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

Dipl.-Phys. Kunhe Li

born in: Guangdong

Date of oral examination: 27.07.2022

Relation between genome organization and its physical properties

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

Abstract

With the rapid development of modern computational techniques, more complex systems have been found to have their global organization principles. In this thesis, we aim to establish a method to systematically unravel chromosome organization principles, which can serve as a general framework for the analysis of 3D genome architecture and other systems.

We start the analysis with crucial physical properties. We compute the contact probability curve for different polymer models and conclude that the asymptotic behavior of the contact probability curve does not depend on the definition of contact. Moreover, the effect of bending rigidity and compartmentalization is examined. The persistence lengths for homogeneous and heterogeneous semi-flexible self-avoiding walks are computed, and it is observed that the persistence length in the heterogeneous case is systematically smaller than in the homogeneous case.

To access genome-wide organizational patterns, experimental nucleosome positioning data for *Candida albicans* are investigated. Specifically, by performing hierarchical clustering on the auto-correlation function of the data, repeated patterns are observed across the entire genome, which supports a classification beyond the typical categories of heterochromatin and euchromatin.

In addition to observing the patterns, we successfully develop a quantitative characterization of intra-chromosomal organizational structure by extracting the inter-nucleosomal potential. These effective potentials capture the interaction between nucleosomes that incorporates the dynamics of related complexes. Moreover, an essential thermodynamic property, namely isothermal compressibility, is computed from the potential. By applying k-means clustering to potential parameters and thermodynamic compressibility, genome-wide clustering result is obtained, and information that leads to the genomic mechanical code is collected.

Finally, we focus on patterns of local structures. The organization principles of the CTCF (abbreviation for nucleotide sequence CCCTC-binding factor) are revealed. The averaged nucleosome frequency near CTCF binding sites is computed, and the corresponding spatial structure is observed for the first time.

Zusammenfassung

Mit der rasanten Entwicklung moderner Computertechniken wird festgestellt, dass komplexe Systeme ihre globalen Organisationsprinzipien haben. Die vorliegende Arbeit zielt darauf ab, eine Methode zur detaillierten Entschlüsselung von Chromosomen-Organisationsprinzipien zu entwickeln, die als allgemeiner Rahmen für die Analyse der 3D-Genomarchitektur und anderer Systeme dienen kann.

Zu diesem Zweck beginnen wir die Analyse mit mehreren entscheidenden Eigenschaften. Wir berechnen die Kontaktwahrscheinlichkeitskurve für verschiedene Polymermodelle und stellen fest, dass das asymptotische Verhalten der Kontaktwahrscheinlichkeitskurve nicht von der Definition des Kontakts abhängt. Darüber hinaus wird der Einfluss von Biegesteifigkeit und Kompartimentierung untersucht. Wir berechnen die Persistenzlängen für homogene und heterogene semi-flexible selbstmeidender Pfad und stellen fest, dass die Persistenzlänge im heterogenen Zustand systematisch kleiner ist als im homogenen Zustand.

Um Zugang zu genomweiten Organisationsmustern zu erhalten, wird eine Vergröberungsmethode auf experimentelle Nukleosomenpositionierungsdaten in *Candida albicans* angewandt. Die Hierarchische Clusteranalyse der Autokorrelationsfunktion der Daten wird verwendet, um konservierte Muster im gesamten Genom zu beobachten und die Klassifikation der Nukleosomenorganisation mit mehr als zwei Zuständen zu unterstützen.

Der nächste Schritt nach der Beobachtung der Muster ist eine quantitative Charakterisierung der intrachromosomalen Organisationsstruktur. Wir haben erfolgreich eine Methode für dieses Ziel entwickelt, indem wir das inter-nukleosomale Potenzial extrahieren. Diese effektiven Potenziale erfassen die Interaktion zwischen Nukleosomen, die die Dynamik der assoziierten Komplexe einbezieht. Darüber hinaus wird eine grundlegende thermodynamische Eigenschaft, nämlich die isothermische Kompressibilität, aus dem Potenzial berechnet. Durch die Anwendung des k-Means-Algorithmus auf die Potenzialparameter und die thermodynamische Kompressibilität wird eine genomweite Klassifikation erreicht und Informationen über den genomischen mechanischen Code erhalten.

Schließlich konzentrieren wir uns auf die Muster der lokalen Strukturen. Die Organisationsprinzipien des CTCF (Nukleotidsequenz CCCTC-Bindefaktor) werden aufgedeckt. Die durchschnittliche Nukleosomendichte in der Nähe von CTCF-Bindungsstellen wird berechnet, und die entsprechende lokale Struktur wird aufgedeckt.

Publications Related to this Thesis

- Jia, J., **Li, K.,** Hofmann, A., & Heermann, D. W. The Effect of Bending Rigidity on Polymers. *Macromolecular Theory and Simulations* **(2019)**, *28*, 1800071

  DOI: 10.1002/mats.201800071

- Mishra, S. K., **Li, K.**, Brauburger, S., Bhattacherjee, A., Oiwa, N. N., & Heermann, D. W. Superstructure detection in nucleosome distribution shows common pattern within a chromosome and within the genome. *Life* **(2022)**, *12*, 541

  DOI: 10.3390/life12040541

- **Li, K.,** Oiwa, N. N., Mishra, S. K., & Heermann, D. W. Inter-nucleosomal potentials from nucleosomal positioning data. *The European Physical Journal E* **(2022)**, *45*, 1–8

  DOI: 10.1140/epje/s10189-022-00185-3

- Oiwa, N. N., **Li, K.**, Cordeiro, C. E., & Heermann, D. W. Prediction and Comparative Analysis of CTCF Binding Sites based on a First Principle Approach. *Physical Biology* **(2022)**, *19*

  DOI: 10.1088/1478-3975/ac5dca

Contents

Acknowledgments

I'm extremely grateful to my advisor Prof. Dieter W. Heermann. This project would not have been possible without his enduring support throughout my doctoral study. His invaluable advice and instructions are indispensable for me to finish my work.

I would like to thank Assoc. Prof. Nestor N. Oiwa for his great suggestions.

I also appreciate the support of my group members, Sujeet K. Mishra, Jiying Jia, Min Chu, and Andreas Hofmann. They helped me a lot in my work.

I am most grateful for the unconditional support from my family.

I'd like to recognize the support of the Institute for Theoretical Physics (ITP) for funding and excellent computational resources.

I am willing to acknowledge funding from the China Scholarship Council (CSC).

Part I

BACKGROUND & FUNDAMENTALS

Chapter 1

Introduction

1.1 Background

Over the last centuries, the methodology for understanding a system has undergone dramatic changes. As our scope of research expands, some basic conditions in classical systems, such as homogenous distribution or negligible fluctuations, are no longer eligible for complex systems. For example, in the research of genomic systems, heterogeneity, non-linearity, and multilayer comprehensive interaction maps must be considered in depth [1, 2].

When a heterogeneous polymer chain, such as a chromosome, is targeted, different locations of the structure may exhibit different bending rigidity, which can significantly affect its functions and lead to different gene expressions and regulations [3]. For some DNA regions, such as enhancers, the bending rigidity strongly influences the contact probability with other regions and alters the three-dimensional genome structure, resulting in different interaction maps [4]. Therefore, what is the impact of heterogeneous bending rigidity? How does the contact probability behave under different definitions? Are there common patterns in different genomic regions? And if such patterns exist, how can they be quantified? These are the pressing questions of the hour.

Our research is based on the continuous discoveries in polymer physics, 3D genome architecture, and computational data analysis.

The baseline is a variety of well-established theoretical models in polymer physics, including investigations of related crucial properties. The classical theoretical models are random walk [5], self-avoiding walk [6, 7], semi-flexible walk [8, 9], etc.

These models have produced great contributions to many subjects, such as materials science, cell biology, and economics [10, 11, 12]. For example, the conditional self-avoidance walk, a derivative of the self-avoiding walk, has been utilized for simulations of protein folding and successfully explained the helical structures inside [13]. Related crucial properties commonly include bond length, radius of gyration, bending rigidity, etc. Statistical parameters such as contact probability are also involved. The contact probability is the central parameter in the analysis of chromatin organization because it coincides with the observations in the chromosome conformation capture (3C) experiments [4]. However, the related mechanism still has not been fully investigated.

On the other hand, along with the emergence of massively parallel sequencing, numerous biological techniques have been developed, which allow us to access genomic systems with a large amount of detail, e.g., sequence, expression rate, critical transcription factors, and so on [14, 15]. However, it is found that the raw sequences do not directly authorize us to obtain gene expression and regulation [16]. The genomic system must be regarded as a 3D genome architecture incorporating multi-factor interactions [17]. In recent years, the discovery of chromosome territories and topologically associating domains has further emphasized this point, leading to the demand for a thorough investigation of its organizational principles [18, 19].

Data analysis is nowadays unavoidable in many research areas, especially in the study of a complex system where the Hamiltonian function is not available. When processing the nucleosome positioning data in the genomic systems, a comprehensive strategy with multiple techniques is required [20], which may involve simulation methods such as pivot algorithm and optimization methods such as reverse Monte Carlo [21, 22]. We also utilize k-means and hierarchical clustering as machine learning methods for identifying the pattern [23, 24].

In this thesis, multiple approaches are proposed to unravel the complex organization principles of genomic systems. To this end, we accomplished a comprehensive investigation of several crucial physical properties and their patterns from the small scale to the coarse-grained scale. We start with the contact probability and bending rigidity of a polymer chain and theoretically characterize their behaviors. Then we detect the organization pattern for the whole genome on a coarse-grained scale. After the consistent pattern is observed, we extract the effective potential representing the inter-nucleosomal interactions to identify the chromosome structure. As a final part, we integrate details of small-scale organization between crucial chromosomal complexes. In this manner, a complete picture of the structure is now raised.

1.2 Scope of This Thesis

In chapter 2, the fundamental knowledge for this thesis is briefly presented. Since the field of our work extends across physics, biology, and computer science, only essential information that is indispensable for the later chapters is covered here. In this chapter, we first focus on polymer models and their properties. After that, we illustrate the chromosome architecture. Then we explain several computational algorithms, paying particular attention to non-linear algorithms and machine learning.

Our results are from chapters 3 to 6. In these chapters, we disentangle the genome organization principles through theories, simulations, models, observations, and classifications.

In chapter 3, we start with one of the most important parameters in current measurements of chromosome conformations, the contact probabilities. The influence of its definition is carefully examined. Our result concludes that the asymptotic behavior of the contact probability curve is preserved under different definitions. In other words, the contact probability does not depend on the definition of the contact range in the limit of infinite contour length. Another crucial parameter that is investigated is bending rigidity. We calculate the corresponding persistence length, as a characteristic parameter for bending rigidity, for both homogeneous and heterogeneous cases. And the influence of heterogeneity is discussed. Additionally, the compartmentalization of the nucleus is also inspected.

In chapter 4, we aim for consistent patterns in the experimental data of nucleosome organization. The coarse-graining technique is applied after observing the highly random signal on the $O(1)$ bp scale in the nucleosome positioning data of *Candida albicans*. On a 5000 bp coarse-graining scale, consistent patterns are found to occur repeatedly throughout all chromosomes. Moreover, hierarchical clustering is applied to the patterns, and a genome-widely conserved clustering result is found.

In chapter 5, we establish a systematic method to derive intra-chromosomal potentials for the whole genome. With the extracted effective potential, we can calculate essential thermodynamical properties and further examine the principles of chromosome organization. In order to obtain the potential, a generalized Lennard-Jones potential is used for parameterization, which is inferred from the calculated mean-field potential. Besides, an intuitive selection strategy is adopted as a robust and highly efficient algorithm to solve the noisy optimization problem in the calculation. After extracting the effective potential, thermodynamic compressibilities are computed, and k-mean clustering is performed for the potentials and the compressibilities. The result allows us to access details for interactions inside chromosome and lead to a genome-wide classification that supports a scheme beyond the typical euchromatin-heterochromatin separation.

In chapter 6, CCCTC transcription factor (CTCF) binding sites are analyzed since they play a primary role in chromatin structure, especially in long-range binding. After predicting the CTCF binding sites through a first principle approach, several patterns are observed, and a prominent spacial structure from the averaged nucleosome density near CTCF binding sites is detected. These results benefit the studies of transcription factors and systematically present a clear local pattern inside the global chromosome architecture.

In chapter 7 we give a summary of our work.

References

[1] Philipp M Diesinger and Dieter W Heermann. "Depletion effects massively change chromatin properties and influence genome folding". In: *Biophysical journal* 97.8 (2009), pp. 2146–2153.

[2] Martijn Zuiddam, Ralf Everaers, and Helmut Schiessel. "Physics behind the mechanical nucleosome positioning code". In: *Physical Review E* 96.5 (2017), p. 052412.

[3] Eran Segal et al. "A genomic code for nucleosome positioning". In: *Nature* 442.7104 (2006), pp. 772–778.

[4] Nynke L Van Berkum et al. "Hi-C: a method to study the three-dimensional architecture of genomes." In: *JoVE (Journal of Visualized Experiments)* 39 (2010), e1869.

[5] Révész Pál. *Random Walk in Random and Non-random Environments*. World Scientific, 1990.

[6] Wyatt Hooper and Alexander R Klotz. "Trapping in self-avoiding walks with nearest-neighbor attraction". In: *Physical Review E* 102.3 (2020), p. 032132.

[7] S Havlin and D Ben-Avraham. "New approach to self-avoiding walks as a critical phenomenon". In: *Journal of Physics A: Mathematical and General* 15.6 (1982), p. L321.

[8] Jan Wilhelm and Erwin Frey. "Radial distribution function of semiflexible polymers". In: *Physical review letters* 77.12 (1996), p. 2581.

[9] Jan Kierfeld et al. "Semiflexible polymers and filaments: From variational problems to fluctuations". In: *AIP Conference Proceedings*. Vol. 1002. 1. American Institute of Physics. 2008, pp. 151–185.

[10] Chase P Broedersz and Fred C MacKintosh. "Modeling semiflexible polymer networks". In: *Reviews of Modern Physics* 86.3 (2014), p. 995.

[11] Guanghui Ping, Guoliang Yang, and Jian-Min Yuan. "Depletion force from macromolecular crowding enhances mechanical stability of protein molecules". In: *Polymer* 47.7 (2006), pp. 2564–2570.

[12] Benoit Mandelbrot. “New methods in statistical economics”. In: *Journal of political economy* 71.5 (1963), pp. 421–440.

[13] Kerson Huang. “CSAW: A dynamical model of protein folding”. In: *arXiv preprint cond-mat/0601244* (2006).

[14] Vladimir B Teif et al. “Genome-wide nucleosome positioning during embryonic stem cell development”. In: *Nature structural & molecular biology* 19.11 (2012), pp. 1185–1192.

[15] Zhijun Duan et al. “A three-dimensional model of the yeast genome”. In: *Nature* 465.7296 (2010), pp. 363–367.

[16] Anton Valouev et al. “A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning”. In: *Genome research* 18.7 (2008), pp. 1051–1063.

[17] Yuwen Ke et al. “3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis”. In: *Cell* 170.2 (2017), pp. 367–381.

[18] Thomas Cremer and Marion Cremer. “Chromosome territories”. In: *Cold Spring Harbor perspectives in biology* 2.3 (2010), a003889.

[19] Jesse R Dixon, David U Gorkin, and Bing Ren. “Chromatin domains: the unit of chromosome organization”. In: *Molecular cell* 62.5 (2016), pp. 668–680.

[20] Kevin Struhl and Eran Segal. “Determinants of nucleosome positioning”. In: *Nature structural & molecular biology* 20.3 (2013), pp. 267–273.

[21] Tom Kennedy. “A faster implementation of the pivot algorithm for self-avoiding walks”. In: *Journal of Statistical Physics* 106.3 (2002), pp. 407–429.

[22] Alexander P Lyubartsev and Aatto Laaksonen. “Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach”. In: *Physical Review E* 52.4 (1995), p. 3730.

[23] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. “A comparative study of efficient initialization methods for the k-means clustering algorithm”. In: *Expert systems with applications* 40.1 (2013), pp. 200–210.

[24] Gabor J Szekely, Maria L Rizzo, et al. “Hierarchical clustering via joint between-within distances: Extending Ward’s minimum variance method”. In: *Journal of classification* 22.2 (2005), pp. 151–184.

Chapter 2

Fundamentals

Our research field extends beyond physics, biology, and computer science. In this chapter, the related basic knowledge is introduced. It is impossible to cover all knowledge about these three subjects. Hence, only the strictly close areas are covered. To highlight the key concepts, we focus on different aspects of each area. The exact contents and their connections to other chapters are listed below.

In section 2.1, polymer models and essential physical properties are stressed. This section aims to elucidate the models and define the properties clearly. Other aspects such as historical evolution are neglected. First, the random walk and self-avoiding walk are demonstrated, which are basic models for the simulations in chapter 3. Meanwhile, the end-to-end distance and radius of gyration are included, which are crucial properties for further calculations. Then, we introduce the bending rigidity by mentioning the semi-flexible chain, especially the worm-like chain. The effects of bending rigidity for polymer models are studied in chapter 3. In addition, the exponential behavior is focused. The exponential behavior is explicitly used in the analysis with contact probability and bending rigidity and forms a foundation for the whole thesis. Later in the section, the log-logistic distribution is related to the nucleosome density distribution in chapter 4 and the Lennard-Jones potential is the basis for our model in chapter 5.

In section 2.2, only the genomic system as a specialization of a complex system is examined. We manage to explain the basic building blocks that construct the chromosome architecture. Therefore, we introduce the chromosome and the nucleosome. Then, we mention the Chromosome Conformation Capture techniques. The experimental data in this thesis is contributed by these techniques.

In section 2.3, related computational methods and their typical algorithms are covered. The aim is to explain the basic procedures of each method. It should be noted that each method has a series of variants and different types of implementations, of which only the most typical one is presented. At the beginning of the section, the Monte Carlo method is demonstrated, then we move on to the reverse Monte Carlo method. The reverse Monte Carlo method is the central part of our algorithm in chapter 5. We introduce the genetic algorithm as the most popular type of evolutionary algorithm. The remarkable structure of the genetic algorithm inspires us to solve the optimization problem in chapter 5. Later, we focus on machine learning techniques. After clarifying the basic concepts of machine learning and defining the distances between vectors, k-means clustering and hierarchical clustering are demonstrated. These two algorithm are utilized for both chapter 4 and chapter 5.

2.1 Polymer Models and Physical Properties

Random Walk

In polymer physics, each basic polymer unit is called a monomer. A monomer is an abstract concept describing a stable complex formed by one molecule or a group of molecules. N monomers can join in a line to form a polymer chain. A random walk or a freely-jointed chain is a polymer chain in which each monomer can freely move around without constraints, i.e., it can occupy any lattice site in discrete space or occupy any position in continuous space. The random walk model also contains the condition of homogeneity, which denotes that all monomers and their connections are identical. This condition ensures that all connections either have the same length or have the same statistics. For a typical random walk, the length of all connections is fixed at a constant l. Since the monomers of a random walk can occupy an arbitrary position, two monomers can overlap and occupy the same space.

If we denote the connection from monomer i to monomer $i+1$ as $\vec{r_i}$, the above mentioned properties of random walk can be described by equation (2.1), where $\langle\cdot\rangle$ refers to the mean value. This equation is valid for all $i = 1, 2...N-1$ with the total number of monomers being N.

$$\langle\vec{r_i}\rangle = \vec{0} \tag{2.1}$$

End-to-End Distance and Radius of Gyration

Several descriptions are widely adopted to account for the physical properties of polymer chains. The most intuitive is the end-to-end distance [1, 2, 3, 4]. The

end-to-end distance is the spatial distance between the first and the last monomer. It is the most important parameter describing the polymer size.

If the position of monomer i is $\vec{P}_i$, the end-to-end vector $\vec{R}_e$ is in equation (2.2).

$$\vec{R}_e = \vec{P}_N - \vec{P}_1 = \sum_{i=1}^{N-1} \vec{r}_i \tag{2.2}$$

The mean value of end-to-end vector is trivial, because $\left\langle \vec{R}_e \right\rangle = \sum_{i=1}^{N-1} \langle \vec{r}_i \rangle = \vec{0}$. The last step is from equation (2.1). There is a related parameter with a non-trivial mean value called the mean squared end-to-end distance. The mean squared end-to-end distance $\langle R_e^2 \rangle$ is in equation (2.3).

$$\left\langle R_e^2 \right\rangle = \left\langle \left(\sum_{i=1}^{N-1} \vec{r}_i \right)^2 \right\rangle = \sum_{i=1}^{N-1} \left\langle \vec{r}_i^{\,2} \right\rangle \tag{2.3}$$

This equation is fulfilled because we have the condition in equation (2.1) and consequently we have $\langle \vec{r}_i \vec{r}_j \rangle = 0$ for $i \neq j$. If the connection length between the monomers are constant $\|\vec{r}_i\| \equiv l$, we further have $\langle R_e^2 \rangle = Nl^2$ with the number of monomer N. Here $\|x\| \equiv l$ denotes the norm of x.

The most common one is the averaged end-to-end distance $\langle R_e \rangle$ which is sometimes just named end-to-end distance R_e. It refers to the averaged value of end-to-end vector norm, i.e., R_e refers to $\left\langle \left\| \vec{R}_e \right\| \right\rangle$. The behavior of R_e is very close to the square root of the mean squared end-to-end distance $\langle R_e^2 \rangle^{1/2}$ [5].

In practice, the end-to-end distance R_e is often easy to be accessed because it can be directly calculated from the end-to-end vector $\vec{R}_e$, which requires only the positions of two monomers $\vec{P}_N$ and $\vec{P}_1$. Due to its convenient accessibility, the end-to-end distance is widely used as a size measurement for polymer chains. On the other hand, the end-to-end distance does not provide information about the details inside the chain, which increases the possibility of generating bias for complex chain models.

Another quantity to describe the chain size, which incorporates the details inside, is the radius of gyration. The squared radius of gyration R_g^2 is defined by equation (2.4). And the radius of gyration can be computed by $R_g = \sqrt{R_g^2}$.

$$R_g^2 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\vec{P}_i - \vec{P}_j)^2 \tag{2.4}$$

For a random walk with constant connection length l, the radius of gyration is proportional to the mean squared end-to-end distance by $R_g^2 \approx \frac{1}{6} Nl^2 = \frac{1}{6} \langle R_e^2 \rangle$. This is to be expected since both quantities describe the same physical property, namely the polymer size.

Self-Avoiding Walk

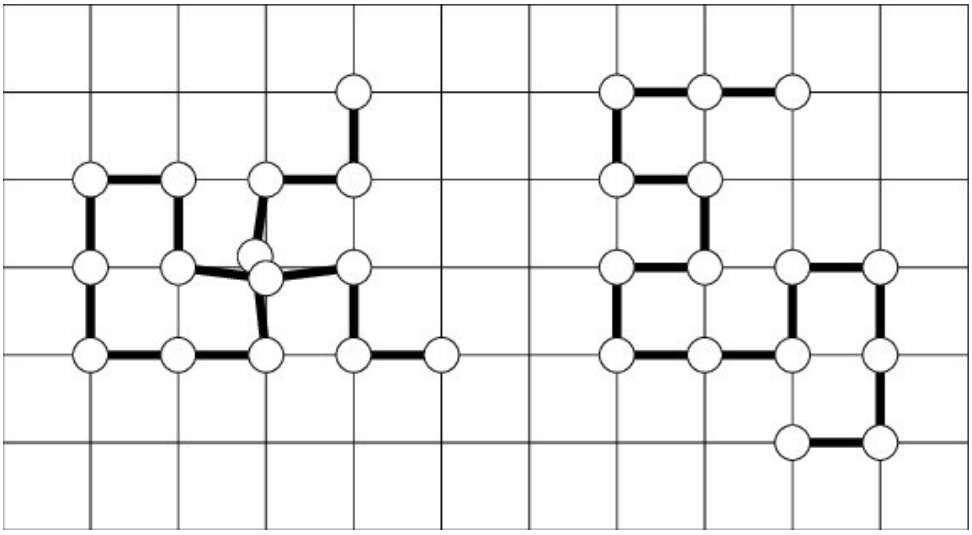


Figure 2.1: Example of a random walk (left) and a self-avoiding walk (right). Only the self-avoiding walk satisfies excluded volume condition. The figure is from [6].

The random walk has succeeded in many areas and has been the basic theoretical model for multiple decades [7]. However, it contains no information about all kinds of interactions inside a natural polymer chain. The most important interaction that is unavoidable is the excluded volume effect. If one extends a random walk to incorporate the excluded volume effect, it is a self-avoiding walk [8, 9].

The excluded volume effect denotes that any two monomers can not be overlapped with each other to occupy the same space. When we add the excluded volume effect into the monomers of the random walk, the monomer becomes a so-called hard-sphere, and the whole model becomes the beads-on-string model. For the beads-on-string model, the potential $U(d)$ between two monomers can be expressed as equation (2.5).

$$U(d) = \begin{cases} \infty, & d < d_0 \\ 0, & d > d_0 \end{cases} \tag{2.5}$$

where d is the distance between two monomers and d_0 is the radius of the hard-sphere. In this way, d_0 also defines the length of the excluded volume effect. In addition, there are other forms of potential that include the excluded volume effect, such as the Lennard-Jones potential.

Semi-Flexible Chain and Other Models

Another model significantly expands the scope of research is the semi-flexible chain model. The semi-flexible chain model incorporates another critical property, bending stiffness. When a chain has no bending stiffness during deformation, it is called a flexible chain. Both the random walk and the self-avoiding walk are flexible chains. The bending stiffness of a semi-flexible chain lies between a flexible chain and a rigid body. Specifically, it allows the chain to resist a certain bending force on a small scale while remaining flexible on a large scale due to entropy. The combined result of the bending stiffness and the entropy exhibits several valuable features [10, 11, 12].

One of the famous models of the semi-flexible chain is the worm-like chain. The worm-like chain is a continuous chain described by its position vector $\vec{r}(s)$. s is the contour distance along the chain. If the total length of the chain is l_n, then $s \in (0, l_n)$. At any point along the chain, a tangent vector $\vec{t}(s)$ can be defined by $\vec{t}(s) = \frac{\partial \vec{r}(s)}{\partial s}$. The bending stiffness can be introduced by the bending energy E_b through equation (2.6).

$$E_b = \frac{1}{2} k_b T \int_0^{l_n} l_p \left(\frac{\partial \vec{t}(s)}{\partial s} \right)^2 \mathrm{d}s \tag{2.6}$$

In this equation, $k_b T$ is the Boltzmann factor, and l_p is a stiffness parameter called persistence length. Here, the persistence length l_p is the central parameter quantifying the bending stiffness, but other similar parameters are possible.

The self-avoiding walk and the semi-flexible chain are two crucial models in this thesis. Apart from these models, models with other characteristics have also been developed. An an example, there are models with complex monomers like polymer rings [13, 14, 15] or compressible soft-spheres [16]. Moreover, the organizational form of models is not restricted to the chain structure. There are also other structures like polymer network [17].

Exponential Behavior

Scale-free is an essential feature that many polymer models have demonstrated. It identifies whether a polymer chain is independent of the scale. In mathematics, if a function takes the form of a power law, it retains its own in the scale transformation. A polymer chain is independent of the scale if its crucial properties obey a power law against the change of its polymer size. For a scale-free polymer chain, the exponent of the power law is a critical parameter that identifies the chain. In other words, a model can be said to have exponential behavior if this type of exponent exists [18].

As an example, for a self-avoiding walk with fixed connection length, the mean squared end-to-end distance $\langle R_e^2 \rangle$ obeyed:

$$\left\langle R_e^2 \right\rangle \propto N^{2\nu} l^2,\ N \to \infty \tag{2.7}$$

Here ν is called the Flory exponent, N is the number of monomers, and l is the length of the connection [19]. For the self-avoiding walk in a three-dimension cubic lattice, the value of ν is estimated to be $\nu = 0.587597(7)$ [20].

The value of ν is invariant even if we switch to another physical parameter such as the squared radius of gyration R_g^2. Therefore the value of ν is of great importance and has become intensively studied. Furthermore, the exponential behavior authorizes us to analyze patterns of a system on different scales.

Additionally, there is also stretched exponential behavior reported in complex systems like glassy disordered systems [21, 22, 23].

Lennard-Jones Potential

If we consider interactions within a polymer chain, it is inevitable to involve potential between monomers. The potential determines whether a monomer is a soft-sphere, a hard-sphere, or something else.

The most commonly used intermolecular potential is the Lennard-Jones potential. It has a simple form but behaves similarly to the interactions between molecules[24].

The normal expression for Lennard-Jones potential is:

$$V_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{2.8}$$

where $V_{LJ}(r)$ is the Lennard-Jones potential with respect to inter-monomer distance r, ϵ is the amplitude parameter, and σ is the scale parameter.

The Lennard-Jones potential has an attractive part and a repulsive part. Its minimum is at $r_m = 2^{1/6}\sigma$. On the left side of the minimum, it is the repulsive part because the first term is overwhelming; on the right side, it is attractive because the second term is overwhelming. If no temperature or other fluctuation term is included, two monomers will eventually rest with a separation of r_m. However, if the temperature is included, it becomes possible for the monomers to escape from the potential well. The energy cost of this process is defined by ϵ, which is the depth of the potential well.

Log-Logistic Distribution

The log-logistic distribution is a continuous probability distribution used in survival analysis, hydrology, and economics. It is also called Fisk distribution [25]. Compared

to the Gaussian distribution, the log-logistic distribution has a longer tail, which gives it a special position for handling non-Gaussian samples.

Its probability distribution function can be expressed as:

$$f(x|\alpha,\beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{\left(1+(x/\alpha)^{\beta}\right)^2} \tag{2.9}$$

In this form, its have two crucial parameters α and β. Here α is the scale parameter, and β is the shape parameter.

2.2 Chromosome Architecture

Chromosome

DNA (DeoxyriboNucleic Acid) is the central molecule in a cell system because it contains the main instructions for the development, survival, and reproduction of an organism. In eukaryotic cells, DNA normally combines with specific proteins called histones to form a complex named chromatin, which could fold into a characteristic structure called chromosome [26].

Electron microscopic observation reveals that chromatin is separated into regions of different brightness corresponding to different densities; the dark regions are called heterochromatin, and the light regions are called euchromatin [27].

Heterochromatin and euchromatin are two main categories of higher-order chromatin structures. Heterochromatin usually has a condensed structure and is inactive for transcription and regulation, whereas euchromatin has a loose structure and is more active [28]. However, in a recent study of nucleosome organization, chromatin was found to have many features that can not be described by the heterochromatin-euchromatin classification. To accurately represent the underlying structure, a better classification is needed [29].

Nowadays, we have more information to understand the chromatin structure. One breakthrough in recent years was chromosome territories. It was observed that different chromosomes tend to uniquely occupy particular regions in the nucleus; these subdomains are called chromosome territories [30]. Moreover, we also observed the topologically associating domain on a smaller scale. Within the topologically associating domains, self-interactions are favored, i.e., the inner sequences interact less frequently with the outer sequences [31, 32, 33].

In addition to the above discoveries, there are also intensive studies of long-range DNA contacts [34, 35] and chromatin properties, e.g. elasticity [36] and flexibility [37, 38].

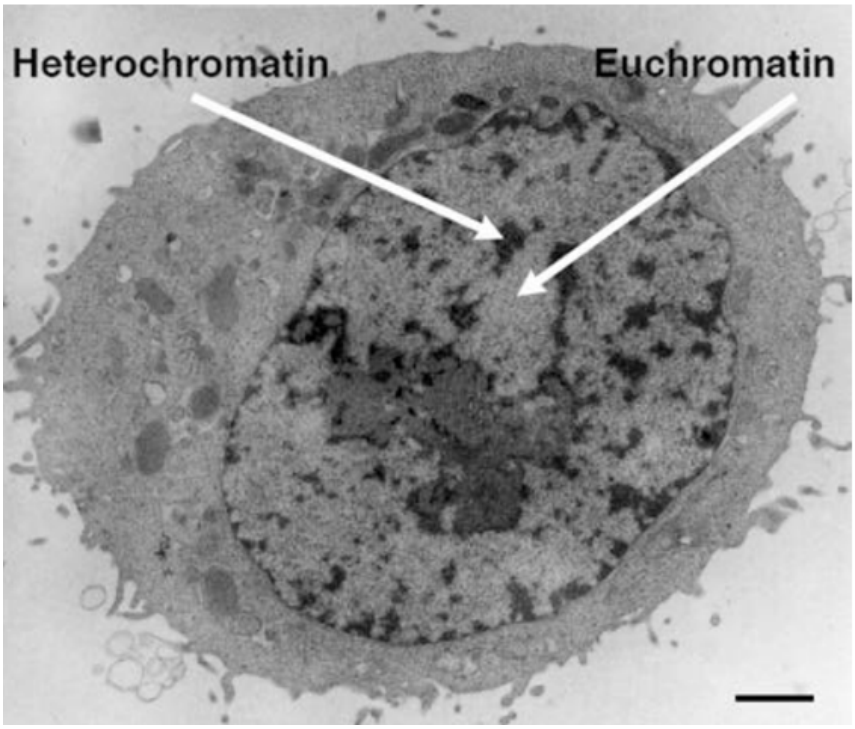


Figure 2.2: Heterochromatin (dark regions) and euchromatin (bright regions) observed by electron microscopy. This figure is from [27].

With all the above advances, chromosomes are no longer viewed as a complex of arbitrary entanglement but rather as a delicately organized structure with specifically selected functionality. Consequently, people have started to focus on genome organization intensively and to regard the chromatin system as a three-dimensional architecture [39, 40, 41, 42].

Nucleosome

The nucleosome is the basic structural unit of the chromosome. A nucleosome consists of approximately 147 base pairs of DNA and a histone octamer formed by eight histone proteins. In a nucleosome, DNA wraps around a histone octamer about 1.65 times [43]. In the nucleus, DNA is often compacted within a nucleosome to allow the genome to fold into a condensed structure. The nucleosome is the elementary factor in the research of chromosome structure [44] because it significantly influences gene expression and most DNA-related processes [45].

Nucleosome occupancy and nucleosome positioning are two important parameters for analyzing nucleosome behavior. Nucleosome occupancy is the parameter quantifying how frequently a DNA sequence wraps around a histone octamer to form

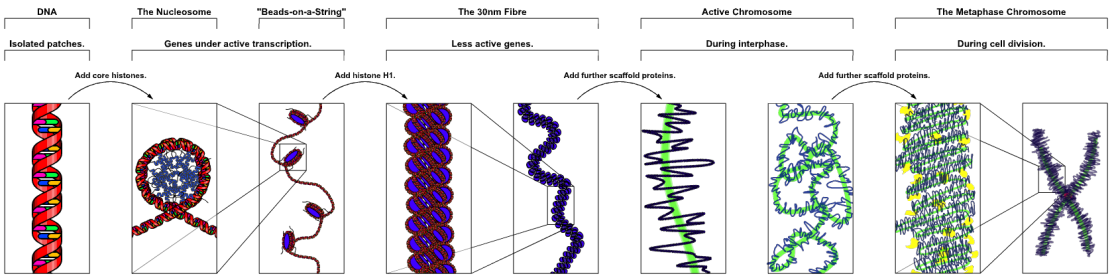


Figure 2.3: The organization of chromatin at different length scales. The figure is from [30].

a nucleosome. Nucleosome positioning indicates where nucleosomes are located in the genomic DNA sequence [45]. There are several well-implemented toolkits for computing the nucleosome positioning from experimental data, e.g., NPS (Nucleosome Positioning from Sequencing) [46], nucleR [47], DANPOS (Dynamic Analysis of Nucleosome Position and Occupancy by Sequencing) [48], and iNPS (improved Nucleosome-Positioning from Sequencing) [49]. There are also platforms for nucleosome positioning prediction [50, 51, 52].

The nucleosome positioning is determined by multiple complex factors [45, 53, 54]. DNA sequence preference is the most common one. Several papers have demonstrated that sequence preference plays a central role in the nucleosome organization [55, 56, 57]. However, recent research found a lack of universal sequence-dictated nucleosome positioning pattern [58], suggesting that characteristic mechanisms within the system may play a significant role [59]. Therefore, researchers are beginning to unravel the genomic organization patterns [60] and propose the organization principle as the mechanical code [61, 62].

Chromosome Conformation Capture Techniques

The continuous discovery in genome architecture is largely related to advances in experimental technology, and the most important one in recent decades has been massively parallel sequencing. Massively parallel sequencing platforms have enabled sequencing of 1 million to 43 billion short reads per instrument run [63]. Massively parallel DNA sequencing has been used to develop a series of conformation capture techniques for chromosomes. The 3C technique (short for Chromosome Conformation Capture) is a technique for detecting the spatial linkage of DNA within chromatin [64]. The Hi-C technique, an extension of 3C, is capable of simultaneously detecting all linkages at the chromosome level and providing us with genome-wide interaction maps [65, 66].

The ChIP-seq technique is massively parallel DNA sequencing combined with

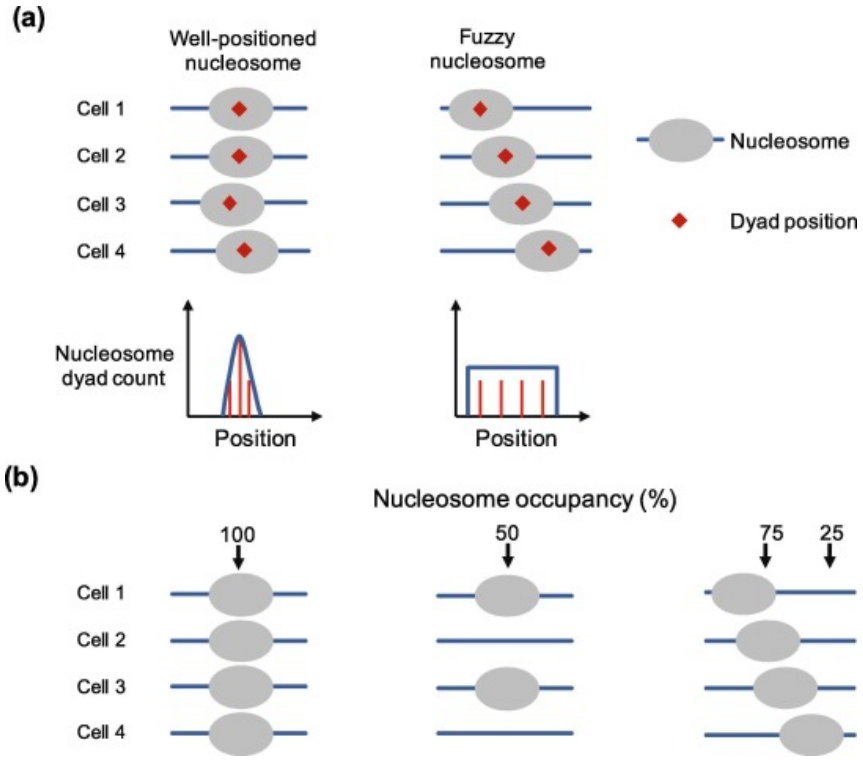


Figure 2.4: The nucleosome positioning and nucleosome occupancy. The figures are from [29].

chromatin immunoprecipitation. It provides access to DNA-protein interactions [67]. In particular, when combined with the micrococcal nuclease digestion, it is the MNase-seq technique. The MNase-seq technique measures specifically the nucleosome occupancy, which grants it an essential place in the research of chromosome architecture [68].

2.3 Computational Techniques

Monte Carlo Method

The best known Markov chain Monte Carlo method is the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm will generate a sample of a given problem defined by its objective function $f(x)$, which can be a specific simple function or an unknown function.

The steps of the algorithm are as follows:

1. Define the problem as objective function $f(x)$. And generate an initial sample

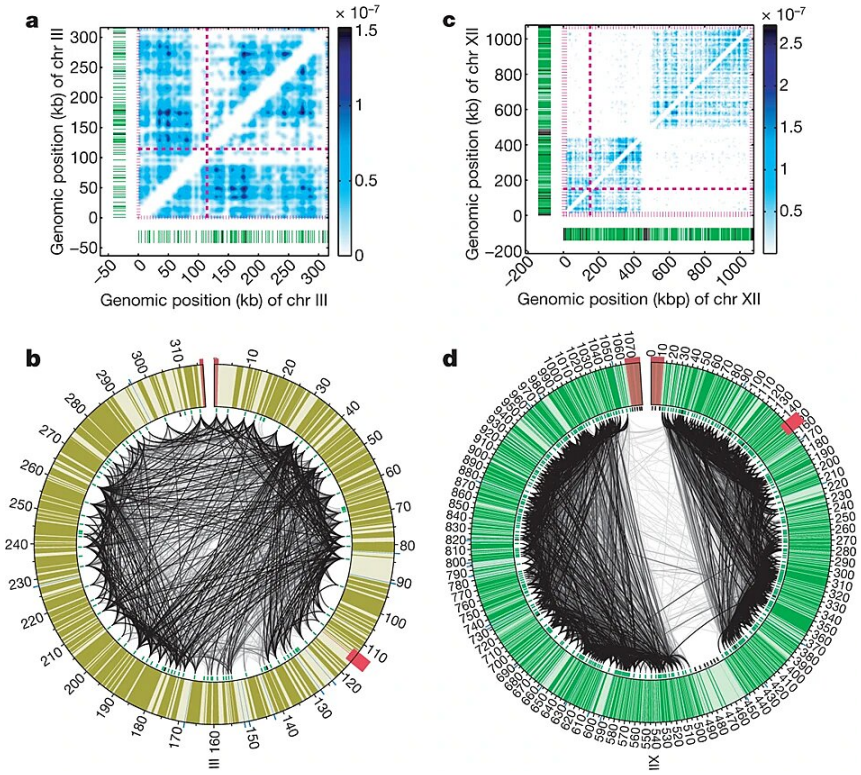


Figure 2.5: Complicated intra-chromosomal interaction maps from the Hi-C experiment. **a** and **c** are the heat maps. **b** and **d** are the Circos diagrams. The figures are from [42].

x_0 either from an arbitrary choice or from prior information. The initial sample now becomes the current sample $x = x_0$.

2. Propose a new sample x' according to the current sample x by a chosen function $g(x'|x)$.

3. Accept or reject the new sample with an acceptance ratio α. α can be computed via equation (2.10). If it is accept $x = x'$ otherwise x remains unchanged.

$$\alpha = \min\left(1, \frac{f(x')\, g(x|x')}{f(x)\, g(x'|x)}\right) \tag{2.10}$$

4. Record the result $x_t = x$ and repeat steps 2 and 3. t is the number of repetitions.

Finally $\{x_0, x_1, x_2, ...\}$ become a sample and the targeted parameters can be estimated through the sample. In order to make the sample independent of the initial

condition, the first part of the sample is usually dropped. The total number of repetitions is the Monte Carlo steps.

Reverse Monte Carlo

The reverse Monte Carlo (RMC) method is a double loop nested Monte Carlo simulation. It is an optimization method specifically designed to find the best conformation of a polymer that satisfies existing constraints.

The RMC method includes an MC simulation in the inner loop and an MC simulation in the outer loop. In the inner loop, a regular MC simulation is performed to calculate the required parameters. In the outer loop, it contains a Markov chain Monte Carlo (MCMC) simulation where each step runs through an entire cycle of the inner loop. After each step, the obtained parameters are inserted into the existing constraints to create a score of goodness. The sign and quantity of this score serve as feedback for the next step of the outer loop. Then the outer loop evolves step by step to reach the optimum [69, 18].

The RMC method has proved successful in many cases, such as sodium chloride solution [69]. However, it has the disadvantage that the algorithm may not converge for a complex system. This problem drives us to design a better algorithm on top of it.

Genetic Algorithm

Genetic algorithm (GA), which belongs to the larger class of evolutionary algorithms, is a family of numerical optimization methods. It is inspired by biological principles, for example, crossover, mutation, and selection. Due to its special process, genetic algorithms can generate high-quality solutions to non-linear optimization problems [70]. The process of genetic algorithm includes the following steps:

1. Initialization: At the beginning, a collection (called population) of the potential solutions (called individuals) to the targeted problem is randomly generated. Normally a population contains several hundred individuals. Each individual is presented with a series of numbers.

2. Selection: For each individual, a score is evaluated via a fitness function. The fitness function evaluates how close an individual is to the target. All individuals are selected according to the value of this function such that the better individual is statistically preferred. After this step, the unselected individuals are deleted.

3. Crossover: The existing individuals are combined to create new individuals. Old individuals are called parents, and a new individual is a child. One

crossover method is to select a pair of parents for each new child and generate a random position for the series of numbers, then create the left part of the child by copying from one of the parents and the right part by copying from another.

4. Mutation: All numbers of individuals have the possibility to change into another random value. The probability is called the mutation ratio, which should be a small number. One mutation method is that it directly goes through each number in the individuals; if the mutation occurs, the number is flipped (in binary case) or randomly changed to a different value (in non-binary case); otherwise, the number remains. After that, the child becomes the new population.

5. Termination: Steps 2-4 are repeated until certain criteria are fulfilled. Typical criteria are, for example, that the best solution remains unchanged over a long period of time or that the allocated computational budget is reached.

The above algorithm shows that GA is a highly non-linear algorithm, which makes it well-known for solving non-linear problems, especially for optimization with non-negligible noise. However, it also has some drawbacks. For example, all its solutions can only be evaluated by comparing them with other known solutions. Therefore, the algorithm will never return a solution as an absolute optimum.

Classification and Clustering in Machine Learning

From the perspective of data analysis, all information about the system, including our observed measurements, forms the data. If we have infinite information about a system, we can know everything directly, and no analysis is required. However, this is impossible in a real situation. The actual situations are usually one of the following three types.

If we have information about essential parameters and organization mechanisms, e.g., we know the Hamiltonian function, we can get the desired information by calculations or simulations via partial differential equations. If we do not know the mechanisms and cannot write down the Hamiltonian but have crucial data, we can perform data analysis techniques or detect their patterns by traditional machine learning methods. If we have very little information, probably only some disorganized data with considerable noise, deep learning might be the best option. Of course, the final choice of strategy depends on the specific problems.

Currently, the genomic system discussed in this thesis is assumably in the second case. Therefore, we focus on the analysis of crucial parameters and the detection of patterns by machine learning.

Classification and clustering are two major subjects in machine learning. Both divide samples according to their measurable properties named features. Clustering aims to group the sample into clusters according to their similarities, and classification aims to assign the samples to labeled classes.

In machine learning, clustering is part of unsupervised learning. Unsupervised learning usually works with unlabelled data and produces results without having prior information. Classification belongs to supervised learning. In supervised learning, we have labeled data and know the final categories[71].

There are a variety of clustering categories, such as connectivity-based clustering, centroid-based clustering, distribution-based clustering, and density-based clustering. In this section, we introduce k-means clustering as a centroid-based clustering. Then we introduce hierarchical clustering, which is also called connectivity-based clustering.

