest forming its own cluster. This suggests a clustering which is influenced by the geometric size and other topological factors.

6.3.4 Discussion and Conclusions

We described an approach for analyzing and grouping molecules from a purely mathematical point of view. However, based on our arguments and the presented example, it

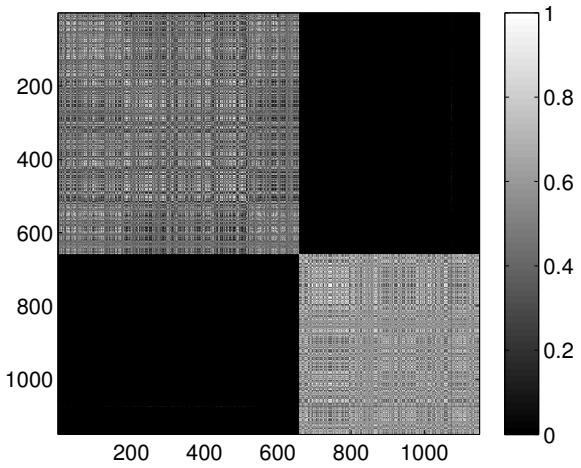

Figure 6.30: Probability of the event when two proteins are assigned to the same class when $k = 5$ classes are requested for the dimension 1 data. The probability is calculated for each pair of proteins by repeating the clustering multiple times and counting how many times the pairs were co-classified.

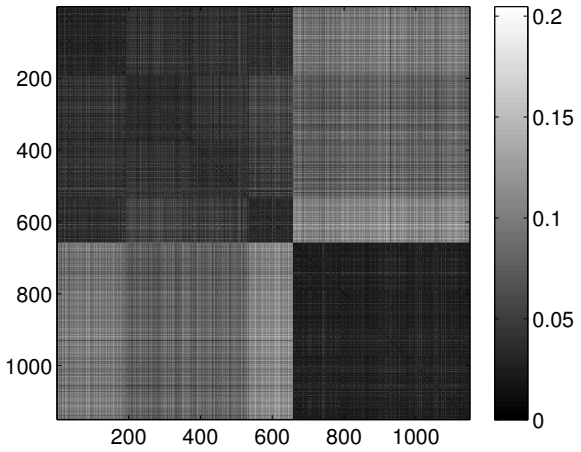

Figure 6.31: Wasserstein distances for the distributions representing the dimension 2 intervals. Each row and column corresponds to a protein-configuration, the ordering of the rows and the columns are the same. A darker shade means small distance while lighter shades implies larger distances. Not the dark blocks on the diagonal, they correspond to protein-groups which are close to each other.

is clear that this simple topological analysis has a much deeper meaning: it considers the topology and the geometry of the molecules within the same mathematical framework. In contrast to the currently available heuristic methods, our approach follows a nice and clear mathematical logic. It has a solid foundation, partially stemming from the field of computational topology and partially based on methods of image processing, the Wasserstein distance being a standard tool in this field. It has been proven in the literature that the distance is a real metric, thus applying a k-means algorithm to find the different, topologically related groups in a given set of proteins is straightforward.

The method can be summarized in a few simple steps: First we analyze the structures and check for the presence of topological features like components, holes and voids, using a

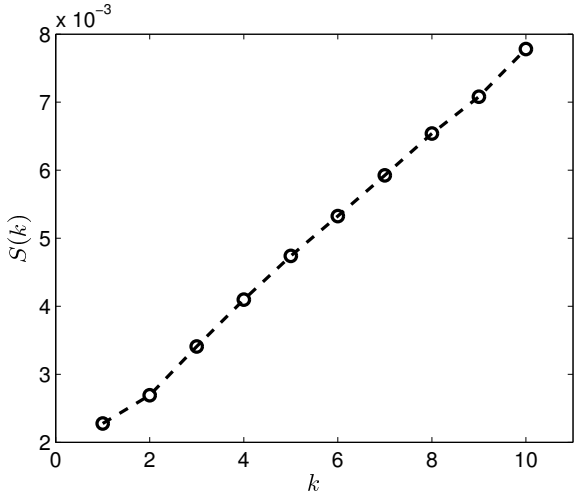

Figure 6.32: The $S(k)$ curve for the clustering of the dimension 2 data, calculated based on equation (6.3.47).

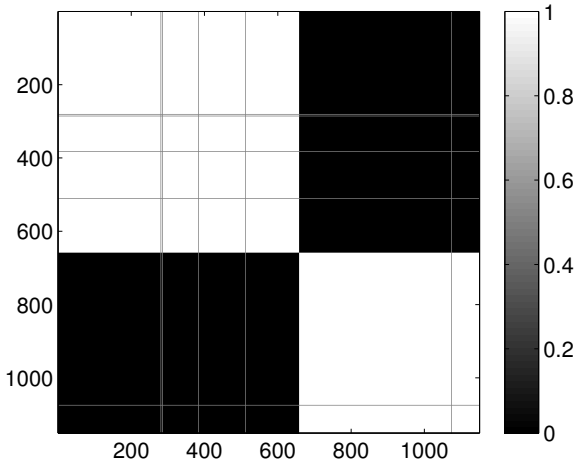

Figure 6.33: Probability of the event when two proteins are assigned to the same class when $k = 2$ classes are requested for the dimension 2 data. The probability is calculated for each pair of proteins by repeating the clustering multiple times and counting how many times the pairs were co-classified.

technique developed in computational topology for arbitrary point-clouds. Then we assess the similarity by statistically comparing the presence or absence of these features in the different molecules.

We presented a test-case, where these groupings are a priori known, however, this knowledge did not constitute an input to our analysis and it was used only for validating the results. Our method was able to reveal the different ensembles with a high precision. As it was demonstrated, a grouping which implies the geometry and size of the proteins is implicitly possible, without having to calculate best alignments.

We mention that the method can be tuned for different scopes by choosing the lower and the upper bounds of the l_c cutoff. For instance, we chose the largest edge in the triangulation as the biggest value for l_c. This leads to a coarse-graining procedure in

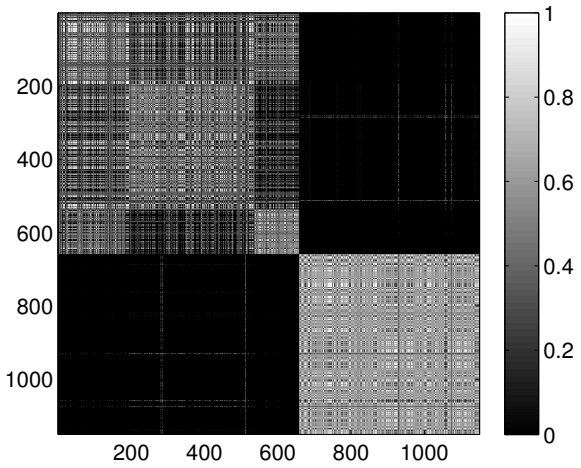

Figure 6.34: Probability of the event when two proteins are assigned to the same class when $k = 5$ classes are requested for the dimension 2 data. The probability is calculated for each pair of proteins by repeating the clustering multiple times and counting how many times the pairs were co-classified.

which, when reaching larger scales, the geometry of the molecule is also encoded. If one uses the largest bond-length as the biggest cutoff, one, in fact, will compare molecular topologies and completely discard the information hidden in the folding of the molecule.

We believe that our method can constitute the basis of a new approach or be a part of a framework which is able to deal with flexibility of chemical structures in terms of similarity and dissimilarity. As there is no unique and well-defined way to classify proteins, we are convinced that such approaches are needed to open up different perspectives for researchers working in the field.

6.4 What Else? Fractal Dimension!

References

This section is adapted from the manuscript submitted for publication,

- G. Máté and D.W. Heermann, *Persistence Intervals of Fractals* , Physica A, under peer review (2013).

Fractals surround us. They are present in nature [269], in the structure of our society [270, 271], in our economic systems [272, 273] and in all kinds of technologies we deal with day-by-day [274–276]. Their common property is their self-similar nature. Investigating them on different scales leads to the same conclusions.

One way to characterize these fractal structures is through the calculation of their dimension. For regular, exactly self-similar fractals this can be done analytically and there are various ways to estimate the dimension numerically for less regular cases. Perhaps the most wide-spread numerical method is the calculation of the Minkowski-Bouligand or box-count dimension [277]. Felix Hausdorff gave a rigorous definition of the dimension of an object [278] and it has been observed that these different definitions yield the same values, specially for fractals satisfying the open set condition [279]. By the dimension of an object, of course, we mean the "fractal dimension" and not the dimension of the space in which the object is embedded.

Fractals are complex structures, their investigation boomed with the development of high-speed computers and, in fact, the analysis of many of their aspects would be almost impossible without the aid of these machines. As computational power is becoming less and less an issue, methods which in the past seemed unfeasible might actually revolutionize research in different fields. Since storing large topological complexes in the memory of the computers is not a problem anymore, some of these methods are offered by computational topology.

In this paper we intend to take a look at the relation between the dimension of fractal-like objects and their topological fingerprints encoded as persistence intervals of Betti numbers. The latter concept is a relatively new development in computational topology [239].

6.4.1 Persistence Intervals

Topology can be investigated by calculating so-called topological invariants. Topological invariants fix certain topological features of the topological space under investigation. The invariants we will focus on are the number of connected components, the number of holes and the number of voids in the investigated object. Connected component in this context means parts which in some way are connected to each other. For example, a regular ball has a single connected component (since it is a single piece), no holes and a single void (inside). A piece of paper also has a single connected component, no holes and no voids. If we would tear the paper in two, the system composed from the two (now separated) pieces of paper would have two connected components (since the two pieces are not connected to each other anymore), no holes and no voids. If we would take a pencil and would poke a hole in one of the papers we would end up with a system with two connected components, one hole and no void. In the field of topology the mentioned topological invariants correspond

to the so-called *Betti numbers* [252]. The number of connected components is the 0th Betti number, the number of holes is the 1st Betti number while the number of voids is the 2nd Betti number. This enumeration, of course, can be extended with the number of 4, 5, 6, etc. dimensional holes in higher dimensional spaces corresponding to 3rd, 4th, 5th, etc. Betti number.

A Barcode Representation of Topology

Since we want to look at fractality, we are interested how the topology of the object is defined on different scales. Therefore we resort to the following abstraction: We will look to the number of holes on different scales. In order to achieve this, we will think about our objects as a dense (if it is the case) set of points and we adopt the following procedure:

- we calculate the Delaunay triangulation of the points
- we define a distance scale l
- we connect all point pairs that are connected by an edge in the triangulation and are closer then l
- we calculate the Betti numbers for the obtained structure

We repeat these steps for all possible values of l up to a l_{max} maximal value and we "record" the formed connected components, holes and voids for each value of l. l_{max} may be defined as the longest edge in the triangulation. This construction procedure is analogous with the building of the so-called "alpha-complexes" [260].

At this point it is a valid question what is counted as a hole and how can voids form when all we do is connecting points with lines. The answer lies in the definition of topological building blocks which are points, lines, triangles, tetrahedrons and their higher dimensional analogs. According to this, triangles do not count as holes, instead they constitute faces. Similarly, the space enclosed by a tetrahedron is not counted as void.

For understanding the definition of a hole, imagine again a ball. If we perforate the membrane of the ball, in theory we could stretch the membrane out to a sheet, therefore a ball with a perforation is homeomorphic to a plane, that is, a single hole on a surface of a ball is in fact not a hole. If we perforate the ball again, this object will be homeomorphic with a plane with a hole on it. One needs to take into account these effects when counting holes.

After we scan the system with the procedure described above we know the numbers of connected components, holes and voids for each value of l. The acquired information can be summarized in a diagram in the following way:

- each instance of connected component, hole and void will be represented by a bar on a different diagram
- the start point of each bar will correspond to the value of l at which the corresponding instance came into existence
- the end point of a bar will correspond to the value of l at which the corresponding instance ceased to exist

The bars for connected components are somewhat special as connected components unite as l increases. This process can be viewed as one of the connected components embeds the other one. Accordingly, the bar of the embedded component will end at the point where the component was embedded while the bar of the embedder component will continue until the letter will be embedded in another component. The role of embedded and embedder is arbitrary. It is easy to see that one of the bars for connected components will persist even at the highest values of l as there will always be at least one connected component, thus this bar can be neglected as it does not carry any information.