Distance Between Vectors

Accurately measuring the similarity between two functions or two samples is a fundamental problem that often arises in data analysis. The most intuitive method is to consider them as two multi-dimensional vectors and compute the distance between them, namely the norm of the difference. There are a variety of methods to calculate the norm for a vector. Here we present two basic ones, the p-norm and the cosine distance.

The distance $d_p(a, b)$ between two functions a and b according to the definition of the p-norm is:

$$d_p(a, b) = \|a - b\|_p = \left(\sum_{i=1}^{d} |a_i - b_i|^p \right)^{1/p} \tag{2.11}$$

In the equation, d is the dimension of a and b, and $\|\cdot\|_p$ denotes the p-norm. The most famous distance, the Euclidean distance, is equivalent to the 2-norm $\|\cdot\|_2$ with $p = 2$.

The cosine distance $d_{cos}(a, b)$ of function a and b is defined as:

$$d_{cos}(a, b) = 1 - \frac{a \cdot b}{\|a\|_2 \|b\|_2} \tag{2.12}$$

Rigorously, the cosine similarity d_{cs} should be the second term of equation (2.12), and its relation to the cosine distance is $d_{cos} = 1 - d_{cs}$.

K-Means Clustering

The k-means clustering is one of the most famous clustering algorithms, which is easy to execute and has excellent performance. In the k-means algorithm, the number of clusters k is predefined according to the requirements of the target problem. After that, the k-mean algorithm can be described as follows:

1. Initialization: k initial points $\{p_n, n = 1, 2, 3...k\}$ are generated. Normally they are randomly picked from the entire feature space. The k points are defined as initial centroids. Each p_n defines a cluster μ_n.

2. Assignment: All data points x_i are assigned to their nearest centroid. Data point $x_i \in \mu_n$ if $d(x_i, p_n) \leq d(x_i, p_m)$ for all $m = 1, 2, 3...k$, where $d(a, b)$ is the distance between a and b.

3. Update: Each centroid is updated to be the point that minimizes the distances to all points in the cluster. The new centroids p_n are:

$$p_n = \frac{1}{|\mu_n|} \sum_{x_i \in \mu_n} x_i \tag{2.13}$$

   where $|\mu_n|$ is the number of points in cluster μ_n.

4. Termination: Steps 2 and 3 are repeated until the termination condition is fulfilled. Termination conditions can be that all clusters μ_n do not change or the distance between the new and the old centroids is smaller than a threshold θ.

In k-means clustering, the definition of distance plays an important role. If the Euclidean distance (2-norm) is applied, i.e. the distance metric is $d(x) = \|x\|_2$, it is a standard k-means algorithm. If the cosine distance $d(x) = d_{cos}(x)$ is applied, it is spherical k-means.

Hierarchical Clustering

Hierarchical clustering is connectivity-based clustering. It has the distinct advantage of utilizing all valid measurements. Hierarchical clustering involves a variety of algorithms. For example, agglomerative hierarchical clustering starts from a distance matrix for all samples and recursively merges pairs of clusters until only one cluster remains. During this process, a dendrogram is generated. By cutting the dendrogram at a certain level, clustering is achieved.

In agglomerative hierarchical clustering, the calculation of the distance matrix is the central part of the algorithm. The distance matrix is the matrix that contains

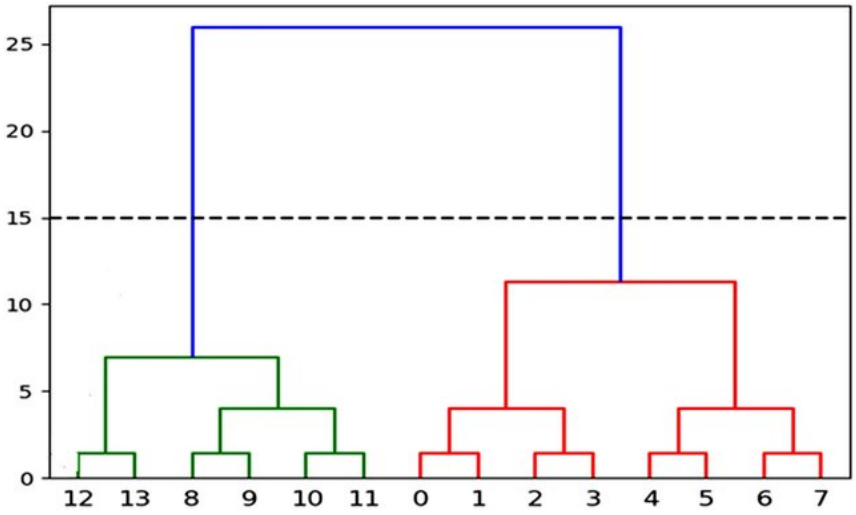


Figure 2.6: Example of a dendrogram illustrating the hierarchical clustering result. The x-axis shows the indices of the samples; the y-axis shows the branch length of the dendrogram. Cutting at the horizontal line gives a result with 2 clusters. The figure is from [72].

all distances between all pairs of samples or clusters. Apparently, the definition of distances has a strong influence on the result. The definition of distances includes the definition of the distance between a pair of samples, i.e., the metric, and the definition of the "distance" between a pair of clusters, which is the so-called linkage criteria. Possible alternatives for metrics can be the Euclidean distance and the cosine distance. There are also a lot of possible alternatives for linkage criteria. For example, in single-linkage clustering, the distance $d(\mu, \nu)$ between clusters μ and ν is $d(\mu, \nu) = \min\{d(x, y) : x \in \mu, y \in \nu\}$, where $d(x, y)$ is the distance between samples x and y, and in complete-linkage clustering, the distance is $d(\mu, \nu) = \max\{d(x, y) : x \in \mu, y \in \nu\}$ [73].

In this thesis, we adopt Ward's method, also called Ward's clustering or Ward's minimum variance method, because it is less susceptible to noise and outliers.

The metric of Ward's method is Euclidean distance $\|\cdot\|_2$. At the core of Ward's method, a parameter called the sum of squares is defined. The sum of squares S_μ for a cluster μ is:

$$S_\mu = \sum_{x \in \mu} \|x - m_\mu\|_2^2 \tag{2.14}$$

where m_μ is the center for the cluster μ. And m_μ is:

$$m_\mu = \frac{1}{|\mu|} \sum_{x \in \mu} x \tag{2.15}$$

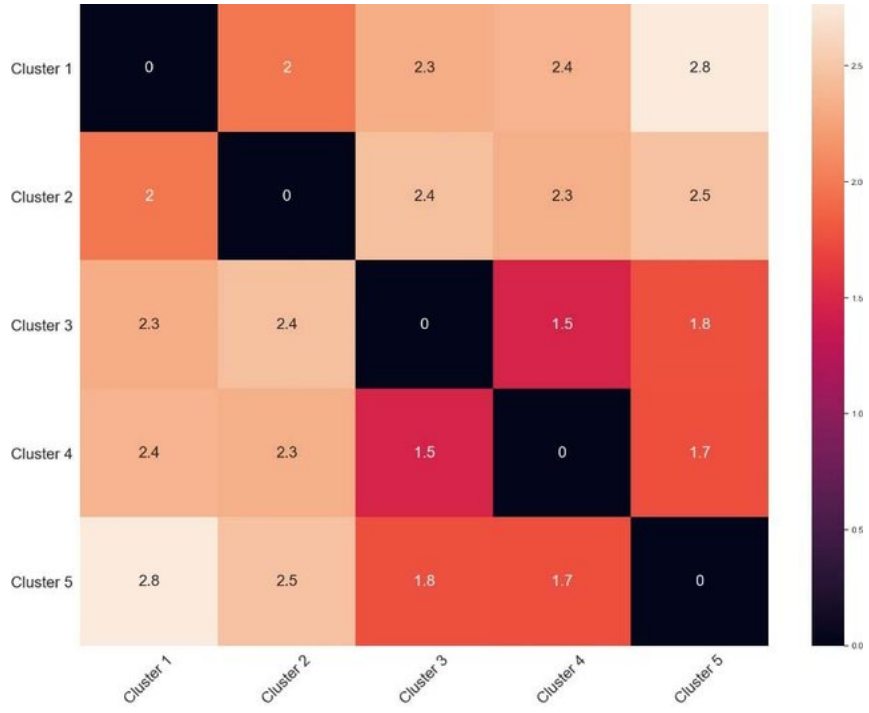


Figure 2.7: Example of a distance matrix. The entries are the distances between pairs of clusters. The figure is from [74].

where $|\mu|$ is the number of samples in cluster μ. On top of the above definition, the linkage criteria for Ward's method is

$$d(\mu,\nu)^2 = S_{\mu\cup\nu} - S_\mu - S_\nu = \frac{|\mu||\nu|}{|\mu|+|\nu|}\|m_\mu - m_\nu\|_2^2 \tag{2.16}$$

In this equation, if μ and ν each contains only one sample, $d(\mu,\nu)$ is equal to $\frac{\sqrt{2}}{2}$ of the Euclidean distance between those two samples [75].

It is more convenient to write the recursive form, which is equation (2.17). In the equation, a new cluster $\mu\cup\nu$ is created and its distance to an old cluster τ can be computed through old distances $d(\mu,\tau)$, $d(\nu,\tau)$, and $d(\mu,\nu)$. Besides, the denominator T is $T = |\mu| + |\nu| + |\tau|$.

$$d(\mu\cup\nu,\tau)^2 = \frac{|\mu|+|\tau|}{T}d(\mu,\tau)^2 + \frac{|\nu|+|\tau|}{T}d(\nu,\tau)^2 - \frac{|\tau|}{T}d(\mu,\nu)^2 \tag{2.17}$$

Equation (2.17) and equation (2.16) are equivalent in calculating the distances. However, in the computational implementation, equation (2.16) requires that the

initial distances between pairs of samples are $\frac{\sqrt{2}}{2}$ of the Euclidean distance, while equation (2.17) only requires that the initial distances are proportional to the Euclidean distance with the same constant.

Now the algorithm of Ward's method can be expressed as follows:

1. Prepare the Euclidean distance matrix for all samples and regard each sample as a new cluster with only one element.

2. Find the minimum in the distance matrix and the corresponding pair (or pairs) of clusters.

3. Merge the found clusters into one cluster to form a clustering and compute the level of the clustering. For found clusters μ and ν, the level of clustering is $L(\mu \cup \nu) = d(\mu, \nu)$, and $d(\mu, \nu)$ can be computed by the linkage criteria in equation (2.17).

4. Compute the distances between the new cluster $\mu \cup \nu$ and all old clusters. Then update the distance matrix.

5. Repeat steps 2-4 until only one cluster exists.

After the iterations are completed, a dendrogram can be generated by setting the branch length of the dendrogram to be half of the clustering level $L(\mu \cup \nu)$. If the dendrogram is cut at a distance according to certain conditions, e.g., the required number of clusters, clustering from the Ward method is accomplished.

References

[1] Jacob Mazur. "Distribution function of the end-to-end distances of linear polymers with excluded volume effects". In: *Journal of research of the National Bureau of Standards. Section A, Physics and chemistry* 69.4 (1965), p. 355.

[2] DS McKenzie. "The end-to-end length distribution of self-avoiding walks". In: *Journal of Physics A: Mathematical, Nuclear and General* 6.3 (1973), p. 338.

[3] Victor Dotsenko. "Distribution function of the endpoint fluctuations of one-dimensional directed polymers in a random potential". In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.02 (2013), P02012.

[4] S Redner. "Distribution functions in the interior of polymer chains". In: *Journal of Physics A: Mathematical and General* 13.11 (1980), p. 3525.

[5] R Everaers, IS Graham, and MJ Zuckermann. "End-to-end distance distributions and asymptotic behaviour of self-avoiding walks in two and three dimensions". In: *Journal of Physics A: Mathematical and General* 28.5 (1995), p. 1271.

[6] Gerard T Barkema. "Project II: Monte Carlo Simulation of Lattice Polymer Models". In: *Instructions for this project* (2010).

[7] Révész Pál. *Random Walk in Random and Non-random Environments*. World Scientific, 1990.

[8] Wyatt Hooper and Alexander R Klotz. "Trapping in self-avoiding walks with nearest-neighbor attraction". In: *Physical Review E* 102.3 (2020), p. 032132.

[9] S Havlin and D Ben-Avraham. "New approach to self-avoiding walks as a critical phenomenon". In: *Journal of Physics A: Mathematical and General* 15.6 (1982), p. L321.

[10] Jan Wilhelm and Erwin Frey. "Radial distribution function of semiflexible polymers". In: *Physical review letters* 77.12 (1996), p. 2581.

[11] JK Bhattacharjee, D Thirumalai, and JD Bryngelson. "Distribution function of the end-to-end distance of semiflexible polymers". In: *arXiv preprint cond-mat/9709345* (1997).

[12] Jan Kierfeld et al. "Semiflexible polymers and filaments: From variational problems to fluctuations". In: *AIP Conference Proceedings*. Vol. 1002. 1. American Institute of Physics. 2008, pp. 151–185.

[13] Thomas Vettorel, Alexander Y Grosberg, and Kurt Kremer. "Statistics of polymer rings in the melt: a numerical simulation study". In: *Physical biology* 6.2 (2009), p. 025013.

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

[15] Miriam Fritsche and Dieter W Heermann. "Confinement driven spatial organization of semiflexible ring polymers: Implications for biopolymer packaging". In: *Soft Matter* 7.15 (2011), pp. 6906–6913.

[16] Giuseppe D'Adamo, Andrea Pelissetto, and Carlo Pierleoni. "Polymers as compressible soft spheres". In: *The Journal of Chemical Physics* 136.22 (2012), p. 224905.

[17] Chase P Broedersz and Fred C MacKintosh. "Modeling semiflexible polymer networks". In: *Reviews of Modern Physics* 86.3 (2014), p. 995.

[18] Kurt Binder and Dieter Heermann. *Monte Carlo simulation in statistical physics*. Springer, Berlin, Heidelberg, 2010.

[19] Paul J Flory. *Principles of polymer chemistry*. Cornell university press, 1953.

[20] Nathan Clisby. "Accurate estimate of the critical exponent ν for self-avoiding walks via a fast implementation of the pivot algorithm". In: *Physical review letters* 104.5 (2010), p. 055702.

[21] Bingyu Cui, Rico Milkus, and Alessio Zaccone. "The relation between stretched-exponential relaxation and the vibrational density of states in glassy disordered systems". In: *Physics Letters A* 381.5 (2017), pp. 446–451.

[22] Katarzyna Górska et al. "The stretched exponential behavior and its underlying dynamics. The phenomenological approach". In: *Fractional Calculus and Applied Analysis* 20.1 (2017), pp. 260–283.

[23] Joseph D Paulsen and Sidney R Nagel. "A model for approximately stretched-exponential relaxation with continuously varying stretching exponents". In: *Journal of Statistical Physics* 167.3-4 (2017), pp. 749–762.

[24] Simon Stephan, Jens Staubach, and Hans Hasse. "Review and comparison of equations of state for the Lennard-Jones fluid". In: *Fluid Phase Equilibria* 523 (2020), p. 112772.

[25] Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons, 2003.

[26] Tanmoy Mondal et al. “Characterization of the RNA content of chromatin”. In: *Genome research* 20.7 (2010), pp. 899–907.

[27] Leanne De Koning. “Chromatin assembly factors and heterochromatin organization during cell prolifertaion, tumorigenesis and in quiescence”. PhD thesis. Université Pierre et Marie Curie-Paris VI, 2009.

[28] Y Murakami. “Heterochromatin and Euchromatin”. In: *Encyclopedia of Systems Biology* (2013), pp. 881–884.

[29] Ashish Kumar Singh and Felix Mueller-Planitz. “Nucleosome positioning and spacing: from mechanism to function”. In: *Journal of Molecular Biology* (2021), p. 166847.

[30] Thomas Cremer and Christoph Cremer. “Chromosome territories, nuclear architecture and gene regulation in mammalian cells”. In: *Nature reviews genetics* 2.4 (2001), pp. 292–301.

[31] Jesse R Dixon, David U Gorkin, and Bing Ren. “Chromatin domains: the unit of chromosome organization”. In: *Molecular cell* 62.5 (2016), pp. 668–680.

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

[33] Yanxiao Zhang et al. “Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells”. In: *Nature genetics* 51.9 (2019), pp. 1380–1388.

[34] Wouter de Laat. “Long-range DNA contacts: romance in the nucleus?” In: *Current opinion in cell biology* 19.3 (2007), pp. 317–320.

[35] Yanli Wang et al. “The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions”. In: *Genome biology* 19.1 (2018), pp. 1–12.

[36] Elena F Koslover et al. “Local geometry and elasticity in compact chromatin structure”. In: *Biophysical journal* 99.12 (2010), pp. 3941–3950.

[37] Kerstin Bystricky et al. “Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques”. In: *Proceedings of the National Academy of Sciences* 101.47 (2004), pp. 16495–16500.

[38] Philipp M Diesinger and Dieter W Heermann. “Depletion effects massively change chromatin properties and influence genome folding”. In: *Biophysical journal* 97.8 (2009), pp. 2146–2153.

[39] Peter Meister et al. “Visualizing yeast chromosomes and nuclear architecture”. In: *Methods in enzymology* 470 (2010), pp. 535–567.

[40] Yuwen Ke et al. “3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis”. In: *Cell* 170.2 (2017), pp. 367–381.

[41] Viviana I Risca et al. “Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping”. In: *Nature* 541.7636 (2017), pp. 237–241.

[42] Zhijun Duan et al. “A three-dimensional model of the yeast genome”. In: *Nature* 465.7296 (2010), pp. 363–367.

[43] Jakob Bohr and Kasper Olsen. “The size of the nucleosome”. In: *arXiv preprint arXiv:1102.0761* (2011).

[44] Andrew Routh, Sara Sandin, and Daniela Rhodes. “Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure”. In: *Proceedings of the National Academy of Sciences* 105.26 (2008), pp. 8872–8877.

[45] Kevin Struhl and Eran Segal. “Determinants of nucleosome positioning”. In: *Nature structural & molecular biology* 20.3 (2013), pp. 267–273.

[46] Robert Schöpflin et al. “Modeling nucleosome position distributions from experimental nucleosome positioning maps”. In: *Bioinformatics* 29.19 (2013), pp. 2380–2386.

[47] Oscar Flores and Modesto Orozco. “nucleR: a package for non-parametric nucleosome positioning”. In: *Bioinformatics* 27.15 (2011), pp. 2149–2150.

[48] Kaifu Chen et al. “DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing”. In: *Genome research* 23.2 (2013), pp. 341–351.

[49] Weizhong Chen et al. “Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data”. In: *Nature communications* 5.1 (2014), pp. 1–14.

[50] Bader A Alharbi et al. “nuMap: a web platform for accurate prediction of nucleosome positioning”. In: *Genomics, proteomics & bioinformatics* 12.5 (2014), pp. 249–253.

[51] Guo-Cheng Yuan et al. “Genome-scale identification of nucleosome positions in S. cerevisiae”. In: *Science* 309.5734 (2005), pp. 626–630.

[52] Vladimir B Teif et al. “Genome-wide nucleosome positioning during embryonic stem cell development”. In: *Nature structural & molecular biology* 19.11 (2012), pp. 1185–1192.

[53] Răzvan V Chereji and David J Clark. "Major determinants of nucleosome positioning". In: *Biophysical journal* 114.10 (2018), pp. 2279–2289.

[54] Marta Radman-Livaja and Oliver J Rando. "Nucleosome positioning: how is it established, and why does it matter?" In: *Developmental biology* 339.2 (2010), pp. 258–266.

[55] Andrew Travers et al. "The DNA sequence-dependence of nucleosome positioning in vivo and in vitro". In: *Journal of Biomolecular Structure and Dynamics* 27.6 (2010), pp. 713–724.

[56] Yin Shen et al. "A map of the cis-regulatory sequences in the mouse genome". In: *Nature* 488.7409 (2012), pp. 116–120.

[57] Assaf Weiner et al. "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging". In: *Genome research* 20.1 (2010), pp. 90–100.

[58] Anton Valouev et al. "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning". In: *Genome research* 18.7 (2008), pp. 1051–1063.

[59] Amanda L Hughes and Oliver J Rando. "Mechanisms underlying nucleosome positioning in vivo". In: *Annual review of biophysics* 43 (2014), pp. 41–63.

[60] Weizhong Chen et al. "Inter-nucleosomal communication between histone modifications for nucleosome phasing". In: *PLoS computational biology* 14.9 (2018), e1006416.

[61] Eran Segal et al. "A genomic code for nucleosome positioning". In: *Nature* 442.7104 (2006), pp. 772–778.

[62] Aakash Basu et al. "Measuring DNA mechanics on the genome scale". In: *Nature* 589.7842 (2021), pp. 462–467.

[63] Tracy Tucker, Marco Marra, and Jan M Friedman. "Massively parallel sequencing: the next big thing in genetic medicine". In: *The American Journal of Human Genetics* 85.2 (2009), pp. 142–154.

[64] Elzo De Wit and Wouter De Laat. "A decade of 3C technologies: insights into nuclear organization". In: *Genes & development* 26.1 (2012), pp. 11–24.

[65] Nynke L Van Berkum et al. "Hi-C: a method to study the three-dimensional architecture of genomes." In: *JoVE (Journal of Visualized Experiments)* 39 (2010), e1869.

[66] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *science* 326.5950 (2009), pp. 289–293.

[67] Isiaka Ibrahim Muhammad et al. "RNA-seq and ChIP-seq as complementary approaches for comprehension of plant transcriptional regulatory mechanism". In: *International journal of molecular sciences* 21.1 (2020), p. 167.

[68] David C Klein and Sarah J Hainer. "Genomic methods in profiling DNA accessibility and factor localization". In: *Chromosome Research* 28.1 (2020), pp. 69–85.

[69] Alexander P Lyubartsev and Aatto Laaksonen. "Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach". In: *Physical Review E* 52.4 (1995), p. 3730.

[70] Firas Gerges, Germain Zouein, and Danielle Azar. "Genetic algorithms with local optima handling to solve sudoku puzzles". In: *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*. 2018, pp. 19–22.

[71] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[72] Sara Kadhum Idrees and Ali Kadhum Idrees. "New fog computing enabled lossless EEG data compression scheme in IoT networks". In: *Journal of Ambient Intelligence and Humanized Computing* (2021), pp. 1–14.

[73] Gabor J Szekely, Maria L Rizzo, et al. "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method". In: *Journal of classification* 22.2 (2005), pp. 151–184.

[74] Valentas Gružauskas et al. "Application of multivariate time series cluster analysis to regional socioeconomic indicators of municipalities". In: *Real Estate Management and Valuation* 29.3 (2021), pp. 39–51.

[75] Annie Murphy Paul. *The cult of personality testing: How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstand ourselves*. Simon and Schuster, 2010.

Part II

RESULTS

Chapter 3

The Effect of Bending Rigidity on Polymers

References

This chapter is from:

- Jia, J., **Li, K.,** Hofmann, A., & Heermann, D. W. (2019). The effect of bending rigidity on polymers.

Kunhe Li and Jiying Jia developed the models and methods. Kunhe Li implemented the simulation program, generated the conformations, calculated the contact probability, and analyzed the effect of bending rigidity for both homogeneous and heterogeneous cases. Jiying Jia calculated the structure factor, estimated the persistence length, and analyzed the effects of confinements. Jiying Jia and Dieter W. Heermann drafted the manuscript and reviewer response. Andreas Hofmann supported the analysis and corrected the manuscript. Dieter W. Heermann supervised the project.

Macromolecular
Theory and Simulations
www.mts-journal.de

The Effect of Bending Rigidity on Polymers

Jiying Jia, Kunhe Li, Andreas Hofmann, and Dieter W. Heermann*

The conformations of chromatin are influenced by many factors. In the regulation of gene expression the bending rigidity of the chromatin polymer and its heterogeneity play an important role for the possible conformations. To elucidate this, the effect of bending rigidity as well as its heterogeneity on various polymer properties is investigated. In the context of chromatin organization, the contact probability is an important measure. It is analyzed whether there is any ambiguity in the definition of a contact. The results show that the contact probability does not depend on the range of contact in the limit of a large contour length between monomers. Further, the persistence length as a function of the bending rigidity is computed in the homogeneous and heterogeneous cases. The persistence length is systematically smaller in the heterogeneous case. Chromosomes are confined by each other in the nucleus and by looking at specific loci, the environment changes much more slowly than the local chromatin part. In conjunction with bending rigidity, polymers in rectangular confinements with several aspect ratios are simulated. Due to the spiraling behavior when the box size is small enough, an oscillation in the contact probability and the orientational correlation function is found.

1. Introduction

Contact probabilities are at center stage in current measurements of chromosome conformations.[1,2] In these experiments one measures the number of self-contacts of chromosomes as well as the inter-chromosomal contacts as these give topological information on the organization of chromosomes in space. Since the chromosomes are confined in the nucleus the question of the packaging and its influence on the intra-chromosomal contacts arises. Further, how is all of this influenced by the stiffness of chromatin?

Of vital importance to the biological function is the packaging of chromosomes.[3,4] First in line is the packing of DNA with the help of histone proteins to form the beads-on-string chain.[5,6] A further packaging is the 30 nm fiber (chromatin) and the packaging of the fiber into the nucleus.[7] Packing on the scale beyond 30 nm is mainly achieved by the dynamic formation of loops and higher order loop structures (loops of loops).[8] These build up local compartments of varying density which in turn build up to chromosome territories.[9,10] Thus the fiber cannot be assumed to be in free space and the kind of contacts that the chain can have with itself is largely influenced by two factors. First is the kind of local confinement the fiber finds itself in and second the bending rigidity of the fiber which for example is controlled by chromosome remodeling.[11,12]

The compartmentalization of the nucleus (such as in human cells) implies a confinement of the chain that is rather symmetric. For *Escherichia coli*, on the other hand, the confinement is rectangularly shaped and this confinement influences the interaction.[13] How does shape influence the contact probabilities?

There are several factors influencing the bending rigidity of a chromosome. For human chromosomes the existence of nucleosomes and their distribution along the backbone of the chain[14–16] imply a distribution of bending rigidity along the fiber. A further factor is the repulsion of the histone tails, that is, methylation.[17] Furthermore, histone H1 depletion has a great influence on the flexibility of the chain.[18,19]

All in all, chromatin is not totally flexible, that is, chromatin is a semiflexible polymer fiber with a distribution of bending rigidity. Moreover, in general, chromatin is heterogeneous, which means that the polymer could have different distributions of bending rigidity in different parts. This could originate from the genome sequence or the distribution of nucleosomes along the backbone.[20] Clearly this heterogeneity itself plays a significant role in chromatin organization as well[21] and has thus an influence on the contact probability.

Chromatin undergoes structural transformation, that is, conformational changes, to carry out biological functions properly. For a long chromatin chain, the overall conformation changes only slowly, while at smaller scale it changes much faster and is confined in a narrower space compared to the overall volume the chromatin occupies. Thus, in this paper, when dealing with polymers in confinement, we focus on rather short chains and investigate the role of bending rigidity and the size and aspect ratio of the confining volume.

The paper is organized as follows. In Section 2, we describe the two polymer models we are using and how we implement the heterogeneity of the bending rigidity with different distributions. The heterogeneity of polymer has been modeled in various ways.[22,23] Our model implements it via the variance

J. Jia, K. Li, A. Hofmann
Institute for Theoretical Physics
Philosophenweg 12, 69120 Heidelberg, Germany
E-mail: jia@thphys.uni-heidelberg.de
Prof. D. W. Heermann
Institute for Theoretical Physics
Philosophenweg 19, 69120 Heidelberg, Germany
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/mats.201800071.

DOI: 10.1002/mats.201800071

 © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

of the bending parameter κ along the chain. In Section 3 we present the results on our main questions. In Section 3.1 we address the question: What is a contact? This question arises both in the context of lattice polymers as well as in continuum. Further, how does the bending rigidity, especially its distribution, affect the contact? In Section 3.2, we focus on properties like persistence length and the structure factor and in Section 3.3, we address the question how the linear semi-flexible chain organizes in confinements of different sizes and shapes. We further investigate the influence of the heterogeneity of bending rigidity on this organization. For the second question, Fritsche[24] and Ostermeir[25] have studied the spatial organization of homogeneous stiff ring polymers in rectangular and weak spherical confinement separately. 2D linear semi-flexible polymers in confined space have been investigated by Liu.[26] Here we study how the stiffness and its distribution affect the conformation of linear polymers in 3D rectangular confinement with different sizes and aspect ratios. Finally, in Section 4, we present our conclusions.

2. The Model

The polymer model we employ is based on the self-avoiding walk. The bending rigidity is introduced as in the Kratky-Porod model, or the worm-like chain model in continuum. In the Kratky-Porod model, where the torsional energy is absent, the origin of stiffness of a polymer is the intrinsic bending energy H_b, which is the sum of energies of successive segments:

$$H_b = -\sum_{i=1}^{N-2} \kappa_i \mathbf{u_i} \cdot \mathbf{u_{i+1}} \tag{1}$$

where κ_i is the stiffness parameter, $\mathbf{u_i}$ is the normalized bond vector and N is the number of monomers. To model a heterogeneous chromatin chain having a variable bending rigidity along the chain, κ_i can be set to obey a distribution of interest, while $\kappa_i = \kappa$ for a homogeneous chromatin chain. When studying heterogeneous chains, we assume that κ_i obeys the Gaussian distribution with mean value $\langle\kappa\rangle$ and standard deviation σ.

The continuous version of the Kratky-Porod model is the worm-like chain model, where the persistence length l_p is defined through the exponential decay of the orientational correlation function:

$$\langle \mathbf{u}(s_1+s)\cdot\mathbf{u}(s_1)\rangle = \langle\cos\theta(s)\rangle = e^{-s/l_p} \tag{2}$$

Here $\mathbf{u}(s) = \dfrac{\partial \mathbf{r}(s)}{\partial s}$ is the unit tangent vector to the chain at contour distance s, and $\mathbf{r}(s)$ is the position vector along the chain. Although chains in a dense melt or at the Θ-point in solution behave like ideal chains without excluded volume effect, as the worm-like chain does, recently it was shown that the orientational correlation function for chains in these conditions shows a power law decay $s^{-3/2}$ instead of the above exponential decay for certain range of contour length $1 \ll s \ll N$.[27,28] For real chains Hsu et al.[29] have shown that the standard definition of persistence length does not describe the local "intrinsic" stiffness either, with $\langle\cos\theta(s)\rangle \approx s^{-\beta}$ for $1 \ll s \ll N$, β being a different power law exponent $\beta = 2(1-\nu) \approx 0.824$. However, the exponential decay fits well at short length scales s for simple linear chains without a complex architecture such as side chains, and it is capable of approximating the stiffness parameter κ fairly. In free space and for the homogeneous chain, the stiffness parameter $\kappa_i = \kappa$ is actually related to l_p defined in Equation (2) via $l_p \approx \kappa\overline{l_b}$ (where energy is measured in the units of k_BT). The deviation of l_p results from the discretization of the continuous worm like chain which makes l_p slightly smaller than κ, and the self-avoiding effect, which makes l_p larger compared to random walk. But the latter is negligible when κ is large enough. $\overline{l_b}$ is the averaged bond length. For the heterogeneous chain, the average persistence length over the entire chain is determined by the distribution ($\langle\kappa\rangle$ and σ for Gaussian distribution), which will be discussed in Section 3.2.

In this paper we use two models to perform the Monte Carlo simulation and study the questions defined in Section 1. First, when simulating very long chains in order to investigate the key question on the definition of a contact we employ a pivot algorithm based on the original idea of Sokal and Kennedy[30,31] in continuous space. There have been several applications of the continuous pivot algorithm in different polymer models. Adamo and Pelissetto[32] have implemented the off-lattice pivot algorithm to study the impact of the thickness of monomers, that is, the effectiveness of the excluded volume interaction, on the asymptotic behavior of polymer chains. Also, a continuous pivot algorithm with narrower choice of pivot angles is used to study the effects of macromolecular crowding on protein stability.[33] Horwath, Clisby, and Virnau[34] use the standard implementation of the pivot algorithm to investigate knots in finite memory walks where the excluded volume effects are considered only at short length scales.

In this algorithm, a pivot with a random pivot point on the chain and a random symmetry matrix is carried out at each Monte Carlo move, producing a global conformation change of the chain. This algorithm is highly efficient in that it reduces remarkably the relaxation time to reach the equilibrium state and de-correlates conformations much faster compared to algorithms based on local moves. Kennedy[31] proposed a faster implementation of the existing pivot algorithm for self-avoiding walks on a lattice, requiring a time $O(N_b^q)$ per accepted pivot with $q < 0.85$ for a 3D lattice instead of $O(N_b)$ for other pivot algorithms. N_b is the number of bonds ($N_b = N - 1$ for a linear chain). We extended this faster on-lattice pivot algorithm into a continuous one, each monomer being a hard sphere of radius $r = 0.4$. Furthermore, the bending energy is also implemented to simulate long semiflexible chains.

The second model is a lattice polymer model, specifically, we are using the Bond Fluctuation Model (BFM)[35] to simulate short linear chains of size up the $N = 160$ in cubic and rectangular confinement. The local "L6" move is used at each Monte Carlo move. These conformations are correlated due to the local moves. We calculate the autocorrelation time τ_{int} following the routine outlined in Sokal[30] based on the radius of gyration. We took conformations into account that are separated at least $2\tau_{int}$ Monte Carlo steps.[36] About 10 000-15 000 independent conformations were generated for each parameter set.

The autocorrelation time τ_{int} for longer and highly stiff chains can be extremely high. A combination of the local "L26" move

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

and a pivot move are employed within the BFM to simulate longer and stiffer bottle-brush polymers owing to the reduction of relaxation and autocorrelation time.[37] For polymer chains in confinement, the local moves are kind of indispensable because of the high rejection rate of global moves in finite space. In our case, the BFM with "L6" move is adequate to simulate short chains in cubic and rectangular confinement.

3. Results

3.1. Contact Probability of the Semi- and Flexible Chains

Does the contact probability as a function of the contour distances depend on the definition of what a contact is? On a lattice we can define a contact if two monomers occupy nearest neighbor sites. But then, we could also define a contact taking place at next nearest neighbor sites. In continuum we need to define a distance such that whenever two monomers are within the defined distance this would count as a contact. We refer to this defined distance as the cut-off distance d_c. **Figure 1**A shows the number of monomers which are in contact with monomer i (solid point). The number of contacts depends on the value of the cut-off distance d_c. Specifically, the contact probability with d_c fixed is calculated as follows: if the distance d_{ij} between monomer i and monomer j is smaller than d_c, then the contribution to the contact probability is $p_\kappa(|i-j|) = 1$ (κ is the stiffness parameter of the chain), otherwise $p_\kappa(|i-j|) = 0$. The entire contact probability is the average over all pairs of i, j and sufficient independent conformations in equilibrium: $P_\kappa(s) = \langle\langle p_\kappa(|i-j|)\rangle_{|i-j|=s}\rangle_c$ where s is the contour distance between monomers.

To establish the asymptotic behavior of the contact probability $P_\kappa(s)$ with respect to contour length s, we simulated long chains (see Figure 1B). In Figure 1B the result for a semiflexible homogeneous chain $N_b = 10000$, $\kappa = 10$ using the continuous pivot algorithm is shown. The contact probability of this chain is calculated with three cut-off distances $d_c = 1.1$, 2.7, 4.1. Only the range $s < 1000$ is shown since $P_\kappa(s)$ for bigger s has rather large statistical fluctuations. The results prove that how we define the cut-off distance d_c for the contact of monomers does not change the asymptotic behavior of the contact probability $P_\kappa(s)$. As long as the value of d_c is not too large compared to the persistence length, $P_\kappa(s)$ has a similar structure over all length scales: a minimum (only for relatively large κ, discussed later) when s is small, and the same power law decay (roughly $s^{-2.2}$) when $s \gg l_p$ (c.f. inset). Shown in the inset are the ratios of contact probabilities for $d_c = 2.7$ and 4.1 over $P_\kappa(s)$ for $d_c = 1.1$. When s is large enough, the ratios level out, showing that the contact probabilities have the same asymptotic behavior only with different prefactors.

The influence of bending rigidity and its distribution on the contact probability $P_\kappa(s)$ (the cut-off distance d_c is set to be 1.1) is shown in **Figure 2**, where results for chains of homogeneous (panel A) and heterogeneous stiffness (panel B) are presented. In panel A, the contact probabilities of semiflexible chains exhibit a drop in the range of small s compared to the flexible chain (blue line). This is because of the fact that bending energy contributes to the parallel of successive chain segments, inducing larger separation between monomers than flexible chains. A local minimum exists if κ is large enough (roughly $\kappa > 3$). When $s \gg 1$, $P_{\kappa=0}(s)$ shows the asymptotic behavior of $P_{\kappa=0}(s) \approx s^{\gamma_0}$ for flexible chains. The exponent γ_0 is approximated by $3(1-3\nu) \approx -2.3$ if the monomers are considered as particles independently distributed in space, or more precisely by $(3+\theta)\nu \approx -2.2$ where θ is a parameter about 0.70[38,39] and $\nu \approx 0.588$ is the critical exponent of self-avoiding walk. For $\kappa > 0$, in the region $l_p \ll s < 1000$, the exponent γ_0 deviates from this value as can be seen from the inset of Figure 2, instead it is a larger exponent for $\kappa > 0$. Nevertheless, this does not mean that for even larger s the semiflexible chains have a different exponent from the flexible chain, since $P_{\kappa>0}(s)$ is leaning down when s grows to 1000.

Figure 2B shows the contact probabilities of chains with Gaussian distributed stiffness parameter κ_i. The mean values of $\kappa_i = 2$, 5, 10 are same as for the homogeneous chains. The corresponding standard deviations are $\sigma = 1$, 2, 3, 1. Comparison of Figure 2A and Figure 2B reveals that the heterogeneous chains have more contact in the small s region than the homogeneous chains with the same averaged stiffness parameter. What is more, for the Gaussian distribution of κ_i we studied, the contact probability increases with the standard deviation σ. This means that κ_i smaller than $\langle\kappa\rangle$ has more influence on the conformation than κ_i which is larger than $\langle\kappa\rangle$. In other words, the heterogeneity flexibilizes the chain. Nonetheless, in length

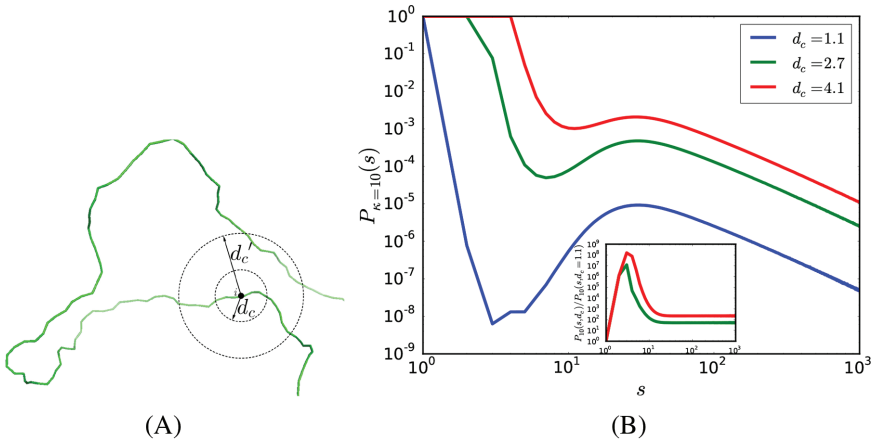


Figure 1. Panel A shows the definition of a contact within the cut-off distance d_c. All the other monomers inside the dashed circle (sphere in 3D) with radius d_c are in contact with monomer i (solid point). Certainly if the cut-off distance d_c is larger, there are potentially more monomers contributing as contacts. Panel B shows the contact probabilities for different cut-off distances d_c. The results are for a chain of length $N_b = 10000$, $\kappa = 10$, bond length $l_b = 1$ and the radius of hard sphere representing one monomer of $r = 0.4$. These results were obtained using the continuous pivot algorithm. Only the range $s < 1000$ is shown. Different d_c only affect $P_\kappa(s)$ in the range of small s, while the asymptotic power law decay behavior of $P_\kappa(s)$ is recovered as is shown in the inset, where we plot the ratios of the probabilities. Other values of polymer length N_b and κ give similar results.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

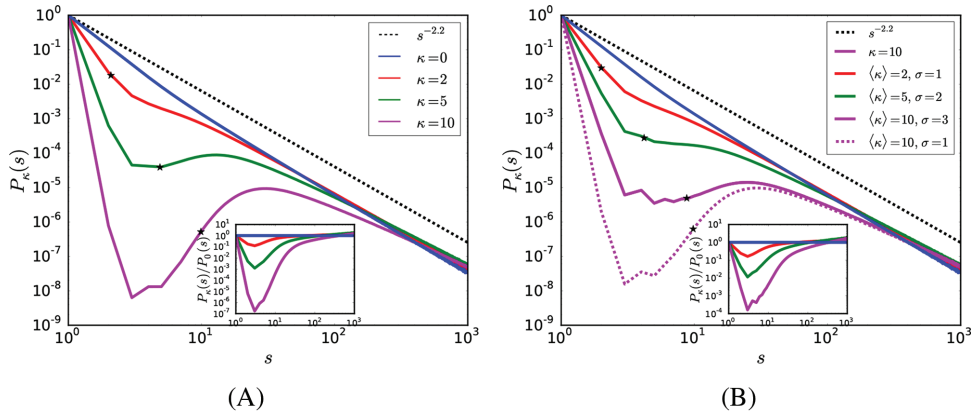


Figure 2. Contact probability $P_\kappa(s)$ for flexible and semiflexible chains with homogeneous and heterogeneous stiffness. The homogeneous chains have the rigidity parameter $\kappa = 2, 5, 10$ (panel A). In the heterogeneous chains each κ_i is sampled from a Gaussian distribution with mean values $\langle\kappa\rangle = 2, 5, 10, 10$ and corresponding standard deviation $\sigma = 1, 2, 3, 1$ (panel B). Only the region $s < 1000$ is showed since $P_\kappa(s)$ for bigger s has large statistical fluctuations. The black dashed line in both figures is the power law $s^{-2.2}$ which is the predicted asymptotic behavior for the self-avoiding walk.[38] The star points indicate the values of persistence length l_p in these cases. The chains $N_b = 10\,000$ are simulated using the pivot algorithm. For the homogeneous chains (panel A), when $s < l_p$, the bending energy that tends to align neighboring bond vectors prevails over the entropy, therefore $P_\kappa(s)$ shows a drop compared to the flexible chain ($\kappa = 0$), and a minimum exists if κ is large enough. In the range $l_p \ll s < 1000$, $P_{\kappa\neq 0}(s)$ shows a power law decay with an exponent slightly larger than the flexible chain. The persistence lengths extracted from the orientational correlation function are $l_p = 2.09, 4.88, 9.90$ for $\kappa = 2, 5, 10$. For the heterogeneous chains (panel B), the contact probabilities drop less in the small contour length range compared to the homogeneous chains due to the heterogeneity of stiffness along the chain even though the mean values are the same ($l_p = 2.00, 4.20, 8.77, 9.80$ for the four cases). Nevertheless, they have similar asymptotic behavior in the range $l_p \ll s < 1000$.

scale much larger than the persistence length, they have the same contact probability despite the heterogeneity.

Hence, the bending rigidity and its heterogeneity mainly exert influence on contact probability in the region where s is smaller than the persistence length l_p for the polymers in free space, while the asymptotic behavior of $P_\kappa(s)$ are similar. For semiflexible chains in finite space, the bending rigidity not only leads to a drop of contact probability in region $s < l_p$, but also introduces an oscillation for $s > l_p$, as will be shown later.

3.2. Persistence Length and Structure Factor

For the worm-like chain without the excluded volume effect, the mean square end-to-end distance is:

$$\langle R_e^2\rangle = 2l_pL\left[1-\frac{l_p}{L}(1-e^{-L/l_p})\right] \tag{3}$$

where the persistence length l_p is defined by $\langle\cos(\theta(s))\rangle = \exp(-s/l_p)$. In the limit $L \gg l_p$, $\langle R_e^2\rangle = 2l_pL \propto N_b$.

For most of the real chain systems except those in the melt condition or in the θ-solvent where polymers act like ideal chains, the excluded volume effect leads to chain swelling, resulting in different scaling exponent $\nu \approx 0.588$ for the end-to-end distance and radius of gyration according to the renormalization group method, or $\nu = 0.6$ by the Flory approximation:

$$\langle R_e^2\rangle = C_eN_b^{2\nu}, \langle R_g^2\rangle = C_gN_b^{2\nu}, N_b \to \infty \tag{4}$$

where C_e, C_g are related to the persistence length.

However, despite the existence of excluded volume effects, Equation (3) does validate itself in semiflexible real chain systems for limited length scale determined by the persistence length l_p.[40] In fact, there are two regimes where Equation (3) does apply: the first one is $s \leqslant l_p$ where the chain behaves like a rod; the second one is $l_p \ll s < s^*$ where the chain can be viewed as ideal since monomers can hardly "collide" and consequently the excluded volume effects are negligible. The value of s^* depends on the persistence length l_p as $s^* \propto l_p^3$ according to the Flory argument, or numerically $s^* \propto l_p^{2.5}$.[41]

There are several ways to determine the persistence length. The traditional one is defined through the exponential decay of the orientational correlation function $\langle\cos\theta(s)\rangle$ (Equation (2)). Although for both random walk and self-avoiding walk the orientational correlation function shows a power law decay behavior[27–29] at a large length scale $s > s^*$, this stays a good estimator considering that it can recover the stiffness parameter κ (Figure 4) and that l_p should not depend on the polymer length (**Figure 3** blue solid line). Another way is to calculate l_p from Equation (3) or simply $\langle R_e^2\rangle/2N_bl_b$ when L is large enough. Clearly for real chains this is not reliable since $\langle R_e^2\rangle \approx N_b^{2\nu}$ due to the excluded volume effect, thus $\langle R_e^2\rangle/2N_bl_b$ would increase with N_b (Figure 3 green solid line). On the other side, $\langle R_e^2\rangle/2N_b^{2\nu}l_b$ does not give reliable results either as shown in Figure 3 (red solid line) because when N is not very large the stiffness weakens the excluded volume effect. Shown in **Figure 4** is the dependence of persistence length l_p on the bending rigidity parameter κ and its distribution, in which l_p is extracted by fitting the exponential decay to the orientational correlation function. The values of persistence length for homogeneous $N_b = 1000$ chains are represented by blue open circles, the linear fitting of which has a slope equal to 1, suggesting the relation $l_p \approx \kappa\langle l_b\rangle(l_b = 1)$.