The diagram compiled in the previously described way will be a barcode-representation of the topology of the system in which each bar represents the interval of l over which the corresponding topological feature persists (persistence interval). An example for such a barcode can be seen on Figure 6.35. This representation was developed by Carlsson and his collaborators [239] and a very good review of their work can be found in [240]. There are more and more software available for calculating this representation. In the present study we used two of them: Perseus [280] and Dionysus [266].

Using the concept of the persistence intervals, the authors of [281] defined a *P.H. dimension* (persistent homology dimension) based on two variables: $(b+d)/2$ and $arcsec(d/b)$, where b is the start-point (or birth-point) of a given interval and d is the end-point (death-point) of the same interval. Here, in turn, we want to analyze the relation of the persistence intervals to the fractal dimension of objects, and are seeking a simple and intuitive explanation for this relation.

6.4.2 Calculating the Dimension from Persistence Intervals

If we assume that the topology is indeed encoded in the barcodes introduced in the previous subsection, it is natural to think that these barcodes should also represent fractality when it comes to objects presenting this property. Indeed, for certain special fractal objects, like the Sierpinski gasket [282] show in Figure 6.36, the fractal dimension d_F can directly be extracted from these barcodes. What is so special about the Sierpinski gasket is that it is exactly self-similar. This implies that the missing parts (the holes) of the fractal scale similarly as the volume of the fractal. This makes the calculation of the dimension easy and the following logic applies to any fractal which has the mentioned property.

Starting from the simple scaling relation between the volume V of a regular solid object and its characteristic diameter l, given as $V \sim l^d$, known from regular geometric spaces, d being the dimension of the space, it can be easily shown that a generic fractal dimension can be given as

$$d_F = \frac{\ln[m]}{\ln[s]}, \tag{6.4.48}$$

where s is the scale constriction factor and m is the multiplicity, that is, the number of the self-similar copies on the smaller scale. For instance, for the Sierpinksi gasket $m = 3$ and $s = 2$ (the gasket is composed of three copies of itself scaled so that edges on the small-scale are half of the edges of the original), therefore, the fractal dimension of the gasket is $\ln 3/\ln 2$ (roughly 1.585).

In general, to approximate the fractal dimension from the barcodes, let us consider the natural logarithm of the number of the bars for holes of a given scale as a function of the natural logarithm of the endpoint of these bars. Note that the value of the end-point of a bar is proportional to the diameter of the corresponding hole. This function, in fact, characterizes how the object is scaling. To understand the reasoning behind this, let us now concentrate on regular two dimensional fractals, noting that the logic generalizes to

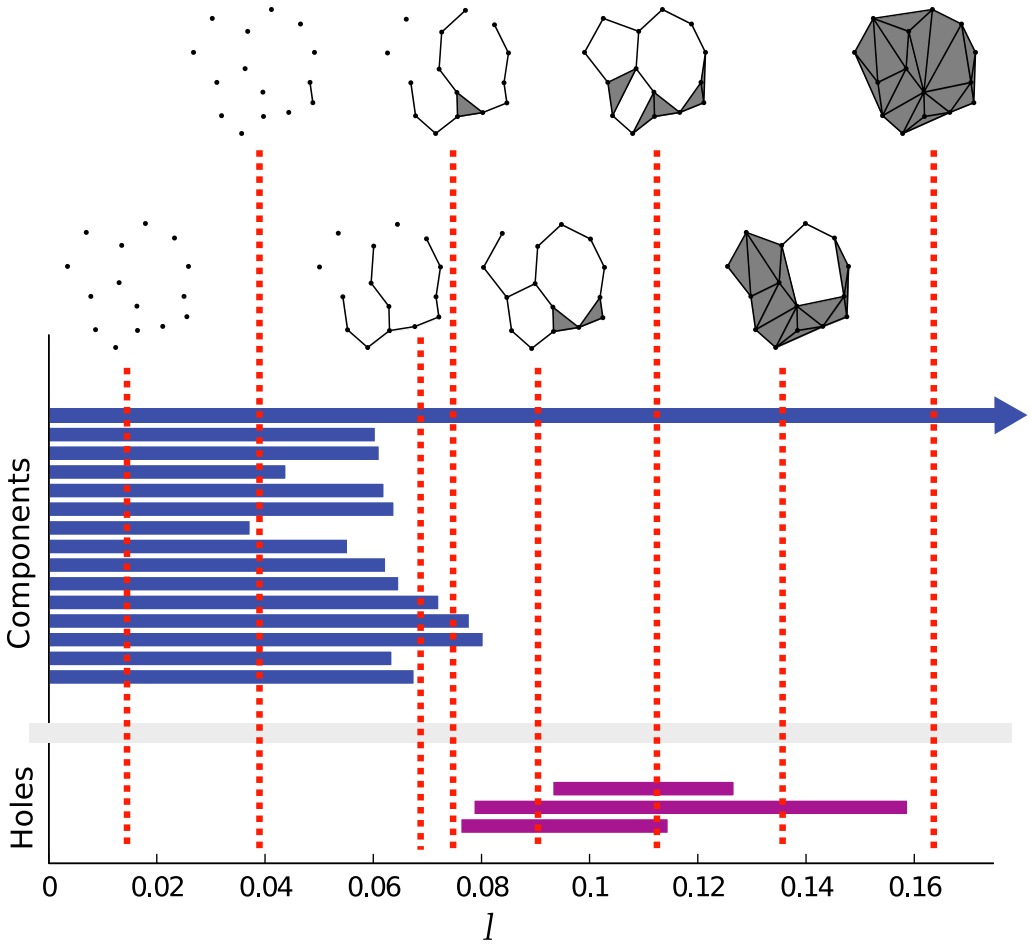

Figure 6.35: Barcodes for a particular set of points. The process of connecting points is also illustrated in the upper part of the figure. Blue bars represent persistence intervals of the 0th Betti number while purple bars represent intervals of the 1st Betti number. Note that for any non-empty set of points there is an infinitely long bar for the components as there is always a single component when everything is connected, however this bar carries no information, it can even be removed from the diagram.

any dimension. As we pointed out in the introduction we will treat these objects as dense sets of points. Since they are regular self similar structures, holes on a given scale will be the same in any particular instance of this class. This would mean, that in the barcode plot we would have as many bars of equal length as the number of the holes on that scale. Let us denote by $n(\varepsilon)$ the number of the bars which have a length of ε. On the other hand, again because of self similarity, both ε and $n(\varepsilon)$ changes according to a power law when going from one scale to the other. For instance, in the case of the Sierpinski gasket, ε on a given scale is always the half (as $s = 2$) of its value on the next larger scale, while $n(\varepsilon)$ is always tripled ($m = 3$). This, in general, means that the logarithms of these quantities define a linear function which can be written as

$$\ln [n(\varepsilon)] = \alpha \ln [\varepsilon] + c, \tag{6.4.49}$$

where α is the slope and c is a constant. In Figure 6.37 we plotted this relationship for the Sierpinski gasket.

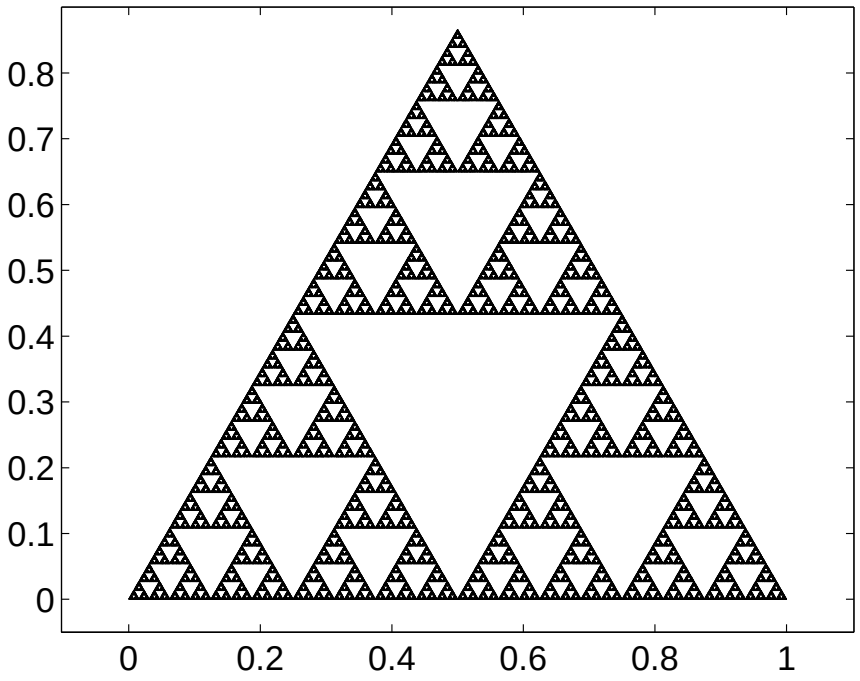

Figure 6.36: The Sierpinski gasket is built by recursively replacing an equilateral triangle with three copies of itself having an edge-length of half of the original triangle's edge-length.

Consider now two different scales ε_i and ε_j so that $\varepsilon_i < \varepsilon_j$. This implies that $n(\varepsilon_i) > n(\varepsilon_j)$ as smaller scales are multiple copies of larger scales. In this case we can write the slope as

$$\alpha = \frac{\ln\left[n(\varepsilon_i)\right] - \ln\left[n(\varepsilon_j)\right]}{\ln\left[\varepsilon_i\right] - \ln\left[\varepsilon_j\right]} = -\frac{\ln\left[n(\varepsilon_j)/n(\varepsilon_i)\right]}{\ln\left[\varepsilon_i/\varepsilon_j\right]}. \tag{6.4.50}$$

Note now that the arguments of the logarithms are in fact some power of the multiplicity ($m^x = n(\varepsilon_j)/n(\varepsilon_i)$) and the same power of the scaling factor ($s^x = \varepsilon_i/\varepsilon_j$). The exponent x depends on the steps between the scales ε_i and ε_j. In fact, $x = j - i$. Replacing the arguments in (6.4.50), we get the following form:

$$\alpha = -\frac{\ln\left[m^x\right]}{\ln\left[s^x\right]} = -\frac{\ln\left[m\right]}{\ln\left[s\right]}, \tag{6.4.51}$$

which is nothing else but the fractal dimension defined in equation (6.4.48), that is, $d_F = -\alpha$. Therefore by fitting (6.4.49), we in fact are estimating the fractal dimension of the system under investigation. Indeed, the fit on Figure 6.37 gives and estimate of $d_F = 1.579$ for the dimension of the Sierpinski gasket, which is very close to the 1.585 (rounded) theoretical value.

Here we note, that it is possible to estimate how consistent the scaling is (to which extent does the object behave like a fractal) by calculating the goodness of the fit.

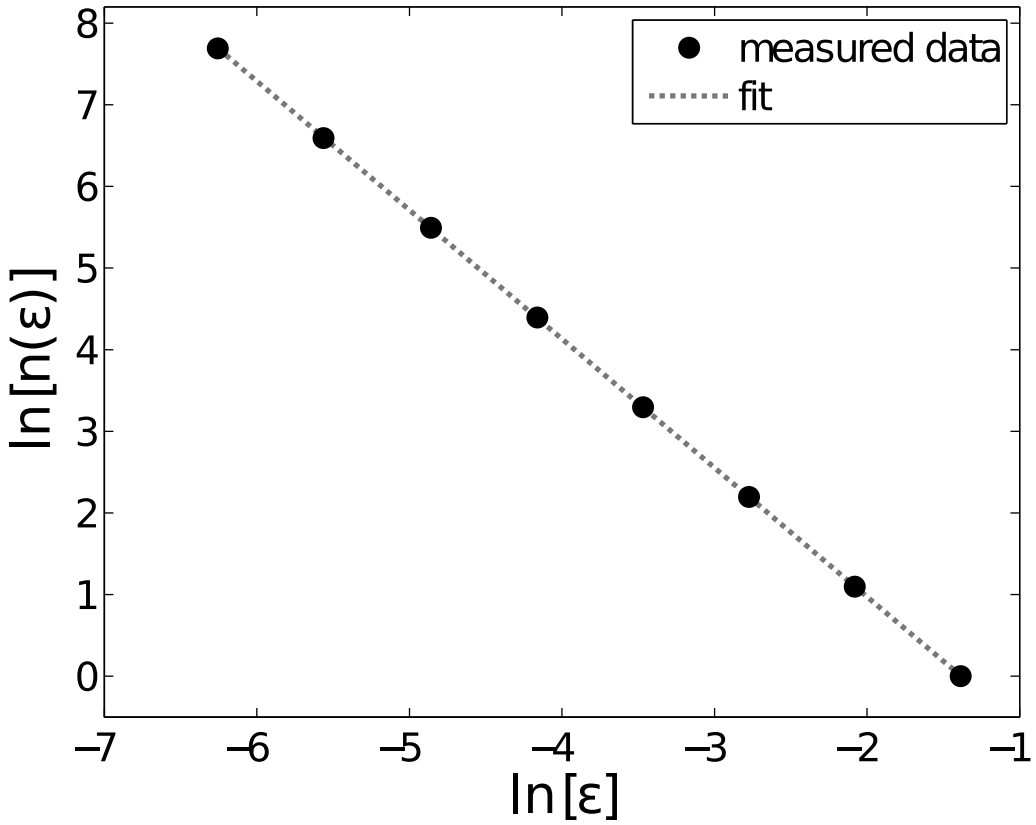

Figure 6.37: The plot presents the natural logarithm of the number of bars of a given length as a function of the natural logarithm of the length for the Sierpinski gasket. The dashed gray line is a linear fit to the points.