The heterogeneous chains have various stiffness parameter κ_i, hence persistence length l_{p_i}, along the backbone. We are interested in how the average persistence length over the chain would change due to the heterogeneity. As mentioned above,

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

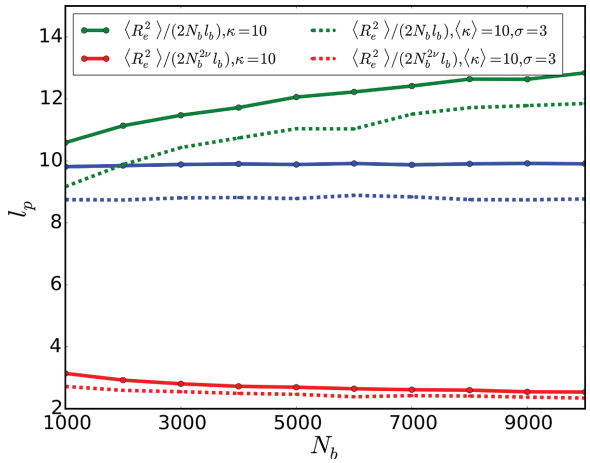


Figure 3. Shown is the persistence length l_p for homogeneous chains (solid lines) and heterogeneous chains (dashed lines) calculated from 1) blue lines: the exponential fit to the orientational correlation function (see Equation (2)); 2) green lines: $\langle R_e^2\rangle / 2N_b l_b$; 3) red lines: $\langle R_e^2\rangle / 2N_b^{2\nu} l_b$. The average bending rigidity parameter is $\langle\kappa\rangle = 10$, and $\sigma = 3$ for the heterogeneous case.

we assume that κ_i obeys the Gaussian distribution with mean value $\langle\kappa\rangle$ and standard deviation σ. Considering the exponential decay of orientational correlation function in Equation (2), when $s = 1$ the persistence length is roughly approximated by $l_p \approx -1/\ln\langle\cos\theta\rangle$. For homogeneous chains, $\langle\cos\theta\rangle \approx \exp(-1/\kappa)$, while for heterogeneous chains,

$$\langle\cos\theta\rangle \approx \langle\exp(-1/\kappa)\rangle = \int_0^\infty \exp(-1/\kappa) f(\kappa) d\kappa \tag{5}$$

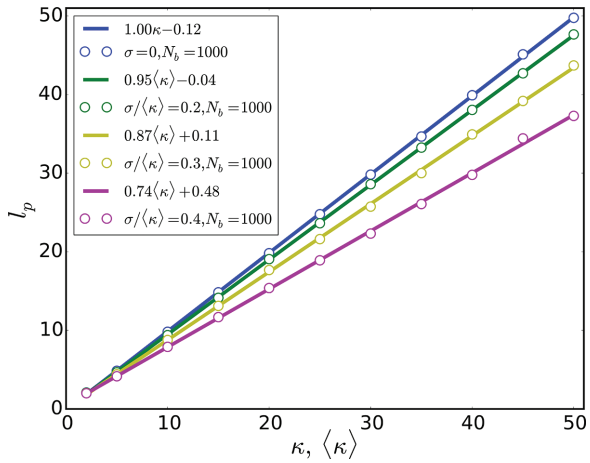


Figure 4. Dependence of average persistence length l_p on the bending rigidity parameter κ and its distribution with different standard deviation $\sigma = 0$, 0.2, 0.3, 0.4. The relation between l_p and $\langle\kappa\rangle$ can be considered linear, with slope dependent on σ. The open circles indicate values of l_p, while the solid lines are the linear fitting results. l_p is extracted by fitting an exponential decay to the orientational correlation function. The bond length is $l_b = 1$ in the continuous pivot algorithm. Results for $N_b = 100$ and 500 are not shown as they are almost on top of the data for $N_b = 1000$. The results show for homogeneous chains that the relation between persistence length and stiffness parameter is $l_p \approx \kappa\langle l_b\rangle$. In the heterogeneous case the average persistence length would be smaller with increasing σ.

where $f(\kappa)$ is the distribution function of the stiffness parameter. The integration starts from 0 because we do not make allowances for negative stiffness parameter κ. The consequential bias from Gaussian distribution is negligible when we take the standard deviation $\sigma \leqslant \langle\kappa\rangle/3$. Obviously on the right-hand side of this equation the smaller values of κ contribute more to the integration, leading to a smaller $\langle\cos\theta\rangle$ and hence a smaller l_p compared to the homogeneous case. The dashed lines in Figure 3 and open circles in Figure 4 show the simulation results of the average persistence length for heterogeneous chains. These chains have smaller persistence lengths and end-to-end distances, which leads to the conclusion that the heterogeneity flexibilizes the chain.

Experimentally the persistence length l_p is usually calculated from the structure factor $S(q)$ which can be measured by the neutron scattering experiments, q is the wavenumber. The structure factor $S(q)$ is defined as:

$$S(q) = \frac{1}{N^2}\left\langle \sum_{i=1}^{N}\sum_{j=1}^{N} \exp[i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)]\right\rangle \tag{6}$$

The semiflexible chain behaves rod-like at a small length scale $s < l_p$, and recovers the self-avoiding property at a much larger length scale $s^* \gg l_p$. This indicates the existence of two regimes in $S(q)$: the self-avoiding regime and the rod-like regime. In addition, as discussed above, in the region $l_p \ll s < s^*$, the semiflexible chain behaves more like a random walk since the excluded volume effect can be ignored. Thus there are several corresponding features for these different regimes in the structure factor $S(q)$. When q is quite small, $S(q) \approx 1 - q^2\langle R_g^2\rangle/3$, which is the Guinier regime. In the region $q > 1/R_g^2\rangle$, $S(q)$ shows the self-avoiding regime: $S(q) \propto q^{-1/\nu}$. Then the crossover from self-avoiding region to random walk region occurs at $qR^* = 1 (R^* \propto l_p^2)$,[40] where $S(q)$ changes to $S(q) \propto q^{-2}$. When $ql_p > 1$, $S(q)$ exhibits the rod-like property $S(q) \propto q^{-1}$. In the Kratky plot $qS(q)$ (**Figure 5**), the rod-like region is the "Holtzer plateau." Therefore the persistence length l_p can be approximated from the onset of the horizontal region in the Kratky plot. Based on the above discussion, there should be three crossovers for $S(q)$,[40,42] but not all the crossover can be seen clearly in the $S(q)$-plot. The Guinier, self-avoiding and rod-like regimes are present in Figure 5, while the random walk regime is hidden. The Gaussian random walk regime can be visible only when the persistence length l_p is large enough.[40] The structure factor of homogeneous $\kappa = 5$, 10 chains and heterogeneous chain with $\langle\kappa\rangle = 10$, $\sigma = 3$ are shown in Figure 5. The latter has a smaller persistence length, hence its structure factor is shifted compared to the homogeneous $\kappa = 10$ chain.

3.3. The Chain in Confinement

There have been studies on semiflexible linear and ring polymers under different kinds of confinements, for example in a spherical capsule,[43] in a channel and in a cavity,[44] in a cylinder,[45] and in rectangles.[24,26] Here we will use the bond fluctuation model to explore different aspects of the structure of semiflexible chains in cubic and rectangular confinement,

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

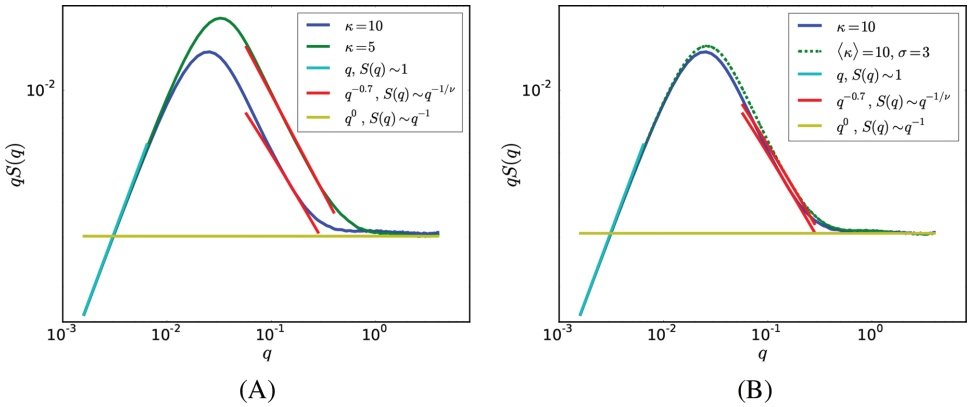


Figure 5. Kratky log-log plot of $qS(q)$ versus q for homogeneous and heterogeneous chains. The chain length is $N_b = 1000$. In panel A, the green and blue lines are for homogeneous chains with $\kappa = 5$ and $\kappa = 10$. Three regimes can clearly be seen: the Guinier regime $S(q) \approx 1$ when $q \ll 1$ (cyan line), the self-avoiding regime $qS(q) \propto q^{1-1/\nu}$ (red lines), the rod-like regime $qS(q) \propto q^0$ (yellow line). The random walk regime is absent because the persistence length l_p or κ is not large enough. Panel B compares the structure factors of homogeneous chain and heterogeneous chain with the same $\langle\kappa\rangle = 10$. The latter has a smaller persistence length, hence its structure factor is shifted compared to the homogeneous chain.

including the contact probability, the ordering of chain segments and the orientational correlation function.

Within finite space, the conformations of semiflexible chains depend on the persistence length l_p and the linear dimension a of the enveloping space, resulting in a "shape transition."[26,43,44] When $l_p \ll a$, chain segments are randomly orientated (**Figure 6**A,B), although at length scales smaller than l_p, they are more ordered due to the bending rigidity. However, when the persistence length l_p is comparable to or larger than the linear dimension a, the chain has to adopt an ordering (Figure 6 C,D as a consequence of the competition between confinement, bending energy and entropy.

The significant difference between the contact probability of polymers in symmetric confinement and in free space is that for the former it does not drop when $s > l_p$. Instead, considering for example a cubic box with side length $a = 40$, $P_\kappa(s)$ levels off after $s > s_c$ (**Figure 7**a,b) where s_c depends on the box size and κ, even when κ is not zero, corresponding to the conformations before the "shape transition" occurs. In this case, $s_c < N/2$, which means that monomers that are separated by $N/2$ or more monomers actually have the same probability to contact each other, suggesting that the maximum distance of monomers has been reached in the finite box.[46,47] Within this space, chains with small κ do not form spirals, while for larger κ, the semiflexible chains begin to spiral but the spirals are not regularly organized in size and direction (Figure 6A). As the space becomes smaller and κ is larger, the "shape transition" condition is satisfied, the semiflexible chain will organize into spirals to accommodate itself in the finite space (Figure 6C). The formation of these spirals leads to an oscillation of the contact probability for the length scale larger than the size of spirals (Figure 7c–f).

In the left column of Figure 7 are the contact probabilities for homogeneous $N = 160$ chains with $\kappa = 0, 3, 7, 9$ confined in cubic boxes of side length $a = 40, 25, 16$. The right column shows contact probabilities of corresponding heterogeneous chains, with standard deviation $\sigma = 1$ or 3. In Sections 3.1 and 3.2, we have mentioned that the heterogeneity flexibilizes the chain and induces more contact at a length scale smaller than the persistence length. Here, when in confinement, the heterogeneity

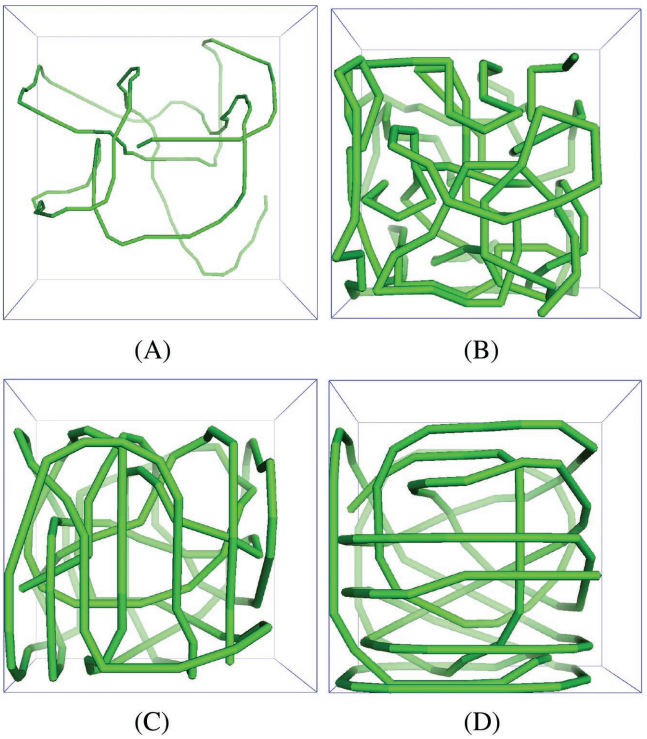


Figure 6. Typical conformations of $N = 160$ in a cubic box with side length a and bending rigidity parameter κ: A) $a = 40$, $\kappa = 9$, $l_p \approx 24.3$; B) $a = 16$, $\kappa = 2$, $l_p \approx 5.4$; C) $a = 16$, $\kappa = 9$, $l_p \approx 24.3$; D) $a = 16$, $\kappa = 20$, $l_p \approx 54$. The persistence length l_p here refers to the value when no confinement is imposed, and is roughly approximated by $l_p \approx \langle l_b \rangle \kappa \approx 2.7\kappa$. When l_p is smaller than the box, the chain forms spirals but they are randomly ordered (see A, B). When l_p is comparable to or larger than the box size, the spirally chain has to arrange itself in an orderly way (see C, D). Meanwhile, Figures C,D show that the conformations do not differ significantly as l_p/a becomes even bigger.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

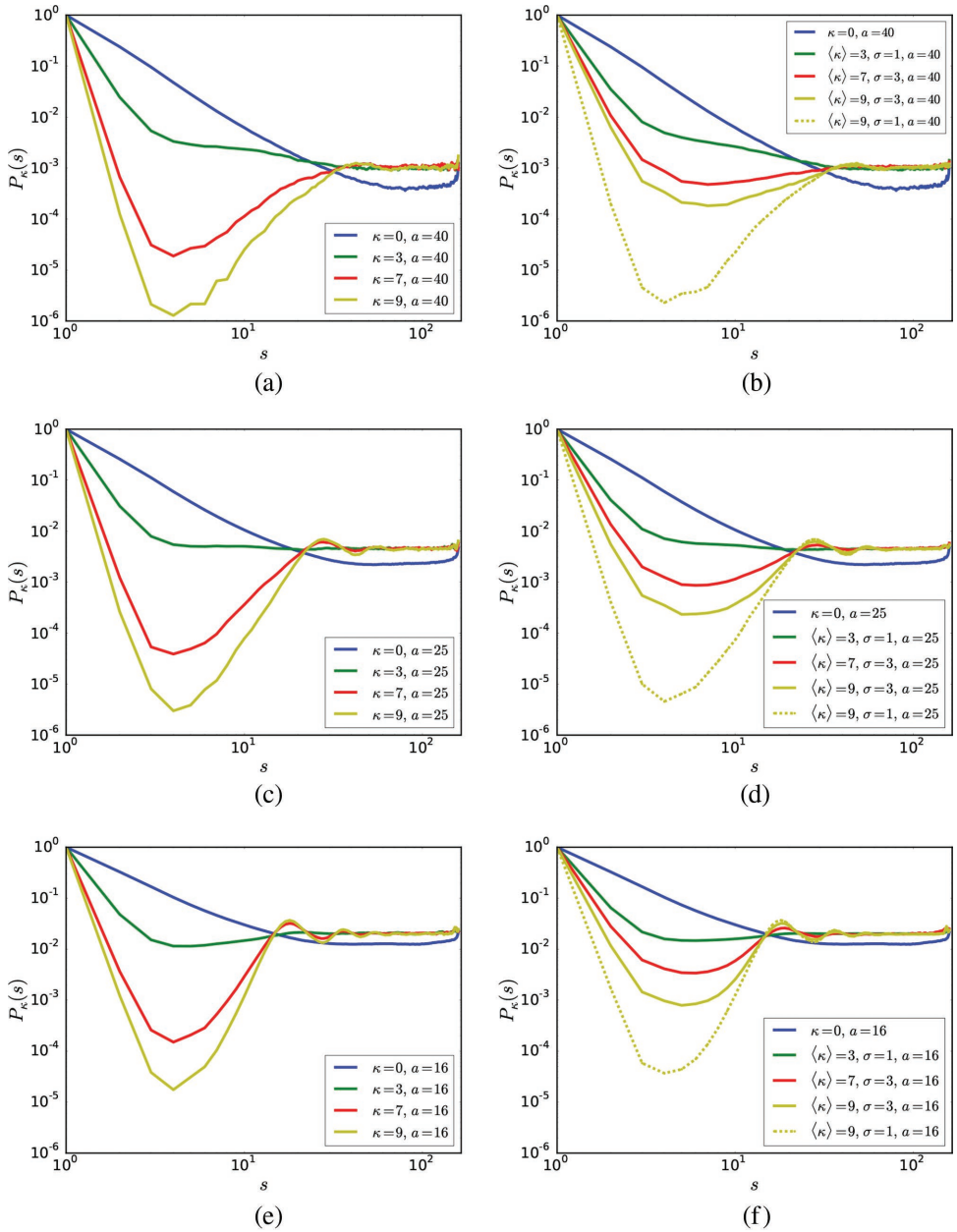


Figure 7. Contact probability $P_\kappa(s)$ for $N = 160$ homogeneous (left column) and heterogeneous (right column) chains with different bending rigidity $\langle\kappa\rangle = 0, 3, 7, 9$ in different sizes of cubic boxes: a,b) a = 40, c,d) a = 25, e,f) a = 16. When the space is finite but not too narrow, $P_\kappa(s)$ begins to level off after $s > s_c$. When κ is larger and the box size is smaller, the chain has to spiral, hence oscillations in $P_\kappa(s)$ appear in large s regime. The heterogeneity induces more contact and weakens oscillation in $P_\kappa(s)$.

could also weaken the oscillation of contact probability in the large s regime because of the enhanced flexibility (Figure 7d,f).

For the chains in asymmetric space, the contact probability exhibits a slightly different behavior in the region $s > s_c$ compared to the symmetric case. Shown in **Figure 8** are the contact probabilities for flexible and semiflexible chains ($N = 160$) in rectangle boxes with different aspect ratios. As the box is elongated, the oscillation in $P_\kappa(s)$ is distorted for semiflexible chains. The volume of the boxes is about 4000. While in a symmetric box, the spirals of the chain have the same radius in all directions on average, in rectangle boxes, the spirals are also elongated, like ellipsoids. The local minimum part is smaller than cubic box case since the space in this direction is narrower and the monomers have higher probability of contact.

3.4. The Effect of Bending Rigidity

Here we investigate how the shape of the confinement affects the packing of a semiflexible chain. We choose rectangular

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

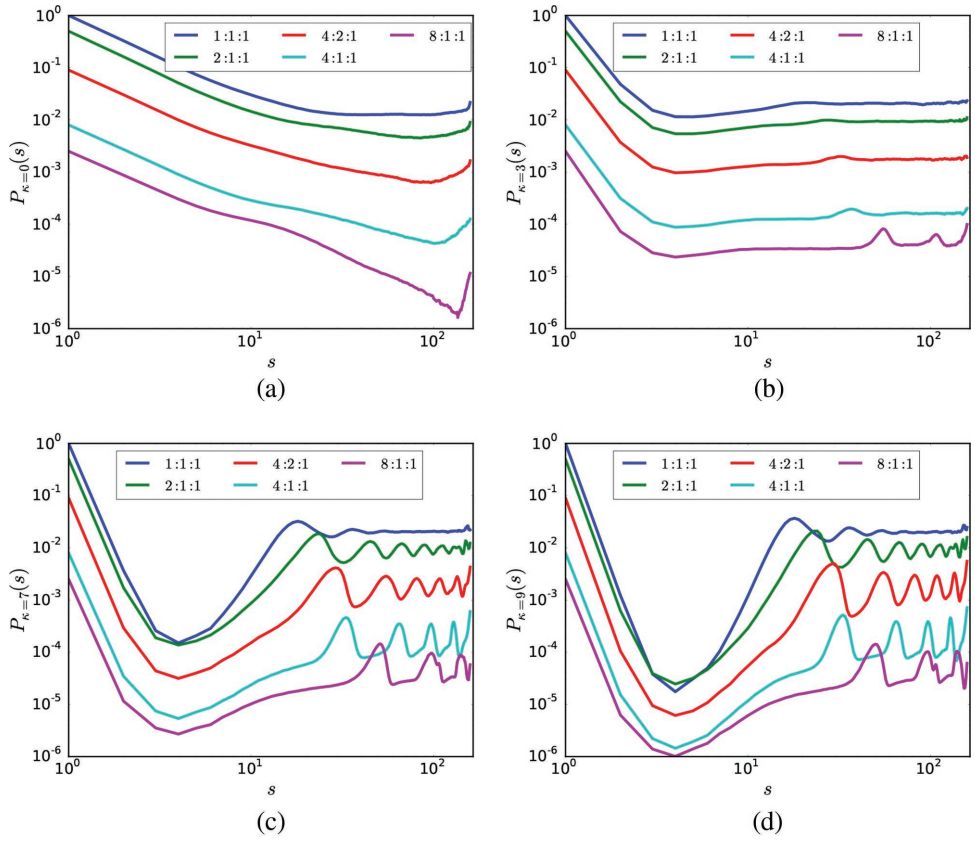


Figure 8. The contact probability of chains in rectangle boxes with different aspect ratios. In each figure, the probability functions are shifted vertically in order to have a better view of them. a) $\kappa = 0$; b) $\kappa = 3$; c) $\kappa = 7$; d) $\kappa = 9$.

boxes of different aspect ratios but the same volume: $a : b : c$ = 1:1:1, 2:1:1, 4:1:1, 8:1:1, 4:2:1. Fritsche et al.[24] showed that a semiflexible ring polymer prefers the long axis of the surrounding envelope. This conclusion holds for semiflexible linear chains as well. **Figure 9** shows conformations for $N = 160$ and $\kappa = 9$ where the chain is confined in selected boxes.

To quantify the ordering of chain segments, we use the order parameter S following[24] which is defined as:

$$S = \frac{1}{N-1}\left\langle \sum_{i=1}^{N-1}\left(\frac{3}{2}\cos^2\theta_i - \frac{1}{2}\right)\right\rangle \tag{7}$$

where θ_i is the angle between chain segment $\mathbf{u}_i$ and the local direction $\mathbf{n}$ of the confined geometry of interest. In the rectangular confinement, $\mathbf{n}$ has three choices, which are parallel to the three sides, namely, $\mathbf{n}_x = (1,0,0), \mathbf{n}_y = (0,1,0)$ and $\mathbf{n}_z = (0,0,1)$. Thus, we have three order parameters S_x, S_y, S_z, each ranging from –0.5 to 1. If the chain segments $\mathbf{u}_i$ have no orientational preference along a given direction $\mathbf{n}$, the order parameter would be $S = 0$, whereas the chain with all $\mathbf{u}_i$ parallel to $\mathbf{n}$ gives $S = 1$, and chain with all $\mathbf{u}_i$ perpendicular to $\mathbf{n}$ has $S = -0.5$. Therefore, $S < 0$ means that the chain segments have a tendency to be organized perpendicularly to $\mathbf{n}$, $S > 0$ indicates the tendency of being parallel to $\mathbf{n}$.

Figure 10 shows the order parameters S_x, S_y, S_z for chains of different bending rigidity κ in rectangular boxes of different aspect ratios. In the cubic case, the chain segments have no orientational preference, S_x, S_y, S_z are almost zero, both for the flexible and semiflexible chains. When the box has a longer side (x direction), S_x is positive which means chain segments tend to be parallel to the x direction, while S_y, S_z are smaller than 0. Note that even for the flexible chain ($\kappa = 0$) in a rectangular box, Figure 10 shows a positive S_x, this is mainly due to the artificial lattice setting. When the chains are stiffer, we get a larger S_x and hence smaller S_y and S_z. This means that bending rigidity makes the chain order itself along the longer axis in a rectangle confinement.

3.5. The Orientation of Bond Vectors

Since bending rigidity forces the chain to order itself along the long side of the confinement and to form spirals, the orientational correlation function $\langle \cos(\theta(s)) \rangle$ also behaves differently from the exponential decay or power law decay in large contour length regime[29] in free space. Instead, the orientational correlation function shows an oscillation due to the existence of spirals. This correlation function for different confining geometries has been studied extensively both by simulations and experiments.[48,49]

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

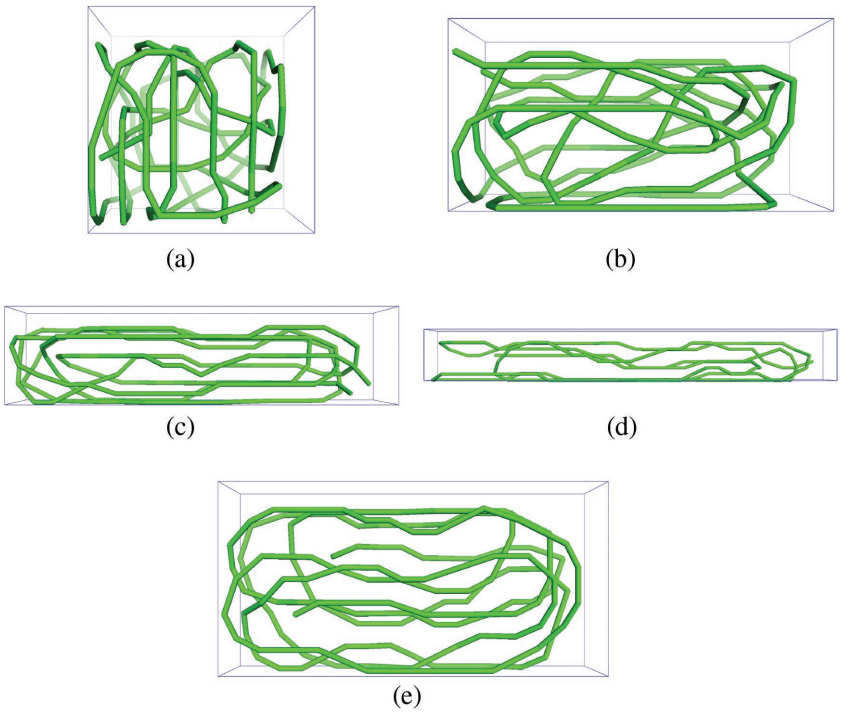


Figure 9. Chain ($N = 160$, $\kappa = 9$) conformations in rectangular boxes of different aspect ratios but the same volume $V \approx 4000$. The aspect ratios are: a) 1:1:1, b) 2:1:1, c) 4:1:1, d) 8:1:1, e) 4:2:1. To minimize the free energy in the narrow space, the semiflexible chain would stretch along the long axis, and spirals around the shortest axis.

Figure 11A shows the orientational correlation functions for $N = 160$ with different bending rigidity $\kappa = 0$, 2, 3, 5, 7, 9 confined in a cubic box with side length $a = 16$. When the chain becomes stiffer, the spirals get larger and better ordered. As a result, the oscillations in this function are more pronounced. Another fact from this figure is that the periodicity of the function does not change with the value of κ. This may imply that it may be determined by the box size, which will be discussed later in this section. In Figure 11B we compare the orientational correlation function of the heterogeneous chains with the homogeneous one. The difference is more identifiable for the $\sigma = 3$ case (red line). This means that the heterogeneity weakens the oscillation by enhancing the flexibility.

Figure 12 shows the orientational correlation functions for $N = 160$ and $\kappa = 9$ in cubical boxes of different sizes. As the space becomes narrower, the chain has more spirals and they are more orderly, thus the amplitude and frequency of the oscillation in $\langle \cos(\theta(s)) \rangle$ get larger.

Liu[26] studied the form of the orientational correlation function $\langle \cos\theta(s) \rangle$ in a 2D square confinement for a worm-like chain, concluding that the leading contribution to $\langle \cos\theta(s) \rangle$ is $e^{-\frac{s}{l_e}} \cos\frac{s}{d}$, where l_e is the effective persistence length, d is linearly related to the size of box a.

Figure 13 shows the fitting of the orientational correlation function to $e^{-\frac{s}{l_e}} \cos\frac{s}{d}$ for chain lengths of $N = 20$, 40, 80, 160 and bending rigidity parameter $\kappa = 9$. The fitting values of l_e and d are listed in the caption. The side lengths of the cubic boxes for these four chains are $a =$ 8, 10, 13, 16. We have roughly the same ratios of a/d : 8/1.526 = 5.24, 10/1.883 = 5.31, 13/2.375 = 5.47, 16/2.863 = 5.59, which means that d is almost proportional to the box size.

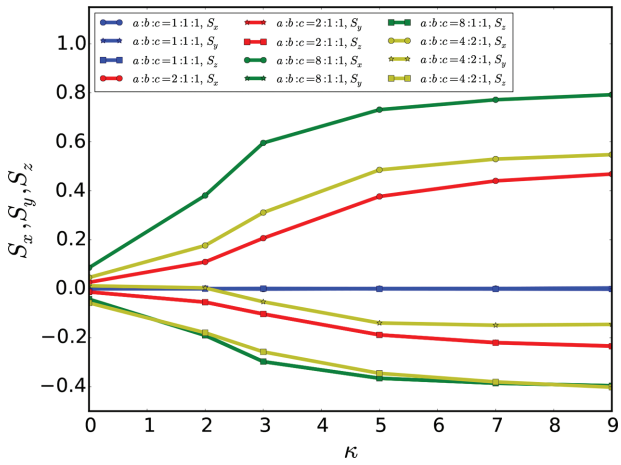


Figure 10. The order parameter S_x, S_y, S_z for chains of different bending rigidities κ in rectangular boxes of different aspect ratios: 1:1:1 (cubic), 2:1:1, 8:1:1, 4:2:1. A positive S means that the chain segments are more parallel to the corresponding axis. Bending rigidity makes the semiflexible chain order itself along the longer axis; therefore, S_x increases with κ for each aspect ratio except the cubic one.

4. Conclusion

In this paper we stressed the contact definition of monomers and the invariability of asymptotic behavior when different cut-off distances and bending rigidity come into play. At very large length scale, the contact probability for linear chains in free space will exhibit the same power law decay $P(s) \approx s^{-2.2}$, with different coefficients when the cut-off distance for contact is involved.

Second, we investigated how the bending rigidity influences the conformations of a linear chain under geometric confinement, represented here by means of cubic and rectangle boxes. The bending potential reshapes the chain due to the competitive interplay of stiffness, entropy and confinement. Moreover, there exists a "shape transition" from overall randomness to orderliness when the persistence length is comparable to the size of confinement. One measure that can reflect the impact of bending rigidity and confinement is the contact probability. The contact probability of a flexible or semiflexible chain in sufficient small confinement

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

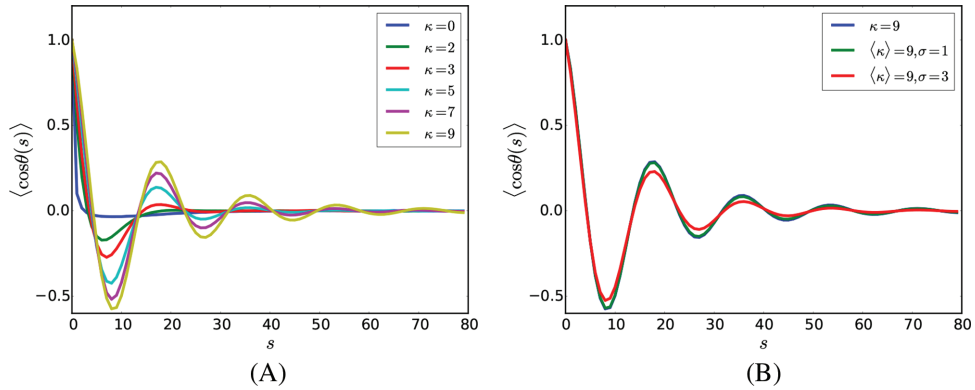


Figure 11. Orientational correlation function for different bending rigidity parameters and distributions. Panel A shows the orientational correlation function $\langle\cos\theta(s)\rangle$ for chains with $N = 160$ and bending rigidities $\kappa(= 0, 2, 3, 5, 7, 9)$ in boxes of size $a = 16$. Only the range $s \leqslant 80$ is shown. As κ becomes larger, the spirals in the finite space are more ordered, therefore $\langle\cos\theta(s)\rangle$ has larger oscillations. Panel B compares the correlation function of the heterogeneous chains with homogeneous one. The oscillation is weakened due to the heterogeneity.

has a plateau region in the large contour length region, as opposed to the power law decay in free space. Moreover, if the bending rigidity is big enough compared to the size of the confinement, this plateau region will turn into an oscillation (Figure 7), which indicates the existence of spirals formed by the chain.

The ordering of the chain according to the shape of geometric confinement can also be studied by constraining the chain into rectangle boxes of different aspect ratios. An order parameter S is defined to quantify the ordering of chain segments. It is shown that the semiflexible chain preferably chooses the long direction of the boxes to order the segments. The orientational correlation function $\langle\cos\theta(s)\rangle$ of bond vectors is also dramatically changed due to the bending rigidity and confinement, and the oscillation in it serves as a direct evidence of the formation of spirals. The leading term of the analytical expression of $\langle\cos\theta(s)\rangle$ consists of two parts: the first one is the exponential decay term that gives the effective persistence length of the semiflexible chain in confinement; the second part is a cosine function which determines the period of the oscillation mentioned above. This period is dependent on the box size.

It has been pointed out that the 3D organization of chromosomes is tightly coupled to the mechano-genomic code.[50] Our study shows that the modulation of the bending rigidity may be part of the mechano-genomic code regulating the contact probability and thus the 3D organization. It remains to decipher the mechano-genomic code. Here, one of the leading contender is the nucleosomal organization. Nucleosomes contribute due to their steric repulsion and their absence alone to the bending rigidity.

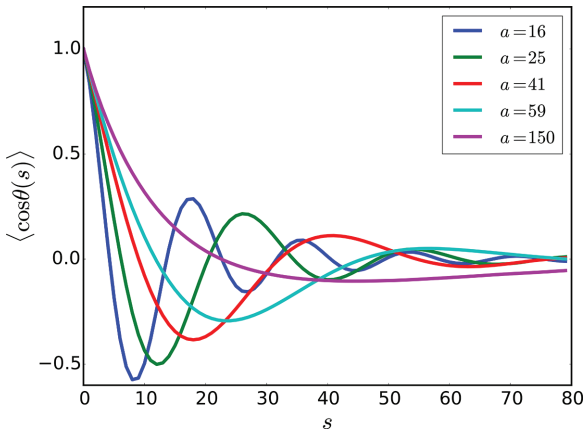


Figure 12. The orientational correlation function $\langle\cos\theta(s)\rangle$ for chains of size $N = 160$ and $\kappa = 9$ in different cubical boxes with side lengths $a = 16$, 25, 41, 59, 150. Only the range $s \leqslant 80$ is shown. As the space becomes narrower, the chain has to spiral more orderly, which contributes to the oscillation in $\langle\cos\theta(s)\rangle$.

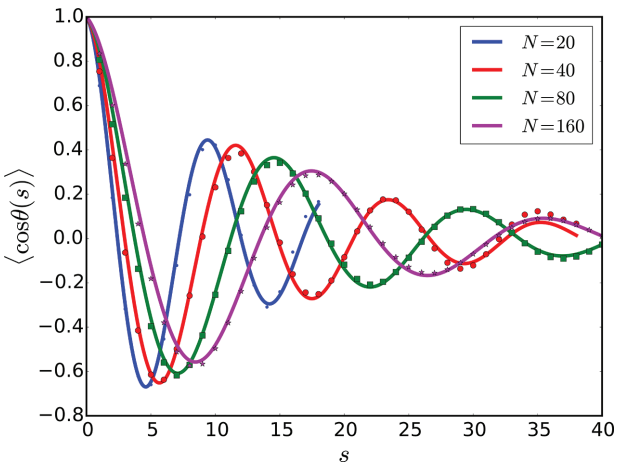


Figure 13. Fitting of the leading term $e^{-\frac{s}{l_e}}\cos\frac{s}{d}$ to the orientational correlation data for different chain lengths $N = 20$, 40, 80, 160 with bending rigidity parameter $\kappa = 9$. The points are data calculated from Monte Carlo simulations, and the solid lines are the curves fitted to corresponding points. Only the range $s \leqslant 40$ is shown. The fitting parameters are: 1) $N = 20$, $l_e = 11.71$, $d = 1.526$; 2) $N = 40$, $l_e = 13.51$, $d = 1.883$; 3) $N = 80$, $l_e = 14.61$, $d = 2.375$; 4) $N = 160$, $l_e = 14.92$, $d = 2.863$.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

A.H. and D.W.H. would like to acknowledge funding from a grant by the International Human Frontier Science Program Organization (RGP0014/2014). J.J. would like to acknowledge funding from the China Scholarship Council (CSC NO.201506210082).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

bending rigidity, confinement, contact probability, heterogeneity, polymer

Received: December 13, 2018
Revised: February 4, 2019
Published online:

[1] J. Dekker, *Trends Biochem. Sci.* **2003**, *28*, 277.
[2] E. Lieberman-Aiden, N. 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, J. Dekker, *Science* **2009**, *326*, 289.
[3] G. M. Cooper, *The Cell: A Molecular Approach*, Sinauer Associates, Sunderland, MA **2000**.
[4] D. W. Heermann, *Curr. Opin. Cell Biol.* **2011**, *23*, 332.
[5] A. T. Annunziato, *Nat. Educ.* **2008**, *1*, 26.
[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, 4th ed., Garland Science, New York, NY **2002**.
[7] H. Schiessel, W. M. Gelbart, R. Bruinsma, *Biophys. J.* **2001**, *80*, 1940.
[8] M. Bohn, D. W. Heermann, R. Van Driel, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2007**, *76*, 1.
[9] S. Goetze, J. Mateos-Langerak, H. J. Gierman, W. de Leeuw, O. Giromus, M. H. G. Indemans, J. Koster, V. Ondrej, R. Versteeg, R. van Driel, *Mol. Cell. Biol.* **2007**, *27*, 4475.
[10] T. Cremer, C. Cremer, *Nat. Rev. Genet.* **2001**, *2*, 292.
[11] A. H. B. De Vries, B. E. Krenn, R. Van Driel, V. Subramaniam, J. S. Kanger, *Nano Lett.* **2007**, *7*, 1424.
[12] V. B. Teif, K. Rippe, *Nucleic Acids Res.* **2009**, *37*, 5641.
[13] S. Jun, A. Arnold, B. Y. Ha, *Phys. Rev. Lett.* **2007**, *98*, 1.
[14] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K Moore, J. Wang, J. Widom, *Nature* **2006**, *442*, 772.
[15] G. Langst, V. B. Teif, K. Rippe, *Genome Organ. Funct. cell Nucl.* **2012**, *111*.
[16] K. Struhl, E. Segal, *Nat. Struct. Mol. Biol.* **2013**, *20*, 267.
[17] J. Y. Lee, T. H. Lee, *Biochim. Biophys. Acta - Proteins Proteomics* **2012**, *1824*, 974.
[18] J. Langowski, D. W. Heermann, *Semin. Cell Dev. Biol.* **2007**, *18*, 659.
[19] P. M. Diesinger, S. Kunkel, J. Langowski, D. W. Heermann, *Biophys. J.* **2010**, *99*, 2995.
[20] C. Vaillant, B. Audit, C. Thermes, A. Arnéodo, *Eur. Phys. J. E* **2006**, *19*, 263.
[21] A. J. Varschavsky, V. V. Bakayev, G. P. Georgiev, *Nucleic Acids Res.* **1976**, *3*, 477.
[22] D. Bensimon, D. Dohmi, M. Mézard, *Europhys. Lett.* **1998**, *42*, 97.
[23] D. Bratko, A. K. Chakraborty, E. I. Shakhnovich, *J. Chem. Phys.* **1997**, *106*, 1264.
[24] M. Fritsche, D. W. Heermann, *Soft Matter* **2011**, *7*, 6906.
[25] K. Ostermeir, K. Alim, E. Frey, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2010**, *81*, 061802.
[26] Y. Liu, B. Chakraborty, *Phys. Biol.* **2008**, *5*, 026004.
[27] A. N. Semenov, *Macromolecules* **2010**, *43*, 9139.
[28] D. Shirvanyants, S. Panyukov, Q. Liao, M. Rubinstein, *Macromolecules* **2008**, *41*, 1475.
[29] H. P. Hsu, W. Paul, K. Binder, *Macromolecules* **2010**, *43*, 3094.
[30] A. D. Sokal, *Funct. Integr.* **1996**, *361*, 131.
[31] T. Kennedy, *J. Stat. Phys.* **2002**, *106*, 407.
[32] G. D'Adamo, A. Pelissetto, *J. Phys. Condens. Matter* **2017**, *29*, 1.
[33] G. Ping, G. Yang, J. M. Yuan, *Polymer* **2006**, *47*, 2564.
[34] E. Horwath, N. Clisby, P. Virnau, *J. Phys. Conf. Ser.* **2016**, *750*, 012010.
[35] I. Carmesin, K. Kremer, *Macromolecules* **1988**, *21*, 2819.
[36] W. Janke, *Quantum* **2002**, *10*, 423.
[37] H. P. Hsu, W. Paul, *Comput. Phys. Commun.* **2011**, *182*, 2115.
[38] J. des Cloizeaux, *J. Phys.* **1980**, *41*, 223.
[39] S. Redner, *J. Phys. Math. Gen.* **1980**, *13*, 3525.
[40] H.-P. Hsu, W. Paul, K. Binder, *Polym. Sci. Ser. C* **2013**, *55*, 39.
[41] H.-P. Hsu, K. Binder, *J. Chem. Phys.* **2012**, *136*, 024901.
[42] H.-P. Hsu, *J. Chem. Phys.* **2014**, *141*, 164903.
[43] P. Cifra, T. Bleha, *Macromol. Symp.* **2010**, *296*, 336.
[44] P. Cifra, T. Bleha, *Eur. Phys. J. E. Soft Matter* **2010**, *32*, 273.
[45] P. Vázquez-Montejo, Z. McDargh, M. Deserno, J. Guven, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2015**, *91*, 063203.
[46] A. M. Chiariello, C. Annunziatella, S. Bianco, A. Esposito, M. Nicodemi, *Sci. Rep.* **2016**, *6*, 1.
[47] J. Dekker, B. van Steensel, in *Handbook of Systems Biology* (Eds: M. Walhout, M. Vidal, J. Dekker), Academic Press, Cambridge, MA **2013**, Ch. 7.
[48] E. Werner, F. Persson, F. Westerlund, J. O. Tegenfeldt, B. Mehlig, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2012**, *86*, 041802.
[49] P. Cifra, Z. Benková, T. Bleha, *Faraday Discuss.* **2008**, *139*, 377.
[50] C. Uhler, G. V. Shivashankar, *Bioarchitecture* **2016**, *6*, 76.

 © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Supporting Information

The Effect of Bending Rigidity on Polymer

Jiying Jia, Kunhe Li, Andreas Hofmann and Dieter W. Heermann*

Below is the information of simulations for each parameter set:

For the data in Figure 1:

The pivot algorithm was used to calculate the contact probability with different cut-off distance with the following sampling:

$N_b = 10000, \kappa = 10$

$d_c = 1.1$	simulated once, 6,730,000 independent conformations were used.
$d_c = 2.7$	simulated once, 1,460,000 independent conformations were used.
$d_c = 4.1$	simulated once, 1,460,000 independent conformatons were used.

For the data in Figure 2:

The pivot algorithm was used to calculate the contact probability with different bending rigidities and the following sampling:

$N_b = 10000$

$\kappa = 0$	simulated once, 3,070,000 independent conformations were used.
$\kappa = 2$	simulated once, 5,010,000 independent conformations were used.
$\kappa = 5$	simulated once, 5,220,000 independent conformations were used.
$\kappa = 10$	simulated once, 6,730,000 independent conformations were used.
$\langle\kappa\rangle = 2, \sigma = 1$	simulated once, 1,900,000 independent conformations were used.
$\langle\kappa\rangle = 5, \sigma = 2$	simulated once, 2,960,000 independent conformations were used.
$\langle\kappa\rangle = 10, \sigma = 3$	simulated once, 4,300,000 independent conformations were used.
$\langle\kappa\rangle = 10, \sigma = 1$	simulated once, 4,390,000 independent conformations were used.

For the data in Figure 3:

The persistence length versus N_b with the following sampling:

	exponential fitting method	using the end to end distance
$N_b = 1000, \kappa = 10$	simul once, 15,000 confs	simul once, 50,000 confs
$N_b = 2000 \sim 10000, \kappa = 10, (\sigma = 3)$	simul once, 5,000 confs	simul once, 50,000 confs

For the data in Figure 4:

The persistence length versus κ with the following sampling:

$\kappa = 2, 5, 10, 20$	simul once, 15,000 confs
$\kappa = 15, 25, 30, 35, 40, 45, 50$	simul once, 10,000 confs
$\sigma/\langle\kappa\rangle = 0.2, 0.3, 0.4$, $\kappa = 2 \sim 50$	for each $\langle\kappa\rangle$ and σ, simul 10 times (different sample of κ_i from the Gaussian distribution), each has 10,000 confs.

For the data in Figure 5:

The structure factor with the following sampling:

$N_b = 1000, \kappa = 5, 10$ and $\langle\kappa\rangle = 10, \sigma = 3$	simul once, 15,000 confs were used.

For the data in Figure 7:

The contact probability of chains in cubic boxes. The bond fluctuation model was used with the following sampling:

$N = 160$.

$\kappa = 0, 3, 7, 9, a = 16, 25, 40$	for each κ and a, simulated once, 15,000 confs were used.

For heterogeneous chains, contact probability was averaged over different runs with the following sampling:

For same $\langle k\rangle$ and σ.

$a = 16$ $\langle k\rangle = 7, \sigma = 3; \langle k\rangle = 9, \sigma = 1, 3$	simul 10 times(different sample of $\kappa_i$from the Gaussian distribution, each has about 13,000 confs
$a = 25, 40$, for all $\langle k\rangle$ and σ	simul 3 times, each has 15,000 confs.
$a = 16$, $\langle k\rangle = 3, \sigma = 1$	simul 20 times, each has 15,000 confs.

For the data in Figure 9:

The order parameter in rectangular boxes of different aspect ratios. The bond fluctuation model was used with the following sampling:

$N = 160$

For each κ and aspect ratio, the system was simulated once, 15,000 independent conformations were used.

For the data in Figure 10:

The orientational correlation function for different κ.

$N = 160, a = 16$.

$\kappa = 0, 2, 3, 5, 7, 9$	simul once, 15,000 confs
$\langle\kappa\rangle = 9, \sigma = 1, 3$	simul 10 times, each has 15,000 confs.

For the data in Figure 11:

The orientational correlation function for different a.

$N = 160, k = 9$.

For each $a = 16, 25, 41, 59, 150$, the system was simulated once, 15,000 independent conformations were used.

For the data in Figure 12:

Fitting of the orientational correlation function.

For each pair of N and a, the system was simulated once, 15,000 independent conformations were used.