6.4.3 Application

One may, of course, ask whether an approach based on the persistence intervals is useful or not. Therefore, we demonstrate the potential of the method with another simple example. Applying the proposed method enables the investigation of scaling in different lower, "non-native" dimensions and this is what we seek to illustrate. Let us consider now the object on Figure 6.38. The object is constructed of short segments, embedded in the two dimensional space, arranged in a structure resembling concentric circles, also suggesting a given degree of randomness.

Although the object is embedded in the two dimensional space, it is composed of short, disconnected segments. Therefore, instead of looking at holes in the object, we will analyze the persistence intervals of the 0th Betti number, which describe how components relate to each other.

Since the object is at least seemingly random, our approach does not apply directly. As Figure 6.39 shows, trying to plot the relation between ε and $n(\varepsilon)$ does not give us a clear indication that there is any scaling in the object. However, we notice that there is a certain grouping around certain values of ε.

The spread around given values of ε may and actually does stem from the randomness present in the object. Taking into account that the smallest scales may be influenced by overlapping because of the perceived randomness and large scales are distorted because of the finite size of the object, we will consider ε values only for the middle, well-pronounced

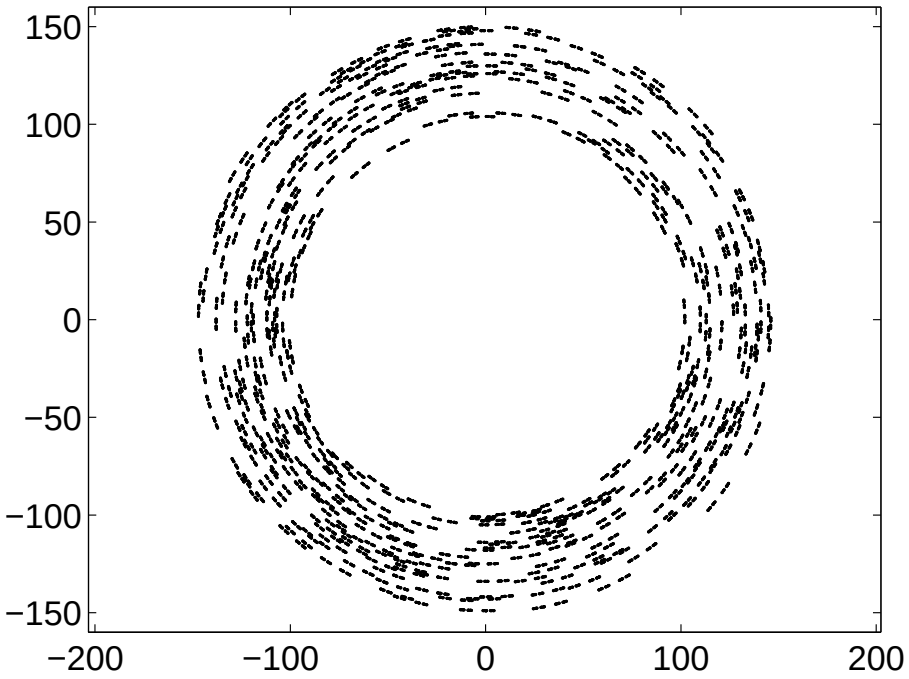

Figure 6.38: A two-dimensional object, composed of a set of short segments arranged in a circular structure.

six peaks of the histogram of ε presented also in Figure 6.39.

Summing up $n(\varepsilon)$ under the different peaks separately and plotting the sums against the mean of the peaks (Figure 6.40) gives us a familiar picture. The slope of the fit corresponds to a fractal dimension of 0.6292 which is very close to the $\ln 2/\ln 3$ dimension of the Cantor set [283] ($\ln 2/\ln 3 \simeq 0.6309$). And indeed, the object is in fact composed of fifty copies of a Cantor set "bent" to a semi circle, as shown on Figure 6.41, with radii ranging 100 to 150 units and rotated with angles randomly drawn form the interval $[-\pi, \pi]$.

This is just a simple illustrative example which demonstrates the applicability and the usefulness of the approach. There are, of course, plenty of situations where these kind of calculations may come in handy. For instance, imagine a hyper-sphere whose very edge is decorated with holes sized according to a power-law. Realizing that the boundary of the object exhibits fractal-like behavior might be extremely difficult because of the high-dimensionality of the space.

6.4.4 Discussion and Conclusions

We introduced a novel method for investigating fractality. The approach is based on calculating Betti numbers, a class of useful topological invariants, on different scales. We showed that analyzing how the Betti numbers change as the scale changes reveals information regarding the fractal nature of the investigated objects. The method is particularly useful when self-similarity is a characteristic "hidden" in a lower dimension of the object, like in the presented example.

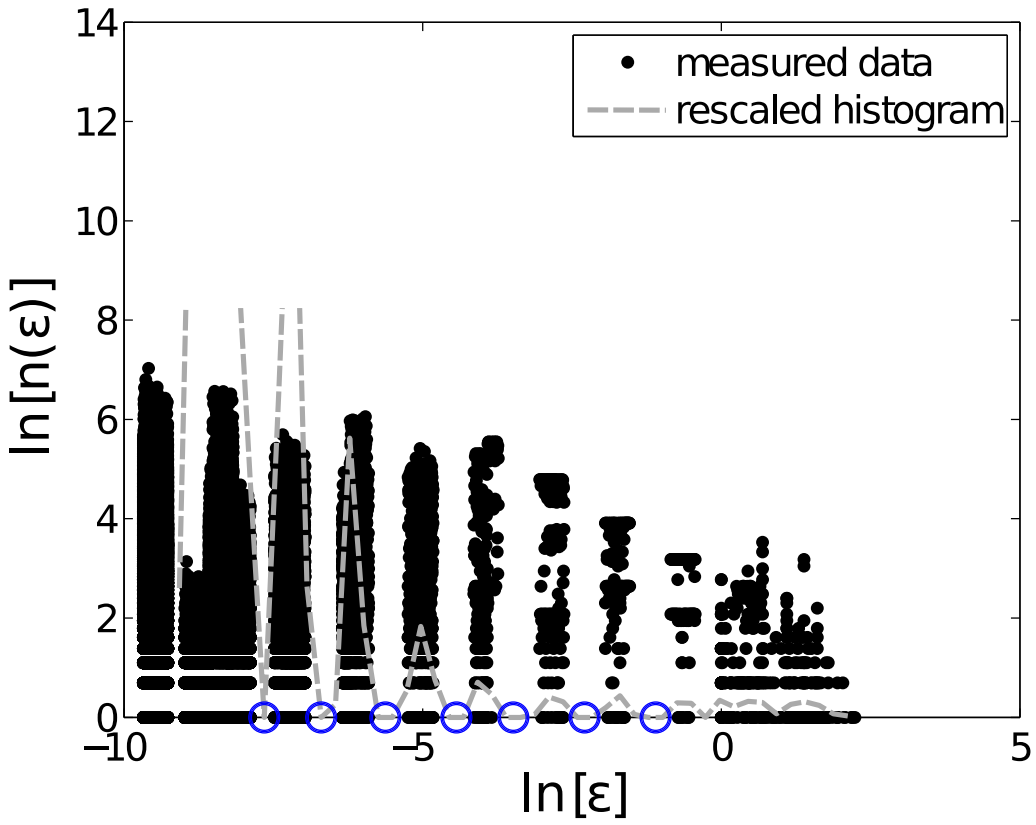

Figure 6.39: The plot presents the natural logarithm of the number of bars the 0th Betti number of a given length as a function of the natural logarithm of their length for the object presented in Figure 6.38. The dashed gray line is a scaled histogram of $\ln(\varepsilon)$, presented with the scope to show that points group around certain values of $\ln(\varepsilon)$. The circles along the horizontal axis indicate positions of local minima of the histogram. These minima will serve as values of ε according to which we sort certain $n(\varepsilon)$ values to one peak or another.

Although the fractal dimension of the objects dimension can be estimated only for a given class of fractals in which holes scale as the object itself, the method is able to reveal scaling on any type of objects, that is, it is able to estimate the fractal dimension of the holes for any object. Moreover, possessing knowledge about the scaling of the volume of the convex hull might enable us to estimate the fractal dimension of any object, however, this needs further investigation.

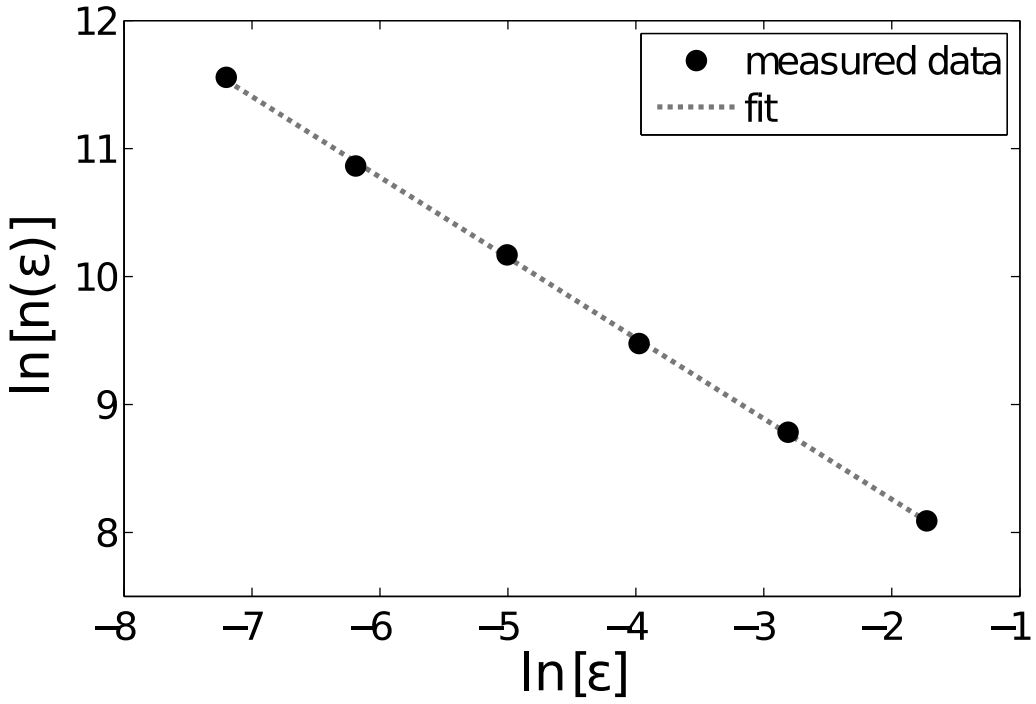

Figure 6.40: Data calculated by averaging data under the different peaks of Figure 6.39. The dashed gray line is a liner fit to the logarithms.