Chapter 4

Superstructure Detection in Nucleosome Distribution shows Common Pattern within a Chromosome and within the Genome

References

This chapter is from:

- Mishra, S. K., **Li, K.**, Brauburger, S., Bhattacherjee, A., Oiwa, N. N., & Heermann, D. W. (2021). Superstructure detection in nucleosome distribution shows common pattern within a chromosome and within the genome.

All authors were involved in the conception, processing of the data, analysis, and drafting of the manuscript. Kunhe Li and Sujeet Kumar Mishra have a shared co-first authorship. Simon Brauburger initialized the analysis. Kunhe Li, Sujeet Kumar Mishra, Nestor Norio Oiwa, and Dieter W. Heermann analyzed the patterns and completed the classification together. Dieter W. Heermann and Nestor Norio Oiwa supervised the work.

life

MDPI

Article

Superstructure Detection in Nucleosome Distribution Shows Common Pattern within a Chromosome and within the Genome

Sujeet Kumar Mishra [1,2], Kunhe Li [1], Simon Brauburger [1], Arnab Bhattacherjee [2], Nestor Norio Oiwa [1,3] and Dieter W. Heermann [1,*]

[1] Institute for Theoretical Physics, Heidelberg University, D-69120 Heidelberg, Germany; sujeetsankrityan@gmail.com (S.K.M.); li@thphys.uni-heidelberg.de (K.L.); brauburger@stud.uni-heidelberg.de (S.B.); oiwa@thphys.uni-heidelberg.de (N.N.O.)
[2] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India; arnab@jnu.ac.in
[3] Department of Basic Science, Universidade Federal Fluminense, Nova Friburgo 28625-650, Brazil
* Correspondence: heermann@tphys.uni-heidelberg.de

Abstract: Nucleosome positioning plays an important role in crucial biological processes such as replication, transcription, and gene regulation. It has been widely used to predict the genome's function and chromatin organisation. So far, the studies of patterns in nucleosome positioning have been limited to transcription start sites, CTCFs binding sites, and some promoter and loci regions. The genome-wide organisational pattern remains unknown. We have developed a theoretical model to coarse-grain nucleosome positioning data in order to obtain patterns in their distribution. Using hierarchical clustering on the auto-correlation function of this coarse-grained nucleosome positioning data, a genome-wide clustering is obtained for *Candida albicans*. The clustering shows the existence beyond hetero- and eu-chromatin inside the chromosomes. These non-trivial clusterings correspond to different nucleosome distributions and gene densities governing differential gene expression patterns. Moreover, these distribution patterns inside the chromosome appeared to be conserved throughout the genome and within species. The pipeline of the coarse grain nucleosome positioning sequence to identify underlying genomic organisation used in our study is novel, and the classifications obtained are unique and consistent.

Keywords: chromatin; nucleosome positioning; nucleosome distribution; heterochromatin; euchromatin; structure classification

check for updates

Citation: Mishra, S.K.; Li, K.; Brauburger, S.; Bhattacherjee, A.; Oiwa, N.N.; Heermann, D.W. Superstructure Detection in Nucleosome Distribution Shows Common Pattern within a Chromosome and within the Genome. *Life* **2022**, *12*, 541. https://doi.org/10.3390/life12040541

Academic Editor: Eva Bartova

Received: 26 January 2022
Accepted: 23 March 2022
Published: 6 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The genomes of all higher eukaryotes are organised in different structures on multi-length scales [1,2]. Of these organisational structures, the chromosome is the biggest one, being observable under a normal light microscope. The smallest organisational structure, one level above the double helix DNA, is the nucleosome where 147 base pairs (bp) of DNA are wrapped 1.65 times around a histone octamer [3–5]. The arrays of nucleosomes organise to form the chromatin fibre, which folds into two mutually excluded structural domains, namely "heterochromatin" and "euchromatin". The "heterochromatin" regions are enriched with inactive/repressive genes and are usually positioned closer to the periphery of the nucleus. The "euchromatin" regions contain transcriptionally active chromatin [3,6,7], which are genes located in the interior of the nucleus. The hierarchical packaging of chromatin renders the genome a very compact conformation that provides controlled accessibility of the regulatory DNA sequences (genes) by other DNA-binding proteins (DBPs) [8,9]. Thus, the chromatin organisation is tightly linked to gene regulation and warrants detailed investigation. Various experimental techniques have been developed to probe the hierarchical chromatin organisation at different length scales. For instance,

the "chromatin conformation capture" experiment (e.g., 3C and HiC) [2,10,11] captures the organisation of chromatin in a kbp to Mbp length scale, revealing the formation of topologically associated domains (TADs) [12] and chromatin loops [13,14]. Further characterisation of the chromatin fibre at the length scale of genes (∼kbp) is achieved by the Micro-C technique that captures the intra-chromatin interactions at a resolution of ∼100 bp within an organisation module called chromosomal interaction domains (CIDs) [15,16]. CIDs are much smaller but still similar to TADs. These structural organisations are strongly regulated by the nucleosome positions, length of linker regions, and presence of nucleosome-depleted regions (NDR) across the chromosome [17].

The term "nucleosome positioning" refers to the location of nucleosomes along the sequence of genomic DNA. Nucleosome positioning is determined by several factors, including DNA sequence [18,19], DNA-binding proteins [20,21], nucleosome remodelers [22–24], RNA polymerases [25], and more. Although nucleosome positioning is a dynamic process, the sequence-based mapping approach identifies its position only in a cell- and time-averaged manner. The technology of micrococcal nuclease (MNase) digestion combined with high-throughput sequencing (MNase-seq) [26] is a powerful method to map the genome-wide distribution of nucleosome positioning and its occupancy. The resulting occupancy maps are ensemble averages of heterogeneous cell populations and may also be influenced by titration [27]. However, it is necessary to retrieve the cell-specific features from the population average to reveal the mechanism of nucleosome organisation and its translocation along the genome. Zhang et al. has developed an algorithm called "Nucleosome Positioning from Sequencing" (NPS) to predict accurate nucleosome positioning from the MNase-seq data, which was later improved to iNPS (improved NPS) [28]. The nucleosome positioning here is considered as an average static picture where they implicitly consider the nucleosome dynamics in the form of snapshots at different time- and cell-averages. This nucleosome positioning provides the frequency of its occurrence from which peaks are annotated to obtain possible nucleosome location along the sequence. In short, the nucleosome positioning data from iNPS are simply the most probable nucleosome position along the chromosome. Furthermore, extensive studies have been performed to recognise nucleosome positioning patterns around CTCFs, transcription start sites (TSSs), exons and introns, promoter and loci regions locally. For instance, a typical nucleosome distribution around TSSs indicates nucleosome depletion, resulting in a nucleosome-free region (NFR), whereas the nucleosomes downstream of TSS are equally spaced [29]. A similar observation around CTCF is obtained: an array of well-positioned nucleosomes flank the sites occupied by the insulator binding protein CTCF across the human genome [30]. Despite the efforts, the global picture of nucleosome positioning remains elusive until a recent study that has reported three types of nucleosomal arrangement by analyzing the nucleosome spacing and phasing in a genome [31]. The evenly spaced nucleosomes in the array are termed as a regular array and irregular otherwise. At a given genomic location in the cell population, nucleosomes may also assume similar positions and are referred to as phased arrays. The phased-regular nucleosome arrays, being most prominent, are the hallmark of chromatin and found to be conserved from yeast to mammals. These phased-regular nucleosome arrays are mostly found near the promoter regions of transcribed genes in the yeast genome and near the binding sites of high-affinity DBPs in higher eukaryotes. However, the findings have limited applicability only at local regions of the chromatin fibre and provide absolutely no information about the nucleosome organisation along a complete chromosome or genome.

We used a theoretical approach to obtain a novel classification of segments across the chromosome based on the similarity in nucleosome patterns. The nucleosome positioning data are used as inputs that are systematically coarse-grained to analyze their auto-correlation function to search for any pattern. The results are processed using hierarchical clustering techniques to investigate if there exists any unique pattern of nucleosome. Our results suggest that the positions and occupancy of nucleosomes in a chromosome are not random; rather, they reveal distinct patterns of distribution within a chromosome. Interestingly, the patterns appear to be conserved within the genome as well and are in

agreement with the previous study that has reported three distinct nucleosome organisations across the genome. Furthermore, at the chromosome level, our approach could capture a few unique patterns in the range of the ∼50 kbp length scale, which repeatedly occur throughout the chromosomes, indicating they might play a crucial role in regulating gene networks at a more local scale. The study underpins the nucleosome positioning architecture inside a genome that can provide insights into the genome organisation (c.f. Figure 1) not known before.

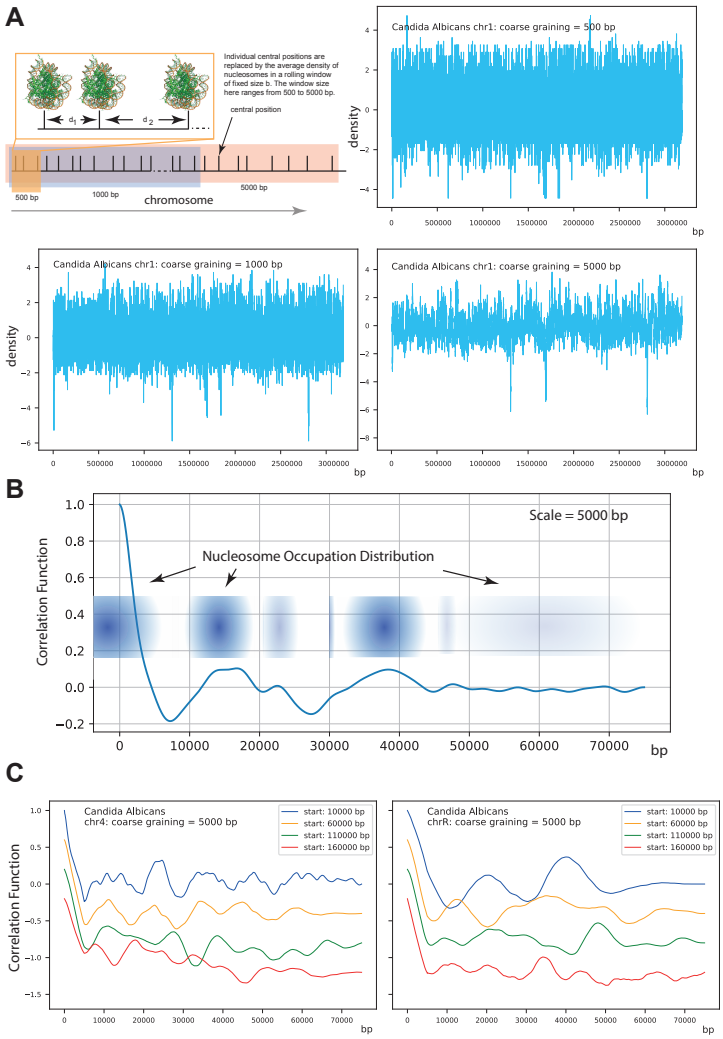


Figure 1. (**A**) shows the performed coarse-graining procedure and results for coarse-graining lengths L of 500 bp, 1000 bp, and 5000 bp. More structure is visible as b is increased. Going up even further washes out the structure. This is typical for systems with an intrinsic length scale. (**B**) shows the correlation among the coarse-grained super nucleosomes. The structure is that of a system exhibiting short range-order that is liquid-like with first and second nearest neighbor peaks. If there is no order or correlation, then the correlation function would be constant. On the other hand, if one would see strong regular peaks, this would indicate a regular ordering with the peak distances giving the preferred distance between the coarse-grained nucleosomes. The oscillatory characteristic with a larger first peak and smaller second peak indicates that two coarse-grained nucleosomes are on average located within a distance from the origin to the first peak and a second coarse-grained nucleosome at the distance indicated by the second peak. Since the peaks are decreasing, this ordering diminishes, much like the local ordering in a liquid. On larger scales larger than 50,000 bp, there is no order, i.e., there is no correlation. (**C**) shows for two chromosomes how the structure differs within as well as among chromosomes. The parameter start indicates from where in the chromosomes the structure was computed. One can see that the structure varies within a chromosome; nevertheless, common structures are found.

1.1. Data

The technology of micrococcal nuclease (MNase) digestion combined with high-throughput sequencing (MNase-seq) [26] is used to map the distribution of nucleosome occupancy genome-wide. In order to map the MNase-seq data to nucleosome positioning data, several programs were developed, such as NPS [32], nucleR [33], and DANPOS [34]. A nucleosome sequencing profile is generated to depict nucleosome distribution in wave-form where nucleosome peaks are detected. The improved nucleosome-positioning algorithm (iNPS) can be applied to identify peaks and correctly detect nucleosome positions [28]. One possible output of the iNPS algorithm is in the binary format, with 1s representing a nucleosome being present and 0s for the nucleosome-free regions or linker regions.

The genome-wide study of the species is a challenging task due to its large sequence size, which needs theoretical expertise and computational power. For our study, we have chosen *Candida albicans* as a simple completely sequenced organism [35] that is small enough to be computationally viable. Furthermore, *C. albicans* allows for similar mechanisms that are found in eucaryotes. Indeed, epigenetic mechanisms across animals, plants, and fungi include DNA methylation as a common epigenetic signalling mechanism, and it is present in *C. albicans*. A putative histone H1 has been identified [36]. Whereas these are technical decisions, we also wanted to select a species that should have a clinical prevalence. It consists of eight sets of chromosome pairs whose complete genome sequence is available. The raw data of the MNase-seq are available from the Gene Expression Omnibus (GSM1542419) and were measured by Puri et al. [37]. We also accessed the processed iNPS data in the NucMap database by Zhao et al. [38].

1.2. Methods

To obtain a consistent classification of the nucleosomal positioning data in genome-wide classes, we perform the following steps (explained in more detail below the list):

1. Each chromosome is divided into segments of 75 kbp of length.
2. For every chromosome, the positioning data are coarse-grained.
3. The coarse-grained nucleosome positioning data are used to calculate auto-correlation functions over the different sections.
4. A distance matrix is calculated over all the auto-correlation function data.
5. These segments are clustered. Various distance matrix and clustering algorithms are used to generalize the results.

1.2.1. Genome Section Classification

In order to extract the global pattern for areas in a genome, the whole genome is separated into sections with equal length. The section length L is an important scale parameter and needs to be properly set. L should not be too large to avoid all features from different areas bounded together. At the same time, L also should not be too small; otherwise, the global structure is flooded by the subtle differences and becomes a pattern for only a single nucleosome. The single nucleosome wrapping length L_n can be used as a lower bound for the choice of L. However, to obtain a relevant structure, we require that $L >> L_n$. Considering the nucleosome length L_n is about 147 bp [3,4], L is chosen to be 50 kbp. Additionally, to avoid boundary effects, for each section, a 12.5 kbp intersection on both sides with its neighbor is added. Hence, the total section length L is 75 kbp. This binning is applied to each chromosome. Chr. 2 for example, with a length of 2,231,883 bp, is separated into 44 sections.

1.2.2. Coarse Graining

The idea of coarse graining is an established ansatz and tool in physics to describe complex systems on a scale that allows identifying structure. Typically, the structure appears as a collective phenomenon among smaller entities. The idea is to eliminate degrees of freedom, i.e., find a representation of the system on a larger time or space scale, iteratively moving to larger scales without changing the system. Over the last few years,

coarse graining has emerged as a way to model large complex systems and has successfully been applied to other biomolecules such as proteins [39].

After the whole genome is separated into sections, coarse graining is applied for each section. The method we implemented for coarse graining is the rolling mean method [40]. This method takes a window with a certain size (e.g., $b = 5$ kbp), computes the averaged value of the nucleosome positioning inside the window, and moves the window to the following location. After this value is computed for each location, coarse-grained data on the scale of the window size are returned. Here, Python pandas.DataFrame.rolling [41] is used to obtain the coarse-graining. To exclude the effect of telomeres, discrete ends of the sections and incorporation of the window size and offset was chosen to be at least

$$\text{offset} \geq \text{window size}/2 \tag{1}$$

1.2.3. Auto-Correlation Function Calculation

An auto-correlation function is a well-known approach in physics and pattern recognition, capturing the inner interaction pattern inside the data [40]. Particularly for structures that are liquid-like, the auto-correlation function, or in this context the radial distribution function, identifies typical length scales and patterns.

For each section j, it is applied on all the coarse-grained data ρ_j. The normalized auto-correlation function $C^j(\tau)$ with respect to distance τ for section j is:

$$C^{\alpha,j}(\tau) = \frac{E[(\rho_i^{\alpha,j} - \mu^{\alpha,j})(\rho_{i+\tau}^{\alpha,j} - \mu^{\alpha,j})]}{(\sigma^{\alpha,j})^2} \tag{2}$$

where $\rho_i^{\alpha,j}$ is the data at position i within the section j of chromosome α. $E(\dots)$ is the mean of everything in the parentheses over all indices i. μ^j is the mean of ρ and σ^j is the variance for the section j. Thus, associated with each section j is the function $C^{\alpha,j}(\tau)$ of chromosome α; hence, at the end, we will have N functions $C^{\alpha,j}(\tau)$ where N is the section number for the particular chromosome.

1.2.4. Distance Matrix Calculation

To classify the functions, a similarity measure is applied, and a resulting distance matrix is computed. The distance matrix is a square matrix containing the pairwise distances between all the elements available in the dataset, measuring the proximity between the correlation functions. Interpreting the functions as high-dimensional vectors, we use the p-norm to define the distance d_p between two functions:

$$d_p(a,b) = \|a-b\|_p = \left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{1/p} \tag{3}$$

where a and b are the functions in the form of vectors. For $p = 2$, the p-norm corresponds to the Euclidean distance.

1.2.5. Clustering

To identify the unique nucleosome organisation or distribution function, there is a need to cluster the sections together on the basis of similarity among them. We used a clustering approach, i.e., hierarchical clustering [42]. This is an unsupervised algorithm that groups similar objects into groups called clusters. It uses a distance matrix to identify the two closest clusters first and then merge the two most similar clusters. This iterative process continues until the clusters are merged to get distinct clusters in a hierarchical manner.

Hierarchical clustering builds a hierarchy of clusters using two methods: agglomerative and divisive algorithms. We used the former, i.e., the Ward method [43], where each observation starts in its own cluster and pairs of clusters are merged, moving up the hierarchy.

1.2.6. Statistical Distributions Fitting

Fitting of the distributions was performed using the scipy stats package [44] under Python.

2. Results

The first indication of non-trivial ordering is given by the distribution of the nucleosome positioning data. The binary nucleosome positioning data for all chromosomes of *Candida albicans* (NucMap database [38]) are subjected to the described coarse graining and then analyzed (see the histogram of densities in the Supplementary Information Figures S1–S3, and Tables S1 and S2). The genome-wide normalised nucleosome density shows a non-Gaussian behaviour with a slight negative skew. Overall, a log-logistic distribution gives the best consistent fit for all chromosomes compared to a normal distribution on the same bin size and rolling average for all chromosomes.

Recall that each chromosome is divided into chunks of 75 kbp with 25 kbp overlapping on each side. The auto-correlation of each chunk is obtained on the coarse-grained nucleosome positioning data. The respective correlation function of each section for all chromosomes are shown in Figure 2 and in detail in Supplementary Information (Figures S9–S16). Shown are the correlation functions on the coarse-grained scale as well as a further smoothing to make the features that are common among a class more apparent (see below). The colour bar indicates the class. Even though there are variations within a class, certain common features are seen. These features are the first and second peak structure, the height of the peaks, and how long a structure persists. Recall that the zero line indicates that there is no correlation; i.e., there, the structure is that of a gas or an unordered behaviour. The first peak indicates an increased probability to find a coarse-grained nucleosome at the distance of the peak position, and the same applies to the second and additional peaks. If these peaks are of similar height, then there is a stronger long-range ordering. A particular example showing similar heights up to a third peak is in section 12 of chromosome no. 3 (see Supplementary Information Figure S11), while section 6 shows a drop in the peak heights. Nevertheless, due to the overall similarity, these fall into the same class.

With diminishing height, the likelihood of the ordering and the strictness of ordering vanishes. Notice that for some of the sections (within one class), many sub-peaks or side-peaks exist, indicating possible sub-orderings. An example on the more extreme side is chromosome 3 and sections such as 3, 5, 16, etc. Overall, the short-range order is much less pronounced. The orange smoothed line indicates that in this class, the salient feature is a smoothly decreasing function indicating a different kind of order than for the class with sections 0, 8 and 12, etc.

Even looking at the correlation functions without the indicated class mapping shows that there are universal features beyond fluctuations. Within a class, a more or less pronounced ordering feature is visible. Comparing the different correlation data between the chromosomes, these become apparent.

These observations can be proven more rigorously by applying similarity measures between the correlation functions. Figure 2 shows the resulting distance matrix between all chromosomes and all sections (the individual results are shown in the Supplementary Information Figures S4–S6). Shown is the distance matrix after reordering on the basis of similarity between sections. The colour indicates the similarity between the correlation functions. Notice the patterns that emerge from the sorting of the data into classes.

These classes, represented by different colours, are shown in the dendrogram. These classes were obtained by hierarchical clustering. In the lower part of the figure on the left are the typical correlation functions representing the corresponding class with its colour code. The orange-coloured class shows a fairly regular pattern and closely spaced ordering on a short scale, such as tightly packed heterochromatin, whereas the light blue class has lost the regularity and shows a less stringent regular but still pronounced pattern on a slightly larger scale. The blue-coloured class shows a rather very irregular pattern compared to the other two classes and corresponds more to euchromatin.

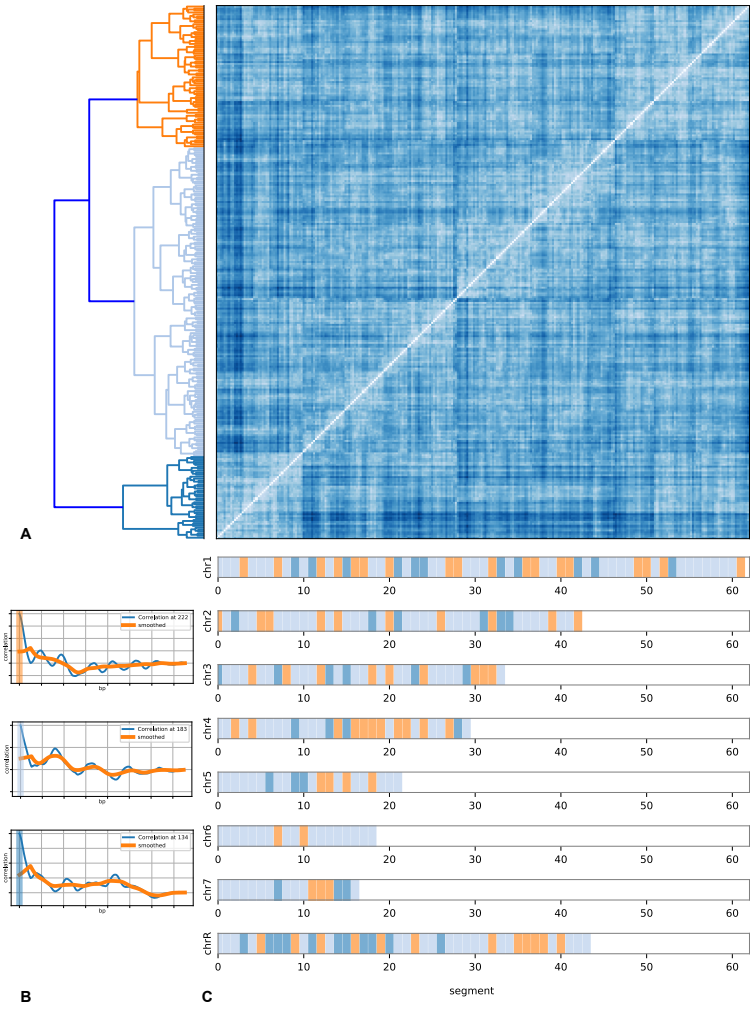


Figure 2. (**A**) shows the genome-wide distance matrix between the correlation functions between segments of size 75 kbp. Hierarchical clustering was applied to identify common patterns. The matrix was sorted according to the patterns. The left side shows the clustering. (**B**) shows the coarse-grained nucleosomal density correlation functions of *Candida albicans* at 5 kb coarse graining. (**C**) shows the genome-wide distribution of segments with colours corresponding to the classification. White space is due to not all chromosomes having the same length. The pattern classification was done genome-wide to yield three main patterns. These three patterns were assigned colours, and the segments of each chromosome corresponding to one of the three patterns are marked. The orange-coloured pattern is characterised by a closely and fairly regularly spaced ordering similar to the tightly packed heterochromatin. The dark and light-coloured blue patterns have lost the regularity and the longer range of the order and thus correspond more to euchromatin. However, note that both these two classes have a huge variety of subclasses. This is not surprising in the sense that one would expect a larger variety of not so ordered patterns in one dimension than for ordered patterns in one dimension.

These observations are consistent with the typical classification from microscopy data into hetero- and euchromatin. The data show that the orange and light blue classes can be mapped on heterochromatin. Thus, the blue-coloured class is euchromatin. The data also show that still, within any of these classes, the features have many sub-features that we salvaged for the larger patterns to allow a "coarse-grained" view on the ordering of the nucleosomes. These sub-features compose elaborated chromatin states such as solenoid [45], zig-zag ribbon [46], or other structures [47], which demand a cross-correlation analysis with CTCF binding sites [48], CpG island position [49], and other data.

Notice that this partitioning into classes is genome-wide. A consistent classification can be established. This is shown in the mapping of the positions of the section to the chromosomes. Notice that, as expected, not a random mixture of the three colours emerges but rather a clear pattern. The larger chromosomes appear to have more internal structuring compared to the smaller chromosomes that are more homogeneous in their internal structure. The partitioning into a clear pattern, genome-wide is not limited to species *Candida albicans*, but the pipeline is generalised and can be used for any species in which the whole genome has been sequenced.

3. Discussion

The structural organisation of the genome depends on the patterns of nucleosome positioning and their distribution in the genome. At a higher scale, the nucleosome positioning distribution varies across the chromosomes, which appear to be conserved along the entire genome. The classification of the chromosomes into segments of the distinct nucleosomal distribution shown here is in line with earlier studies. Although two major classifications of the chromosomal region as heterochromatin and euchromatin are suggested, we find that their organisations can be further subdivided. Nucleosomes can be well-positioned to form phased and unphased arrays consisting of regularly spaced nucleosomes or can be fuzzy to form irregular arrays of nucleosomes. The three distinct nucleosome distribution patterns along the genome obtained in our result are in agreement with this study. Moreover, further classification of nucleosomal distribution is obtained along each chromosome. Around five to seven different nucleosome distribution patterns are observed for all chromosomes. However, for the entire genome, three patterns are found to be conserved.

We have analysed the effect for different $p = 2, 7$ in the p-norm on the outcome of the clustering of similar correlation functions, and the outcome comes to be similar for all p. For high p values, some of the clusters split into further clusters. In addition, the cosine similarity norm was tested for further verification, yielding similar clustering (see Supplementary Information Video S1). This rules out that the clustering is an artifact of the model and its architecture.

Around five patterns of chromosomal organisation are obtained for each chromosome by analysing the nucleosome positioning data distribution. These patterns obtained are generally coincident with gene densities and lead to the distinct spatial organisation of genomic DNA. The genome's hierarchical structure–function relationship [12] is governed by chromatin domains and their higher-order folding. The formation of chromatin boundaries and associated TADs are controlled by the nucleosome distribution patterns. Recent studies by Wiese et al. [16] suggested that domain formation and genome organisation can be predicted with nucleosome positioning only. Pulivarty et al. [50] primarily focused on nucleosome studies, which are limited to a very local individual promoter and enhancer but can be a more general mechanism by which cells can regulate the accessibility of the genome during development at different scales. After an extensive analysis of nucleosome positioning data, the way of organisation of nucleosomal distribution patterns is found to be different at different scales and for different chromosomes. The distinct patterns obtained from our calculation correspond to different ways of nucleosome positioning and may control domain formation and genome organisation in the cell. However, the three distinct patterns of nucleosome organisation that appeared to be conserved in the genome show the global consistency of distribution patterns inside the genome. The consistency in different kinds of distinct patterns observed in the genome corresponds to identical gene densities and similar expression regions for specific locations inside the cell.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/life12040541/s1, Figures S1–S27: Nucleosome Correlation Data; Tables S1 and S2: Nucleosome Correlation Data; Video S1: Coarse Graining.

Author Contributions: All authors were involved in the conception, processing of the data, analysis and draftting of the manuscript. D.W.H. and N.N.O. supervised the work. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. Kunhe Li would like to acknowledge funding by the Chinese Scholarship Council (CSC). Sujeet Kumar Mishra would like to acknowledge funding by the India government Ministry of Science and Technology, Department of Biotechnology (DBT)-Interdisciplinary Research Center for Scientific Computing (IWR) PhD program.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Oluwadare, O.; Highsmith, M.; Cheng, J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol. Proced. Online* **2019**, *21*, 7. [CrossRef]
2. Jerkovic, I.; Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 511–528. [CrossRef]
3. Routh, A.; Sandin, S.; Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 8872–8877. [CrossRef]
4. Bohr, J.; Olsen, K. The size of the nucleosome. *arXiv* **2012**, arXiv:1102.0761.
5. Staneva, D.; Georgieva, M.; Miloshev, G. Kluyveromyces lactis genome harbours a functional linker histone encoding gene. *FEMS Yeast Res.* **2016**, *16*, fow034. [CrossRef] [PubMed]
6. Bohn, M.; Diesinger, P.; Kaufmann, R.; Weiland, Y.; Müller, P.; Gunkel, M.; Ketteler, A.; Lemmer, P.; Hausmann, M.; Heermann, D.; et al. Localization Microscopy Reveals Expression-Dependent Parameters of Chromatin Nanostructure. *Biophys. J.* **2010**, *99*, 1358–1367. [CrossRef]
7. Tchasovnikarova, I.A.; Kingston, R.E. Beyond the Histone Code: A Physical Map of Chromatin States. *Mol. Cell* **2018**, *69*, 5–7. [CrossRef] [PubMed]
8. Struhl, K.; Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **2013**, *20*, 267–273. [CrossRef] [PubMed]
9. Hilbert, L.; Sato, Y.; Kuznetsova, K.; Bianucci, T.; Kimura, H.; Jülicher, F.; Honigmann, A.; Zaburdaev, V.; Vastenhouw, N.L. Transcription organizes euchromatin via microphase separation. *Nat. Commun.* **2021**, *12*, 1360. [CrossRef] [PubMed]
10. Dekker, J.; Rippe, K.; Dekker, M.; Kleckner, N. Capturing Chromosome Conformation. *Science* **2002**, *295*, 1306–1311. [CrossRef] [PubMed]
11. van Berkum, N.L.; Lieberman-Aiden, E.; Williams, L.; Imakaev, M.; Gnirke, A.; Mirny, L.A.; Dekker, J.; Lander, E.S. Hi-C: A method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **2010**, *6*, 1869. [CrossRef]
12. Beagan, J.A.; Phillips-Cremins, J.E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **2020**, *52*, 8–16. [CrossRef] [PubMed]
13. Ghavi-Helm, Y.; Jankowski, A.; Meiers, S.; Viales, R.R.; Korbel, J.O.; Furlong, E.E.M. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **2019**, *51*, 1272–1282. [CrossRef] [PubMed]
14. Nora, E.P.; Lajoie, B.R.; Schulz, E.G.; Giorgetti, L.; Okamoto, I.; Servant, N.; Piolot, T.; van Berkum, N.L.; Meisig, J.; Sedat, J.; et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **2012**, *485*, 381–385. [CrossRef]
15. Quentin, S.; Frédéric, B.; Giacomo, C. Principles of genome folding into topologically associating domains. *Sci. Adv.* **2022**, *5*, eaaw1668. [CrossRef]
16. Wiese, O.; Marenduzzo, D.; Brackley, C.A. Nucleosome positions alone can be used to predict domains in yeast chromosomes. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 17307. [CrossRef] [PubMed]
17. Kharerin, H.; Bai, L. Thermodynamic modeling of genome-wide nucleosome depleted regions in yeast. *PLoS Comput. Biol.* **2021**, *17*, e1008560. [CrossRef] [PubMed]
18. Drew, H.R.; Travers, A.A. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **1985**, *186*, 773–790. [CrossRef]
19. Chung, H.R.; Vingron, M. Sequence-dependent Nucleosome Positioning. *J. Mol. Biol.* **2009**, *386*, 1411–1422. [CrossRef]
20. Parmar, J.J.; Marko, J.F.; Padinhateeri, R. Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence. *Nucleic Acids Res.* **2014**, *42*, 128–136. [CrossRef] [PubMed]

21. Angermayr, M.; Oechsner, U.; Bandlow, W. Reb1p-dependent DNA bending effects nucleosome positioning and constitutive transcription at the yeast profilin promoter. *J. Biol. Chem.* **2003**, *278*, 17918–17926. [CrossRef]
22. Rippe, K.; Schrader, A.; Riede, P.; Strohner, R.; Lehmann, E.; Längst, G. DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15635–15640. [CrossRef] [PubMed]
23. Wippo, C.J.; Israel, L.; Watanabe, S.; Hochheimer, A.; Peterson, C.L.; Korber, P. The RSC chromatin remodelling enzyme has a unique role in directing the accurate positioning of nucleosomes. *EMBO J.* **2011**, *30*, 1277–1288. [CrossRef]
24. Shim, Y.S.; Choi, Y.; Kang, K.; Cho, K.; Oh, S.; Lee, J.; Grewal, S.I.; Lee, D. Hrp3 controls nucleosome positioning to suppress non-coding transcription in eu-and heterochromatin. *EMBO J.* **2012**, *31*, 4375–4387. [CrossRef] [PubMed]
25. Helbo, A.S.; Lay, F.D.; Jones, P.A.; Liang, G.; Grønbæk, K. Nucleosome Positioning and NDR Structure at RNA Polymerase III Promoters. *Sci. Rep.* **2017**, *7*, 41947. [CrossRef]
26. Klein, D.C.; Hainer, S.J. Genomic methods in profiling DNA accessibility and factor localization. *Chromosome Res.* **2020**, *28*, 69–85. [CrossRef] [PubMed]
27. Mieczkowski, J.; Cook, A.; Bowman, S.K.; Mueller, B.; Alver, B.H.; Kundu, S.; Deaton, A.M.; Urban, J.A.; Larschan, E.; Park, P.J.; et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **2016**, *7*, 11485. [CrossRef]
28. Chen, W.; Liu, Y.; Zhu, S.; Green, C.D.; Wei, G.; Han, J.D.J. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat. Commun.* **2014**, *5*, 4909. [CrossRef]
29. Georgakilas, G.K.; Perdikopanis, N.; Hatzigeorgiou, A. Solving the transcription start site identification problem with ADAPT-CAGE: A Machine Learning algorithm for the analysis of CAGE data. *Sci. Rep.* **2020**, *10*, 877. [CrossRef] [PubMed]
30. Oiwa, N.N.; Cordeiro, C.E.; Heermann, D.W. The Electronic Behavior of Zinc-Finger Protein Binding Sites in the Context of the DNA Extended Ladder Model. *Front. Phys.* **2016**, *4*, 13. [CrossRef]
31. Singh, A.K.; Mueller-Planitz, F. Nucleosome positioning and spacing: From mechanism to function. *J. Mol. Biol.* **2021**, *433*, 166847. [CrossRef] [PubMed]
32. Schöpflin, R.; Teif, V.B.; Müller, O.; Weinberg, C.; Rippe, K.; Wedemann, G. Modeling nucleosome position distributions from experimental nucleosome positioning maps. *Bioinformatics* **2013**, *29*, 2380–2386. [CrossRef]
33. Flores, O.; Orozco, M. nucleR: A package for non-parametric nucleosome positioning. *Bioinformatics* **2011**, *27*, 2149–2150. [CrossRef] [PubMed]
34. Chen, K.; Xi, Y.; Pan, X.; Li, Z.; Kaestner, K.; Tyler, J.; Dent, S.; He, X.; Li, W. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **2013**, *23*, 341–351. [CrossRef]
35. Price, R.J.; Weindling, E.; Berman, J.; Buscaino, A. Chromatin Profiling of the Repetitive and Nonrepetitive Genomes of the Human Fungal Pathogen Candida albicans. *mBio* **2019**, *10*, e01376-19. [CrossRef] [PubMed]
36. Skrzypek, M.S.; Binkley, J.; Binkley, G.; Miyasato, S.R.; Simison, M.; Sherlock, G. The Candida Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* **2017**, *45*, gkw924. [CrossRef]
37. Puri, S.; Lai, W.K.M.; Rizzo, J.M.; Buck, M.J.; Edgerton, M. Iron-responsive chromatin remodelling and MAPK signalling enhance adhesion in Candida albicans. *Mol. Microbiol.* **2014**, *93*, 291–305. [CrossRef] [PubMed]
38. Zhao, Y.; Wang, J.; Liang, F.; Liu, Y.; Wang, Q.; Zhang, H.; Jiang, M.; Zhang, Z.; Zhao, W.; Bao, Y.; et al. NucMap: A database of genome-wide nucleosome positioning map across species. *Nucleic Acids Res.* **2019**, *47*, D163–D169. [CrossRef] [PubMed]
39. Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A.E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898–7936. [CrossRef]
40. Reichl, L.E. *A Modern Course in Statistical Physics*, 4th ed.; Wiley: Hoboken, NJ, USA, 2016.
41. Pandas Development Team T. Pandas-Dev/Pandas: Pandas. 2020. Available online: https://doi.org/10.5281/zenodo.3509134 (accessed on 12 February 2022).
42. Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2005.
43. Everitt, B.S.; Landau, S.; Leese, M. *Cluster Analysis*, 4th ed.; Oxford University Press: Oxford, UK, 2001.
44. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
45. Finch, J.T.; Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 1897–1901. [CrossRef] [PubMed]
46. Diesinger, P.M.; Kunkel, S.; Langowski, J.; Heermann, D.W. Histone depletion facilitates chromatin loops on the kilobasepair scale. *Biophys. J.* **2010**, *99*, 2995–3001. [CrossRef] [PubMed]
47. Williams, S.P.; Athey, B.D.; Muglia, L.J.; Schappe, R.S.; Gough, A.H.; Langmore, J.P. Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophys. J.* **1986**, *49*, 233–248. [CrossRef]
48. Norio Oiwa, N.; Li, K.; Cordeiro, C.E.; Heermann, D.W. Prediction and Comparative Analysis of CTCF Binding Sites based on a First Principle Approach. *arXiv* **2021**, arXiv:2110.10508.

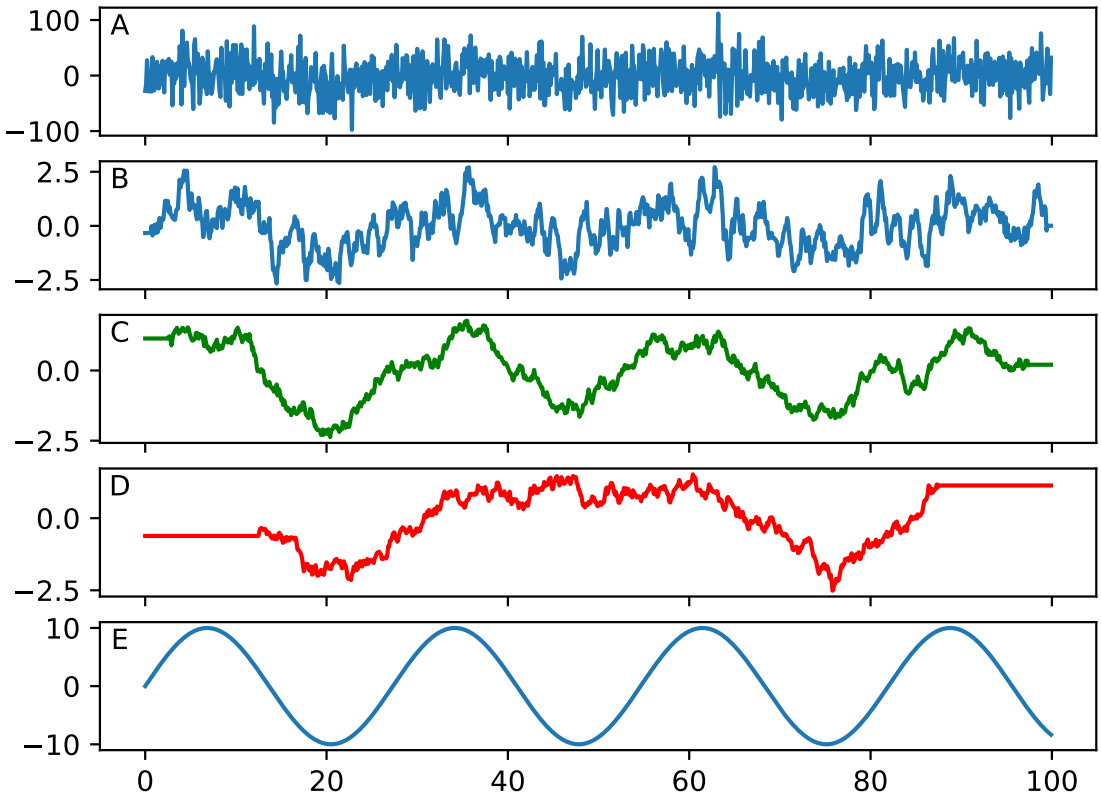
49. Gardiner-Garden, M.; Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **1987**, *196*, 261–282. [CrossRef]
50. Pulivarthy, S.R.; Lion, M.; Kuzu, G.; Matthews, A.G.W.; Borowsky, M.L.; Morris, J.; Kingston, R.E.; Dennis, J.H.; Tolstorukov, M.Y.; Oettinger, M.A. Regulated large-scale nucleosome density patterns and precise nucleosome positioning correlate with V(D)J recombination. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 201605543. [CrossRef] [PubMed]

Supporting information

1.1 Coarse-Graining

Coarse-graining is a procedure that has successfully been developed and applied to critical phenomena in physics. The basic idea is that each system has a fundamental length scale on which the physical interactions play out. While there are interactions such as excluded volume interaction or Van-der-Waals interactions on a short scale, these all add up to the relevant scale given by the typical correlation length of the system. If the correlations are small, such as in a gas where the constituents particles almost never interact then the fundamental interactions determine the physical scale. For more dense system, there is a scale, the correlation length, on which the system needs to be described.

The coarse-graining procedure is demonstrated in Figure. Panel A shows a noisy signal based on the data shown in panel E. For panels B to D we increase the coarse-graining length L from 10 to 50 and to 250. The first coarse-graining step shown in panel B already recovers some aspects of the underlying data. The second coarse-graining length $L = 50$ essentially has recovered the underlying structure while for $L = 250$ the signal is too much washed out.



1.2 Nucleosome Density

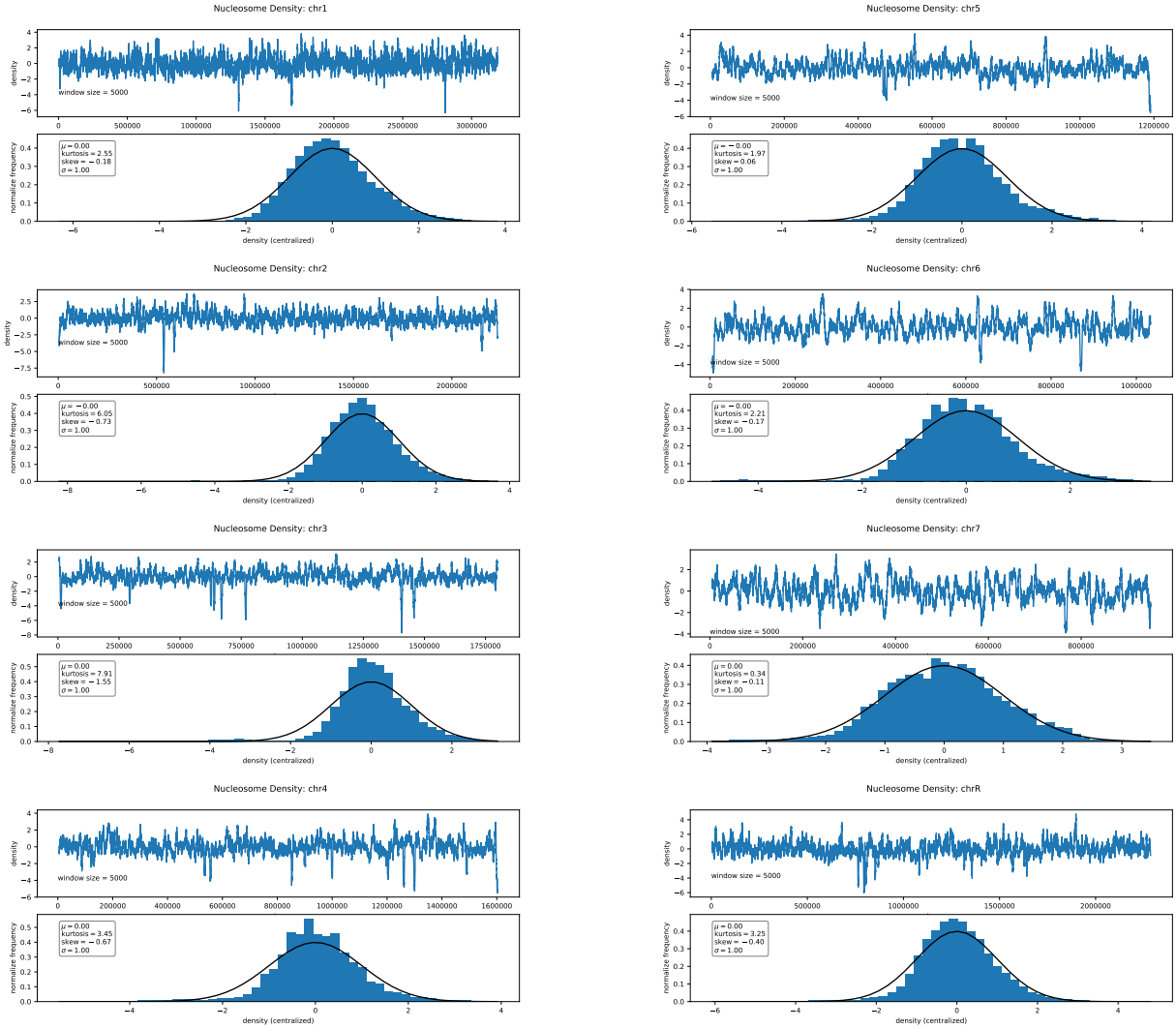


Figure S1: Shown is the nucleosomal density after applying a rolling average with a window size of 5000 of all of the chromosomes (upper panels). The lower panels show the corresponding histogram of the densities with a bin size of 50. The black line is the fit with a gaussian distribution.

1.2.1 Nucleosome Density at $b = 2500$

Chromosome	Distribution	chi_square	D_statistic
chr1	fisk	1.675740e+05	0.026701
chr1	norm	4.029705e+05	0.047346
chr2	fisk	2.215488e+05	0.034197
chr2	norm	4.888579e+05	0.049294
chr3	fisk	2.315703e+05	0.038608
chr3	norm	1.174085e+06	0.083966
chr4	fisk	1.530916e+05	0.034538
chr4	norm	6.824904e+05	0.070372
chr5	fisk	9.322028e+04	0.030918
chr5	norm	2.783759e+05	0.056306
chr6	fisk	1.280710e+05	0.037654
chr6	norm	2.753396e+05	0.052656
chr7	fisk	5.100021e+04	0.032512
chr7	norm	7.679109e+04	0.031660
chrR	fisk	2.258400e+05	0.033580
chrR	norm	6.023527e+05	0.054608

Table S1: The Fisk distribution, also known as the log-logistic distribution gives the best consistent fit. The fit was done for the bin size of 50 and the rolling average of size 5000. Statistical Kolmogorov-Smirnov test for goodness of fit was done using SciPy.org scipy.stats.kstest function **?**. The D statistic is the absolute max distance (supremum) between the CDFs of the two samples. All results show small values D values corresponding to p-values close to 1, the log-logistic distribution may explain the data.

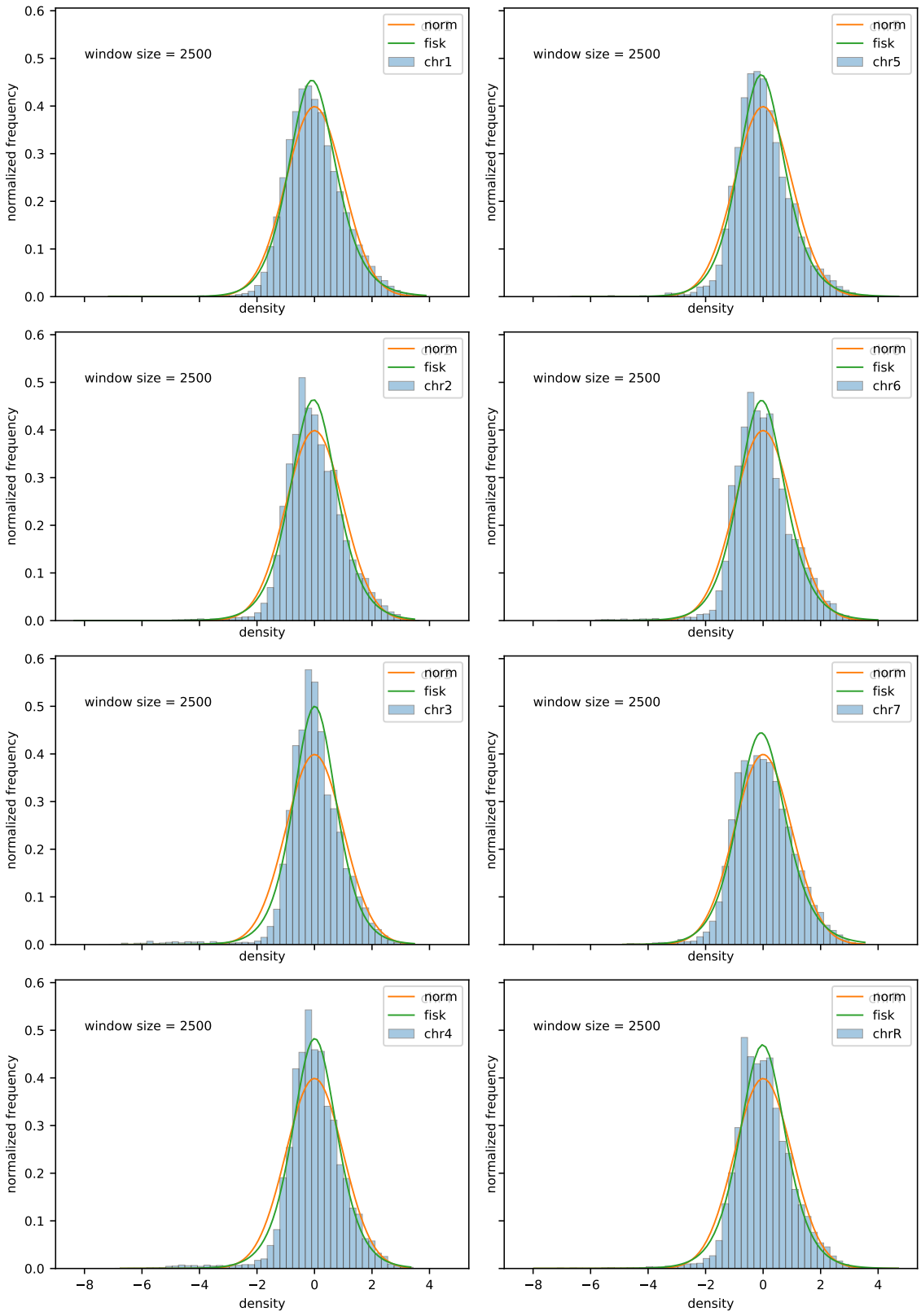


Figure S2: Normalized nucleosome density distributions for all of the chromosomes. The data shows the non-gaussian behavior (red line). For comparison a fit to a Log-logistic distribution is shown yielding a much better consistent fit. The bin size was 50 and the rolling average of size 2500 was used.

1.2.2 Nucleosome Density at $b = 5000$

Chromosome	Distribution	chi_square	D_statistic
chr1	fisk	1.678610e+05	0.021809
chr1	norm	4.310806e+05	0.040372
chr2	fisk	1.011539e+05	0.024922
chr2	norm	4.873082e+05	0.048215
chr3	fisk	2.078474e+05	0.038179
chr3	norm	1.362080e+06	0.094966
chr4	fisk	9.270418e+04	0.027728
chr4	norm	6.014198e+05	0.069622
chr5	fisk	4.004712e+04	0.020815
chr5	norm	1.715085e+05	0.048451
chr6	fisk	1.347119e+04	0.020806
chr6	norm	1.609603e+05	0.038205
chr7	fisk	1.682594e+04	0.022172
chr7	norm	1.636277e+04	0.016584
chrR	fisk	9.955245e+04	0.025810
chrR	norm	4.967464e+05	0.052443

Table S2: The Fisk distribution, also known as the log-logistic distribution gives the best consistent fit. The fit was done for the bin size of 50 and the rolling average of size 5000. Statistical Kolmogorov-Smirnov test for goodness of fit was done using SciPy.org scipy.stats.kstest function **?**. The D statistic is the absolute max distance (supremum) between the CDFs of the two samples. As the all the results show small values D values corresponding to p-values close to 1, the log-logistic distribution may explain the data.

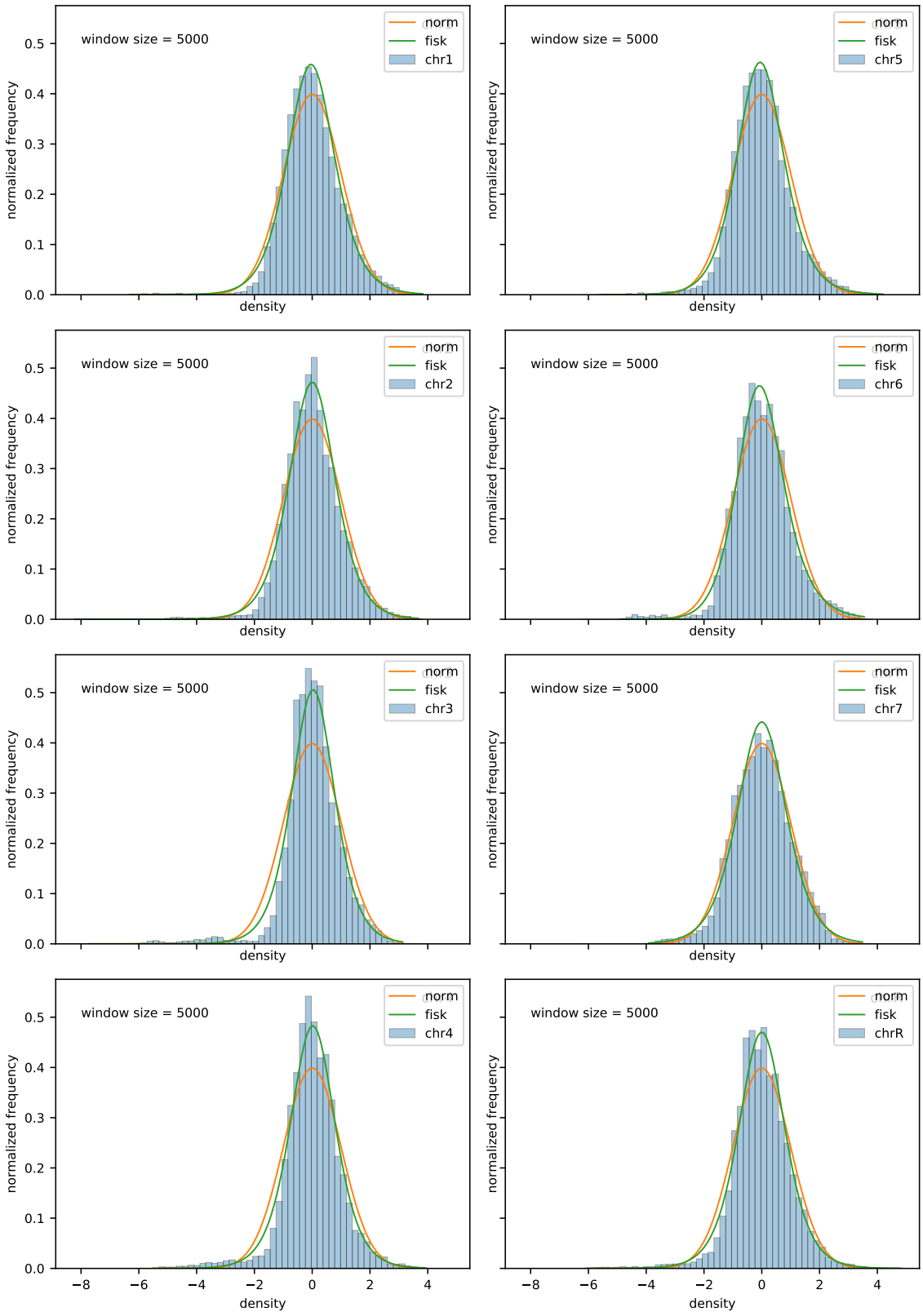


Figure S3: Normalized nucleosome density distributions for all of the chromosomes. The data shows the non-gaussian behavior (red line). For comparison a fit to a Log-logistic distribution is shown yielding a much better consistent fit. The bin size was 50 and the rolling average of size 5000 was used.

1.3 Distance Matrix for Individual Chromosomes

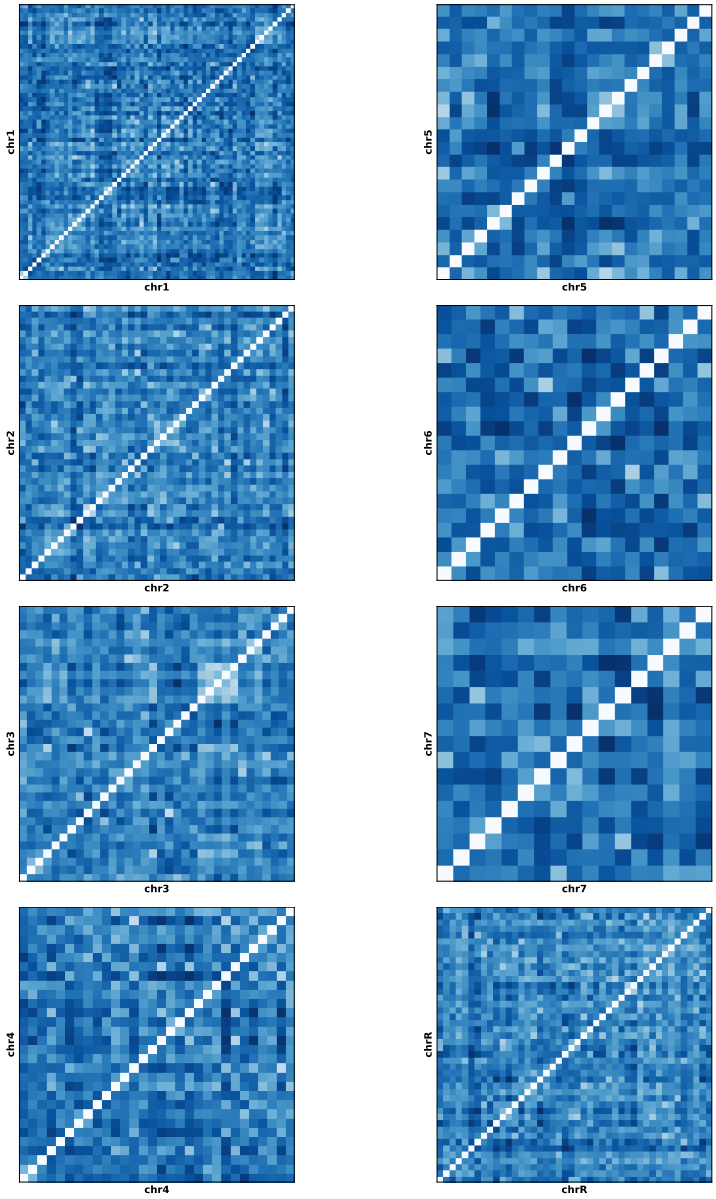


Figure S4: Shown are the distance matrices for all chromosomes. Distance refers to the distance between two correlation functions as measured by the euclidean distance (np.linalg.norm(x-y,ord=norm), with norm = 2. The ordering along the axes corresponds to the coarse-grained sections. The rolling average was of size 5000.

1.4 Clustering for Individual Chromosomes

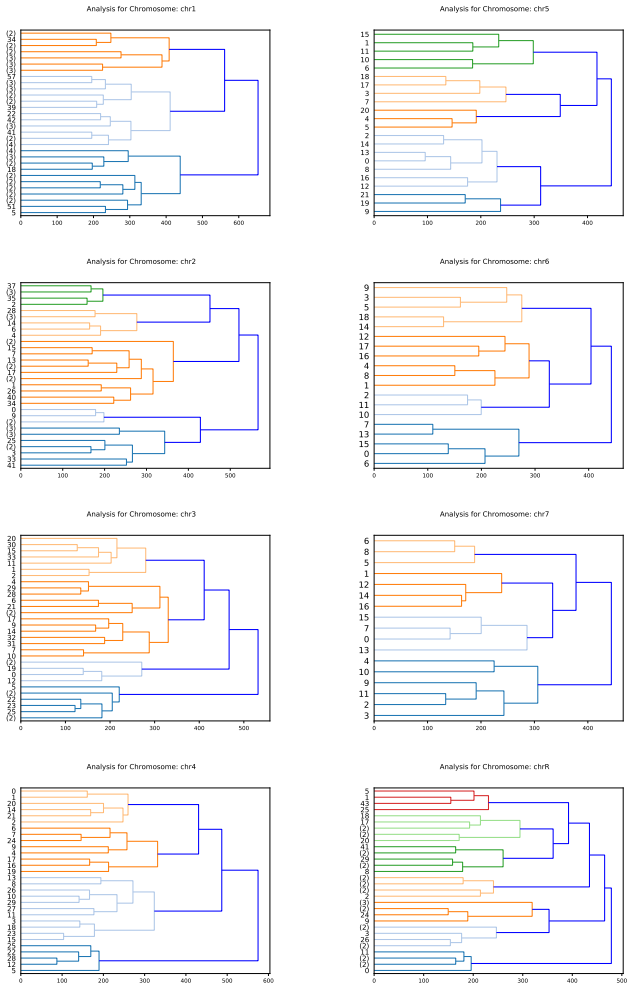


Figure S5: Shown are the dendrograms resulting from the distance matrices for all chromosomes. Results are for the hierarchical clustering on the individual chromosome. The Ward distance was used for the variance minimization algorithm used by SciPy **?**. The labels correspond to the distance matrix entries. Labels in parentheses give the number of labels corresponding to the leave. The rolling average was of size 5000. Labels in parentheses give the number of labels corresponding to the leaf.

1.5 Distance Matrix and Clustering for Individual Chromosomes

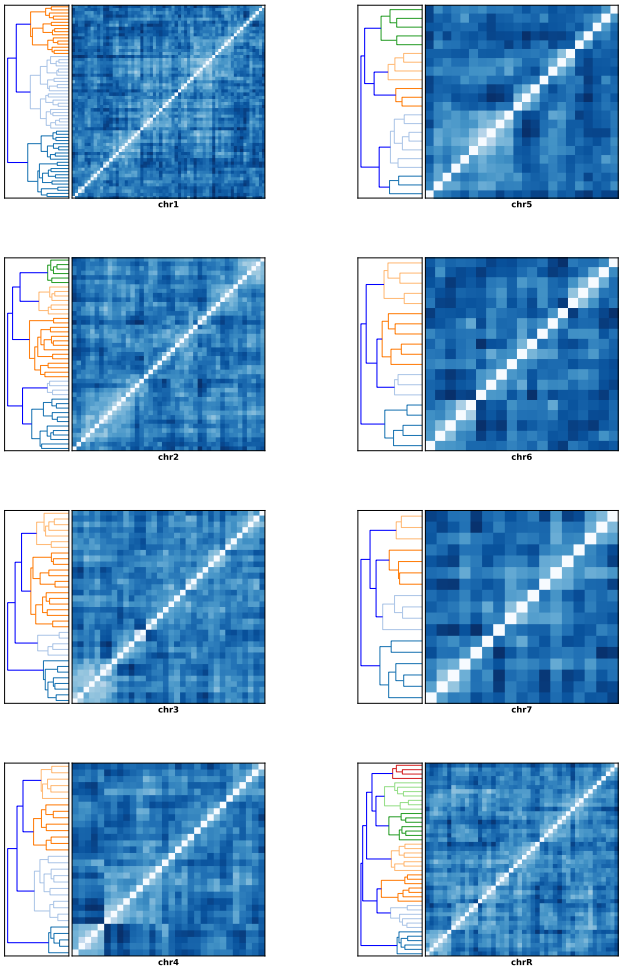


Figure S6: Shown are the distance matrices and corresponding dendrograms for all chromosomes. The matrix entries are sorted to correspond to the identified clusters. The rolling average was of size 5000.

1.6 Cluster Pattern in Chromosomes

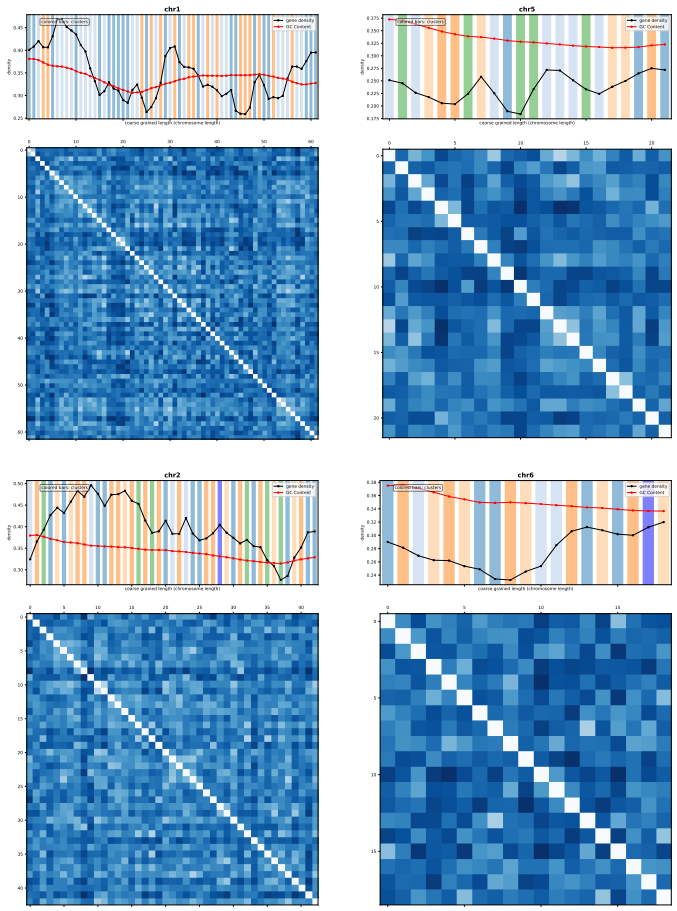


Figure S7: Part 1: Shown are the distance matrices and corresponding mapping of the pattern on the chromosomes. The matrix entries correspond to the positions on the chromosome. The rolling average was of size 5000.

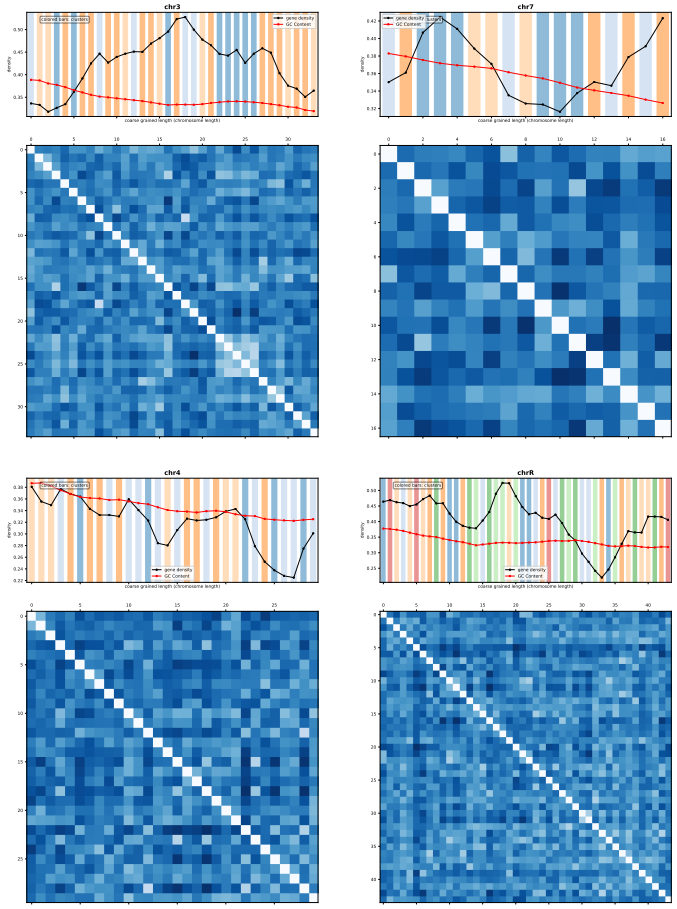


Figure S8: Part 2: Shown are the distance matrices and corresponding mapping of the pattern on the chromosomes. The matrix entries correspond to the positions on the chromosome. The rolling average was of size 5000.

1.7 Correspondence between Pattern and Correlation Function within individual Chromosome

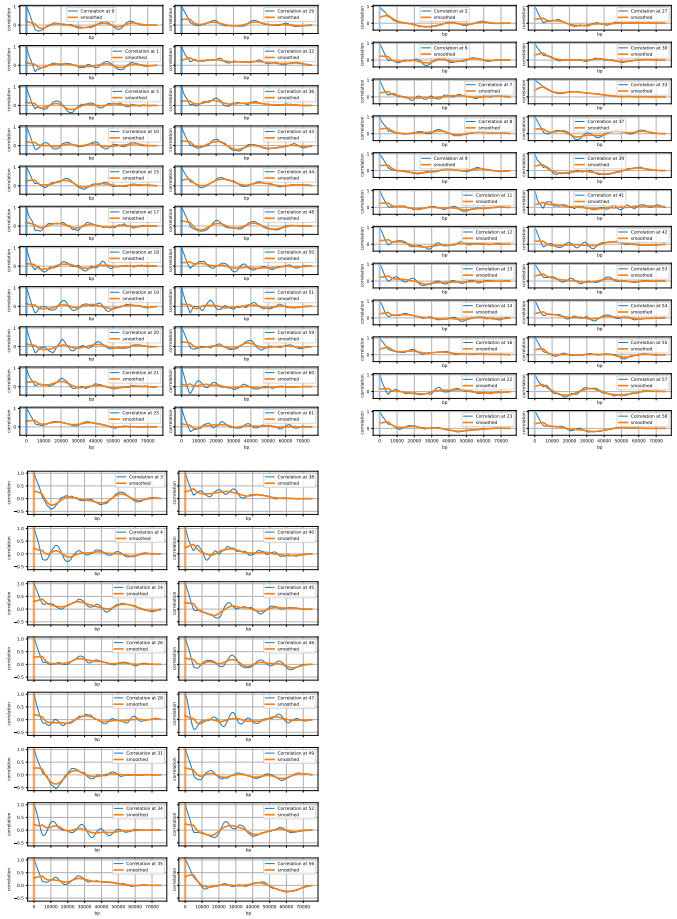


Figure S9: Shown are the correlation functions and the corresponding mapping of the pattern on the chromosome 1. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

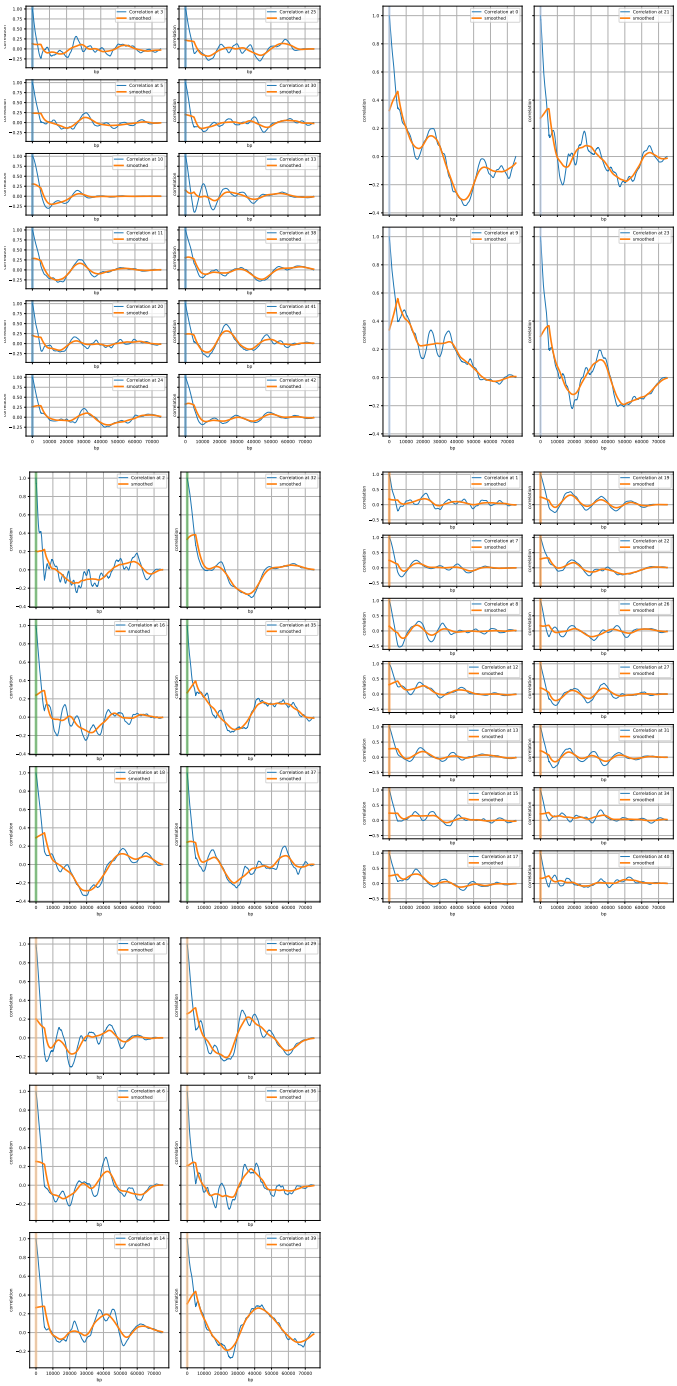


Figure S10: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 2. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

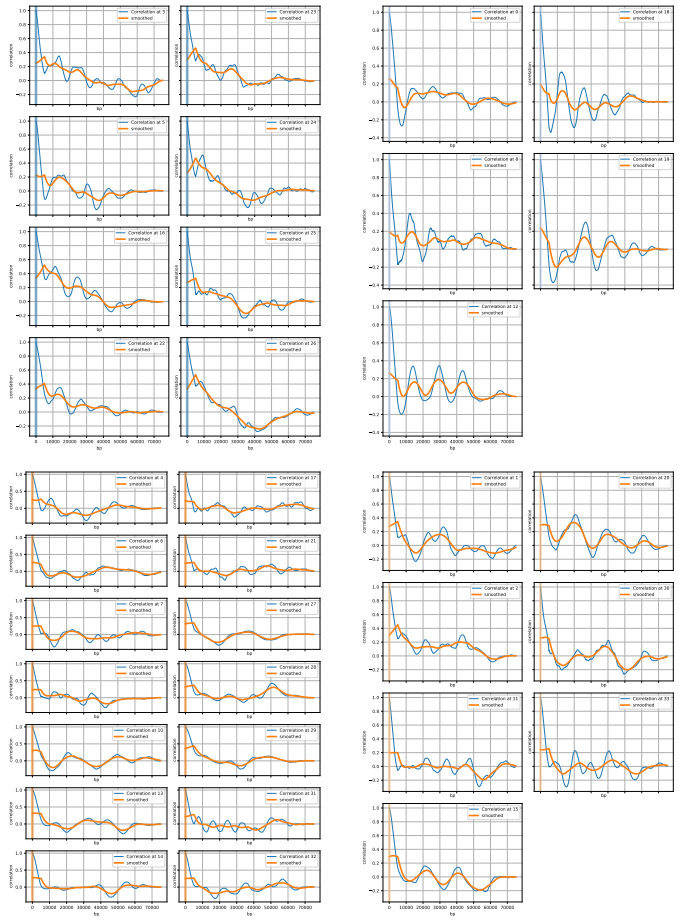


Figure S11: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 3. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

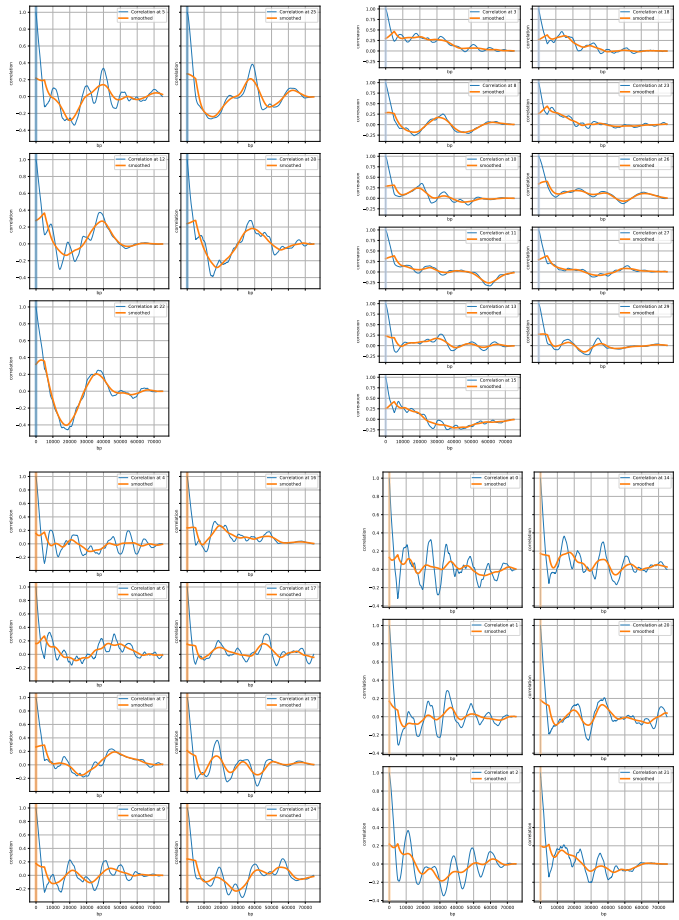


Figure S12: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 4. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

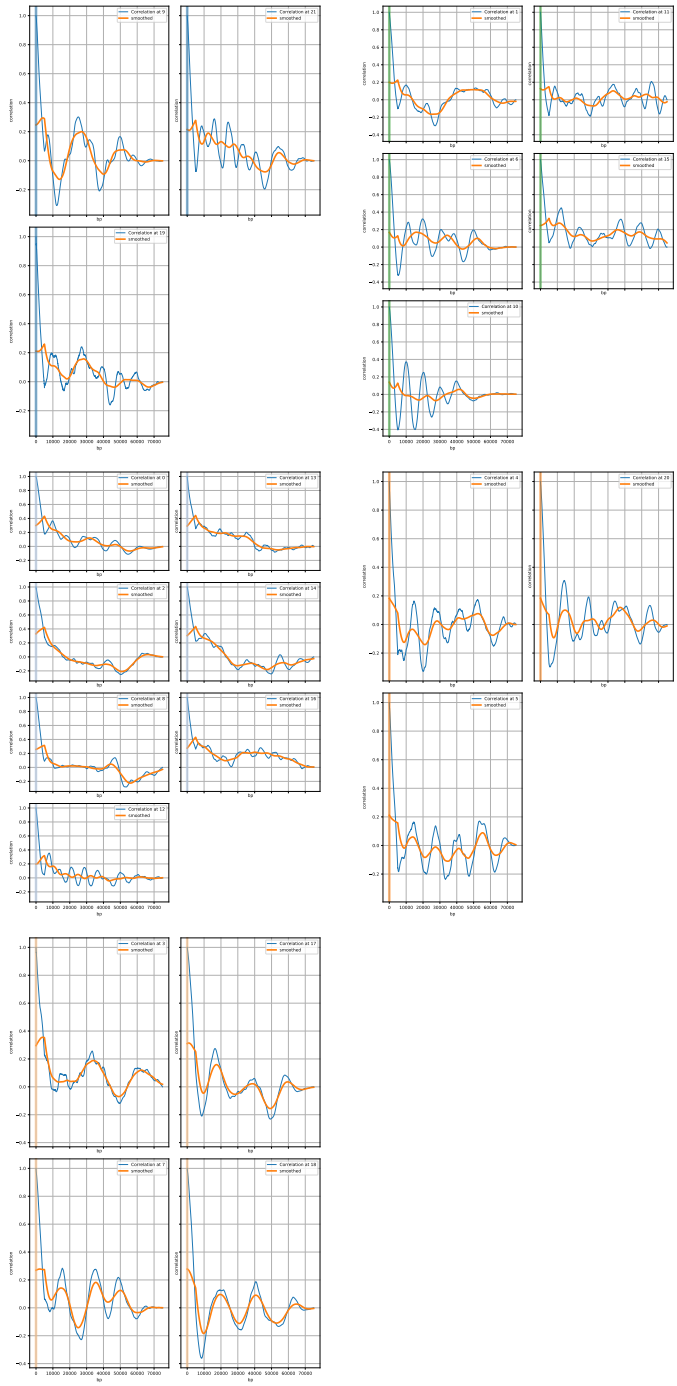


Figure S13: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 5. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

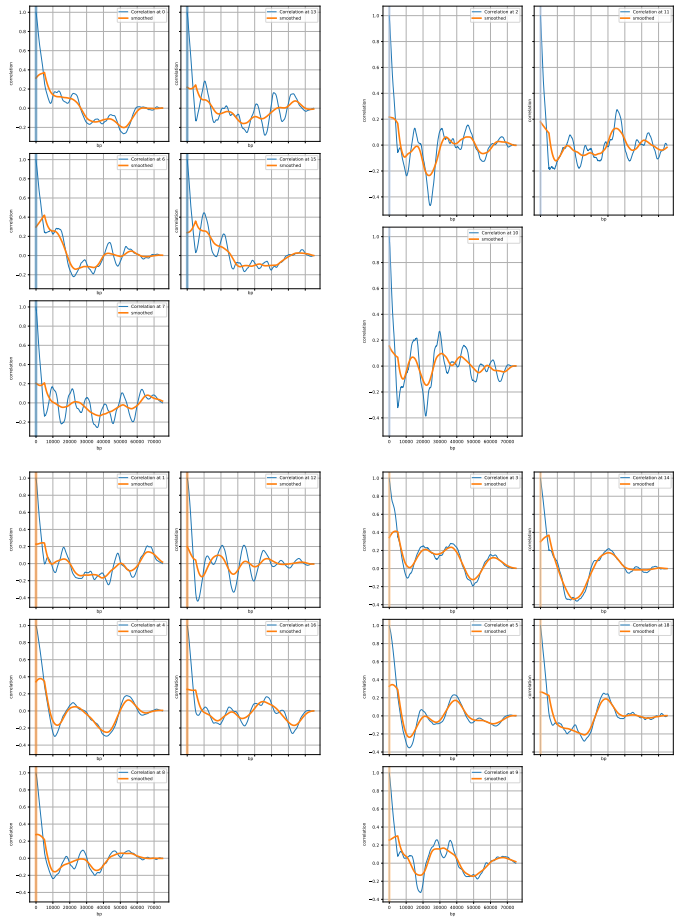


Figure S14: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 6. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

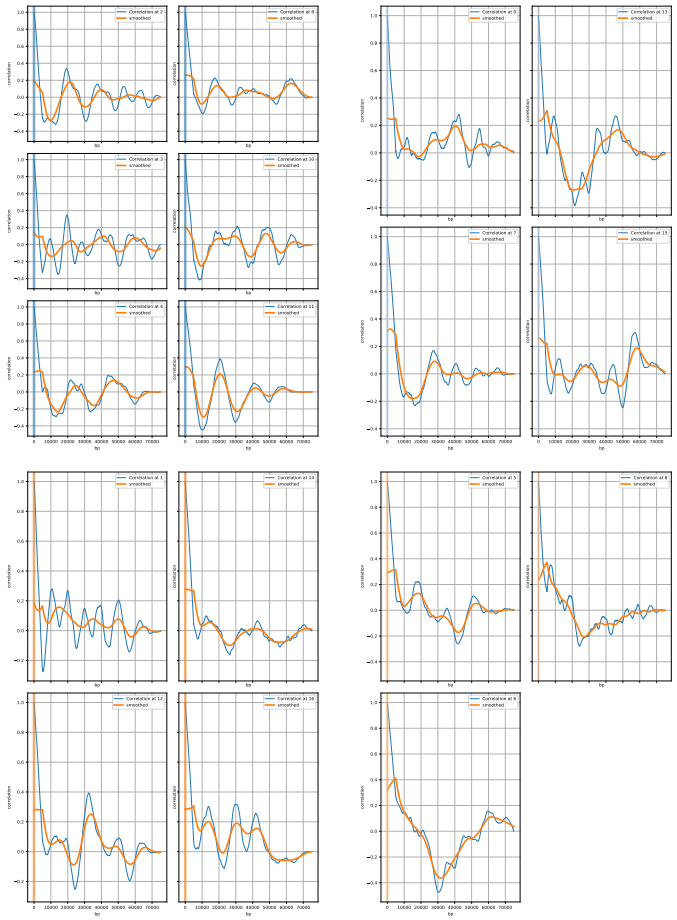


Figure S15: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 7. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

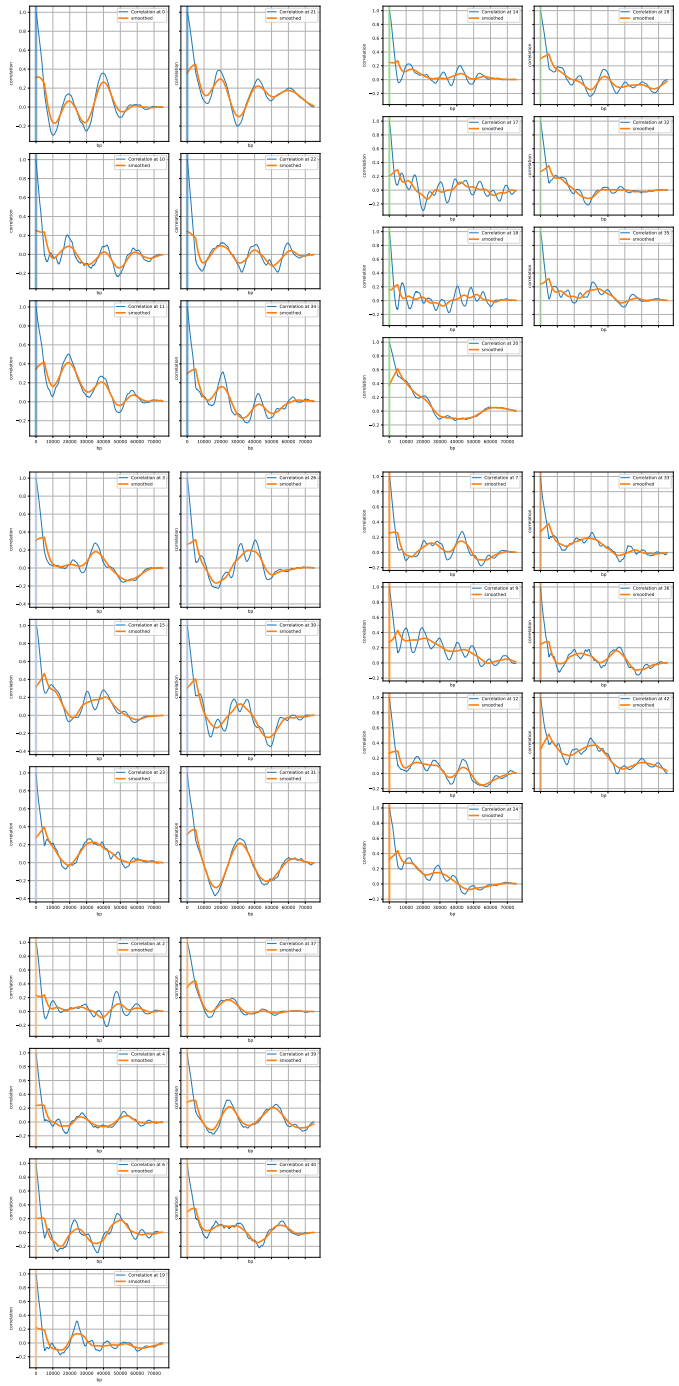


Figure S16: Shown are the correlation functions corresponding mapping of the pattern on the chromosome R. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

1.8 Genome-Wide Distance Matrix

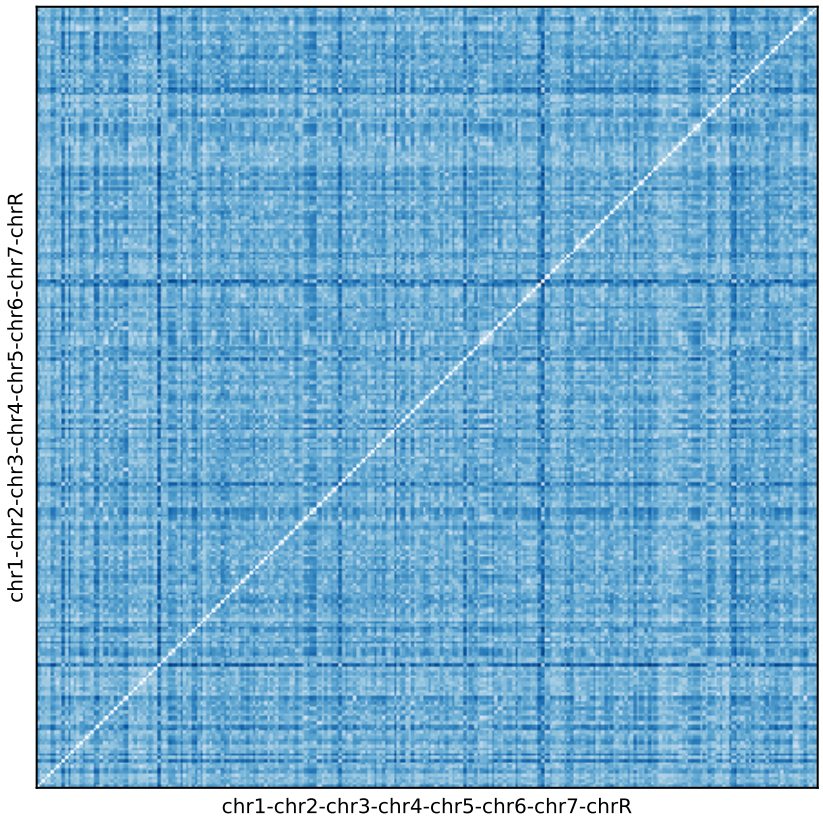


Figure S17: Shown is the genome-wide distance matrice. Distance refers to the distance between two correlation functions as measured by the euclidean distance (np.linalg.norm(x-y,ord=norm), with norm = 2. The ordering along the axes corresponds to the coarse-grained sections. The rolling average was of size 5000.

1.9 Genome-Wide Clustering

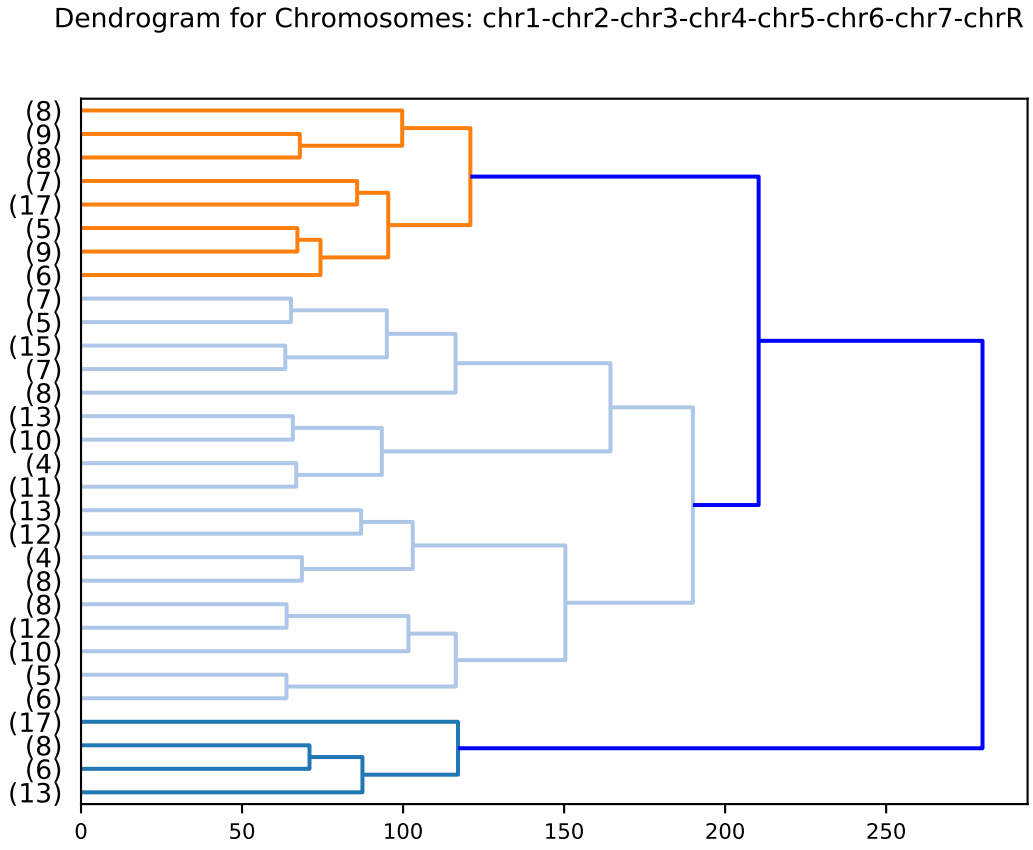


Figure S18: Shown is the dendrograms resulting from the genome-wide distance matrix. Results are for the hierarchical clustering on the individual chromosome. The Ward distance was used for the variance minimization algorithm used by SciPy **?**. The labels correspond to the distance matrix entries. Labels in parentheses give the number of labels corresponding to the leave. The rolling average was of size 5000.