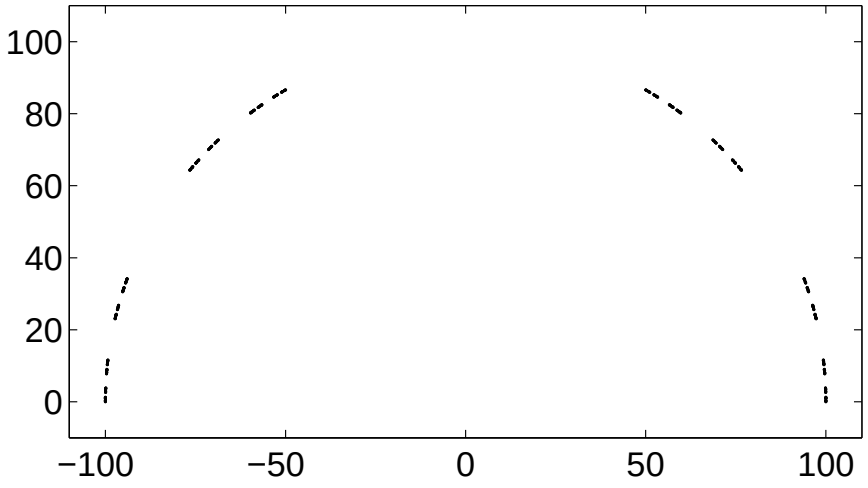

Figure 6.41: A Cantor set "bent" to a semicircle.

Chapter 7

Summary

7.1 Main Results

The topics and the results presented in this thesis are gathered around two main concepts: (i) modelling microscopy images and (ii) using persistence diagrams as fingerprints of topology. While they are somewhat unrelated topics, they meet in a common denominator worded by the scope of this thesis, which is to develop theoretical and computational ideas useful in analysing the spatial organization of the chromatin fibre inside the nucleus eukaryote cells.

We described microscopy images by characteristic distributions. In some cases, these distributions featured easily observed quantities like distances or local densities. In other situations, they represented the image as a whole.

Markov random fields are notably useful in the latter context. They often specialize in representing very complex structures by means of local interactions. This is very suggestive behaviour, as in physics we usually have a distance-dependent coupling among components of systems. In the simplest cases we may resume to not more than the effect of excluded volumes. It is not a surprise that Markov random fields show resemblance with physical systems as, in fact, they emerged from interdisciplinary applications of statistical mechanics models.

We used the inertia provided by generic tools developed in Markov random fields and modelled confocal microscopy images by learning the parameters of a generalized Potts model from image data. Our generalisation consisted of the introduction of interactions between different spin states. This allows a smooth spatial transition between different states and overcomes the formation of similar state domains which is characteristic for ferromagnetic model. Since we wanted to represent different levels of chromatin condensation with different states of the spin variables, observing a smooth transition was a desired behaviour. While the spatial density distribution of the nucleus is rather inhomogeneous, it does not change abruptly. Our model proved to be able to capture the proper distributions.

Besides being useful in modelling the images, and thus being able to generate artificial samples with desired properties, the generalization of the Potts model proved to be very

interesting from a statistical mechanics point of view. We extensively studied the thermodynamic properties of the model both by analytic calculations and computer simulations. Our analytical predictions were supported and extended by the results of the simulations. Although we constrained the parameter space, we found that the model is capable to exhibit different interesting phases. We showed that while for certain combinations of the parameters the model functioned like a ferromagnet, for other values it behaved like an antiferromagnet with highly degenerate ground state. Intermediate phases were also found. Furthermore, we demonstrated, that the transition temperature is shifting, depending on the parameters of the model. It is worth to mention that the generalization contains models like the original Potts model or the XY model as special cases.

Little bit towards the more traditional side of image processing, we characterized images by the distribution of the intensities and the different momenta of this distribution. We argued that assuming stoichiometry, the intensity levels should be proportional to the amount of chromatin in the case of uniform labelling of the fibre. While the aim was to quantify differences between samples of plant cells grown under different environmental factors, we argued that observed differences might also stem from different experimental conditions like slight deviations in microscope gain. This might be especially the case if images are thresholded and the resulted binary images is subjected to cluster analysis. We showed how to correct for these differences and motivated our approach by probability theory arguments. As for the result of the analysis itself, we were able to detect significant differences even after corrections were applied.

Following the thread of image analysis, we investigated the effects of ionizing radiation over the spatial organization of the chromatin. As a result of the irradiation, the DNA fibre may be damaged and in the most serious cases both strands of the DNA may break. This is known to result in a spatial reorganization of the chromatin in order to give access to the repair mechanisms.

Super resolution images of non irradiated and irradiated HeLa cells were obtained by localization microscopy techniques. This type of microscopy provides the data in form of two dimensional points marking the position of projections of different chromatin labels. Samples irradiated with different doses were analysed by means of graph theoretical approaches. A network of localized points were built by connecting each point with its nearest neighbours defined by the Delaunay triangulation. Quantities characterizing the spatial structure of the constructed network were determined. Comparing the quantities calculated for non-irradiated and irradiated samples, we found that although overall the chromatin may compactify to a certain degree, the local neighbouring properties of the localized points do not change. For instance, the distribution of the edge-lengths belong to the same family for irradiated and non irradiated samples. At the same time, we found that heterochromatic regions, marked separately, open up open irradiation. Furthermore, assuring a long enough healing time after irradiation, we observe a recover of the heterochromatin from the mentioned opened state. Our findings are in agreement with other experiments observing structural changes caused by the presence of the double strand breaks.

A similar graph theoretical analysis was performed to assess spatial relationships between nuclear pore complexes and lamin B receptor proteins. In a preprocessing step of the images, obtained by Structured Illumination Microscopy, positions of the pore complexes and the receptors were determined separately. Since the pore complexes and the receptors are located in and along the nuclear envelope, we calculated the radial distribution along the surface mesh defined by the positions of the pores and receptor proteins. We observed

a short-range, fluid-like ordering between the pore-pore or receptor-receptor pairs. This order disappears for pore-receptor pairs. Furthermore, we found that while the position of pores are seemingly uncorrelated with the position of receptors, lamin B proteins prefer to be located in the spatial proximity of the pores. This observation can be explained by the reduced mobility of pore complexes as they are embedded in the nuclear envelope.

On the other front, we laid the basis of assessing similarity of molecules based on calculating a set of topological invariants known as persistence intervals. These intervals indicate geometric scales over which certain topological features, like holes and voids, persist. We presented three different approaches. They differ not only in the way they compare the persistence diagrams, but also by the topological space representation of the molecules. The first and the second approach uses the same object representation, the Vietoris-Rips complex, while in the third method we represented objects by alpha complexes.

Our first approach defined the similarity of proteins based on the Haussdorff distance between the sets of persistence intervals. We showed that while the method correlates with the geometric similarity, calculated as the Jaccard measure of the volumes, high similarity values were assigned not only to geometrically similar pairs but also to rather different geometric configurations of the same protein. Nevertheless, the Haussdorff distance represents a "worst-case" measure and it is probably not the most appealing choice. Furthermore, it is a distance, not a similarity, and its difficult to interpret the meaning of the value of a given distance.

Our second approach is an improvement over the first one. Here we defined the similarity as the average of the best Jaccard indexes. We rigorously proved that this average is a similarity relation in its mathematical sense, that is, a value close to one means high similarity, while a value closer to zero indicates a reduced similarity. Applying this measure on real world data, we saw that while it still correlates with the geometric similarity, which is an expected behaviour, it is able to match not only different foldings of the same molecule but also structures composed of similar molecular motifs.

We approach the problem from a different angle in our last method. Here we represent the proteins by the distribution of the end-points of the persistence intervals. Then we define the similarity as the Wasserstein distance of the distributions. We apply the approach on an ensemble of proteins, calculating the distance between all of the pairs. By clustering the fully connected graph of proteins, we are able to classify the different instances with a high precision.

We remarked that devising a similarity between chemical structures based on persistence diagrams assumes that the diagrams indeed represent the topology. We supported this assumption by showing that there a simple and intuitive relation between the structure of the diagrams and the fractal dimension of certain objects. We were able to numerically estimate fractal dimensions of the Sierpinski gasket and a specific Cantor set using this relation with high accuracy.

7.2 Outlook

Besides the many unexplored thoughts regarding possible approaches to certain problems, sometimes independent from the content of this thesis, the developing of the presented methods and models generated plenty of interesting ideas. Unfortunately, time is not limitless, or at least not for Ph.D. projects. Nevertheless, there are a few aspects of the presented works that worth to be further examined.

One of them is to extend the analysis of the generalized Potts model. There are many approaches documented in the literature which might help us gain a deeper insight in the behaviour of the model. Devising an alternative Landau theory, performing finite size scaling analysis or applying renormalization [284–286] are just some of the possible approaches which may help us understand the nature of the occurring phase transitions. We can define alternative order parameters to capture the essence of the mixed phases. We should investigate if frustration could arise for certain combinations of model parameters. If so, than the system needs special treatment in this cases. Extending the parameter space might be another possible way to continue. While recently we started performing a finite size scaling analysis, other mentioned approaches should also be carried out.

The studied version of the model brakes the symmetry with respect to the exchange of the spins which are in the states represented by the first and the last rows of the coupling matrix J with any other state. Correcting the lack of this symmetry would render the $q = 3$ case equivalent with the original Potts model. However, for $q > 3$ the number of the degenerated ground states of the antiferromagnetic phase would increase. Without the symmetry, the switching from the first state to the last state (or the other way around) is less likely then switching from the first state to the second state or from the last state to the next to last state, in case if $b > 0$. This behaviour, for instance, blocks the model to form sharp boundaries between very dark pixels and very bright ones, when consecutive states correspond to image intensities. While mean-field treatment of the model is more difficult because of the missing symmetry, the implication of its presence should be studied if introduced.

When we look at the evolution of the Monte-Carlo sampling, consecutive samples create the impression of the existence of a "flow" in the systems, especially when neighbouring states are represented by a colour code, for instance grey values, such that strongly coupled states have the smallest differences in shades. Even though Monte-Carlo dynamics are not physical, such flows and other motifs may arise as a result of local interactions in a similar manner as waves and other patterns form in cellular neural networks [287–289]. Investigating this patterns may lead to exciting findings and may shed light on further potential applications of the model. We also mention here that a quasi-physical implementation could be studied on a *CNN computer* [290, 291].

Other applications may be found in more interdisciplinary topics. Although the original Potts model is established in fields like economy, sociology or other social sciences [177, 292, 293], critics often emphasise the lack of the connection between the model and the behaviour of real world systems. We believe that the presented generalization has several advantages over the original model in such applications. Most importantly, it is much more flexible and given enough data, parameters can be properly learned, thus ensuring a better suitability for modelling complex behaviour. Another possibility is offered by evolutionary game theory of spatially structured populations where individuals of a community interact with their neighbours to develop optimal strategies in the posed situations [294].

Regarding the image processing side of the model, there are a few aspects to mention. The first one considers the learning of the parameters. While, according to our tests, learning converges and gives rather good estimates, it is also relatively slow. The wide variety of available learning procedures offers us some flexibility and other learning approaches should be tested.

On the other hand, it is important to mention that obviously, there is a connection between the density distribution seen on microscopy and the arrangement of the chromatin fibre. Therefore, we should also be able to find a correspondence between our model with

parameters learned from data and polymer models which are able to generate ensembles reproducing spatial proximity maps of genes seen in experiments like Hi-C [295]. However, this first needs a careful investigation of the generic relation of polymer and lattice models.

The analysis of the localization data should also be extended. Presently, we are engaging in applying topological approaches to detect repeating motifs of point sets. On the other hand, we also analysed the Voronoi tessellation defined by the localized points. Correlations of the area and characteristic diameter of the cells were calculated for pairs of cells separated by a given distance. No significant linear correlation was found between cells other than nearest neighbours. However, there might be a higher order relation, thus the problem should be further examined.

Last, we would like to mention that the similarity measures presented in this thesis can also be enhanced. Chemical information, for instance, can be built in the calculations. This could be done, for example, by replacing the used topological representation with a complex which is built strictly on the frame defined by the bond-connections between atoms.

While this list could be continued, the suggested possibilities and challenges are numerous enough to provide an accurate picture about what else can, and hopefully will be done in the context of the presented methods. The author of this thesis had a great time working on each of the projects and is highly motivated to carry on with them.

Acknowledgments

I would like to thank my doctoral advisor, Prof. Dieter W. Heermann for guiding me through my Doctoral studies, offering constant support and sharing his knowledge.

I am grateful for all the exciting discussions I had with Prof. Michael Hausmann, Prof. Ronald Dickman and Prof. Christoph Schnörr.

I owe many thanks to Dr. Mariliis Tark-Dame and to Dr. Lindsay Shopland for their kindness for explaining biology and sharing their experimental data.

I would like to thank Prof. Michael Hausmann, Prof. Christoph Schnörr and Prof. Norbert Herrmann for accepting to be members of my examination committee.