1.10 Genome-Wide Distance Matrix and Clustering

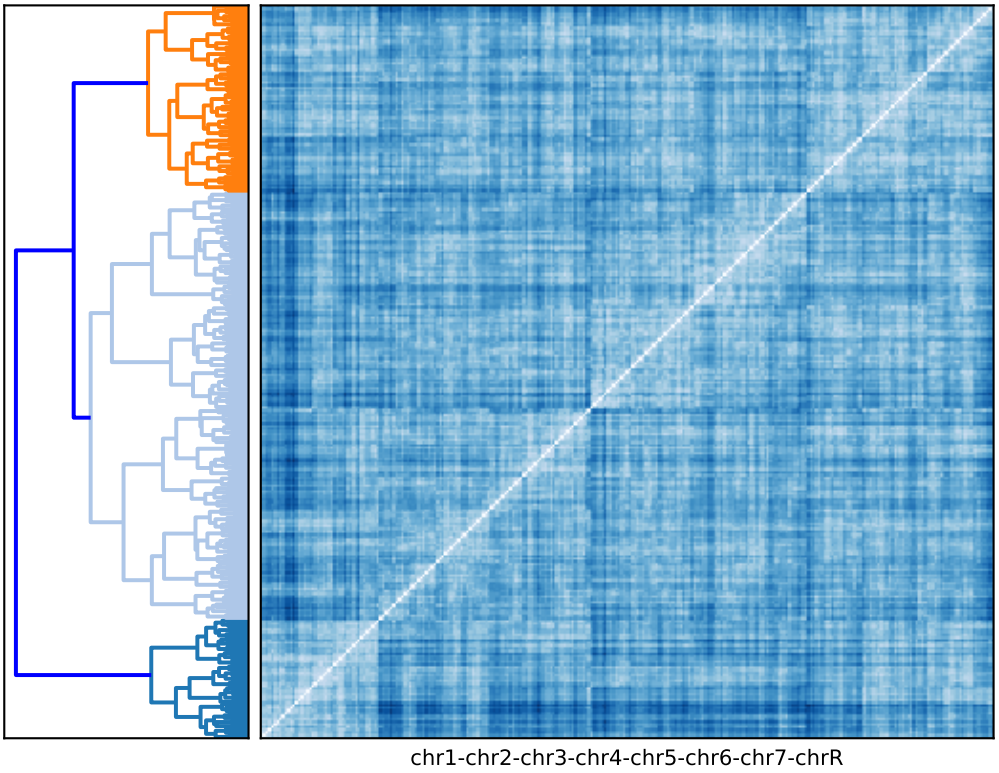


Figure S19: Shown is the genome-wide distance matrix and the corresponding dendrogram. The matrix entries are sorted to correspond to the identified clusters. The rolling average was of size 5000.

1.11 Correspondence between Pattern and Correlation Function Genome-Wide

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 1

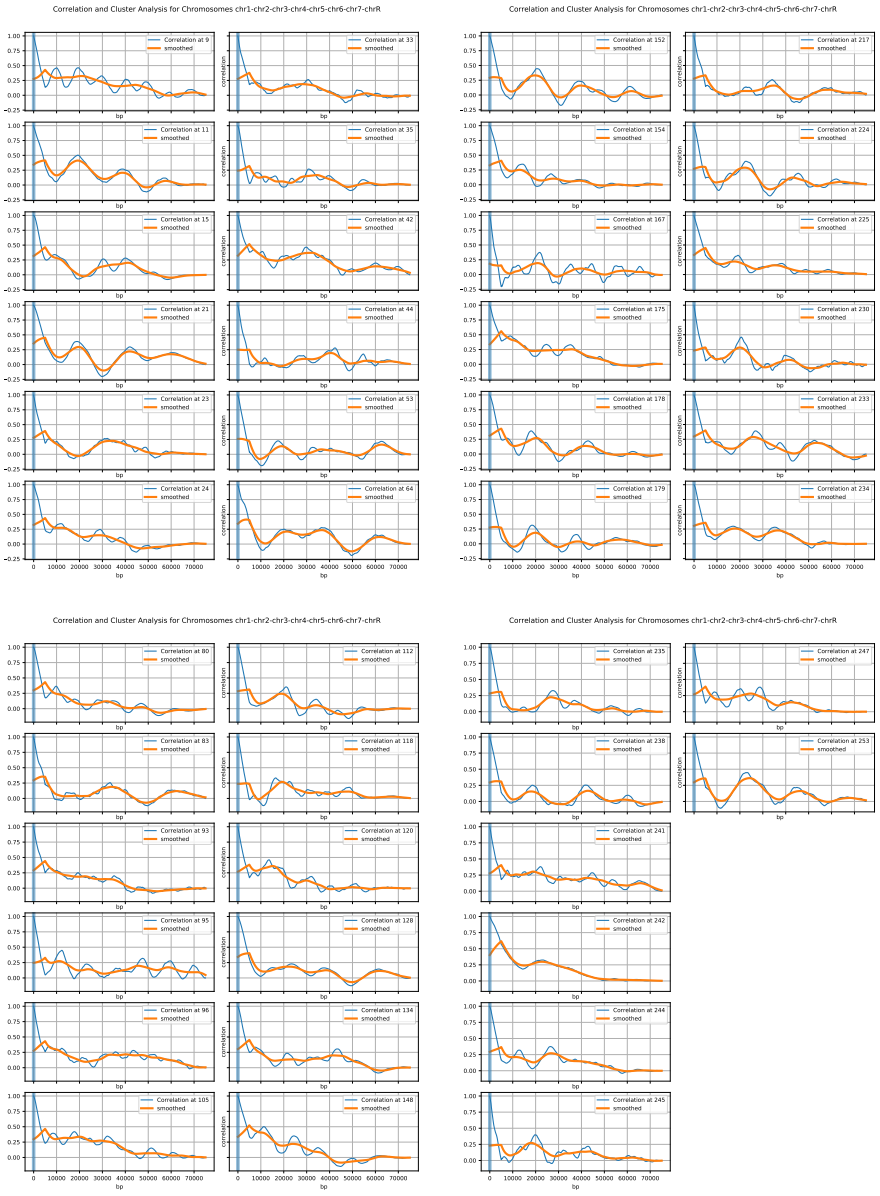


Figure S20: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (1)

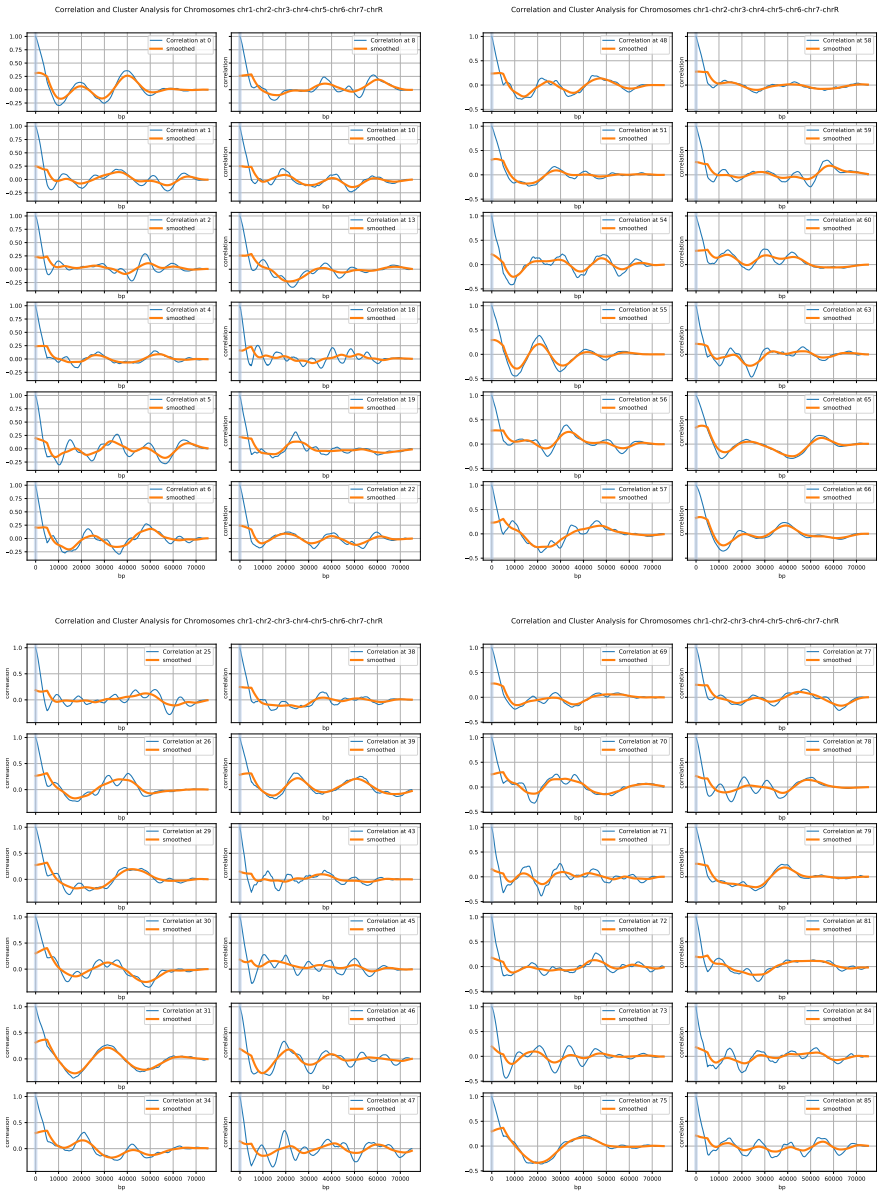


Figure S21: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (2)

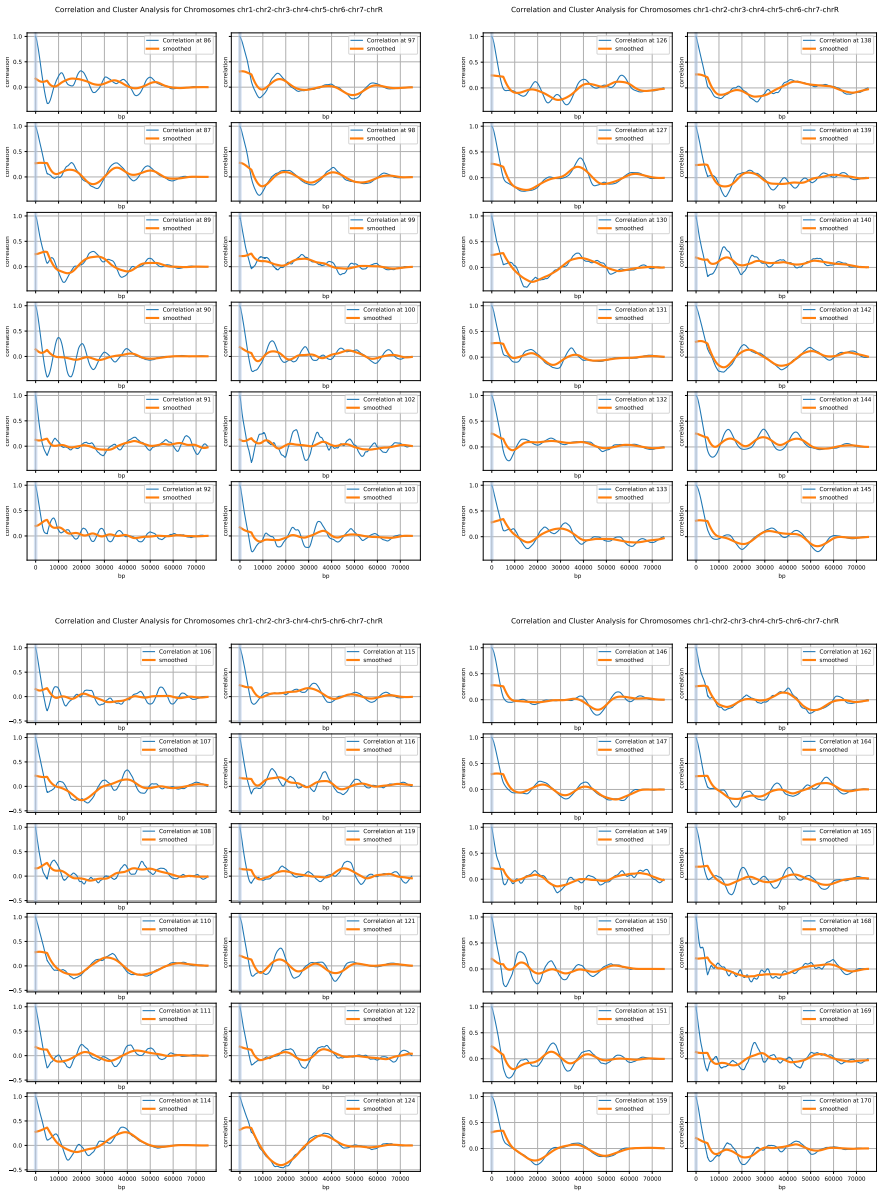


Figure S22: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (3)

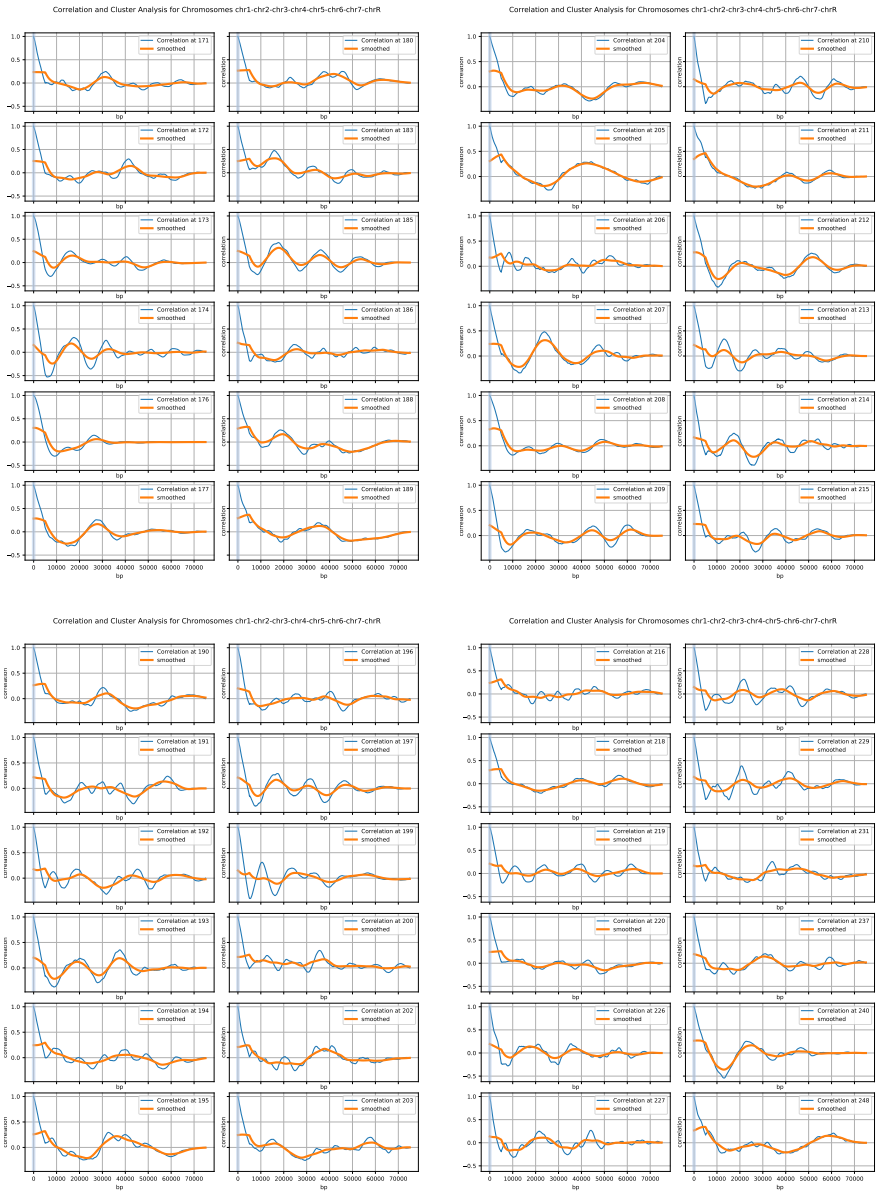


Figure S23: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (4)

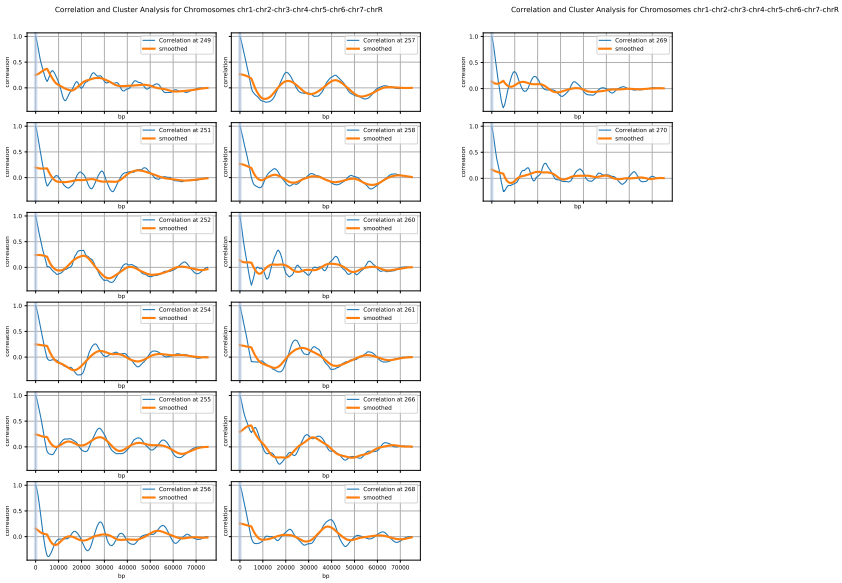


Figure S24: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 3 (1)

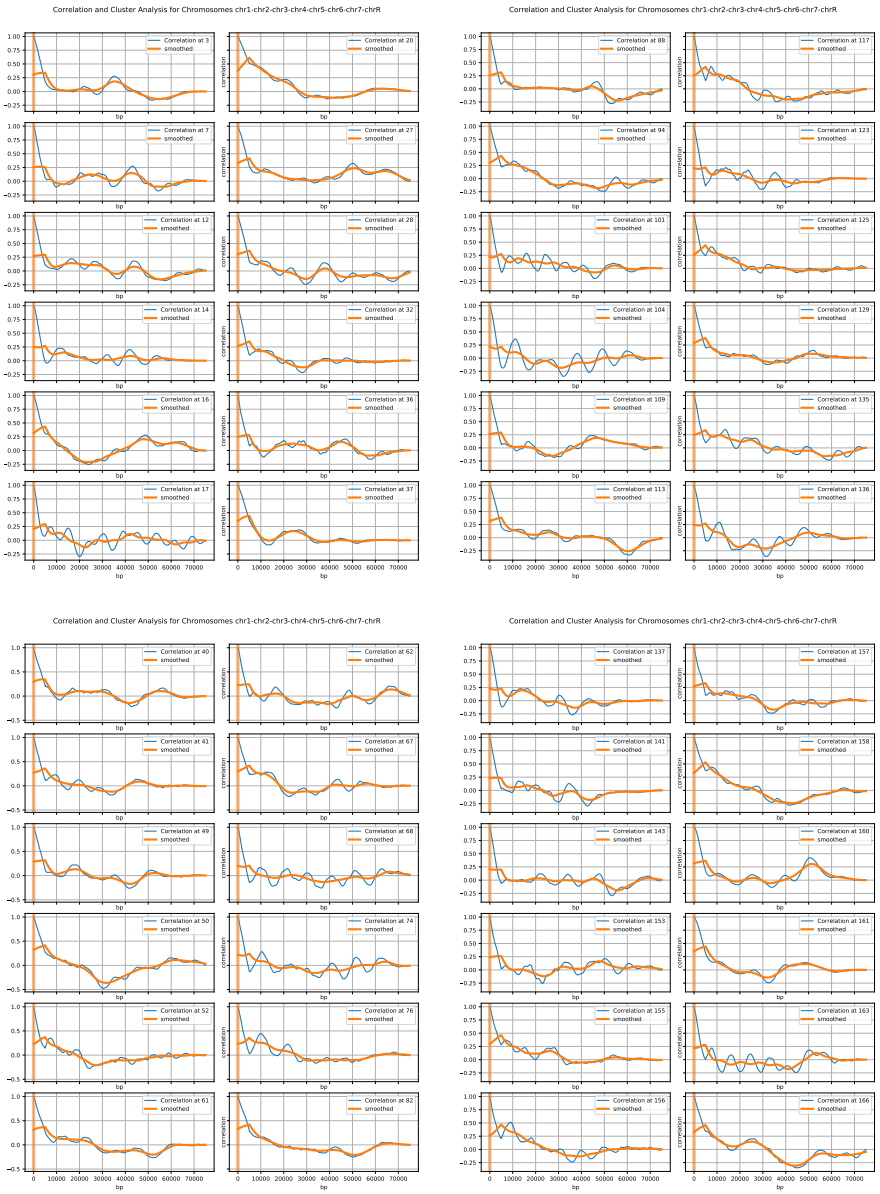


Figure S25: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 3 (2)

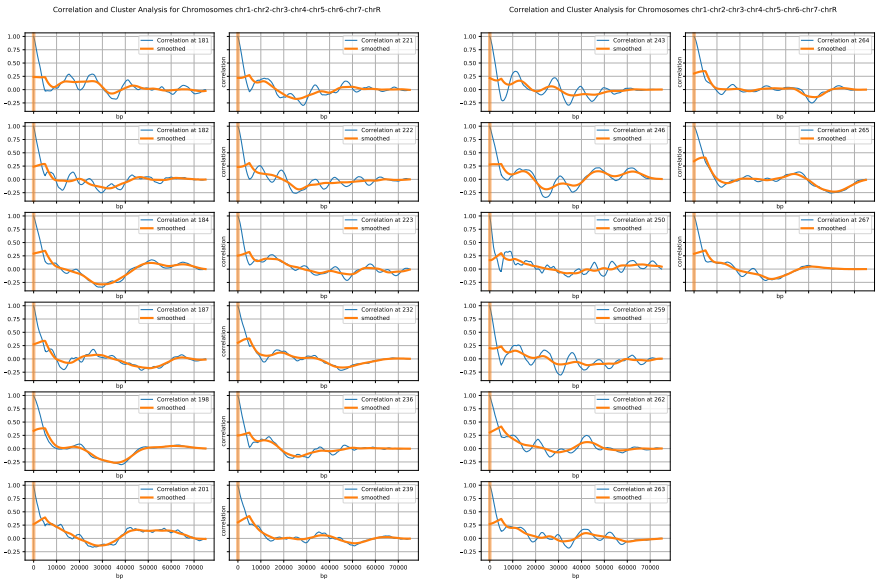


Figure S26: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

1.12 Comparison of Different Metrics

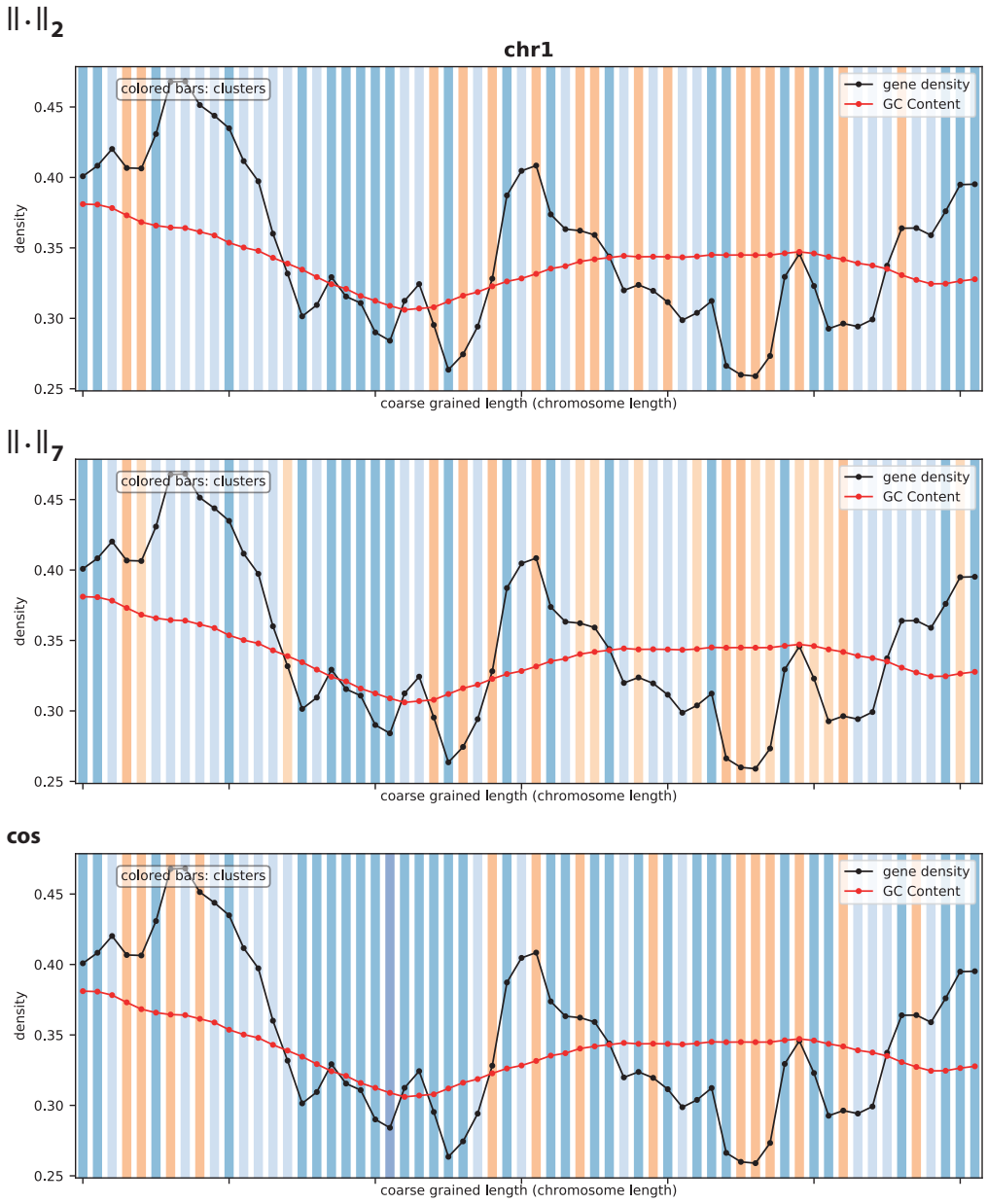


Figure S27: The upper panel shows the classification of the structures with respect to the euclidean distance $\|.\|_2$ while the middle one shows the result for $\|.\|_7$. Note that $\|.\|_7$ shows a further subdivision of the orange colored regions. Otherwise, the structure is stable against the two metrics for the distance between two correlation functions. The black line shows the gene density and the red line the GC content. The lower panel shows the application of the cosine similarity measure. While there are differences between the different metric, overall, a stable pattern is observed. What is remarkable is that the similarity measure shows less variation within certain domains than the other measures.

Chapter 5

Inter-Nucleosomal Potentials from Nucleosomal Positioning Data

References

This chapter is from:

- **Li, K.,** Oiwa, N. N., Mishra, S. K., & Heermann, D. W. (2021). Intra-Chromosomal Potentials from Nucleosomal Positioning Data.

Kunhe Li implemented the algorithm, generated the effective potentials, computed the compressibilities, and performed the classification. Kunhe Li, Nestor Norio Oiwa, Sujeet Kumar Mishra, and Dieter W. Heermann analyzed the results together. Kunhe Li and Dieter W. Heermann drafted the manuscript. Dieter W. Heermann supervised the work. All authors were involved in interpretation, proofreading, and addressing reviewer comments.

Eur. Phys. J. E (2022)45:33
https://doi.org/10.1140/epje/s10189-022-00185-3

THE EUROPEAN PHYSICAL JOURNAL E

Regular Article - Living Systems

Inter-nucleosomal potentials from nucleosomal positioning data

Kunhe Li[1], Nestor Norio Oiwa[2], Sujeet Kumar Mishra[3], and Dieter W. Heermann[1,a]

[1] Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, D-69120 Heidelberg, Germany
[2] Department of Basic Science, Universidade Federal Fluminense, Rua Doutor Sílvio Henrique Braune 22, Centro, Nova Friburgo 28625-650, Brazil
[3] Center for Computational Biology and Bioinformatics, School of Computational and Integrative Sciences (SCIS) Jawaharlal Nehru University, New Delhi, India

Received 7 January 2022 / Accepted 17 March 2022
© The Author(s) 2022

Abstract No systematic method exists to derive inter-nucleosomal potentials between nucleosomes along a chromosome consistently across a given genome. Such potentials can yield information on nucleosomal ordering, thermal as well as mechanical properties of chromosomes. Thus, indirectly, they shed light on a possible mechanical genomic code along a chromosome. To develop a method yielding effective inter-nucleosomal potentials between nucleosomes, a generalized Lennard-Jones potential for the parameterization is developed based on nucleosomal positioning data. This approach eliminates some of the problems that the underlying nucleosomal positioning data have, rendering the extraction difficult on the individual nucleosomal level. Furthermore, patterns on which to base a classification along a chromosome appear on larger domains, such as hetero- and euchromatin. An intuitive selection strategy for the noisy optimization problem is employed to derive effective exponents for the generalized potential. The method is tested on the *Candida albicans* genome. Applying k-means clustering based on potential parameters and thermodynamic compressibilities, a genome-wide clustering of nucleosome sequences is obtained for *C. albicans*. This clustering shows that a chromosome beyond the classical dichotomic categories of hetero- and euchromatin is more feature-rich.

1 Introduction

The organization of a complex system such as the nucleosome organization and with it the three-dimensional organization of a chromosome is influenced by hundreds of factors from DNA sequence, nucleosome remodelers to transcription factors [1]. Each of these factors influences not only the chemical environment but also the mechanical properties of the chromatin fiber such as the bending rigidity. Since the chromatin fiber is a heteropolymer, the bending rigidity is not a constant along the backbone [2]. Changing the bending rigidity by a more compact packing of the nucleosomes, for example, by a microphase separation [3,4] changing the order parameter and packing, has an influence on the loop structure of a chromosome and hence on regulation [5].

It has long been speculated that there must be something like a mechanical code (a comprehensive map determining shapes of DNA and mechanical properties) on top of the genetic code [6,7]. This mechanical code stems from the organizational structure of the nucleosomes since elasticity is a direct result of interatomic interaction. A tighter packing gives rise to more steric repulsion and hence higher bending rigidity. This in turn leads a reduced possibility for distal interactions, i.e., looping, hence controlling the three-dimensional organizational structure. And, there is more and more evidence surfacing that there is a richer variety of compactification of nucleosomes beyond the hetero- and euchromatin picture [8–10]. Experimental as well as theoretical work has indicated that indeed there is more than just two [11,12].

In this work, we take the point of view that we can extract larger nucleosomal structure from nucleosomal positioning data by coarse graining.

To reveal the thermodynamic properties and hence give indication on the mechanical code, we move to a larger global scale and ask for nucleosomal distribution patterns along a single chromosome as well as universal pattern between all chromosomes of a given genome. For this, we need to eliminate some of the smaller structures to reveal structure on a coarser level which is also more in line with the local phase separation picture [13].

There are at least two main directions that can be chosen. Physically, it is possible to start with geometric properties, e.g., the bending rigidity or stiffness, which is already verified to have a significant correlation with the compaction [14,15]. Chemically, it is desirable to

[a] e-mail: heermann@tphys.uni-heidelberg.de (corresponding author)

Published online: 11 April 2022

extract the effective pair-wise potential between single nucleosomes, and essential properties can be calculated subsequently. This allows to compute thermodynamic properties such as the compressibility for all of stretches showing a particular pattern of nucleosome distribution. Eventually, this leads to information on the mechanical properties since it allows to bring in line information on varying compressibilities and along the chromosomes with effective potentials. Furthermore, it also allows to extract the χ-parameter for the Flory–Huggins theory and shed light on the possible thermodynamic state, in particular the microphase separation [16].

2 Methods

2.1 Computational methods

One of the basic techniques to measure the nucleosome activity is the micrococcal nuclease digestion with deep sequencing (MNase-seq) [17]. The method measures the nucleosome occupancy by measuring the frequency of nucleosome-bounded DNA fragments. However, it does not directly identify the nucleosome position, the probabilistic genomic position where each nucleosome is located. In order to map the MNase-seq data to nucleosome positioning data, several programs were developed, such as NPS [18], nucleR [19], DANPOS [20], and iNPS [21] (improved nucleosome positioning from sequencing).

Our starting point is iNPS data for *Candida albicans*. The raw data (MNase-seq) are available from the Gene Expression Omnibus (GSM1542419) [22] and were measured by Puri et al. [23]. We also accessed the processed iNPS data in the NucMap database by Zhao et al. [24].

A section of the raw data is shown in Fig. 1 in panel A indicated by the red line. The areas with value 1 are the nucleosome positions, and the areas with value 0 are voids. This data are noisy due to missing data. Furthermore, on this small scale it is difficult to discern structure.

The goal is to extract potentials from the nucleosomal positioning data. One approach to obtain those is to compute the radial distribution function (RDF) $G(r)$ with respect to the distance r (measured in base pairs)

$$G(r) = \frac{1}{\rho N S_d} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \delta(r - r_{ij}) \tag{1}$$

where ρ is the density, N is the number of nucleosomes, S_d is a dimensional related term, r_{ij} is the distance between two nucleosomes i and j, and $\delta(r-r_{ij})$ is equal to 1 if $r = r_{ij}$ and 0 otherwise.

A chromosome is split into sections of 50,000 bp with 12,500 bp extra intersection at each end with its neighbor. For each section, we calculate the corresponding RDF. The sectioning of the chromosome is such that a substantial overlap between neighboring sections is guaranteed. Thus, the actual boundary position is somewhat fuzzy so that the actual starting position becomes less relevant.

To derive pair potentials from the nucleosomal distribution patterns [25], there are several paths such as the Berg–Harris method [26], Yvon–Born–Green equation [27], and reverse Monte Carlo [28]. We employ an reverse process on the nucleosomal radial distribution function. Its solution is guaranteed to converge by combining the noisy optimization [29,30] with the coarse-graining technique of molecular models, i.e., the reverse Monte Carlo [31,32], and, for example, implemented for the aqueous NaCl solution [28]. We implemented the basic idea with several improvements: most importantly, a generalized Lennard-Jones model for the potential and an intuitive selection strategy (ISS) for the noisy optimization problem are used.

The reverse Monte Carlo (RMC) method is a double loop nested Monte Carlo (MC) simulation. In the inner loop, a standard molecular Monte Carlo simulation is implemented to obtain the desired parameter for a given potential, while for the outer loop a Monte Carlo Markov Chain (MCMC) [33] is employed. A MCMC step proposes a new potential, runs the inner step, compares the computed parameter with the target result, and updates the potential until the tolerance level is reached. The RMC method succeeded in many cases, for example, in NaCl solutions [28]. However, it has the flaw that it has no guarantee to convergence, especially for a complex system. This issue also emerged applying RMC for the nucleosome system. In this circumstance, we have developed two improvements.

The original RMC uses a general potential. This, however, leads to convergence problems. From the computed radial distribution function $G(r)$ (Figure S2) and the related mean-field potential

$$P_{\mathrm{MF}}(r) \propto -\log(G(r)) \,, \tag{2}$$

we can actually observe that the target potential has a type similar to a Lennard Jones potential. Hence, without losing most of the generality, our ansatz is a generalized Lennard-Jones potential

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{\delta} - \left(\frac{\sigma}{r}\right)^{\nu} \right] . \tag{3}$$

Consistent with the Lennard Jones potential, ϵ determines the amplitude, and σ determines the length scale. The parameters δ and ν are the exponents that determine the shape and allow it to preserve most of the generality.

Another modification is substituting the MCMC step in RMC. The MCMC step is intended to solve the optimization problem, i.e., finding the RDF minimizing the differences. However, calculating an RDF from a potential via simulation produces non-negligible noise, especially for a more complex system. Therefore, the MCMC or other methods, e.g., Hill Climbing, Gradient Descent, and Simulated Annealing, have low effi-

ciency or are not converging. Consequently, we use for this non-trivial step a noisy optimization technique (dynamic optimization [30], or optimization with erroneous oracles [34]). The straightforward application is via an evolution strategy [29]. We have modified this to an intuitive selection strategy (ISS). This approach is more stable and well suited for parallel computing. Due to this parallelization, the computational cost is strongly reduced.

The ISS is very straightforward: 1. Execute the MC simulation for each possible potential in low precision, i.e., smaller number of MC steps. 2. Choose the best N candidates according to a selection ratio θ. 3. Increase the number of MC steps to a larger value and repeat the process. Repeating this many times, finally, there will be only one candidate, which is the result.

Note that our model is continuous along the section axis. Hence, basepair preferences of the nucleosomes are not taken into account. To include this, a modified continuous model with preferred attraction sites would be needed or a discrete model on the level of basepairs, since nucleosomes can slide as well as the uncertainty of the data has guided us in our model choice.

2.2 Compressibility

We compute the reduced isothermal compressibilities χ_T^∞ by the block density distribution method [35,36]. In this method, the whole section with size L_0 is separated into M_b blocks. The size of each block is $L = L_0/M_b$. Let N be the number of the nucleosomes in a block. If the distribution of N is $P_{L,L_0}(N)$, its kth moments $\langle N^k\rangle_{L,L_0}$ is given by

$$\langle N^k\rangle_{L,L_0} = \sum_N N^k P_{L,L_0}(N)\ . \tag{4}$$

The summation is over all possible value of N. Then, the reduced isothermal compressibility of a block is

$$\chi_T(L,L_0) = \frac{\langle N^2\rangle_{L,L_0} - \langle N\rangle^2_{L,L_0}}{\langle N\rangle_{L,L_0}}\ . \tag{5}$$

The difference between the finite size $\chi_T(L,L_0)$ and the thermodynamic limit χ_T^∞ is related to boundary effects associated with the finite-size of the subdomains. It takes the form:

$$\chi_T(L,L_0\to\infty) = \chi_T^\infty + \frac{c}{L} + O\left(\frac{1}{L^2}\right)\ . \tag{6}$$

Here c is a constant. Under this circumstance, the reduced isothermal compressibility of block $\chi_T(L,L_0)$ can be extrapolated to compute the reduced isothermal compressibility χ_T^∞ by just taking the limits $L, L_0 \to \infty$. Hence, in the $\chi_T(L,L_0)$ vs. M_b plot, the value at $M_b = 0$ is the result χ_T^∞.

The block density distribution method can compute the compressibility efficiently, but the calculation needs a large amount of conformations. In this paper, after the effective potential is obtained, we generate conformations through a MC simulation of 1,000,000 MCSs for each section.

2.3 Parameters

For the each of the eight chromosomes of the genome, we partitioned the chromosome in sections of 50,000 bp length each. There is a 12,500 bp extra intersection at each end with its neighbor to reduce the boundary effect. Thus, the total length of each section is 75,000 bp including the overlap. For the particle-based Monte Carlo simulation, section i starts from $12,500{+}50,000{\cdot}i$ bp to $12,500 + 50,000(i+1)$ bp, while actually the data are taken from $50,000\cdot i$ bp to $50,000\cdot i + 75,000$ bp. This binning is applied to the whole genome. For example, the length of chr. 2 is 2.231.883 bp [37], and it is separated into 44 sections.

In the one-dimensional Monte Carlo simulation, each monomer represents a nucleosome and occupies a volume equal to the averaged nucleosome length for that section. For every MC step, a random move for each monomer is proposed. It ranges from 0 to λ. The move is rejected or accepted according to the energy difference multiplied by the Boltzmann factor k_BT. In our simulation, k_BT is set to be 1.

The value of λ is chosen to be the smallest value that allows the acceptance rate to be equal to or smaller than 50% on average.

For the differences between the target RDF and the simulated, we used the mean squared residual (MSR)

$$MSR = \frac{1}{(n-p)}\sum(x-\hat{x})^2\ , \tag{7}$$

where p is the number of parameters in the regression (including the intercept). x is the target value, and $\hat{x}$ is an estimator.

For the modified Lennard-Jones potential the domain of σ is $[140, 170]$. It has the unit of one base pair. Inside the ISS, the selecting ratio is 0.25.

2.4 Classification

The resulting potentials from the Monte Carlo with its parameters can be used for clustering approaches such as k-means. Panel C in Fig. 1 shows the obtained values for the exponents as well as on the z-axis the compressibility data. The parameters ν and δ that characterize the short range repulsion and the long-range attraction together with the information on the compressibility are used for a k-means clustering.

3 Results

Effective potentials and classification

The results on the effective potential for *C. albicans* are shown in Fig. 2a. The colors indicate the class according

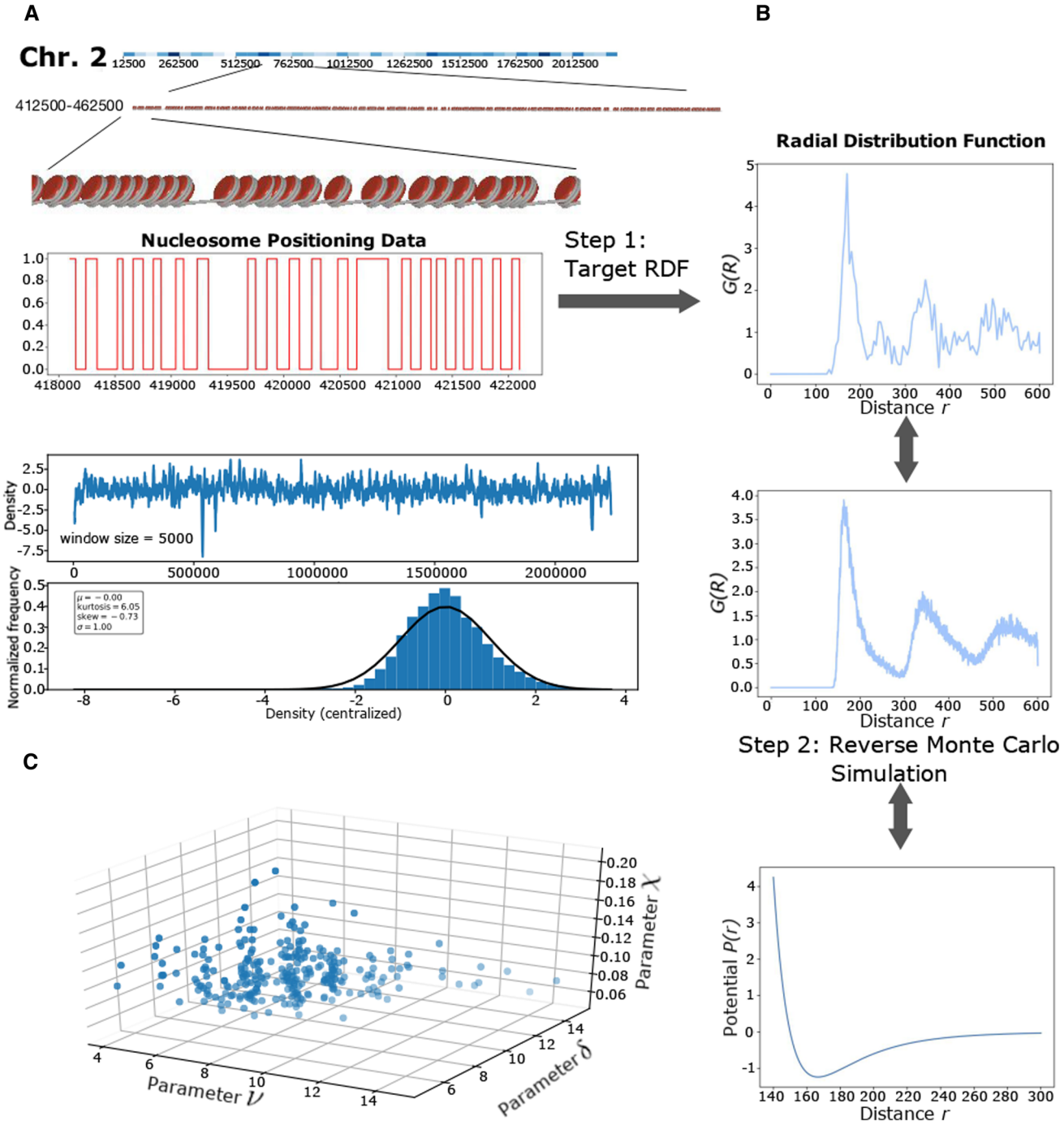


Fig. 1 *Steps to derive inter-nucleosomal potentials from nucleosomal positioning data.* Panel **A** shows schematically the distribution of nucleosomes in a section of chromosome 2 of *C. albicans*. We split the chromosome into sections, typically of size 50,000 bp. The lower part of Panel **A** shows the density after applying a rolling mean averaging with window size 5000 bp, and the typical section size is chosen to be 10 times of this scale. Step 1 takes the red binary data. Based on this data, the radial distribution function (RDF) is computed. This step enables us to obtain a coarse-grained representation of the chromosome that allows for an effective and efficient simulation of a chromosome. There is also a 12,500 bp extra intersection at each end with its neighbor. This resolves the boundaries between the sections. Once the radial distribution is computed, we apply a cut-off to the potential. Using a reverse Monte Carlo simulation, we estimate a potential from the RDF. We employ an intuitive selection strategy, i.e., a noisy optimization technique to find the best fit for the generalized Lennard-Jones exponents (see Panel **C**)

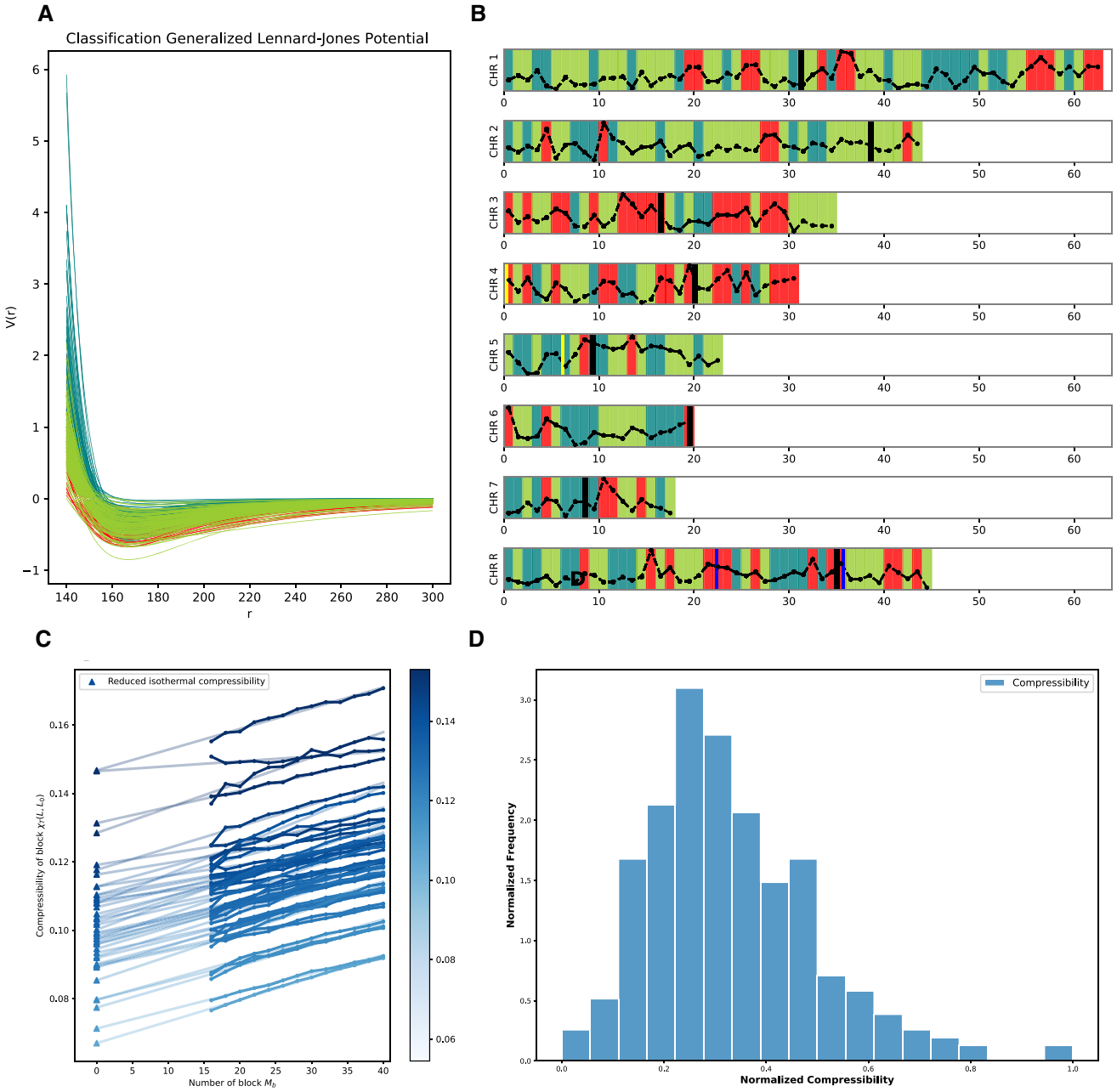


Fig. 2 *Effective pair-potential, genome-wide classification, and compressibility.* **Panel A**: Shown is the result for *C. albicans*. Each chromosome is partitioned into several sections, each containing 50,000 base pairs with two additional 12,500 bp intersections on both sides. The curves are the effective potentials, which quantify the global interaction pattern between nucleosomes. Their coloring is adjusted to be consistent with panel B. **Panel B** shows the classification of the sections based on the pair potentials and compressibilities for the whole genome. This classification is based on a k-means clustering into 3 clusters. They are intentionally classified to be comparable with the classification of heterochromatin, euchromatin, and differently organized. The dashed lines are the compressibility results. The two yellow and the two blue lines mark the position of known characterization. **Panel C**: This panel shows the reduced isothermal compressibility χ_T^∞ employing the block density method. The plot displays the process for chr. 2. The x-axis is the number of blocks M_b. The linked dots are the compressibilities of block $\chi_T(L, L_0)$. By extrapolating their linear regressions, we obtain the intercepts as the compressibility, marked by triangles. **Panel D:** For a better representation of the complex structure, we calculated the distribution of the compressibility $P(\chi_T^\infty)$

to a k-means clustering based on three clusters taking into account the exponents and the compressibility (see Figs. 1 and 2 Panels C and D.) From Fig. 2a, it can be seen that they all share a minimum lying between 160 bp and 180 bp. However, the well depths are falling into different classes. A shallow minimum with a steep repulsive part indicates an area where nucleosomes are loosely bound, corresponding to an irregular array, i.e., with liquid-like structure. A deep minimum with a less steep repulsion leads to a regular array in contrast, i.e., a much more ordered structure. Thus, the section partitions into those that are liquid- and those that are more solid-like in agreement with the classical classification eu- and hetero-chromatin picture, disregarding the nuances of a finer partitioning. However, the classification did no trivially sort the potentials according to the potential minima. Rather, an interplay between attraction, repulsion, and compressibility can be seen. The sorting into classes is more toward how the potential behaves at short distances and a larger distances, whereas in the well part of the potential a substantial criss-crossing can be seen the far ends are much more sorted.

The classification is based on all of the sections of the entire genome. This effectively constraints the pattern to be of a universal genome-wide character. Local variations are subsumed into broader classes filtering out the universal patterns underlying the local variations within a chromosome as well as among the chromosomes.

The resulting coloring of three clusters is shown in Fig. 2b. The coloring of Fig. 2a is adjusted to be consistent with that in panel B. The classification results suggest that there is more than hetero- and euchromatin. At least a further class can be distinguished genome-wide. In the supplementary information, Figure S3 shows a principal component analysis for various given k-means clusterings. Since we cannot employ directly a method such as the elbow method to look for the best classification, the visual inspection partitioning of the clusters in principal component space is used. A classification into three clusters shows the best result. Two clusters show a trivial partition while for a larger number of clusters a significant overlap is seen. Indeed, already in the first experiments it was noticed that within hetero- and euchromatin variations exist [38].

The result of the classification into three classes mapped to their original genomic location is shown in panel B of Fig. 2. Also shown in the figure are the results for the compressibility. The compressibilities themselves are shown in panel C and D. In Fig. 2c, we show the results from the block density method for all sections in chr. 2. Each line presents one section. The linked dots are the reduced isothermal compressibility of block $\chi_T(L, L_0)$ with respect to the number of blocks M_b. The straight lines are the corresponding linear regression results for the extrapolation to the thermodynamic limit. The triangles mark the intercepts, i.e., the reduced isothermal compressibilities χ_T^∞. All lines are colored according to their χ_T^∞ value. Note that no corrections for the scaling are necessary as the extrapolation proportional to $1/L$ is consistent with the data.

The distribution of the extrapolated compressibility values for the whole genome (for *C. albicans*) is shown in Fig. 2d. The distribution is clearly non-gaussian. The obtained extrapolated values are used for the classification and shown in panel B. A high value of compressibility is associated with a few location along the chromosomes. Marked by the thick black line is the location of the centromeres. Four further markers from gene expression results confirmed by three experimental groups [23,39,40] are also included. They have measured the expression for those genes in different conditions, especially in different iron concentrations, and they concluded that in our circumstance, the two blue marked regions were suppressed while the yellow marked regions were not suppressed. Both results are compatible with the classification. The sections that are classified as heterochromatin are indeed consistent with the deeper wells of the potentials while the euchromatic region is in general associated with more shallow wells of the potentials (Fig. 2).

4 Conclusion

Based on the nucleosomal positioning data, the extraction of effective potentials is possible for an entire genome. If this information is supplemented with thermodynamic information in terms of compressibility, i.e., density fluctuations, a genome-wide consistent classification in sections is possible. The classification into the classes shows that at least three different classes must exist. Hence, beyond hetero- and euchromatin a third kind of ordering is necessary. The grouping of the exponents of the generalized Lennard-Jones potential may suggest that there may be more than three classes. However, the principal component analysis of the parameters into two dimensions shows that at least for this projection three is the best decomposition into classes.

Positioning data and simulations of the fluctuations of the positioning data should incorporate such effects as nucleation of hetero-chromatic regions. Thus, in a consistent manner the classification into more or less ordered regions is possible. Beyond this classification, having the information on the coarse-grained potentials, this approach allows for the modeling of chromosomes as hetero-polymers with inter-nucleosomal interactions. If this is further augmented with inter-chromosomal information derived from chromosomal conformation capture methods, a consistent framework for the simulation of chromosomes with the effective potentials is possible. This then allows to look for the mechanics, i.e., the mechanical code. Having the information on the potentials enables the modeling of the nucleosomes as effective disks such that the steric interactions together with the density fluctuations yield information on the stiffness of the particular section and thus on its bending rigidity.

One aspect of the ordering and stiffness of segments that is not yet covered by the approach are methylation effects. However, this can in principal be incorporated if a consistent set of experimental data would be available for a particular genome. This would add a further dimension for the classification.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1140/epje/s10189-022-00185-3.

Acknowledgements This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster). The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK), and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. Kunhe Li would to acknowledge funding by the Chinese Scholarship Council (CSC). Sujeet Kumar Mishra would like to acknowledge funding by the India government Department of Biotechnology (DBT)-Interdisciplinary Research Center for Scientific Computing (IWR) PhD program.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. K. Struhl, E. Segal, Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. **20**(3), 267–273 (2013)
2. R. Kumar, A. Chaudhuri, R. Kapri, Sequencing of semiflexible polymers of varying bending rigidity using patterned pores. J. Chem. Phys. **148**(16), 164901 (2018)
3. M. Conte, L. Fiorillo, S. Bianco, A.M. Chiariello, A. Esposito, M. Nicodemi, Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. Nature Commun. **11**(1), 3289 (2020)
4. S.E. Farr, E.J. Woods, J.A. Joseph, A. Garaizar, R. Collepardo-Guevara, Nucleosome plasticity is a critical element of chromatin liquid-liquid phase separation and multivalent nucleosome interactions. Nature Commun. **12**(1), 2883 (2021)
5. Y. Ghavi-Helm, A. Jankowski, S. Meiers, R.R. Viales, J.O. Korbel, E.E.M. Furlong, Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature Genet. **51**(8), 1272–1282 (2019)
6. M. Zuiddam, R. Everaers, H. Schiessel, Physics behind the mechanical nucleosome positioning code. Phys. Rev. E **96**(5), 052412 (2017)
7. A. Basu, D. G. Bobrovnikov, B. Cieza, Z. Qureshi, T. Ha. Deciphering the mechanical code of genome and epigenome. *bioRxiv*, page 2020.08.22.262352, 01 2020
8. A. Routh, S. Sandin, D. Rhodes, Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. Proc. National Acad. Sci. United States of America **105**(26), 8872–8877 (2008)
9. M. Bohn, P. Diesinger, R. Kaufmann, Y. Weiland, P. Müller, M. Gunkel, A. Ketteler, P. Lemmer, M. Hausmann, D. Heermann, C. Cremer, Localization microscopy reveals expression-dependent parameters of chromatin nanostructure. Biophys. J. **99**, 1358–1367 (2010)
10. I.A. Tchasovnikarova, R.E. Kingston, Beyond the histone code: a physical map of chromatin states. Molecular Cell **69**(1), 5–7 (2018)
11. J. Liu, M. Ali, Q. Zhou, Establishment and evolution of heterochromatin. Ann. New York Acad. Sci. **1476**(1), 59–77 (2020)
12. L. Hilbert, Y. Sato, K. Kuznetsova, T. Bianucci, H. Kimura, F. Jülicher, A. Honigmann, V. Zaburdaev, N.L. Vastenhouw, Transcription organizes euchromatin via microphase separation. Nature Commun. **12**(1), 1360 (2021)
13. P.B. Singh, S.N. Belyakin, P.P. Laktionov, Biology and physics of heterochromatin-like domains/complexes. Cells **9**(8), 1881 (2020)
14. S. Eran, F.-M. Yvonne, C. Lingyi, T. AnnChristine, F. Yair, I.K. Moore, J.-P.Z. Wang, W. Jonathan, A genomic code for nucleosome positioning. Nature **442**(7104), 772–778 (2006). https://doi.org/10.1038/nature04979
15. M.G. Poirier, S. Eroglu, J.F. Marko, The bending rigidity of mitotic chromosomes. Molecular Biol. Cell **13**(6), 2170–2179 (2002)
16. B.A. Gibson, L.K. Doolittle, M.W.G. Schneider, L.E. Jensen, N. Gamarra, L. Henry, D.W. Gerlich, S. Redding, M.K. Rosen, Organization of chromatin by intrinsic and regulated phase separation. Cell **179**(2), 470-484.e21 (2019)
17. D.C. Klein, S.J. Hainer, Genomic methods in profiling dna accessibility and factor localization. Chromosome Res. **28**(1), 69–85 (2020)
18. R. Schöpflin, V.B. Teif, O. Müller, C. Weinberg, K. Rippe, G. Wedemann, Modeling nucleosome position distributions from experimental nucleosome positioning maps. Bioinformatics **29**(19), 2380–2386 (2013)
19. O. Flores, M. Orozco, nucler: a package for non-parametric nucleosome positioning. Bioinformatics **27**(15), 2149–2150 (2011)
20. K. Chen, Y. Xi, X. Pan, Z. Li, K. Kaestner, J. Tyler, S. Dent, X. He, W. Li, Danpos: dynamic analysis of nucle-

osome position and occupancy by sequencing. Genome Res. **23**(2), 341–351 (2013)
21. W. Chen, Y. Liu, S. Zhu, C.D. Green, G. Wei, J.-D.J. Han, Improved nucleosome-positioning algorithm inps for accurate nucleosome positioning from sequencing data. Nature Commun. **5**(1), 1–14 (2014)
22. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse55819
23. S. Puri, W.K.M. Lai, J.M. Rizzo, M.J. Buck, M. Edgerton, Iron-responsive chromatin remodelling and mapk signalling enhance adhesion in c andida albicans. Molecular Microbiol. **93**(2), 291–305 (2014)
24. Y. Zhao, J. Wang, F. Liang, Y. Liu, Q. Wang, H. Zhang, M. Jiang, Z. Zhang, W. Zhao, Y. Bao et al., Nucmap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Res. **47**(D1), D163–D169 (2019)
25. S. K. Mishra, K. Li, S. Brauburger, A. Bhattacherjee, N.N. Oiwa, D.W. Heermann. Superstructure detection in nucleosome distribution shows common pattern within a chromosome and within the genome. preprint, November 2021
26. M. Shimoji, Relation between pair potentials and radial distribution functions in liquid metals and alloys. Adv. Phys. **16**(64), 705–716 (1967)
27. H.M. Cho, J.-W. Chu, Inversion of radial distribution functions to pair forces by solving the yvon-born-green equation iteratively. J. Chem. Phys. **131**(13), 134107 (2009)
28. A.P. Lyubartsev, A. Laaksonen, Calculation of effective interaction potentials from radial distribution functions: a reverse monte carlo approach. Phys. Rev. E **52**(4), 3730 (1995)
29. D.V. Arnold, *Noisy optimization with evolution strategies*, vol. 8 (Springer, Berlin, 2012)
30. J.M. McNamara, A classification of dynamic optimization problems in fluctuating environments. Evolut. Ecol. Res. **2**(4), 457–471 (2000)
31. A. Lyubartsev, A. Mirzoev, L.J. Chen, A. Laaksonen, Systematic coarse-graining of molecular models by the newton inversion method. Faraday Discuss. **144**, 43–56 (2010)
32. K. Binder, D.W. Heermann, *Monte carlo simulation in statistical physics*, first edition. (Springer-Verlag, Berlin, 1988)
33. K. Binder, D.W. Heermann, *Monte Carlo simulation in statistical physics* (Springer, Berlin, 2010)
34. Y. Singer, J. Vondrák, Information-theoretic lower bounds for convex optimization with erroneous oracles. Adv. Neural Inf. Process. Syst. **28**, 3204–3212 (2015)
35. M. Heidari, K. Kremer, R. Potestio, R. Cortes-Huerto, Fluctuations, finite-size effects and the thermodynamic limit in computer simulations: revisiting the spatial block analysis method. Entropy **20**(4), 222 (2018)
36. M. Rovere, D.W. Heermann, K. Binder, Block density distribution function analysis of two-dimensional Lennard-Jones fluids. EPL (Europhy. Lett.) **6**(7), 585 (1988)
37. S.H. Rangwala, A. Kuznetsov, V. Ananiev, A. Asztalos, E. Borodin, V. Evgeniev, V. Joukov, V. Lotov, R. Pannu, D. Rudnev et al., Accessing ncbi data using the ncbi sequence viewer and genome data viewer (gdv). Genome Res. **31**(1), 159–169 (2021)
38. R.C. Allshire, H.D. Madhani, Ten principles of heterochromatin formation and function. Nature Rev. Molecular Cell Biol. **19**(4), 229–244 (2018)
39. C. Chen, K. Pande, S.D. French, B.B. Tuch, S.M. Noble, An iron homeostasis regulatory circuit with reciprocal roles in candida albicans commensalism and pathogenesis. Cell Host Microbe **10**(2), 118–135 (2011)
40. C.-Y. Lan, G. Rodarte, L.A. Murillo, T. Jones, R.W. Davis, J. Dungan, G. Newport, N. Agabian, Regulatory networks affected by iron availability in candida albicans. Molecular Microbiol. **53**(5), 1451–1469 (2004)

Supplemental Information

The source code for the program is available at https://github.com/mdscolour/reverseMC.

The genome-wide effective potential data as well as the corresponding compressibility is available at the following DOI link: https://doi.org/10.11588/data/H3KPEU.

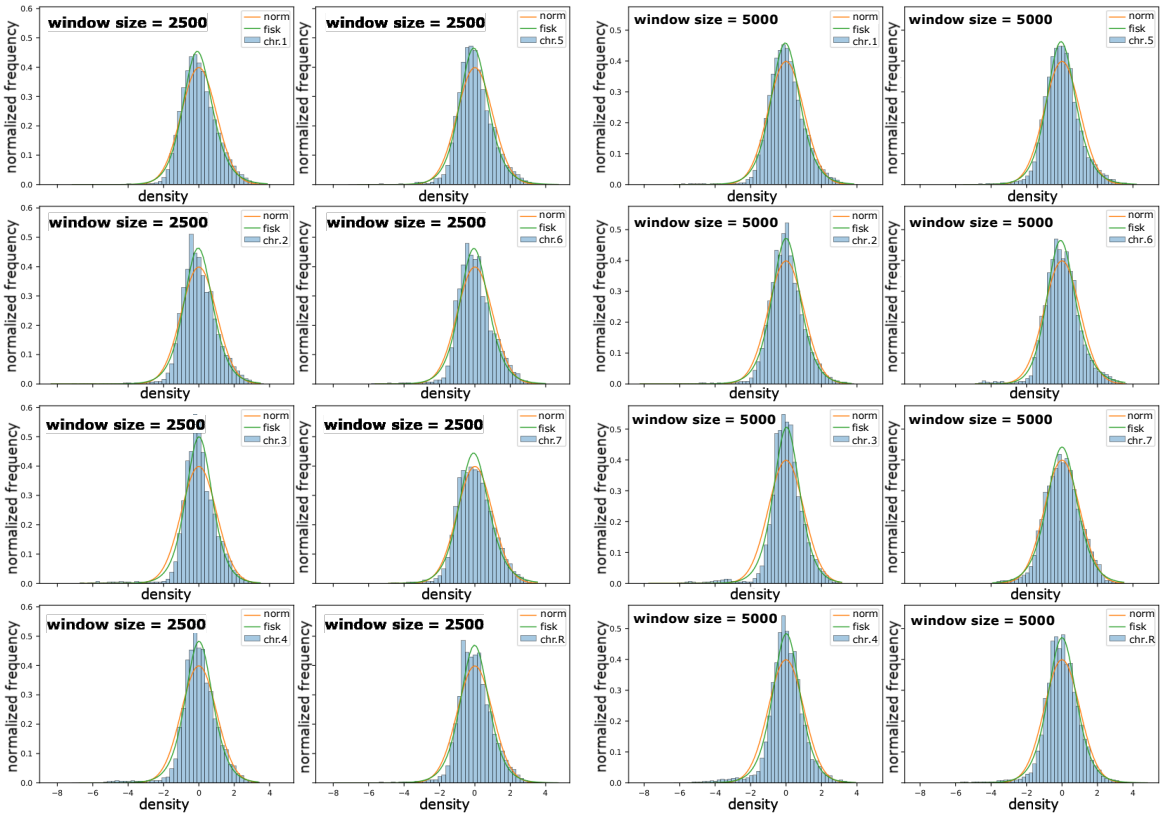


Figure S1: **Nucleosomal density after coarse-graining** The nucleosomal density distribution appears to be close to a Fisk distribution, i.e. is a log-logistic distribution. Shown are the results for a window size of 2500 and 5000. Several window sizes are examined and the 5000 bp length is the most suitable coarse-graining scale. Hence the typical section length is chosen to be 50000 bp.

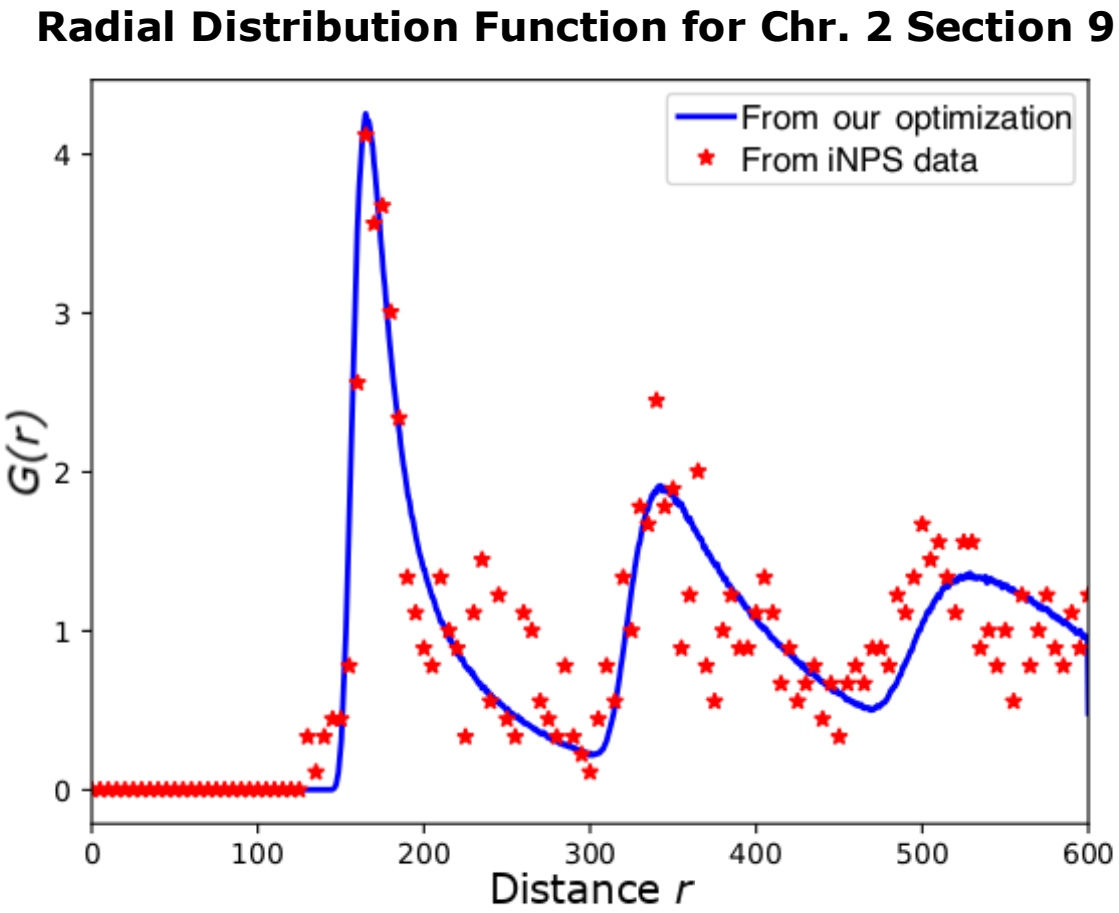


Figure S2: **iNPS data and resulting radial distribution function for chr. 2 section 9** Red stars show the radial distribution function (RDF) data calculated from experimental iNPS data. The blue curve is the estimated result for the effective potential at the same area by implementing An MC simulation. The RDF is computed from a total of 150000 MC steps.

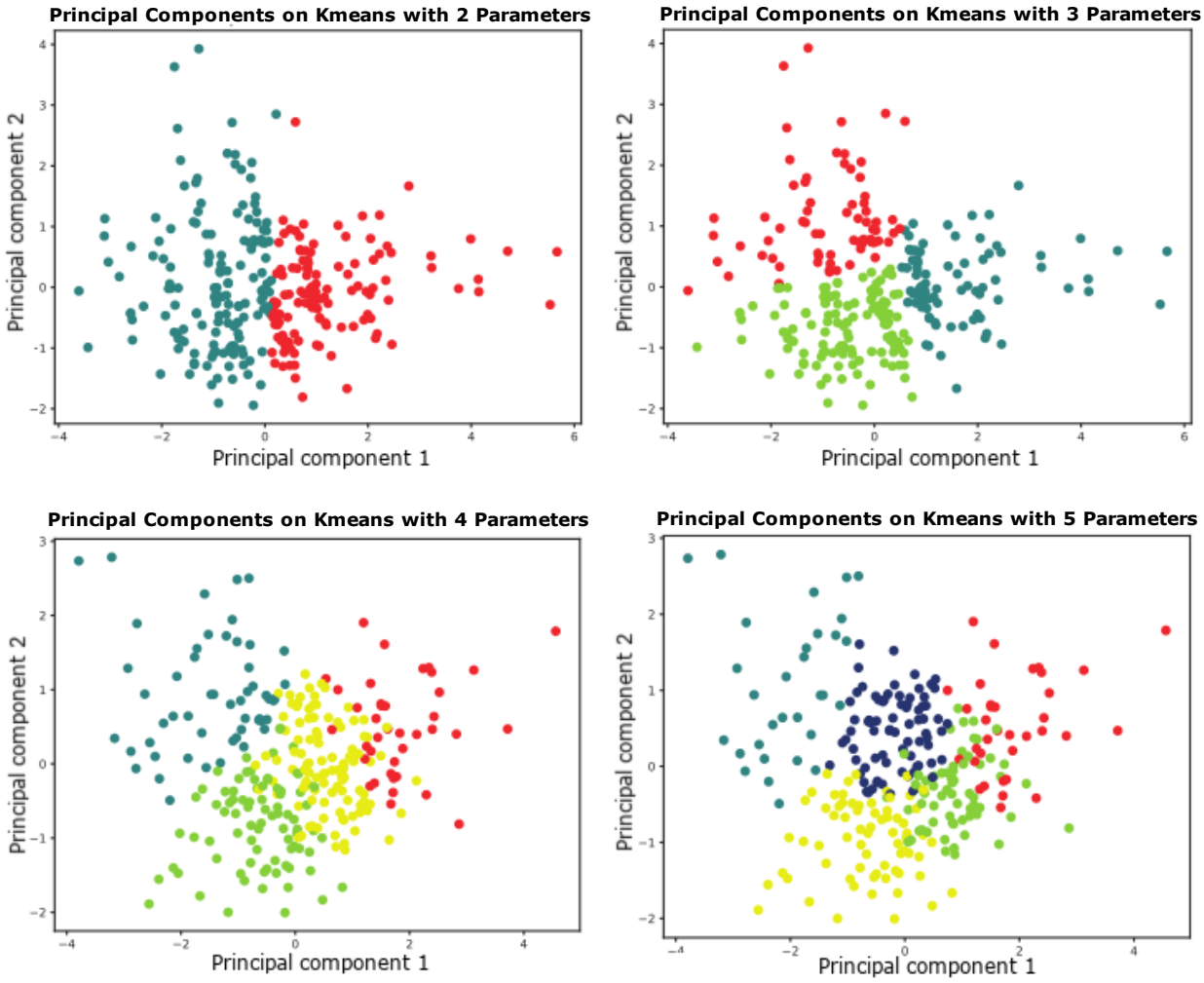


Figure S3: **PCA on K-means** Shown are the principal component analysis for a two-dimensional projection of the k-means clustering with various given cluster numbers. Rather than using the elbow or similar methods to find the optimum number of clusters, we have chosen to visually detect the best number of clusters. From the visual inspection we see that two clusters trivially separate into two clusters. The cluster separate non-trivially for three clusters whereas, above three clusters there is always a non-negligible overlap between the clusters. The parameters for the k-means clustering were ν, the potential minimum and the compressibility χ within the section.

Chapter 6

Prediction and Comparative Analysis of CTCF Binding Sites based on a First Principle Approach

References

This chapter is from:

- Oiwa, N. N., **Li, K.**, Cordeiro, C. E., & Heermann, D. W. (2021). Prediction and Comparative Analysis of CTCF Binding Sites based on a First Principle Approach.

Nestor Norio Oiwa validated the CTCF binding sites and calculated the exponents of the power-law decays. Kunhe Li detected the nucleosome positioning pattern around CTCF binding sites. Claudette E. Cordeiro supported the work. Dieter W. Heermann supervised the work. All authors were involved in drafting the manuscript, proofreading, and addressing reviewer comments.

Physical Biology

PAPER

Prediction and comparative analysis of CTCF binding sites based on a first principle approach

RECEIVED
28 October 2021

REVISED
28 February 2022

ACCEPTED FOR PUBLICATION
9 March 2022

PUBLISHED
6 April 2022

Nestor Norio Oiwa[1,2], **Kunhe Li**[2], **Claudette E Cordeiro**[3] **and Dieter W Heermann**[2,*]

[1] Department of Basic Science, Universidade Federal Fluminense, Rua Doutor Sílvio Henrique Braune 22, Centro, 28625-650 Nova Friburgo, Brazil
[2] Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, D-69120 Heidelberg, Germany
[3] Department of Physics, Universidade Federal Fluminense, Avenida Atlântica s/n, Gragoatá, 24210-346 Niterói, Brazil
[*] Author to whom any correspondence should be addressed.

E-mail: heermann@tphys.uni-heidelberg.de

Keywords: CTCF binding sites, Cys_2His_2 zinc finger, extended ladder model

Supplementary material for this article is available online

Abstract

We calculated the patterns for the CCCTC transcription factor (CTCF) binding sites across many genomes on a first principle approach. The validation of the first principle method was done on the human as well as on the mouse genome. The predicted human CTCF binding sites are consistent with the consensus sequence, ChIP-seq data for the K562 cell, nucleosome positions for IMR90 cell as well as the CTCF binding sites in the mouse HOXA gene. The analysis of *Homo sapiens*, *Mus musculus*, *Sus scrofa*, *Capra hircus* and *Drosophila melanogaster* whole genomes shows: binding sites are organized in cluster-like groups, where two consecutive sites obey a power-law with coefficient ranging from 0.3292 ± 0.0068 to 0.5409 ± 0.0064; the distance between these groups varies from 18.08 ± 0.52 kbp to 42.1 ± 2.0 kbp. The genome of *Aedes aegypti* does not show a power law, but 19.9% of binding sites are 144 ± 4 and 287 ± 5 bp distant of each other. We run negative tests, confirming the under-representation of CTCF binding sites in *Caenorhabditis elegans*, Plasmodium falciparum and *Arabidopsis thaliana* complete genomes.

1. Introduction

In mammals the primary insulator is the nucleotide sequence CCCTC-binding factor (CTCF), a protein with 10 Cys_2His_2 and one C_2HC zinc finger and the major eukaryotic DNA-protein binding motifs [1–4] (cf figure 1(b)). These transcription factors are characterized by 3 to 29 zinc finger (ZF) units [5, 6], each composed by one zinc ion linking two cysteines at the end of two β-sheets and two histeines in the C-terminal of one α helix [7, 8]. Chromatin immunoprecipitation assays with DNA microarray indicate at least 13 804 actives binding sites [2] and Xie *et al* [3] reports a minimum of 15 000 binding sites for CTCF, using chromatin immunoprecipitation assay with massively parallel DNA sequencing (ChIP-seq). Chen *et al* [4] estimates 326 840 possible sites along the human genome, combining the data from 38 cell lines. Despite CTCF relevance, the quality is poor in 20% to 30% of the available data due to limitations of the experimental apparatus and the algorithms for localizing binding site [2, 4, 9]. Same mistakes are made, adding false binding sites and making impossible in see the structure of the CTCF distribution. In this paper we present a new method to finding CTCF binding sites based on the interaction of the zinc finger and the electronic cloud of the nucleotide π-orbital of the double DNA (dDNA). This is a first principle approach method, because we compute the local electron density of states using electron–nucleotide interactions along the genome [10] (referee 1: item 5). This quantum mechanic charge transport description of the nucleotide, typical in semiconductor physics, adds a new layer of information beyond traditional four letter nucleotide

© 2022 IOP Publishing Ltd

genomics. In this way, we overcome the limitations of previous works, unveiling a power law along CTCF biding sites in many complete genomes.

The workflow and organization of the article is illustrated in figure 1(a). First of all, we collect 23 experimentally detected CTCF-DNA binding site (see supplementary material S1 (https://stacks.iop.org/PB/19/036005/mmedia)). Then, we study the electronic cloud of the nucleotide π-orbital using [10]. This analysis extends the usual nucleotide alignment based on hydrogen bonds, adding information about the electronic behavior in CTCF binding sites as ground state, highest occupied orbital (HOMO) and lowest unoccupied orbital (LUMO) (see S2). Once we establish a pattern based on our electronic nucleotide alignment, we apply it over a complete genome in multiple genomes (see S3). We validate our putative CTCF binding sites with the consensus sequence [2, 11–13], ubiquitous ChIP-seq K562 data [4, 14], MNase-seq of IMR90 cell with improved nucleosome positioning (iNPS) [15–17] and the cluster HOXA [18]. After corroboration of our putative CTCFbs, we study the distribution of CTCFbs over the complete human, mouse, pig, goat, fruit fly and *Aedes aegypti* (mosquito) genomes. We use the complete *Caenorhabditis elegans*, *Plasmodium falciparum* and *Arabidopsis thaliana* genomes as negative controls. We report cluster-like structures for the CTCF distribution in multiple species. Finally, we discuss the limitations of our method as well as ChIP-seq data.

2. Method

2.1. CTCF samples

In order to establish an electronic nucleotide pattern, we consider 23 experimentally confirmed CTCF binding sites, figure 1(b). Detailed descriptions about these CTCFbs are in supplementary material S1.

The nucleotide sequences in figure 1(b) are fasta or gbk files extracted from the GenBank reference map [19]. We do not use the original sequences from the articles, because the literature only publishes the binding site nucleotides. This is insufficient for π-orbitals. We are not restricted just to the nucleotides of the consensus CTCF motif. The electronic nucleotide description of nucleotide π-orbitals considers the effects of the surround of the core 20-mers. Electrons can easily hop for 16.8 (AT rich sequences) or 25 Å (CG rich) [20], which comprehend at least 5 to 8 bp of the surrounding nucleotides over the core 20-mer binding site. Results with transcription factor specificity protein 1 (SP1) and early growth response protein 1 (EGR1) [10] show the existence of HOMO and LUMO surrounding binding site. Similar phenomena happen for CTCF as we will report in this work, although the biological function of HOMO and LUMO is unknown yet. We can easily find the selected binding sites in the GenBank reference map with the same SP1, EGR1, initiator element (Inr), Goldberg-Hogness box (TATA box) and other expected genomics features. All selected binding sites must be experimentally confirmed for multiple methods.

2.2. Nucleotide alignment using local electronic density of states

The starting point in our method is the quantum description of the nucleotide π-orbitals along the genome, considering three terms in the Hamiltonian (equation (1) of the supplementary material S2): electron–electron, electron–nucleotide displacement field and electron–nucleotide interactions. The first term is just the free electron along the base pairs. The electron–nucleotide displacement field is the π-orbital response with its own nucleotide. The electron–nucleotide interaction between two base pairs are represented by the Morse potential and anharmonic spring. This technique combines DNA melting [21] with the extended ladder model [22, 23]. When we diagonalize the proposed Hamiltonian in eigenvalues and eigenvectors, the nucleotide π-orbitals along DNA is described as local density of states (LDOS) of the ground state, holes (nucleotides in the valence band without free electron) and highest occupied orbital (HOMO) along with lowest unoccupied orbitals (LUMO), beyond the usual four letters nucleotide alignments [10]. The computation of LDOS is detailed in supplementary material S2.

In the context of charge transport, the valence band is the energy levels of the electrons between the ground state and HOMO. The conduction band are the energy levels of the electrons beyond LUMO. Since we have one free electron per nucleotide in the extended ladder model, the valence band will be completely filled and the conduction band will be empty. The ground state electrons are the least mobile, while the HOMO electrons are the most movable ones and they may hop from HOMO to LUMO. In this work, the nucleotides with ground state electrons, marked in yellow in figure 1(c), are actually the nucleotides with at least 10% of probability of finding the degenerated ground state electrons. The difference between HOMO and LUMO is absent in conductors, while electric insulators present wide gaps. The gap in the extended ladder model [10] gives a semi-conductor characteristic for the double helix.

The common tools for four letters nucleotide alignment are useful for early alignments of the CTCF samples mentioned in the previous section [24, 25], but these drafts should be reevaluated since they do not consider the electronic features pointed in this article. So, we perform a second nucleotide alignment, considering simultaneously adenine (A), citosine (C), guanine (G), timine (T), ground states (yellow), HOMO (orange) and LUMO (red bordered boxes), figure 1(c). This second alignment in figure 1(c) is made manually.

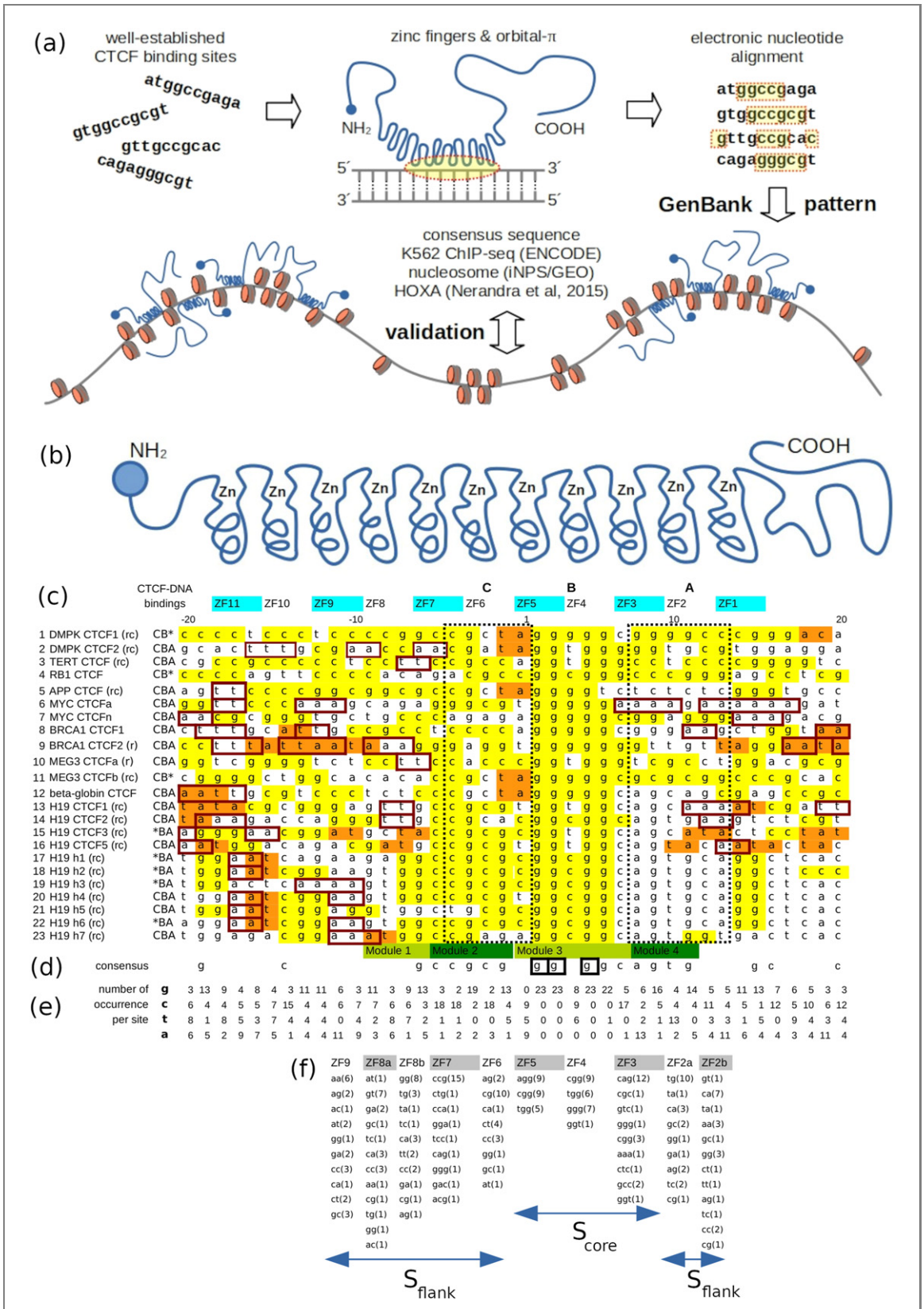


Figure 1. (a) Our workflow. (b) Sketch of the molecular structure of the CCCTC-binding factor (CTCF) with 11 zinc fingers (ZF). Panel (c) shows the electronic nucleotide alignment for CTCF. We indicate the reading direction in reverse and complementary strands with r and c in the parentheses, respectively. Nucleotides with at least 10% of probability in finding the degenerated ground state electrons are in yellow. The highest occupied (HOMO) and the lowest unoccupied molecular orbital (LUMO) nucleotides with at least 10% of probability in localizing one electron are respectively indicated by orange and red bordered boxes. The eleven CTCF zinc finger positions are marked by the succession of blue and white boxes. The CTCF-DNA binding patterns are indicated by the letters CBA. We indicate respectively the binding positions A and C of ZF2 and ZF6 using dashed boxes. B is about the CTCF-DNA binding of ZF4 and ZF5. We tag respectively by *BA and CB*, when the CTCF miss ZF6 or ZF2 bindings. We remark that there is no nucleotide position zero, in line with [26]. (d) The consensus sequence is the simple majority (number of alignment nucleotides $\geqslant 12$). The black bordered boxes indicate the guanines that appear in all studied CTCF binding sites. The four modulus in [13] are indicated in light and dark green. (e) Is number of nucleotide occurrence per site. (f) Is the nucleotide motifs along the columns with the number of repetitions in parentheses.

We cannot ignore the symmetries of the genetic code, since CTCF read dDNA in four directions in function of complementary and reflection symmetries. So, the charge patterns of the tips of ZF and the LDOS of DNA chains must be evaluated in the direct or positive strand and direct reading (from 5′ to 3′), in the direct strand and reverse reading (from 3′ to 5′), complementary or negative strand with direct reading and complementary strand with reverse reading.

2.3. Pattern identification

We divide the prediction technique in two parts. In the first part of the technique, we scan the contiguous sequences (contigs), looking for the electronic distribution patterns as a very specific ground state positions in guanines and absence of HOMO and LUMO around CTCFbs. These electronic patterns, figure 1(c), are described further in the text (section 3.1, consensus sequence). Then, we consider the number of nucleotide occurrence and the motifs in figures 1(e) and (f). Since the length and the number of the binding sites is small in figure 1(c), we do not use any algorithm for motifs detection and classification. We arrange the nucleotides manually. Indeed, there are only four and three observed motifs in the ZF4 and ZF5 triplets, figure 1(f). The number of motifs is reduced in the middle of the binding site, but large in the flanking region. So, we divide the nucleotides in two sets: S_{core} and S_{flank}.

In the core of the CTCFBs, we define the geometric average probability $P_{\text{core}}(S_{\text{core}}) = \left[\prod_k P(S_k)\right]^{1/3}$ where $S_{\text{core}} = \cup_k S_k$, $k = \{\text{ZF3, ZF4, ZF5}\}$, and $P(S_k)$ is the probability of occurrence of the motif S_k, figure 1(f). We have a cubic root in P_{core}, because we are analyzing the patterns of 3 zinc fingers. After extensive tests localizing the listed figure 1(c) in GenBank flat files, we conclude that a minimum of 9.0% for P_{core} is required for a valid DNA-CTCF binding.

In the region flanking the core, we define a probability $P_{\text{flank}}(S_{\text{flank}}) = \frac{1}{2}\left[\prod_k P(S_k)\right]^{1/7} + \frac{1}{2}\left[\prod_i P(S_i)\right]^{1/15}$ where $S_{\text{flank}} = \cup_k S_k$, $k = \{\text{ZF2a, ZF2b, ZF6, ZF7, ZF8a, ZF8b, ZF9}\}$, $P(S_k)$ is the probability of occurrence of the motif S_k, and $P(S_i)$ is the probability of the nucleotide occurrence S_i in the position i, $i = -11, \ldots, -1, 10, \ldots, 13$. The first term $\left[\prod_k P(S_k)\right]^{1/7}$ in P_{flank} guarantees the detection of nucleotide sequences listed in figure 1(f), and we have 7th root in the expression since we are considering seven elements in S_k. However, there are considerable variation in S_{flank}, comparing with P_{core}. If we restrict the motifs just in figure 1(f), we will miss valid CTCFbs. So, we introduce $\left[\prod_i P(S_i)\right]^{1/15}$ in P_{flank}. We decompose the flanking sequence in their 15 nucleotides, $S_i = \{a, t, c, g\}$. Then, we estimate the geometric average probability associated with the occurrence of each particular nucleotide S_i along the binding site, figure 1(e). Our tests show that the probability of a valid CTCFbs P_{flank} should be bigger than 6.5%.

We illustrate the procedure in the supplementary material S4.

3. Validation

3.1. Consensus sequence

The most striking feature of the alignment of 23 CTCFbs in figure 1(c) is the guanine at the positions 2, 3 and 5, marked with a black box in figure 1(d). Actually, guanines at the position 2 and 5 coincide with the middle nucleotide of the triplet of the ZF4 and ZF5 and the amino acid of tip of these ZF tips are base. So, the positive charged tips of ZF4 and ZF5 bind with the ground state electrons of guanines in position 2 and 5; a similar mechanism is described in [5, 7, 8, 27, 28]. Coarse-grained Monte Carlo simulations confirm this finding [29, 30]. Further, Kim *et al* [2] increases the specificity of their CTCF binding site prediction using these same nucleotides in positions 2 and 5 as well as −4 and 7 (positions 6, 11, 14 and 16 in their article). There is always adsorption of the zinc fingers 4 and 5 by the DNA.