I am grateful to Yang Zhang, Emőke-Ágnes Horvát and Miriam Fritsche for taking the time to proofread parts of this thesis.

I would like to thank Miriam Fritsche, Benoît Knecht, Yang Zhang, Hansjörg "Hansi" Jerabek, Songling Li, Lei Liu, Wei Xiong, Fei Xing, Jörg Eisele, Andreas Hofmann, Christina Schindler, Christoph Feinauer and, our newest group member, Min Chu, for all the cheerful moments and interesting discussions. I certainly learned much from them.

I would also like to thank the members of the RTG 1653 for lifting up the spirits during seminar breaks.

I especially would like to thank Emőke-Ágnes Horvát for the patience, kindness and support, not to mention the many-many useful discussions.

I would like to acknowledge the support from the German Science Foundation (DFG) as a member of the Research Training Group "Spatio/Temporal Probabilistic Graphical Models and Applications in Image Analysis", grant GRK 1653, the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences and the Institute for Theoretical Physics, all at the University of Heidelberg

I am the most grateful for the love, encouragement and support of my parents,
Eugenia and Árpád

Conference/Workshop Participation

I have participated in the following conferences and workshops:

- Internal Workshop of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis", Heidelberg, 2010 [talk]
- Workshop on Graphical Models, Heidelberg, 2010
- 5th Workshop on Monte Carlo Methods, Heidelberg, 2011
- Retreat of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis", St. Martin, 2011 [talk]
- Summer School on Probabilistic Graphical Models, Shanghai, 2011 [talk]
- Summer School on Probabilistic Graphical Models, Shanghai, 2011 [poster]
- Annual Workshop of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis", Heidelberg, 2011
- Annual Colloquium of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, Altleiningen, 2011 [poster]
- Retreat of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis", Trifels in Annweiler, 2012 [talk]
- Annual Workshop of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis" – Discrete Graphical Models and Combinatorial Optimization, Heidelberg, 2012
- German Physical Society Meeting, Regensburg, 2013 [talk]
- German Physical Society Meeting, Regensburg, 2013 [two posters]
- American Physical Society Meeting, Baltimore, 2013 [talk]
- American Physical Society Meeting, Baltimore, 2013 [poster]
- 6th Workshop on Monte Carlo Methods, Heidelberg, 2013 [poster]
- Retreat of the Research Training Group "Spatio/Temporal Graphical Models and Applications in Image Analysis", Trifels in Annweiler, 2013 [talk]

Bibliography

[1] Clive L. N. Ruggles. *Astronomy in Prehistoric Britain and Ireland*. Yale University Press, 1999.

[2] Antonella Galvan, John P.A. Ioannidis, and Tommaso A. Dragani. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in Genetics*, 26(3):132–141, March 2010.

[3] Sekar Kathiresan and Deepak Srivastava. Genetics of human cardiovascular disease. *Cell*, 148(6):1242–1257, March 2012.

[4] Jason D. Cooper, Deborah J. Smyth, Neil M. Walker, Helen Stevens, Oliver S. Burren, Chris Wallace, Christopher Greissl, Elizabeth Ramos-Lopez, Elina Hyppönen, David B. Dunger, Timothy D. Spector, Willem H. Ouwehand, Thomas J. Wang, Klaus Badenhoop, and John A. Todd. Inherited variation in vitamin d genes is associated with predisposition to autoimmune disease type 1. *Diabetes*, 60(5):1624–1631, May 2011. PMID: 21441443.

[5] Raz Somech, Ninette Amariglio, Zvi Spirer, and Gideon Rechavi. Genetic predisposition to infectious pathogens: a review of less familiar variants. *The Pediatric infectious disease journal*, 22(5):457–461, May 2003. PMID: 12792391.

[6] Schuckit Ma. An overview of genetic influences in alcoholism. *Journal of substance abuse treatment*, 36(1):S5–14, January 2009. PMID: 19062348.

[7] Toni-Lee Sterley, Fleur M. Howells, and Vivienne A. Russell. Effects of early life trauma are dependent on genetic predisposition: a rat study. *Behavioral and Brain Functions*, 7(1):11, May 2011. PMID: 21548935.

[8] S. Li, J. H. Zhao, J. Luan, C. Langenberg, R. N. Luben, K. T. Khaw, N. J. Wareham, and R. J. F. Loos. Genetic predisposition to obesity leads to increased risk of type 2 diabetes. *Diabetologia*, 54(4):776–782, April 2011.

[9] Richard I. Morimoto and Ana M. Cuervo. Protein homeostasis and aging: Taking care of proteins from the cradle to the grave. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 64A(2):167–170, February 2009. PMID: 19228787.

[10] Maria N. Starodubtseva. Mechanical properties of cells and ageing. *Ageing Research Reviews*, 10(1):16–25, January 2011.

[11] Judith Campisi. Aging, cellular senescence, and cancer. *Annual Review of Physiology*, 75(1):685–705, 2013. PMID: 23140366.

[12] Cynthia J. Kenyon. The genetics of ageing. *Nature*, 464(7288):504–512, March 2010.

[13] Marc A. Marti-Renom and Leonid A. Mirny. Bridging the resolution gap in structural modeling of 3D genome organization. *PLOS Computational Biology*, 7(7):e1002125, July 2011.

[14] Dieter W. Heermann, Hansjoerg Jerabek, Lei Liu, and Yixue Li. A model for the 3D chromatin architecture of pro and eukaryotes. *Methods*, 58(3):307–314, November 2012.

[15] Benjamin Albert, Isabelle Léger-Silvestre, Christophe Normand, and Olivier Gadal. Nuclear organization and chromatin dynamics in yeast: Biophysical models or biologically driven interactions? *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(6):468–481, June 2012.

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

[17] Tom Sexton, Frédéric Bantignies, and Giacomo Cavalli. Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. *Seminars in Cell & Developmental Biology*, 20(7):849–855, September 2009.

[18] Stephan Kadauke and Gerd A. Blobel. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1):17–25, January 2009.

[19] Anita Göndör and Rolf Ohlsson. Chromosome crosstalk in three dimensions. *Nature*, 461(7261):212–217, September 2009.

[20] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009. PMID: 19815776.

[21] Esperanza Nunez, Xiang-Dong Fu, and Michael G Rosenfeld. Nuclear organization in the 3D space of the nucleus – cause or consequence? *Current Opinion in Genetics & Development*, 19(5):424–436, October 2009.

[22] J. Zlatanova and P. Caiafa. CCCTC-binding factor: to loop or to bridge. *Cellular and Molecular Life Sciences*, 66(10):1647–1660, May 2009.

[23] Dieter W Heermann. Physical nuclear organization: loops and entropy. *Current Opinion in Cell Biology*, 23(3):332–337, June 2011.

[24] Ann Dean. In the loop: long range chromatin interactions and gene regulation. *Briefings in Functional Genomics*, 10(1):3–10, January 2011. PMID: 21258045.

[25] Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, PhilipÂăD. Gregory, Ann Dean, and Gerd A. Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244, June 2012.

[26] Delphine Quénet, James G. McNally, and Yamini Dalal. Through thick and thin: the conundrum of chromatin fibre folding in vivo. *EMBO reports*, 13(11):943–944, November 2012.

[27] Jeffrey C. Hansen. Human mitotic chromosome structure: what happened to the 30-nm fibre? *The EMBO Journal*, 31(7):1621–1623, April 2012.

[28] Hugo Maruyama, Janet C. Harwood, Karen M. Moore, Konrad Paszkiewicz, Samuel C. Durley, Hisanori Fukushima, Haruyuki Atomi, Kunio Takeyasu, and Nicholas A. Kent. An alternative beads-on-a-string chromatin architecture in thermococcus kodakarensis. *EMBO reports*, 14(8):711–717, August 2013.

[29] Davide Marenduzzo, Cristian Micheletti, and Peter R. Cook. Entropy-driven genome organization. *Biophysical Journal*, 90(10):3712–3721, May 2006.

[30] Peter R. Cook and Davide Marenduzzo. Entropic organization of interphase chromosomes. *The Journal of Cell Biology*, 186(6):825–834, September 2009. PMID: 19752020.

[31] Mario Nicodemi and Antonella Prisco. Thermodynamic pathways to genome spatial organization in the cell nucleus. *Biophysical Journal*, 96(6):2168–2177, March 2009.

[32] Manfred Bohn and Dieter W. Heermann. Topological interactions between ring polymers: Implications for chromatin loops. *The Journal of Chemical Physics*, 132(4):044904, January 2010.

[33] Suckjoon Jun and Andrew Wright. Entropy as the driver of chromosome segregation. *Nature Reviews Microbiology*, 8(8):600–607, August 2010.

[34] R. K. Sachs, G. van den Engh, B. Trask, H. Yokota, and J. E. Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences*, 92(7):2710–2714, March 1995. PMID: 7708711.

[35] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002. PMID: 11847345.

[36] Miriam Fritsche, Ras B. Pandey, Barry L. Farmer, and Dieter W. Heermann. Variation in structure of a protein (H2AX) with knowledge-based interactions. *PLoS ONE*, 8(5):e64507, May 2013.

[37] Yang Zhang and Dieter W Heermann. DNA double-strand breaks: linking gene expression to chromosome morphology and mobility. *Chromosoma*, August 2013. PMID: 23982751.

[38] Barbara Di Ventura, Benoît Knecht, Helena Andreas, William J. Godinez, Miriam Fritsche, Karl Rohr, Walter Nickel, Dieter W. Heermann, and Victor Sourjik. Chromosome segregation by the escherichia coli min system. *Molecular Systems Biology*, 9(1), September 2013.

[39] Yang Zhang and Dieter W. Heermann. Loops determine the mechanical properties of mitotic chromosomes. *PLoS ONE*, 6(12):e29225, December 2011.

[40] Tamar Schlick and Stephen Neidle. *Innovations in Biomolecular Modeling and Simulations*. Royal Society of Chemistry, 2012.

[41] Nikolay Korolev, Yanping Fan, Alexander P Lyubartsev, and Lars Nordenskiöld. Modelling chromatin structure and dynamics: status and prospects. *Current Opinion in Structural Biology*, 22(2):151–159, April 2012.

[42] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *Biomedical Engineering, IEEE Reviews in*, 2:147–171, 2009.

[43] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. Local binary patterns variants as texture descriptors for medical image analysis. *Artificial Intelligence in Medicine*, 49(2):117–125, June 2010.

[44] Quanli Wang, Jarad Niemi, Chee-Meng Tan, Lingchong You, and Mike West. Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy. *Cytometry Part A*, 77A(1):101–110, 2010.

[45] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, July 2012.

[46] G. Teodoro, T. Pan, T.M. Kurc, Jun Kong, L.A.D. Cooper, N. Podhorszki, S. Klasky, and J.H. Saltz. High-throughput analysis of large microscopy image datasets on CPU-GPU cluster platforms. In *2013 IEEE 27th International Symposium on Parallel Distributed Processing (IPDPS)*, pages 103–114, 2013.

[47] F C Bernstein, T F Koetzle, G J Williams, Jr Meyer, E F, M D Brice, J R Rodgers, O Kennard, T Shimanouchi, and M Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*, 112(3):535–542, May 1977. PMID: 875032.

[48] Yuzhen Ye and Adam Godzik. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, 32(suppl 2):W582–W585, July 2004. PMID: 15215455.

[49] Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19(suppl 2):ii246–ii255, September 2003. PMID: 14534198.

[50] Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson. Flexible protein alignment and hinge detection. *Proteins: Structure, Function, and Bioinformatics*, 48(2):242–256, 2002.

[51] Chan-Yong Park and Chi-Jung Hwang. A study of flexible protein structure alignment using three dimensional local similarities. *The KIPS Transactions:PartB*, 16B(5):359–366, October 2009.

[52] Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, June 2009.

[53] Jairo Rocha, Joan Segura, Richard C. Wilson, and Swagata Dasgupta. Flexible structural protein alignment by a sequence of local transformations. *Bioinformatics*, 25(13):1625–1631, July 2009. PMID: 19417057.

[54] Yu-Shen Liu, Meng Wang, Jean-Claude Paul, and Karthik Ramani. 3DMolNavi: a web-based retrieval and navigation tool for flexible molecular shape comparison. *Bmc Bioinformatics*, 13, May 2012. WOS:000308069900001.

[55] M. N. Nguyen, K. P. Tan, and M. S. Madhusudhan. CLICK–topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Research*, 39(suppl 2):W24–W28, July 2011. PMID: 21602266.

[56] Yu-Shen Liu, Yi Fang, and Karthik Ramani. IDSS: deformation invariant signatures for molecular shape comparison. *BMC Bioinformatics*, 10(1):157, May 2009. PMID: 19463181.

[57] Christoph J. Feinauer, Andreas Hofmann, Sebastian Goldt, Lei Liu, Gabriell Máté, and Dieter W. Heermann. Zinc finger proteins and the 3D organization of chromosomes. In *Advances in Protein Chemistry and Structural Biology*, volume 90, pages 67–117. Elsevier, 2013.

[58] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*, chapter Intorduction to the Cell. In [296], 4 edition, 2002.

[59] Gerald Karp. *Cell and Molecular Biology: Concepts and Experiments*, chapter The Nucleus of a Eukaryotic Cell. Wiley, 6 edition, 2009.

[60] E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science : A Publication of the Protein Society*, 7(4):1029–1038, April 1998. PMID: 9568909 PMCID: PMC2143985.

[61] Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. *Molecular Cell Biology*. New York: W. H. Freeman, 2004.

[62] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8âĂĽÃĚ resolution. *Nature*, 389(6648):251–260, September 1997.

[63] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921):448–453, January 2003.

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

[65] Thomas Dange, Aviva Joseph, and David Grünwald. A perspective of the dynamic structure of the nucleus explored at the single-molecule level. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 19(1):117–129, January 2011. PMID: 20842420.

[66] Mary Hoff. Loopy chromatin brings distant DNA to bear on silencing promoter genes. *PLoS Biol*, 6(12):e313, December 2008.

[67] Bartek Wilczynski, Ya-Hsin Liu, Zhen Xuan Yeo, and Eileen E. M. Furlong. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput Biol*, 8(12):e1002798, December 2012.

[68] Dieter W. Heermann, Hansjoerg Jerabek, Lei Liu, and Yixue Li. A model for the 3D chromatin architecture of pro and eukaryotes. *Methods*, 58(3):307–314, November 2012.

[69] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, and Thomas Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 3(5):e157, April 2005.

[70] E Lukásová, S Kozubek, M Kozubek, M Falk, and J Amrichová. The 3D structure of human chromosomes in cell nuclei. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 10(7):535–548, 2002. PMID: 12498343.

[71] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol*, 4(5):e138, April 2006.

[72] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor Perspectives in Biology*, 2(3):a003889, March 2010. PMID: 20300217.

[73] Shiv I. S. Grewal and Songtao Jia. Heterochromatin revisited. *Nature Reviews Genetics*, 8(1):35–46, January 2007.

[74] Fyodor D Urnov and Alan P Wolffe. Chromatin organisation and human disease. *Expert Opinion on Therapeutic Targets*, 4(5):665–685, October 2000.

[75] Malcolm V Brock, James G Herman, and Stephen B Baylin. Cancer as a manifestation of aberrant chromatin structure. *Cancer journal (Sudbury, Mass.)*, 13(1):3–8, February 2007. PMID: 17464240.

[76] Howard J Worman, Cecilia Ostlund, and Yuexia Wang. Diseases of the nuclear envelope. *Cold Spring Harbor perspectives in biology*, 2(2):a000760, February 2010. PMID: 20182615.

[77] Mira Jakovcevski and Schahram Akbarian. Epigenetic mechanisms in neurological disease. *Nature Medicine*, 18(8):1194–1204, August 2012.

[78] Jaco H. Houtgraaf, Jorie Versmissen, and Wim J. van der Giessen. A concise review of DNA damage checkpoints and repair in mammalian cells. *Cardiovascular Revascularization Medicine*, 7(3):165–172, July 2006.

[79] S Clancy. DNA damage & repair: mechanisms for maintaining DNA integrity. *Nature Education*, 1(1), 2008.

[80] Dana Branzei and Marco Foiani. Regulation of DNA repair throughout the cell cycle. *Nature Reviews Molecular Cell Biology*, 9(4):297–308, April 2008.

[81] T. Helleday, J. Lo, D.C. van Gent, and B.P. Engelward. Dna double-strand break repair: From mechanistic understanding to cancer treatment. *DNA Repair*, 6(7):923 – 935, 2007. Replication Fork Repair Processes.

[82] Elizabeth M. Slayter and Henry S. Slayter. *Light and Electron Microscopy*. Cambridge University Press, 1993.

[83] James Pawley, editor. *Handbook of Biological Confocal Microscopy*, volume 1. Springer, 3 edition, 2006.

[84] Michiel Muller. *Introduction to Confocal Fluorescence Microscopy*. SPIE Press, 2006.

[85] Alberto Diaspro, editor. *Optical Fluorescence Microscopy: From the Spectral to the Nano Dimension*. Springer, 2010.

[86] Jeff W. Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature Methods*, 2(12):910–919, December 2005.

[87] James Pawley, editor. *Handbook of Biological Confocal Microscopy*, chapter Fluorophores for Confocal Microscopy. Volume 1 of Pawley [83], 3 edition, 2006.

[88] Denis Semwogerere and Eric R. Weeks. *Confocal Microscopy*, chapter 66, pages 705–714. CRC Press, 2 edition, 2008.

[89] Yuval Garini, Bart J Vermolen, and Ian T Young. From micro to nano: recent advances in high-resolution microscopy. *Current Opinion in Biotechnology*, 16(1):3–12, February 2005.

[90] Jennifer Lippincott-Schwartz and Suliana Manley. Putting super-resolution fluorescence microscopy to work. *Nature Methods*, 6(1):21–23, January 2009.

[91] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual Review of Biochemistry*, 78(1):993–1016, 2009. PMID: 19489737.

[92] Christoph Cremer, Rainer Kaufmann, Manuel Gunkel, Sebastian Pres, Yanina Weiland, Patrick Müller, Thomas Ruckelshausen, Paul Lemmer, Fania Geiger, Sven Degenhard, Christina Wege, Niels A W Lemmermann, Rafaela Holtappels, Hilmar Strickfaden, and Michael Hausmann. Superresolution imaging of biological nanostructures by spectral precision distance microscopy. *Biotechnology journal*, 6(9):1037–1051, September 2011. PMID: 21910256.

[93] R. K. Pathria. *Statistical Mechanics*, chapter Historical Introduction. In [297], second edition, 1996.

[94] R. K. Pathria. *Statistical Mechanics*, chapter 2. The Canonical Ensemble. In [297], second edition, 1996.

[95] R. K. Pathria. *Statistical Mechanics*, chapter 1. The Statistical Basis of Thermodynamics. In [297], second edition, 1996.

[96] R. K. Pathria. *Statistical Mechanics*, chapter 11. Phase Transitions: Criticality, Universality and Scaling. In [297], second edition, 1996.

[97] Franz Schwabl. *Statistical Mechanics*, chapter 7. Phase Transitions, Scale Invariance, Renormalization Group Theory, and Percolation. Springer, second edition, 2006.

[98] K. Binder. Theory of first-order phase transitions. *Reports on Progress in Physics*, 50(7):783, July 1987.

[99] Stephen Blundell. *Concepts in Thermal Physics*. Oxford University Press, second edition, 2006.

[100] W. Janke. *Ageing and the Glass Transition*, chapter 5. Introduction to Simulation Techniques. Springer, 2007.

[101] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomena*, chapter 3. Critical-Point Exponents and Rigorous Relations Among Them. Oxford University Press, 1971.

[102] Kurt Binder and Dieter W. Heermann. *Monte Carlo simulation in statistical physics: an introduction*, chapter 2. Theoretical Foundations of the Monte Carlo Method. In [103], 2002.

[103] Kurt Binder and Dieter W. Heermann. *Monte Carlo simulation in statistical physics: an introduction*. Springer, 2002.

[104] William D.C. Moebs. A monte carlo simulation of chemical reactions. *Mathematical Biosciences*, 22:113–120, 1974.

[105] Pierre Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, May 1999.

[106] Chaitanya Athale. Monte carlo cell simulations. *Genome Biology*, 3(1):reports2001, December 2001.

[107] Todd D. Little. *The Oxford Handbook of Quantitative Methods in Psychology*. Oxford University Press, March 2013.

[108] Peter Jäckel. *Monte Carlo Methods in Finance*. Wiley, April 2002.

[109] David Vose. *Quantitative risk analysis: a guide to Monte Carlo simulation modelling*. Wiley, Chichester, 1996.

[110] A. Berdondini, M. Bettuzzi, D. Bianconi, R. Brancaccio, F. Casali, S. Cornacchia, A. Flisch, J. Hofmann, N. Lanconelli, M.P. Morigi, A. Pasini, A. Rossi, C. Sauerwein, and M. Simon. Monte carlo optimization of an industrial tomography system. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 580(1):771–773, September 2007.

[111] Mark Chang. *Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies*. CRC Press, October 2010.

[112] J. G. Park, C. H. Kim, M. C. Han, S. H. Jung, J. B. Kim, and J. Moon. Optimization of detection geometry for industrial SPECT by monte carlo simulations. *Journal of Instrumentation*, 8(04):C04006, April 2013.

[113] B. A. Berg. *Markov Chain Monte Carlo: Innovations and Applications*, chapter 1. Introduction to Markov Chain Monte Carlo Simulations and their Statistical Analysis. World Scientific, January 2005.

[114] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087, June 1953.

[115] Silvio R.A. Salinas. *Introduction to Statistical Physics*, chapter 16. Nonequilibrium Phenomena: II Stochastic Methods. Springer, February 2001.

[116] R. K. Pathria. *Statistical Mechanics*, chapter 12. Exact (or Almost Exact) Results for the Various Models. In [297], second edition, 1996.

[117] Carlo Vanderzande. *Lattice Models of Polymers*. Cambridge University Press, April 1998.

[118] Amikam Aharoni. *Introduction to the Theory of Ferromagnetism*, chapter 4. Magnetizaation vs. Temperature. Oxford University Press, 2000.

[119] RJ Baxter and FY Wu. Ising model on a triangular lattice with three-spin interactions. i. the eigenvalue equation. *Australian Journal of Physics*, 27(3):357–368, January 1974.

[120] Luo Zhi-Huan, Loan Mushtaq, Liu Yan, and Lin Jian-Rong. Critical behaviour of the ferromagnetic ising model on a triangular lattice. *Chinese Physics B*, 18(7):2696, July 2009.

[121] Seung-Yeon Kim. Ising model on $L \times L$ square lattice with free boundary conditions up to $L = 19$. *Journal of Physics: Conference Series*, 410(1):012050, February 2013.

[122] Thomas P. Handford, Francisco J. Pérez-Reche, and Sergei N. Taraskin. Zero-temperature random-field ising model on a bilayered bethe lattice. *Physical Review E*, 88(2):022117, August 2013.

[123] J S Valverde, Onofre Rojas, and S M de Souza. Two-dimensional XXZ -ising model on a square-hexagon lattice. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79(4 Pt 1):041101, April 2009. PMID: 19518167.

[124] Dan Simon Zimmerman. A study of the ising model on the hexagonal closed-packed lattice with competing interactions. 1986. Graduate.

[125] J. Bricmont, H. Kesten, J. L. Lebowitz, and R. H. Schonmann. A note on the ising model in high dimensions. *Communications in Mathematical Physics*, 122(4):597–607, December 1989.

[126] Y. Liu and J.P. Dilger. Application of the one- and two-dimensional ising models to studies of cooperativity between ion channels. *Biophysical Journal*, 64(1):26–35, January 1993.

[127] Nikolaos G. Fytas and Víctor Martín-Mayor. Universality in the three-dimensional random-field ising model. *Physical Review Letters*, 110(22):227201, May 2013.

[128] Y. Liu and J.P. Dilger. Application of the one- and two-dimensional ising models to studies of cooperativity between ion channels. *Biophysical Journal*, 64(1):26–35, January 1993.

[129] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117–149, February 1944.

[130] Reinhard Diestel. *Graph Theory*. Springer, January 2005.

[131] Norman L. Biggs, E. Keith Lloyd, and Robin J. Wilson. *Graph Theory 1736-1936*, chapter 1. Paths. Oxford University Press, 1976.

[132] Lowell W. Beineke and Robin J. Wilson. *Topics in Topological Graph Theory*. Cambridge University Press, July 2009.

[133] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[134] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[135] Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, March 2009.

[136] Clark N. Glymour. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press, 2001.

[137] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21, January 2011. PMID: 21281499.

[138] Alessandro Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 38(33):R309, August 2005.

[139] Alejandro Lage-Castellanos, Andrea Pagnani, and Martin Weigt. Statistical mechanics of sparse generalization and graphical model selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(10):P10009, October 2009.

[140] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*, chapter 3. The Bayesian Network Representation. In [133], 2009.

[141] Judea Pearl. *Causality*. Cambridge University Press, September 2009.

[142] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, September 2003.

[143] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*, chapter 3. Undirected Grapahical Models. In [133], 2009.

[144] Ferenc Szalma and Ferenc Iglói. Two-dimensional dilute ising models: Critical behavior near defect lines. *Journal of Statistical Physics*, 95(3-4):759–766, May 1999.

[145] Sander Dommers, Cristian Giardinà, and Remco van der Hofstad. Ising models on power-law random graphs. *Journal of Statistical Physics*, 141(4):638–660, November 2010.

[146] Jack Raymond and K. Y. Michael Wong. Next nearest neighbour ising models on random graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(09):P09007, September 2012.

[147] H. Derin and Howard Elliott. Modeling and segmentation of noisy and textured images using gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(1):39–55, 1987.

[148] R.C. Dubes, A.K. Jain, S.G. Nadabar, and C.C. Chen. MRF model-based algorithms for image segmentation. In *, 10th International Conference on Pattern Recognition, 1990. Proceedings*, volume i, pages 808–814 vol.1, 1990.

[149] Xavier Descombes, Miguel Moctezuma, Henri Maître, and Jean-Paul Rudant. Coastline detection by a markovian segmentation on SAR images. *Signal Processing*, 55(1):123–132, November 1996.

[150] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[151] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Soc., 2010.

[152] Mark Anthony Armstrong. *Basic Topology*. Springer, 1990.

[153] Connie C. Cortez and Peter A. Jones. Chromatin, cancer and drug therapies. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 647(1–2):44–51, December 2008.

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

[155] Travis N. Mavrich, Cizhong Jiang, Ilya P. Ioshikhes, Xiaoyong Li, Bryan J. Venters, Sara J. Zanton, Lynn P. Tomsho, Ji Qi, Robert L. Glaser, Stephan C. Schuster, David S. Gilmour, Istvan Albert, and B. Franklin Pugh. Nucleosome organization in the drosophila genome. *Nature*, 453(7193):358–362, May 2008.

[156] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol*, 4(5):e138, April 2006.

[157] Jane L. Grogan, Markus Mohrs, Brian Harmon, Dee A. Lacy, John W. Sedat, and Richard M. Locksley. Early transcription and silencing of cytokine genes underlie polarization of t helper cell subsets. *Immunity*, 14(3):205–215, March 2001.

[158] Duncan Sproul, Nick Gilbert, and Wendy A. Bickmore. The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics*, 6(10):775–781, October 2005.

[159] W. J. Bodell and J. E. Cleaver. Transient conformation changes in chromatin during excision repair of ultraviolet damage to DNA. *Nucleic Acids Research*, 9(1):203–213, January 1981. PMID: 6259620.

[160] Nicole J. Francis, Robert E. Kingston, and Christopher L. Woodcock. Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574–1577, November 2004. PMID: 15567868.

[161] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43(11):1059–1065, November 2011.

[162] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, October 2012.

[163] Songling Li and Dieter W. Heermann. Using chimaeric expression sequence tag as the reference to identify three-dimensional chromosome contacts. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 20(1):45–53, February 2013. PMID: 23213109 PMCID: PMC3576657.

[164] Pierre-Gilles de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, 1979.

[165] Manfred Bohn, Dieter W. Heermann, and Roel van Driel. Random loop model for long polymers. *Physical Review E*, 76(5):051805, November 2007.

[166] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, February 2012.

[167] Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Morgan & Claypool Publishers, 2011.

[168] Mitsugu Matsushita and Hiroshi Fujikawa. Diffusion-limited growth in bacterial colony formation. *Physica A: Statistical Mechanics and its Applications*, 168(1):498–506, September 1990.

[169] P. Boolchand, G. Lucovsky, J. C. Phillips, and M. F. Thorpe. Self-organization and the physics of glassy networks. *Philosophical Magazine*, 85(32):3823–3838, 2005.

[170] E. Sackmann. Chapter 5 physical basis of self-organization and function of membranes: Physics of vesicles. In R. Lipowsky and E. Sackmann, editor, *Handbook of Biological Physics*, volume Volume 1 of *Structure and Dynamics of Membranes*, pages 213–304. North-Holland, 1995.

[171] Dirk Helbing and Tamás Vicsek. Optimal self-organization. *New Journal of Physics*, 1(1):13, August 1999.

[172] Akira Hasegawa. Self-organization processes in continuous media. *Advances in Physics*, 34(1):1–42, 1985.

[173] Eric Bonabeau, Guy Theraulaz, Jean-Louls Deneubourg, Serge Aron, and Scott Camazine. Self-organization in social insects. *Trends in Ecology & Evolution*, 12(5):188–193, May 1997.

[174] Dmitry Chistilin. Principles of self-organization and sustainable development of the world economy are the basis of global security. In Ali Minai, Dan Braha, and Yaneer Bar-Yam, editors, *Unifying Themes in Complex Systems*, pages 382–389. Springer Berlin Heidelberg, January 2008.

[175] Cheng-Yuan Liou and Jiann-Ming Wu. Self-organization using potts models. *Neural Networks*, 9(4):671–684, June 1996.

[176] Long-Qing Chen and Wei Yang. Computer simulation of the domain dynamics of a quenched system with a large number of nonconserved order parameters: The grain-growth kinetics. *Physical Review B*, 50(21):15752–15756, December 1994.

[177] Jörg Reichardt and Stefan Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93(21):218701, November 2004.

[178] Zoltán Kato and Ting-Chuen Pong. A markov random field image segmentation model for color textured images. *Image and Vision Computing*, 24(10):1103–1114, October 2006.

[179] V. Grau, A. U J Mewes, M. Alcaniz, R. Kikinis, and S.K. Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE Transactions on Medical Imaging*, 23(4):447–458, 2004.

[180] Gilles Celeux, Florence Forbes, and Nathalie Peyrard. EM procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, January 2003.

[181] M Bohn. *Modelling of Interphase Chromosomes: From Genome Function to Spatial Organization*. PhD thesis, University of Heidelberg, 2010.

[182] J. M. Kosterlitz. The critical properties of the two-dimensional xy model. *Journal of Physics C: Solid State Physics*, 7(6):1046, March 1974.

[183] C. F. Caiafa and A. N. Proto. Temperature Estimation in the Two-Dimensional Ising Model. *International Journal of Modern Physics C*, 17(01):29–38, January 2006.

[184] Chuanshu Ji and Lynne Seymour. A consistent model selection procedure for markov random fields based on penalized pseudolikelihood. *The Annals of Applied Probability*, 6(2):423–443, May 1996. Mathematical Reviews number (MathSciNet): MR1398052; Zentralblatt MATH identifier: 0856.62082.

[185] Charles J Geyer. *Markov chain Monte Carlo maximum likelihood*. Defense Technical Information Center, 1992.

[186] Ruslan Salakhutdinov. Learning in markov random fields using tempered transitions. In *Advances in neural information processing systems*, pages 1598–1606, 2009.

[187] F. Y. Wu. The potts model. *Reviews of Modern Physics*, 54(1):235–268, January 1982.

[188] R. J. Baxter, S. B. Kelland, and F. Y. Wu. Equivalence of the potts model or whitney polynomial with an ice-type model. *Journal of Physics A: Mathematical and General*, 9(3):397, March 1976.

[189] Ger Koper and Michal Borkovec. Binding of metal ions to polyelectrolytes and their oligomeric counterparts: An application of a generalized potts model. *The Journal of Physical Chemistry B*, 105(28):6666–6674, July 2001.

[190] Noriyuki Bob Ouchi, James A. Glazier, Jean-Paul Rieu, Arpita Upadhyaya, and Yasuji Sawada. Improving the realism of the cellular potts model in simulations of biological cells. *Physica A: Statistical Mechanics and its Applications*, 329(3–4):451–458, November 2003.

[191] H. W. J. Blöte and R. H. Swendsen. First-order phase transitions and the three-state potts model. *Physical Review Letters*, 43(11):799–802, September 1979.

[192] V. V. Bazhanov and Yu G. Stroganov. Chiral potts model as a descendant of the six-vertex model. *Journal of Statistical Physics*, 59(3-4):799–817, May 1990.

[193] B. Nienhuis, E. K. Riedel, and M. Schick. Magnetic exponents of the two-dimensional q-state potts model. *Journal of Physics A: Mathematical and General*, 13(6):L189, June 1980.

[194] R. J. Baxter. Potts model at the critical temperature. *Journal of Physics C: Solid State Physics*, 6(23):L445, November 1973.

[195] François Graner and James A. Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical Review Letters*, 69(13):2013–2016, September 1992.

[196] L Kullmann, J Kertész, and R.N Mantegna. Identification of clusters of companies in stock indices via potts super-paramagnetic transitions. *Physica A: Statistical Mechanics and its Applications*, 287(3–4):412–419, December 2000.

[197] Oliver Ibe. *Markov Processes for Stochastic Modeling*. Academic Press, September 2008.

[198] F. Y. Wu. Critical point of planar potts models. *Journal of Physics C: Solid State Physics*, 12(17):L645, September 1979.

[199] T. D. Lee and C. N. Yang. Statistical theory of equations of state and phase transitions. II. lattice gas and ising model. *Physical Review*, 87(3):410–419, August 1952.

[200] M. Blume, V. J. Emery, and Robert B. Griffiths. Ising model for the λ transition and phase separation in he^2-he^4 mixtures. *Physical Review A*, 4(3):1071–1077, September 1971.

[201] W. S. Lo and Robert A. Pelcovits. Ising model in a time-dependent magnetic field. *Physical Review A*, 42(12):7471–7474, December 1990.

[202] R. B. Stinchcombe. Ising model in a transverse field. i. basic theory. *Journal of Physics C: Solid State Physics*, 6(15):2459, August 1973.

[203] S. W. Sides, P. A. Rikvold, and M. A. Novotny. Kinetic ising model in an oscillating field: Finite-size scaling at the dynamic phase transition. *Physical Review Letters*, 81(4):834–837, July 1998.

[204] H. G. Ballesteros, L. A. Fernández, V. Martín-Mayor, A. MuÃśoz Sudupe, G. Parisi, and J. J. Ruiz-Lorenzo. Scaling corrections: site percolation and ising model in three dimensions. *Journal of Physics A: Mathematical and General*, 32(1):1, January 1999.

[205] P. M. Chaikin and T. C. Lubensky. *Principles of Condensed Matter Physics*. Cambridge University Press, September 2000.

[206] Michael Plischke and Birger Bergersen. *Equilibrium Statistical Physics*. World Scientific, 2006.

[207] M. Schick and R. B. Griffiths. Antiferromagnetic ordering in the three-state potts model. *Journal of Physics A: Mathematical and General*, 10(12):2123, December 1977.

[208] Fugao Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, March 2001.

[209] D. H. E. Gross, A. Ecker, and X. Z. Zhang. Microcanonical thermodynamics of first order phase transitions studied in the potts model. *Annalen der Physik*, 508(5):446–452, 1996.

[210] Federico Tessadori, Martijn van Zanten, Penka Pavlova, Rachel Clifton, Frédéric Pontvianne, L. Basten Snoek, Frank F. Millenaar, Roeland Kees Schulkes, Roel van Driel, Laurentius A. C. J. Voesenek, Charles Spillane, Craig S. Pikaard, Paul Fransz, and Anton J. M. Peeters. PHYTOCHROME b and HISTONE DEACETYLASE 6 control light-induced chromatin compaction in arabidopsis thaliana. *PLoS Genet*, 5(9):e1000638, September 2009.

[211] Martijn van Zanten, Federico Tessadori, Fionn McLoughlin, Reuben Smith, Frank F. Millenaar, Roel van Driel, Laurentius A. C. J. Voesenek, Anton J. M. Peeters, and Paul Fransz. Photoreceptors CRYTOCHROME2 and phytochrome b control chromatin compaction in arabidopsis. *Plant Physiology*, 154(4):1686–1696, December 2010. PMID: 20935177.

[212] Martijn van Zanten, Federico Tessadori, Anton J. M. Peeters, and Paul Fransz. Shedding light on large-scale chromatin reorganization in arabidopsis thaliana. *Molecular Plant*, 5(3):583–590, May 2012. PMID: 22528207.

[213] Frank Y. Shih. *Image Processing and Mathematical Morphology: Fundamentals and Applications*. CRC Press, September 2010.

[214] A Huisman, Lennert S. Ploeger, Hub F. J. Dullens, Neal Poulin, and William E. Grizzle. Development of 3D chromatin texture analysis using confocal laser scanning microscopy. *Cell Oncol.*, 27, 2005.

[215] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*, chapter 5. DNA Replication, Repair and Recombinaiton. In [296], 4 edition, 2002.

[216] M.J. Kruhlak, A. Celeste, G. Dellaire, O. Fernandez-Capetillo, W.G. Müller, J.G. McNally, D.P. Bazett-Jones, and A. Nussenzweig. Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *The Journal of Cell Biology*, 172(6):823–834, 2006.

[217] David W. Scott. *Multivariate density estimation.* Wiley series in probability and mathematical statistics ; A Wiley Interscience publication. Wiley, New York, 1992.

[218] R. K. Pathria. *Statistical Mechanics*, chapter 14. Fluctuations. In [297], second edition, 1996.

[219] Douglas B. West. *Introduction to Graph Theory (2nd Edition).* Prentice Hall, August 2000.

[220] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nat Phys*, 8(1):32–39, January 2012.

[221] Paolo Bajardi, Chiara Poletto, Jose J. Ramasco, Michele Tizzoni, Vittoria Colizza, and Alessandro Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PLoS ONE*, 6(1):e16591, 01 2011.

[222] Sergey V. Buldyrev, Roni Parshani, Gerald Paul, H. Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, April 2010.

[223] Brian J. O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D. Smith, Emily H. Turner, Ian B. Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M. Akey, Elhanan Borenstein, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Jay Shendure, and Evan E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, May 2012.

[224] Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E. Assmus, Miguel A. Andrade-Navarro, and Erich E. Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling*, 4(189):rs8, September 2011.

[225] B. Delaunay. Sur la sphère vide. *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 6:793–800, 1934.

[226] Franz Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM COMPUTING SURVEYS*, 23(3):345–405, 1991.

[227] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*, chapter 12. Intracellular Compartments and Protein Sorting. In [296], 4 edition, 2002.

[228] William M. Dougherty and Paul C. Goodwin. 3D structured illumination microscopy. volume 7905, pages 790512–790512–9, 2011.

[229] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, 1979.

[230] B. Delaunay. Sur la sphère vide. *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 6:793–800, 1934.

[231] Kieran F. Mulchrone. Application of delaunay triangulation to the nearest neighbour method of strain analysis. *Journal of Structural Geology*, 25(5):689 – 702, 2003.

[232] F. Davoine, M. Antonini, J.-M. Chassery, and M. Barlaud. Fractal image compression based on delaunay triangulation and vector quantization. *Image Processing, IEEE Transactions on*, 5(2):338–346, 1996.

[233] Gabriell Máté, Zoltán Néda, and József Benedek. Spring-block model reveals region-like structures. *PLoS ONE*, 6(2):e16518, 02 2011.

[234] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[235] D. C. Rapaport. *The Art of Molecular Dynamics Simulation.* Cambridge University Press, New York, NY, USA, 1996.

[236] E. W. Dijkstra. A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK*, 1(1):269–271, 1959.

[237] Peter Willett, John M. Barnard, and Geoffrey M. Downs. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 38(6):983–996, July 1998.

[238] David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

[239] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[240] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

[241] R. Tyrrell Rockafellar and Roger J. B. Wets. Set convergence. In *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*, pages 108–147. Springer Berlin Heidelberg, 1998. 10.1007/978-3-642-02431-3_4.

[242] Niu Huang, Brian K. Shoichet, and John J. Irwin. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, 2006.

[243] Vidit Nanda. Perseus: The Persistent Homology Software., 2012. (Date 15.10.2012).

[244] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012.

[245] Justin Balthrop, Stephanie Forrest, M. E. J. Newman, and Matthew M. Williamson. Technological networks and the spread of computer viruses. *Science*, 304(5670):527–529, 2004.

[246] Emőke-Ágnes Horvát, Michael Hanselmann, Fred A. Hamprecht, and Katharina A. Zweig. One plus one makes three (for social networks). *PLoS ONE*, 7(4):e34740, 04 2012.

[247] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*, 10(3):186–198, March 2009.

[248] Laura A. Zager and George C. Verghese. Graph similarity scoring and matching. *Applied Mathematics Letters*, 21(1):86 – 94, 2008.

[249] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[250] Christoph J. Feinauer, Andreas Hofmann, Sebastian Goldt, Lei Liu, Gabriell Máté, and Dieter W. Heermann. Chapter three - zinc finger proteins and the 3d organization of chromosomes. In Rossen Donev, editor, *Organisation of Chromosomes*, volume 90 of *Advances in Protein Chemistry and Structural Biology*, pages 67 – 117. Academic Press, 2013.

[251] Yuefeng Wang, John C. Fisher, Rose Mathew, Li Ou, Steve Otieno, Jack Sublet, Limin Xiao, Jianhan Chen, Martine F. Roussel, and Richard W. Kriwacki. Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21. *Nat Chem Biol*, 7(4):214–221, April 2011.

[252] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.

[253] Wen-Liang Hung and Miin-Shen Yang. Similarity measures of intuitionistic fuzzy sets based on hausdorff distance. *Pattern Recognition Letters*, 25(14):1603–1611, 2004.

[254] Pedro J. Ballester and W. Graham Richards. Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 463(2081):1307–1321, May 2007.

[255] Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, June 2009.

[256] Yu-Shen Liu, Yi Fang, and Karthik Ramani. IDSS: deformation invariant signatures for molecular shape comparison. *BMC Bioinformatics*, 10(1):157, May 2009. PMID: 19463181.

[257] Thomas Fober, Marco Mernberger, Gerhard Klebe, and Eyke Hüllermeier. Graph-based methods for protein structure comparison. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2013.

[258] Kaare Teilum, Johan G. Olsen, and Birthe B. Kragelund. Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences*, 66(14):2231–2247, July 2009.

[259] Jean-Claude Hausmann. On the Vietoris-Rips complexes and a cohomology theory for metric spaces. In *Prospects in topology (Princeton, NJ, 1994)*, volume 138 of *Ann. of Math. Stud.*, pages 175–188. Princeton Univ. Press, Princeton, NJ, 1995.

[260] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *Information Theory, IEEE Transactions on*, 29(4):551–559, July 1983.

[261] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

[262] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, 8:85–108, 1963.

[263] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[264] Cédric Villani. *Topics in optimal transportation*. Graduate studies in mathematics. American Mathematical Society, Providence (R.I.), 2003.

[265] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[266] Dmitriy Morozov. Dionysus – a c++ library for computing persistent homology. http://www.mrzv.org/software/dionysus/, June 2013. [Online; accessed 1-July-2013].

[267] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision, 1998*, pages 59–66, 1998.

[268] Daniel M. Zuckerman. The ensemble protein database. http://www.epdb.pitt.edu/, July 2006. [Online; accessed 1-July-2013].

[269] Benoit B. Mandelbrot. *The fractal geometry of nature*. W.H. Freeman, 1 edition, August 1982.

[270] Marcus J. Hamilton, Bruce T. Milne, Robert S. Walker, Oskar Burger, and James H. Brown. The complex structure of hunter–gatherer social networks. *Proceedings of the Royal Society B: Biological Sciences*, 274(1622):2195–2203, September 2007. PMID: 17609186.

[271] Russell A. Hill, R. Alexander Bentley, and Robin I. M. Dunbar. Network scaling reveals consistent fractal pattern in hierarchical mammalian societies. *Biology Letters*, 4(6):748–751, December 2008. PMID: 18765349.

[272] Misako Takayasu and Hideki Takayasu. Fractals and economics. In Robert A. Meyers, editor, *Complex Systems in Finance and Econometrics*, pages 444–463. Springer New York, 2011.

[273] Thomas Lux and Kiel Institut für Volkswirtschaftslehre. The multi-fractal model of asset returns : its estimation via GMM and its use for volatility forecasting. Economics working paper / Christian-Albrechts-Universität Kiel, Department of Economics 2003,13; 2003,13; 2003,13, Institut für Volkswirtschaftslehre, Kiel, 2003. urn:nbn:de:101:1-200911022352.

[274] G. Caldarelli, R. Marchetti, and L. Pietronero. The fractal properties of internet. *EPL (Europhysics Letters)*, 52(4):386, November 2000.

[275] Yongmei Lu and Junmei Tang. Fractal dimension of a transportation network and its relationship with urban growth: a study of the dallas - fort worth area. *Environment and Planning B: Planning and Design*, 31(6):895–911, 2004.

[276] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, July 2007. PMID: 17586683.

[277] Larry S. Liebovitch and Tibor Toth. A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A*, 141(8–9):386–390, November 1989.

[278] Felix Hausdorff. Dimension und äusseres mass. *Mathematische Annalen*, 79(1-2):157–179, 1918.

[279] Christoph Bandt, Nguyen Hung, and Hui Rao. On the open set condition for self-similar fractals. *Proceedings of the American Mathematical Society*, 134(5):1369–1374, 2006.

[280] Vidit Nanda. Perseus: The Persistent Homology Software. http://www.math.rutgers.edu/~vidit/perseus.html, 2012. [Online; accessed 1-July-2013].

[281] Robert MacPherson and Benjamin Schweinhart. Measuring shape with topology. *Journal of Mathematical Physics*, 53(7), JUL 2012.

[282] M. Fukushima and T. Shima. On a spectral analysis for the sierpinski gasket. *Potential Analysis*, 1(1):1–35, March 1992.

[283] Henry J.S. Smith. On the integration of discontinuous functions. *Proceedings of the London Mathematical Society*, 6(1), 1874.

[284] Claudio Rebbi and Robert H. Swendsen. Monte carlo renormalization-group studies of q-state potts models in two dimensions. *Physical Review B*, 21(9):4094–4107, May 1980.

[285] B. Nienhuis, Eberhard K. Riedel, and M. Schick. First- and second-order phase transitions in potts models: Renormalization-group solution. *Physical Review Letters*, 43(11):737–740, September 1979.

[286] H.W.J Blöte and M.P Nightingale. Critical behaviour of the two-dimensional potts model with a continuous number of states; a finite size scaling analysis. *Physica A: Statistical Mechanics and its Applications*, 112(3):405–465, June 1982.

[287] L.O. Chua, Martin Hasler, George S. Moschytz, and Jacques Neirynck. Autonomous cellular neural networks: a unified paradigm for pattern formation and active wave propagation. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 42(10):559–577, 1995.

[288] H.G. Othmer and L.E. Scriven. Instability and dynamic pattern in cellular networks. *Journal of Theoretical Biology*, 32(3):507–537, September 1971.

[289] Cheng-Hsiung Hsu, Song-Sun Lin, and Wenxian Shen. Traveling Waves in Cellular Neural Networks. *International Journal of Bifurcation and Chaos*, 09(07):1307–1319, July 1999.

[290] T. Sziranyi, J. Zerubia, D. Geldreich, and Z. Kato. Cellular neural network for markov random field image segmentation. In *, 1996 Fourth IEEE International Workshop on Cellular Neural Networks and their Applications, 1996. CNNA-96. Proceedings*, pages 139–144, 1996.

[291] T. Roska and L.O. Chua. The CNN universal machine: an analogic array computer. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 40(3):163–173, 1993.

[292] Andrzej Radosz, Katarzyna Ostasiewicz, Paulina Hetman, Piotr Magnuszewski, and Michal H. Tyc. "social temperature"–relation between binary choice model and ising model. *AIP Conference Proceedings*, 965(1):317–320, December 2007.

[293] Georg Zaklan, Frank Westerhoff, and Dietrich Stauffer. Analysing tax evasion dynamics via the ising model. *Journal of Economic Interaction and Coordination*, 4(1):1–14, June 2009.

[294] Christoph Hauert and György Szabó. Game theory and physics. *American Journal of Physics*, 73(5):405, 2005.

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

[296] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland, 4 edition, 2002.

[297] R. K. Pathria. *Statistical Mechanics*. Butterworth Heineman, Oxford, UK, second edition, 1996.