We do not observe HOMO between −5 to −2 and 2 to 9, and there is no LUMO between −4 to 6. We never observe over-position between ground state and HOMO or LUMO electrons. The core of CTCF-DNA binding sites is a region without mobile electrons and CTCF anchors their zinc fingers in the most stable electrons, i.e. ground state electrons.

Since the electronic alignment considers the charges in the tips of the zinc-finger [10], the eleven ZFs in CTCF reveal more details about the protein-DNA attachment. There are five ZF with well-defined charge motifs: ZF2, ZF4 ZF5, ZF6 and ZF9. The finger tip is acid (negative) for ZF2 and ZF6 as well as base (positive) for ZF4, ZF5 and ZF9. Electrons in the nucleotides will bind the positive tips, and holes in negative ones. We will ignore ZF9, because it is neither in the core binding site nor fundamental for CTCF-DNA binding [29]. ZF4 and ZF5 always bind with the dDNA [29]. We do not find any particular property for ZF3. Thus, we will focus on the binding sites for ZF2 and ZF6 (respectively A and C in figure 1(b)) and ZF4 and ZF5 marked as B in figure 1(c). Instead of three nucleotides of the triplet, we consider five nucleotides in A and C, blue box in figure 1(c), because the CTCF is a flexible molecule and the finger may displace back and forward along the double helix. The site B is the triplets under ZF4 and ZF5. CTCF sometimes misses the binding sites A or C, but it always binds in B. CTCF-DNA binding is successful only if we do not miss A and C sites simultaneously, figure 1(c).

The consensus sequence in figure 1(d) is just the simple majority (number of alignment nucleotides $\geqslant 12$). We avoid the Schneider and Stephens logo, and we use neither the Shannon information content,

Gibbs binding free energy nor position weight matrix for the calculus of the specific-binding free energy [31–34], because we have neither a clear boundary for the binding for the background sequences nor consider the flanking sequences. We get better results circumventing the intricate heuristic weighting factors and scores of the nucleotide alignments or misalignments [35], neural networks [36], and we do not use MNase-seq [37] and ChIP-seq sequences from ENCODE in order to find the motif behind CTCF [14, 38], simplifying the localization process and saving computational time. Despite the over simplification, we have a good matching with the consensus 5′-ccgcgnggnggcag-3′ [2, 11–13]. These nucleotides are divided in four moduli [13]. We use the border between module 2 and 3 as the position of reference. So, the nucleotide in position 1 is at the beginning of the module 3. The nucleotide at the position −1 is the first one before nucleotide in position 1. Following the literature, there is no position zero [26]. The modulus 2 in [13] is related with ZF6, modulus 3 with ZF4 and ZF5 and modulus 4 with ZF2 and ZF3. ZF9 is maybe connected with modulus 1, but the sequence is at the right of ZF9 triplet and the evidence of consensus sequence is too faint for conclusions [2, 13].

3.2. CTCF and ChIP-seq K562 data

Once we identify the electronic nucleotide pattern and establish a criteria for CTCF binding sites, we localize all human CTCFbs along the assembly hg38, table 1. We find 335 088 binding sites. This number is remarkable close to the total cumulative number of 326 840 CTCF binding sites identified by Chen *et al* using data from 38 human cell lines [4].

We compare our predicted CTCFbs to the ChIP-seq K562 ubiquitous binding sites. The 8771 ubiquitous CTCFbs from 5 ENCODE K562 files are described in supplementary material S5. We have 29.8 ± 3.8% of perfect match between our method against experimental data. The median Q2 of the distances between predicted and observed binding sites shows us that 50% of the putative are just at a 473 bp distant from the expected one and 75% of them (third quartile, Q3) are at the maximum 2352 bp. Beyond Q3, we have some huge discrepancies reaching 73 250 bp. As we lay out in the discussion, the discrepancy of the last quartile (25% of data) between our putative CTCF binding sites and those detected by ChIP-seq comes from the limitations of the chromatin immunoprecipitation technique.

We can improve the matching in light of the helical geometry of the dDNA. When we observe the three-dimensional structure of dDNA, there are two possible grooves where the zinc finger will insert into the dDNA to read the π-orbital. The major groove is 22 Å large, while the minor groove has only 11 Å [39]. We expect more CTCFbs in the direct strand and direct reading (from 5′ to 3′) and in complementary strand and reverse reading (from 3′ to 5′), since it is easier for the CTCF to insert into the major groove. We can see in figure 1(c) that we have 21 samples in the major groove and the matching between predicted and ChIP-seq K562 data increases: 34% of binding sites will have a perfect matching, with Q2 = 401 bp, Q3 = 2238 bp and a maximum discrepancy of 60 676. However, we have only 22% of matching, Q2 = 621 bp, Q3 = 2348 and a maximum of 73 246 bp difference for CTCF binding in the minor dDNA groove. Here, we linked the minor groove with the direct reading in the complementary strand and reverse reading in the direct strand. In figure 1(c) BRCA1 CTCF2 and MEG3 CTCFa are in the direct strand and reverse reading, associated with the minor glove. The absence of major and minor groove distinction in our method is obvious when we see the chromosomal average proportions of each reading direction: direct strand and direct reading is 29 ± 1% of the predicted CTCFbs; direct strand and reverse reading has 21 ± 1%; complementary strand and direct reading values 20.8 ± 0.7%; and complementary strand and reverse reading is 29.4 ± 0.7%. The number of direct strand and reverse reading as well as complementary strand and direct reading could be overestimated.

3.3. CTCF and nucleosome

In order to evaluate the coherence of our findings, we study the nucleosome distribution around our putative binding sites. The nucleosome binding sites are localized using an improved nucleosome positioning algorithm (iNPS) over the sample GSM1095279 from Gene Expression Omnibus database, a MNase-seq assay in human IMR90 fetal lung fibroblast cell [15, 17]. iNPS increases the number of detected nucleosomes [16]. A detailed description of iNPS can be found in supplementary material S6. We extract 5968 503 nucleosomes from this sample, covering 658 606 003 bp, resulting in an average nucleosome density ρ of 21% for human genome.

We combine iNPS and our CTCFbs data in figure 2(a). $\rho(i_{\mathrm{nucl}} - i_{\mathrm{ctcf}})$ is the nucleosome density around CTCFbs for the complete human genome, the main variable $i_{\mathrm{nucl}} - i_{\mathrm{ctcf}}$ is the nucleosome position minus the position of CTCF center in the unit of bp. In order to improve the quality of the nucleosome peaks, we consider only the CTCFbs of the major DNA groove, because they are less affected by the helical geometry of dDNA as we discussed in the previous section. The average nucleosome density ρ around CTCF binding site is 30% instead of 21%, mentioned in the last paragraph. ρ is always higher in CTCFbs rich domain. When we look for nucleosome fluctuation around each CTCFbs, we find 7 nucleosomes peaks around CTCF in direction of the N-terminus and 8 nucleosomes in the C-terminus direction, while [4, 40] report 20 nucleosomes around CTCFbs. [4] uses data for nucleosome and CTCF from the same source, GENCODE [41]. [40] uses CTCFbs from

Table 1. L is the genome length, n_{ctcf} is the predicted CTCF binding sites (CTCFbs) and $\langle l_{ctcf} \rangle$ is the chromosomal average of CTCFbs density. The probability distribution $P(\Delta)$ of the difference Δ between two consecutive CTCFbs obeys a scaling law α from 11 bp $\leqslant \Delta \leqslant$ 2000 bp to 16 bp $\leqslant \Delta \leqslant$ 17 000 bp, depending of the considered genome. $P(\Delta)$ follows an exponential decay with typical length λ, when we consider the fitting regions from 2000 bp $\leqslant \Delta \leqslant$ 78 kbp to 9.7 kbp $\leqslant \Delta \leqslant$ 99 kbp.

	L (bp)	n_{ctcf}	$\langle l_{ctcf} \rangle$ (kbp)	α	λ (kbp)
Human all[a]	2814 809 546	331 668	8.8 ± 3.1	0.511 ± 0.014	19.28 ± 0.24
Human centromer	76 305 151	1892	38 ± 17	No structure	
Human variable	14 059 087	1528	9.207[b]	No structure	
Mouse all[c]	2541 456 020	277 027	9.4 ± 1.9	0.3292 ± 0.0068	19.79 ± 0.31
Mouse chromY	82 248 315	2512	32.742[b]	Detailed in the text[d]	
Pig	2389 924 585	316 919	7.9 ± 2.7	0.484 ± 0.013	22.44 ± 0.37
Goat	2462 599 335	264 286	9.7 ± 3.2	0.5409 ± 0.0064	24.24 ± 0.35
Fruit fly	128 506 876	8962	14.4 ± 2.2	0.454 ± 0.012	18.08 ± 0.52
Fruit fly chrom4	1200 662	20	60.033[b]	No structure	
A. aegypti	1195 030 408	39 777	30.043[b]	144 ± 4 bp, 287 ± 5 bp[d]	42.1 ± 2.0
C. elegans	100 272 607	2086	48.2 ± 7.7	No structure	
P. falciparum	23 264 338	46	530 ± 340	No structure	
A. thaliana	116 129 212	1595	72.7 ± 3.6	No structure	

[a]Heterochromatins were excluded.
[b]No standard deviations due to the reduced amount of data.
[c]Chromosome Y is excluded.
[d]Well-defined Δ CTCFbs distances.

UCSC genome browser [42] and the nucleosome positions are predictions using [43]. Since our analysis is done on a genome-wide scale, the relatively remote nucleosomes are considered to be more fluctuated and hard to recognized in the figure; on the other hand, this also demonstrates the prominence of the observed peaks. Each nucleosome in our work includes 185 ± 21 bp, equivalent to the sum of 147 bp necessary to wrap one nucleosome and 38 bp for the linker in agreement with [4, 16, 40]. [4, 16] describe a symmetrical distribution, because they do not consider the CTCF reading direction. Since the reading direction is available in our analysis, we observe asymmetry in the nucleosome positioning around CTCFbs as [40, 44]. The distance between the CTCFbs and the first peak for the C-terminus is shorter than N-terminus as reported in [44], and we have a substantial fluctuation in the position 962 bp with an error of 40 bp, figure 2(a). We should expect this asymmetry in the nucleosome distribution around CTCFbs, because N and C terminus have different structures.

3.4. CTCF and HOXA

Once we have established a reliable protocol for the human genome, we have applied the method to the mouse genome. We find 279 539 CTCF binding sites in the build GRCm38. The whole genomic chromosomal average predicted binding sites $\langle l_{ctcf} \rangle$ for mouse are comparable with the human genome.

We also assess our method for the mouse HOXA gene cluster, composed by 11 genes (A1-7, A9-11 and A13), figure 2(b), in order to reproduce Narendra *et al* findings [18]. We detect all CTCFbs between A5 and A6 (C5|6), A6 and A7 (C6|7), A7 and A9 (C7|9), A10 and A11 (C10|11), before and after A13 reported by [18]. Nevertheless, we should evaluate this statement carefully, since our program detects actually 52 binding sites, while [18] reports just 6. There are many CTCFbs organized in cluster-like groups [4] as C6|7 with two binding sites, C7|9 with 4, C10|11 with two, the CTCFbs before and after A13 with two and four respectively, figure 2(b). The CTCF assays by [18] have a precision around 1 kbp and are unable to find one particular 20 bp long CTCF binding site. In the case of the CTCFbs after A13 gene, the CTCFbs cluster stretches for 3625 bp. The motif based methods in CTCF assays do not consider the local repeats as alternative sites. They choose one of possible sites that may spread for few kbps. Interestingly, the faint signal between the gene A4 and A5, not reported by [18], is positioned in one putative CTCFbs group, detected by our program. This faint signal is a cluster with seven CTCFbs, asterisk in figure 2(b). We also observe, respectively, a mismatch of 2981 and 2795 bp between our results and [18] for C5|6 and C6|7. The source for this displacement is the considered CTCF consensus motifs. In order to test the robustness of our electronic alignment, we do not include their CTCF motifs (supplemental material in [18]) in our 23 samples, figure 1(b).

In the previous section, we show that we get better results considering only CTCFbs in the major dDNA groove. However, the CTCFbs in the minor DNA cannot be completely neglected. Since we have many CTCFbs in figure 2(b). For example, the CTCFbs of C5|6, C6|7 and C10|11 are all in the minor dDNA groove.

The mismatches between our results and [18] around C5|6 and C6|7 give us an idea about the inaccuracy in the positioning i_{ctcf} of our method. The estimate of misplacement is around 3 kbp. However,

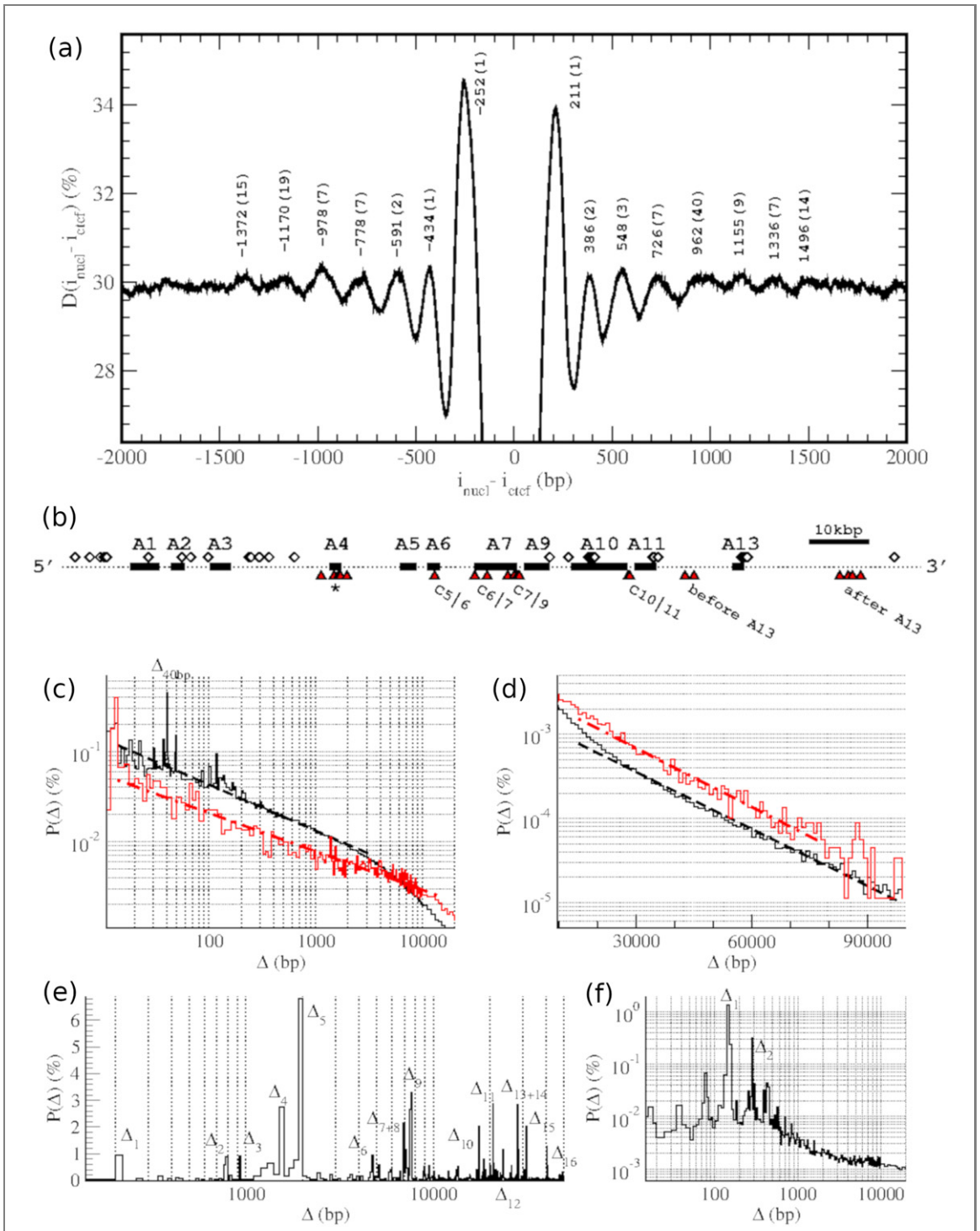


Figure 2. (a) The black line shows the average chromosomal density $\rho(i_{\mathrm{nucl}} - i_{\mathrm{ctcf}})$ of nucleosomes per nucleotide around the predicted CTCF binding site (CTCFbs) i_{ctcf}, excluding the nucleosome in the CTCF position. The exact neighbor nucleosome positions are indicated by numbers above the peaks with the error in the parenthesis. (b) Shows the HOXA genes (black line), the predicted (white diamond) and predicted CTCFbs which are experimentally confirmed (red triangle) from the mouse HOXA gene cluster. The red triangles in C5|6, C6|7, C7|9, C10|11, before and after C13 are the CTCFbs reported in [18]. The asterisk indicates the faint response for CTCF in [18], not reported by the authors. (c) and (d) Are probabilities $P(\Delta)$ in finding the next consecutive putative CTCF binding site in percentage against the distance Δ in base pairs for human (black) and fruit fly (red). (c) The dark dashed and the red dotted dashed lines indicate the power-law for human and *D. melanogaster*, respectively. (d) The exponential fitting for human and fly in semi-log scale are also pointed by dark dashed and red dotted dashed lines. (e) $P(\Delta)$ of mouse chromosome Y with multiple typical $\Delta_1 - \Delta_{16}$ distances. (f) *Aedes aegypti* $P(\Delta)$ with the characteristic 144 ± 4 bp (Δ_1) and 287 ± 5 bp (Δ_2) distances.

the most evident feature in figure 2(b) is the coalescence of the CTCFbs, reported by [4] as clusters of binding sites. However, the concept of cluster demand a Gaussian among CTCFbs distribution and we do not observe such structure. Since our electronic alignment is not limited by poor quality data [9, 45] or absence of the expected 20-mer consensus motif [4], we make more accurate analysis.

4. Results

Instead of a cluster organization for CTCFbs suggested by [4, 12], we implement another evaluation, detecting a power law in $P(\Delta)$, table 1 and figure 2(c), indicating organized structure for CTCFbs. Here, Δ is the distance of two consecutive CTCFbs and $P(\Delta)$ is the probability of finding the next binding site. In humans we adjust α in $P(\Delta) \approx \Delta^{\alpha}$, considering two or three orders of magnitude. The human euchromatic regions have $\alpha = 0.511 \pm 0.014$, fitting within the interval 14 bp $\leqslant \Delta \leqslant$ 2400 bp. The region with a power law in $P(\Delta)$ covers 39.98% of the euchromatic binding sites. Furthermore, the chromosomal average α of mouse values 0.3292 ± 0.0068, covering 43.93% of binding sites, and it is fitted in the interval 20 bp $\leqslant \Delta \leqslant$ 4.1 kbp.

For the region beyond polynomial fitting, $P(\Delta)$ decays exponentially, $P(\Delta) \approx e^{-\Delta/\lambda}$. The characteristic length λ for humans values $\lambda = 19.28 \pm 0.24$ kbp and 15 kbp $\leqslant \Delta \leqslant$ 99 kbp is the exponential adjustment region, comprising 16.36% of binding sites. In the case of the mouse, the genomic λ values 18.06 ± 0.29 kbp with 11 kbp $\leqslant \Delta \leqslant$ 89 kbp, containing 23.77% of CTCFbs.

A similar feature is described for the human K562 CTCF binding sites distribution [4]. However, we cannot compare the power law CTCFbs distribution for the entire genome directly with their cluster analysis [4], since the power law has not a characteristic length by definition. Thus we use a cluster analysis, assuming those CTCFbs to be nearest neighbors that are within 3058 bp and hence in one particular cluster. We choose 3058 bp because this is the median for the complete genome Δ as well as this is close to the upper limit of the power fitting, table 1. Thus, 63.68% of our cluster-like structures can be classified as singletons (isolated CTCFbs), while [4] reports 38.94%. The groups with 2, 3, 4, 5, 6 and more than 6 CTCFbs values respectively 18.09%, 7.08%, 3.45%, 2.11%, 1.26% and 4.26% while [4] indicate 25.09%, 14.60%, 8.79%, 5.22%, 3.10% and 4.26% in their cluster map. Although we have more singletons in our results, we have the same percentage for cluster-like structures with more than 6 CTCFbs reported by [4].

We do not restrict our analysis just to human and mouse. We confirm the existence of cluster-like structures in pig and goat, where we find 316 919 and 264 286 CTCFbs respectively. Both average chromosomal CTCFbs densities $\langle l_{\mathrm{ctcf}} \rangle$ are compatible with the human and mouse, but direct comparison should be avoided because we exclude the heterochromatin in the human genome. α values are 0.484 ± 0.013 and 0.5409 ± 0.0064 for pig and goat respectively. They contain 44.27% (pig) and 41.31% (goat) of the binding sites. Both species have the same regions for α fitting: 14 bp $\leqslant \Delta \leqslant$ 2000 bp, but the domains for λ adjustments are different: 14 kbp $\leqslant \Delta \leqslant$ 99 kbp for pig, covering 13.75% of binding sites; and 18 kbp $\leqslant \Delta \leqslant$ 99 kbp in the case of goat, composing 13.31% of CTCFbs.

$P(\Delta)$ is not limited just to polynomial and exponential fittings. We have many CTCFbs that are 13 bp apart from each other as well. 5.67% of human euchromatin, 6.39% of mouse without chromosome Y, 7.58% of pig and 7.19% of goat binding sites are in the region $0 < \Delta \leqslant$ 13 bp, and $P(\Delta)$ distributions are not uniform. We observe few binding sites with $\Delta = 2$, 5 or 7 bp and the height of $P(\Delta)$ is species dependent. By the way, 0.23%, 0.40%, 0.49% and 0.46% of binding sites are $\Delta = 0$ distance respectively in human, mouse, pig and goat, i.e. the CTCF has multiple binding modes in these sites as mentioned previously.

We also apply our method to the fruit fly and localize 8962 binding sites. Although the genome size is just 5% of mammals, $P(\Delta)$ of *Drosophila melanogaster* resembles mammal with a well-defined power law $\alpha = 0.454 \pm 0.012$ and exponential decay $\lambda = 18.08 \pm 0.52$ kbp. The polynomial and exponential fittings are along 14 bp $\leqslant \Delta \leqslant$ 14 000 bp and 15 kbp $\leqslant \Delta \leqslant$ 78 kbp, covering 62.73% and 29.83%. 5.87% of the binding sites are 13 bp or less distant each other and 0.12% has $\Delta = 0$.

We study the genome of *A. aegypti* and identify 39 777 binding sites. We do not find a power law, but 16.33% and 3.57% of binding sites are respectively 144 ± 5 bp (Δ_1) and 287 ± 5 bp (Δ_2) at a distance of each other, figure 2(e). We remark that we need 146 bp to wrap one nucleosome. 0.37% of sites has multiple binding modes. $P(0 < \Delta \leqslant 13$ bp) is unlike the other genomes, since 1.26% of CTCFbs are just at one bp distance of each other. When we consider a region of 4.5 kbp $\leqslant \Delta \leqslant$ 99 kbp for the exponential fitting, we have $\lambda = 42.1 \pm 2.0$ kbp. The exponential fitting contains 61.47% of CTCFbs.

This odd behavior can be observed in mouse chromosome Y too, where we find 2512 binding sites. The low density of $\langle l_{\mathrm{ctcf}} \rangle = 32\,742$ bp per predicted CTCFbs hides a surprise. This $\langle l_{\mathrm{ctcf}} \rangle$ is just 9% higher than *A. aegypti*, and there is neither a power law nor an exponential decay. The number of binding sites in the region where $0 < \Delta \leqslant$ 13 bp is minimal, is just 0.6%. We do not observe multiple CTCF binding modes, $P(\Delta = 0) = 0$. Although mouse chromosome Y lacks a power law and an exponential decay, 35.51% of binding sites presents well-defined Δ distances: 0.96%, 1.47%, 0.92%, 2.15%, 6.21%, 0.80%, 1.83%, 1.15%, 5.02%, 2.03%, 2.87%, 1.15%, 2.83%,

1.15%, 2.03% and 2.95% of the binding sites are 208 ± 5 (Δ_1), 780 ± 7 (Δ_2), 931 ± 1 (Δ_3), 1539 ± 8 (Δ_4), 1927 ± 2 (Δ_5), 4736 ± 4 (Δ_6), 6949 ± 11 (Δ_7), 7161 ± 22 (Δ_8), 7587 ± 12 (Δ_9), $17\,523 \pm 16$ (Δ_{10}), $20\,862 \pm 21$ (Δ_{11}), $23\,551 \pm 27$ (Δ_{12}), $28\,153 \pm 29$ (Δ_{13}), $28\,460 \pm 21$ (Δ_{14}), $31\,452 \pm 25$ (Δ_{15}) and $40\,565 \pm 91$ bp distance of each other (Δ_{16}) in figure 2(d), respectively.

We test our method for *Plasmodium falciparum* (low unicellular eukaryote) and *Arabidopsis thaliana* (plant), where CTCF is absent [46]. In the case of *P. falciparum*, table 1, the number of CTCFbs spotted by our method is so small that we cannot even build $P(\Delta)$. As a matter of fact, there are only 3 ± 3 CTCF binding sites per chromosome. We have better statistic for *A. thaliana*, table 1, where we detected 1595 CTCFbs. The expected binding sites in the region $0 < \Delta \leqslant 13$ bp is represented by 7.4% of CTCFbs and they are at 6 ± 4 bp distance of each other. We do not report multiple binding modes for these species, $P(\Delta = 0) = 0$ and there is neither a polynomial nor an exponential decay for $P(\Delta)$. These binding sites detected by our method are false positives. They are born from the $P(S_i)$ statistics in P_{flank} from pattern identification and other limitations outlined along this manuscript.

Caenorhabditis elegans is another interesting specimen. Although this worm lost its CTCF gene along the evolution [47], we encounter 2086 binding sites, possible remains of its segmented body past [47]. In the region $0 < \Delta \leqslant 13$ bp, we have 5.27% of the binding sites and there are two sites with multiple binding modes. These values are compatible with mammalian genomes. But we do neither find a power law nor an exponential decay in its CTCFbs distribution. The density of CTCFbs in *C. elegans* is 48.2 ± 7.7 kbp. This $\langle l_{\text{ctcf}} \rangle$ is not far from human centromeric domains (38 ± 7 kbp per CTCFbs, table 1). Here we have $P(0 < \Delta \leqslant 13 \text{ bp}) = 3.3\%$ and 0.2% of sites present multiple binding mode, but we do neither find a power law nor an exponential decay.

One may argue the absence of a power law and exponential decay in $P(\Delta)$ is due to the low density of CTCFbs $\langle l_{\text{ctcf}} \rangle$ in the human centromeric domain or in mouse chromosome Y. However, we have an unusual concentration of binding sites in the human noncentromic and nontelomeric heterochromatin regions (gvar). These domains are: the entire 3q11.2 and 19q12; the initial part of 9q12, 19p12 and Yq12; final part of 1q12, 13p11.2, 16q11.2 and 22p11.2. They have 14 Mbp of the length, represent 44.7% of all heterochromatic CTCFbs and 0.3% of the sites has multiple binding modes. Nevertheless, similarities with euchromatic segments end at this point. We do not observe the $P(0 < \Delta \leqslant 13 \text{ bp})$ distribution of the mammal genomes, but 7.5% of CTCFbs are 10 ± 2 bp distant each other. We do neither observe a power law nor an exponential decay in $P(\Delta)$ too.

Finally, we report just 20 binding sites in the chromosome 4 of fruit fly. But, this number is too small for conclusive results.

5. Discussion

The molecular basis for the four letters alignment is the hydrogen bonds of the nucleotides. The adaptation of the Peyrad–Bishop model of the DNA melting for the transcription factor binding [21] also considers the hydrogen bonds as responsible for the electronic pattern along the genome. Although the Peyrad–Bishop explains successfully the separation of the base pair under the temperature variation in polymerase chain reaction, transcription factors, as EGR1, SP1 and CTCF, do not open the double Helix in their search for binding sites. They scan the dDNA, inserting zinc fingers into the major and minor grooves of DNA and probing for π-orbital electronic patterns [7]. So, the Peyrad–Bishop cannot be applied directly for the search of the transcription factor binding site. However, the nucleotide π-orbitals have successfully been described by the extended ladder model, which interprets dDNA as semiconductor-like material [22, 23]. When we apply the extended ladder to transcription factor binding DNA sequences, patterns as in figure 1(c) appear. Again we emphasize that this semiconductor-like description is *in situ* condition dependent.

The method presented in this work is solvent dependent. The electronic nucleotide alignment using the extended ladder model considers the dDNA in atmosphere, low vacuum or Tris-HCl buffers [10, 20, 48, 49]. There is no consensus about the electronic transport properties of dDNA, since the experimental frameworks change the electronic properties of DNA [49]. Ethylenediaminetetraacetic acid (EDTA) or 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffers may induce an electric insulator effect [50]. However, at room temperature and under tris(hydroxymethyl)-aminomethane and hydrochloride salt (Tris-HCl), a traditional physiological buffer with pH = 7.382 at 37 °C [51], dDNA has a semiconductor like behavior [10, 48, 49]. This buffer may emulate the living HeLa cytosol and nucleus conditions, i.e. an aqueous solution with pH around 7.35 [52]. Under this circumstance, we may adopt the charge transport formalism to the nucleotide analysis.

The charge transport formalism adds a new layer over the nucleotide alignment. We are not restricted just to four letter pattern. As in the three ZFs of EGR1 and SP1 [10], CTCF also anchors ZFs in the nucleotides with the most stable electrons, i.e. the ground state and the lowest occupied π-orbitals. These nucleotides are at the core of the consensus motif. Numerical simulation in [29, 30] also show that the central ZFs are the most relevant nucleotide for the CTCF binding. Although the CTCF binding sites localization is not temperature dependent,

the CTCF-DNA interaction is affected by the thermal fluctuation [29]. Coarse-grained Monte Carlo of multi-Cys_2His_2 (mC_2H_2) zinc finger proteins as EGR1, $TATA_{ZF}$, transcription factor IIIA (TFIIIA) and CTCF shows rotation-coupled sliding, asymmetrical roles of zinc fingers and nucleotide dependency. Furthermore, simulated mC_2H_2 binds just with its central zinc fingers as we observe in CTCF.

When we examine our CTCFbs with those in [4], we observe many mismatches. One source for the predict and experimental ChIP-seq CTCFbs differences is the sample number for nucleotide and electronic pattern in figure 1 and for the statistics of P_{core} and P_{flank} in section pattern identification. The 23 samples do not cover all possibilities, although they catch the most common features. Actually [3, 4], also mention these additional motifs beyond the 20-mer consensus motif, positions −9 to 11 in figure 1(d). Moreover, we are not considering homologous CTCFs [6]. The samples in figure 1(c) belong to mouse and human only, and we are defining one common CTCF pattern for them. Although we may expect a general mechanism from a common arrangement, we may foresee specie depend variations in the electronic pattern.

We also introduce noise when we consider the second criteria $\cup_i S_i$ for S_{flank} in pattern identification, based on the nucleotide occurrence, figure 1(e). This term plays a similar role as the background frequency correction in DNA sequence motifs. Although, this approach adds flexibility, it introduces systematic error in site prediction: the method will consider some false motifs.

CTCF can bind to dDNA in multiple ways as in shown in figure 1(c), but we combined all binding possibilities in one simple binding pattern. Indeed the literature about CTCF motifs does not consider multiple CTCF binding possibilities. However, experimental results [53], numerical simulations [29] and careful charge analysis of the tips of the zinc fingers show many viable binding arrangements. Unfortunately, the sample number in this work is too small for each individual binding configuration. Thus, we joint all, following the literature [4, 13].

The process for positioning the CTCF binding sites in the K562 uses hg38, which is a consensus sequence of nine healthy males [19], while K562 is a tumoral cell from a woman [54]. So, we are using sequences of one person to find the position in the consensus of nine others individuals. Most of sequences will be placed in the correct spot, but we expect discrepancies between these data.

Despite all limitations and criticisms about our method and the ChIP-seq technique, we have 29.8 ± 3.9% of perfect matching and 20.2% of near matching ($\|i_{\text{ctcf}} - i_c\| < 474$ bp, median, Q2), 25% with intermediate misplacing ($474 \leqslant \|i_{\text{ctcf}} - i_c\| < 2376$ bp, third quartile, Q3) and 25% of mismatching bigger than 2374 bp. Surprisingly [9], reports similar result: 55% of successful identification, around 25% with intermediate quality and 20% with poor quality. [9] attributes the poor quality data to the low depth reading in ChIP-seq assays. [4] also reports nearly 30% of CTCFbs without the characteristic 20-mer consensus motif in ChIP-seq data and [2] reports the 20-mer motif in just over 75% of experimentally identified CTCFbs. Moreover, using limited quality data from ENCODE and only five samples of K562 ubiquitous CTCF binding sites do not help us in the evaluation of the electronic nucleotide alignments. Nevertheless, extensive tested and analyzed genome using huge ENCODE data by independent peer as [2, 4, 40] are rare. Otherwise, we may estimate the amount of misleading binding sites captured by our method from the *P. falciparum* and *A. thaliana*, table 1.

There is no CTCF gene for protozoan and plants [46]. So, these binding sites are false positives generated by $P(S_i)$ statistics in P_{flank} in pattern identification. Since we have around one CTCFbs in 9 kbp for mammals (human, mouse, pig and goat), we estimate from 2% to 13% of false positives in our technique considering *P. falciparum* and *A. thaliana* as negative controls. *C. elegans* is not a good negative test. Although this worm lost its CTCF genes along its evolution [46], this organism still hold CTCFbs.

There are three regions for the probability distribution $P(\Delta)$ of the distance Δ of two consecutive CTCF binding sites in human, mouse, pig, goat and fruit fly. In the first region, the binding sites appear in tandem and they are very close to each other, $0 < \Delta \leqslant 13$ bp. The second region starts at 11 bp $\sim$ 20 bp and extends in between 2 kbp to 17 kbp. These are the domains for the power law fitting. The third domain ranges from 2 kbp $\sim$ 15 kbp to 62 kbp $\sim$ 99 kbp, when we have an exponential decay in $P(\Delta)$. Beyond 100 kbp, we have visible structures in optical microscope as the high packed chromatin, coordinated by scaffold proteins in mitotic cells. But, this very large scale organization is not a topic in this paper.

In the $0 < \Delta \leqslant 13$ bp domain, the number of binding sites represents 5.67% to 7.58% of the total. Further, there are always binding sites with multiple reading modes: $0.12\% \leqslant P(\Delta = 0) \leqslant 0.49\%$. Here, we have multiple binding modes due to the molecular CTCF shape variations [29, 53], beyond the different dDNA reading modes due to the symmetries of the genomic code. The upper limit of this region is delimited by the size of the CTCF binding site. The binding site from the position −11 to 13 in figure 1, resulting in a 24 bp of length, is compatible with the literature, where the length values 11 bp $\sim$ 60 bp [4, 13, 55, 56]. However, we need just 4 $\sim$ 5 ZFs for the CTCF-DNA attachment, using just 13 nucleotides. So, it is not surprise that this region end at 13 bp.

We have a power law for Δ beyond 13 bp. This domain ranges from 11 bp $\sim$ 20 bp to 2 kbp $\sim$ 17 kbp,

covering between 39.98% to 62.73% of binding sites. For these distances, CTCF may interact with dDNA as well as other transcription factors due to the N and C-terminals. In human, they are respectively 150 and 265 long amino-acid sequences with distinct highly acid and basic domains [1, 57]. Further, the electronic nucleotide alignment in figure 1(b) shows consistently the presence of LUMOs and HOMOs around a binding site, reinforcing such a possibility. Although the SysZNF database provide insights about the molecular structures of the head and end of homologous CTCFs [6], detailed studies about N and C terminals interaction with DNA are rare and vague, despite experimental results [53].

The CTCF alone is not able to explain the power law. *Aedes aegypti* genome gives us a cue about the CTCF organization in these regions. The characteristic distances of 144 ± 4 bp and 287 ± 5 bp in $P(\Delta)$, table 1 and figure 2(e), reflect the action of the nucleosomes in chromatin. We need 147 bp to wrap one nucleosome core. Moreover, the mouse chromosome Y has a recognizable 208 ± 5 bp distance in $P(\Delta)$, indicating a nucleosome wrapping by 147 bp with linker of 61 bp long. Indeed the mouse chromosome Y distinct distances 780 ± 7 bp, 931 ± 1 bp, $1539\pm$ and 1927 ± 2 bp, figure 2(d), can be also interpreted as a chromatin with respectively 4, 5, 8 and 10 nucleosomes attached in the dDNA with two CTCF in the extremities. The CTCFs of these complexes may connect each other creating small DNA-loops. In the case of Δ ranging from 4736 ± 4 bp to $40\,565 \pm 91$ bp, figure 2(d), we have from 25 to 219 nucleosomes between the binding sites. The presence of nucleosomes around CTCF binding sites is confirmed by [4, 40] as well as in figure 2(a).

The interaction of CTCFs and nucleosomes result in a solenoidal, zig-zag ribbon or other irregular chromatin structures with a polynomial decay in $P(\Delta)$. The distribution of CTCFbs will have a cluster-like appearance, figure 2(e), troubling ChIP-seq procedures [9, 45]. Binding sites in tandem will bring ambiguities in motif alignments used in the ChIP-seq protocol too.

The distance between these cluster-like CTCFbs groups can be examined by the behavior of $P(\Delta)$, when Δ ranges from 2 kbp $\sim$ 15 kbp to 62 kbp $\sim$ 99 kbp. $P(\Delta)$ becomes exponential, because the probability in finding the next CTCFbs after Δ nucleotides is $p(1-p)^{\Delta}$, where p is the probability of occurrence of the CTCFbs. We can approximate this expression as $p\,\mathrm{e}^{-p\Delta}$, since $p \ll 1$. So, we expect an exponential decay in the case of random distribution of CTCFbs. Calling $p = 1/\lambda$, we observe an exponential behavior for $P(\Delta)$, when Δ is bigger than 2 kbp $\sim$ 15 kbp.

We may illustrate the power law and the exponential decay of $P(\Delta)$ in the mouse HOXA gene cluster (cf figure 2(b)). The distance between CTCFbs inside of a cluster-like group never exceed 3058 bp and obeys a power law with $\alpha = 0.3292 \pm 0.0068$ in mouse. Nonetheless, we have a distance around 17 kbp between A4 (*) and C5|6 as well as before A13 and after A13, and $\lambda = 19,79 \pm 0.31$ kbp in table 1.

The number of binding sites is not small in the exponential distances, ranging from 13.31% to 29.83%. In the case of *A. aegypti*, we have 61.47%. The chromatin folding process in these distances cannot be explained just with CTCF and nucleosomes. Multiple different chromosome folding for these Δ distances is mediated by non-histone proteins as cohesin, Ying and Yang 1 (YY1) and others [13, 57].

Moreover, CTCF may skip many binding sites [13]. Monte Carlo simulations show that the depletion of histones along the chromatin has influence over the folding process [58]. This is illustrated in the putative cluster-like binding sites of the genes A10, A11 and A13, indicated by diamonds in figure 2(b), where the binding sites were overlooked by CTCF. The number of binding sites localized by ChIP-seq is usually a fraction of the expected ones, with a chromosomal average of just one in 42 ± 12 human ubiquitous euchromatic CTCF binding sites in the K562 cells.

Finally, we are working with incomplete data. So, direct comparison between species must be done carefully. Major efforts from the community must be done seeking for less fragmented complete sequences. When the number of contigs are large and the size is small, most of them are too short for computing distances between binding sites and the segment number is excessive for handling them individually. The procedures described in this article are not automated yet. So, the manipulation of thousands of contigs is not viable. Furthermore, the many gaps will add noise in the probability distribution $P(\Delta)$ of the distance Δ between two consecutive binding sites. In fact, most of genomes deposited in GenBank are excessively fragmented, even those organized in chromosomes. However, new sequences deposited in GenBank overcome such limitations. The recently reviewed genomes of pig and goat have few gaps (see material), opening new perspectives to unveil the chromosomal organization in the coming years.

6. Conclusions

The CCCTC transcription factor binding sites (CTCFbs) have a characteristic π-orbital nucleotide motif. Mobile electrons are absent in the core of CTCF binding regions, i.e. we do neither observe highest occupied molecular orbitals (HOMO) nor lowest unoccupied molecular orbitals (LUMO) between ZF3 to ZF5. The CTCF may miss ZF2 or ZF6 binding with DNA. But, it cannot miss both simultaneously. There are at least three different ways to CTCF attach to the DNA. Our nucleotide alignment match with those reported in the literature.

We report 335 088 predicted CTCFbs in the whole human genome, using the electronic nucleotide

alignment. When we compare our results with the ubiquitous K562 chromatin immunoprecipitation with massively parallel DNA sequencing data (ChIP-seq), we have 29.8 ± 3.8% of matching. And, 75% of mismatches are with less than 2352 bp distance between the measured one and the predicted from our method. These 2 kbp discrepancies are expected because we use reduced number of experimental sequences for the search of our electronic pattern and the limitations of the extended ladder model. However, larger mismatches (>2 kbp) are due to ChIP-seq assay: insufficient depth of reading, the absence of the 20-mer consensus motif in the ChIP-seq data or even position of multiple CTCF motifs, each one related with one possible binding pattern.

When we combine our predicted CTCFbs and nucleosome positions, we localize 15 nucleosomes flanking CTCFbs as expected. Furthermore, the distribution of nucleosomes around CTCF reveal asymmetry, reflecting the N and C-terminous molecular differences.

We also confirm the experimental results with our theoretical study, detecting all CTCFbs in the mouse HOXA cluster.

We have studied the genomes of *Mus musculus* (mouse), *Sus scrofa* (pig), *Capra hircus* (goat), *Drosophila melanogaster* (fruit fly) and *Aedes aegypti* (mosquito) finding 277 027, 316 919, 264 286, 8982 and 39 777 CTCF binding sites respectively. We also analyzed *Caenorhabditis elegans*, *Plasmodium falciparum* and *Arabidopsis thaliana* as negative controls. Since *C. elegans*, protozoans and plants have no CTCF gene, there are few binding sites as expected.

The CTCFbs distribution along whole genomes of studied mammals and insects, totalizing 11.77 billion nucleotides, may be described as follows: for distances between 11 bp $\sim$ 20 bp and 2 kbp $\sim$ 17 kbp, CTCFbs compose cluster-like groups, where the interval Δ between two consecutive binding sites obeys a power law with a coefficient α varying from 0.3292 ± 0.0068 (mouse) to 0.5409 ± 0.0064 (goat). There is no power law for the *Aedes* genome, but 19.9% of binding sites are at 144 ± 4 and 287 ± 5 bp distance of each other. These cluster-like CTCFbs groups are separated with a typical distance between 18.08 ± 0.52 kbp (fruit fly) to 42.1 ± 2.0 kbp (*Aedes*).

Acknowledgments

The authors wish to thank Lei Liu and Sujeet Kumar Mishra for the discussions about zinc fingers and CTCF. This work is supported by Conselho Nacional de Desenvolvimento Tecnológico e Científico (CNPq), Process Number 248589/2013, Brazil. We acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme Open Access Publishing, by the Baden-Württemberg Ministry of Science, Research and the Arts and by Heidelberg University. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through Grant INST 35/1134-1 FUGG. Kunhe Li would like to acknowledge funding by the Chinese Scholarship Council (CSC). This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1—390900948 (the Heidelberg STRUCTURES Excellence Cluster).

Data availability statement

All CTCF binding sites computed for this work are freely available in Heidelberg Open Research Data (HeiDATA): https://doi.org/10.11588/data/RDISCE.

ORCID iDs

Dieter W Heermann https://orcid.org/0000-0002-3148-8382

References

[1] Klenova E M, Nicolas R H, Paterson H F, Carne A F, , Heath C M, Goodwin G H, Neiman P E and Lobanenkov V V 1993 CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms *Mol. Cell. Biol.* **13** 7612–24

[2] Kim T H, Abdullaev Z K, Smith A D, Ching K A, Loukinov D I, Green R D, Zhang M Q, Lobanenkov V V and Ren B 2007 Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome *Cell* **128** 1231–45

[3] Xie X, Mikkelsen T S, Gnirke A, Lindblad-Toh K, Kellis M and Lander E S 2007 Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites *Proc. Natl Acad. Sci. USA* **104** 7145–50

[4] Chen H, Tian Y, Shu W, Bo X and Wang S 2012 Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome *PLoS One* **7** e41374

[5] Iuchi S 2001 Three classes of C_2H_2 zinc finger proteins *Cell. Mol. Life Sci.* **58** 625–35

[6] Ding G, Lorenz P, Kreutzer M, Li Y and Thiesen H-J 2009 SysZNF: the C_2H_2 zinc finger gene database *Nucleic Acids Res.* **37** D267–73

[7] Wolfe S A, Nekludova L and Pabo C 1999 DNA recognition by Cys_2His_2 zinc finger proteins *Annu. Rev. Biophys. Biomol. Struct.* **3** 183–212

[8] Klug A 2010 The discovery of zinc fingers and their applications in gene regulation and genome manipulation *Annu. Rev. Biochem.* **79** 213–31

[9] Marinov G K, Kundaje A, Park P J and Wold B J 2014 Large-scale quality analysis of published ChIP-seq data *Genes, Genomes, Genet.* **4** 209–23

[10] Oiwa N N, Cordeiro C E and Heermann D W 2016 The electronic behavior of zinc-finger protein binding sites in the context of the DNA extended ladder model *Front. Phys.* **4** 13

[11] Bell A C and Felsenfeld G 2000 Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene *Nature* **405** 482

[12] Essien K, Vigneau S, Apreleva S, Singh L N, Bartolomei M S and Hannenhalli S 2009 CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features *Genome Biol.* **10** R131
[13] Ong C-T and Corces V G 2014 CTCF: an architectural protein bridging genome topology and function *Nat. Rev. Genet.* **15** 234–46
[14] The ENCODE Project Consortium 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project *Nature* **447** 799–816
[15] Edgar R, Domrachev M and Lash A E 2002 Gene expression omnibus: NCBI gene expression and hybridization array data repository *Nucleic Acids Res.* **30** 207–10
[16] Chen W, Liu Y, Zhu S, Green C D, Wei G and Han J-D J 2014 Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data *Nat. Commun.* **5** 4909
[17] Zhao Y *et al* 2019 NucMap: a database of genome-wide nucleosome positioning map across species *Nucleic Acids Res.* **47** D163–9
[18] Narendra V, Rocha P P, An D, Raviram R, Skok J A, Mazzoni E O and Reinberg D 2015 CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation *Science* **347** 1017–21
[19] Benson D A, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman D J, Ostell J and Sayers E W 2013 GenBank *Nucleic Acids Res.* **41** D36–42
[20] Yoo K-H, Ha D H, Lee J-O, Park J W, Kim J, Kim J J, Lee H-Y, Kawai T and Choi H Y 2001 Electrical conduction through poly(dA)–poly(dT) and poly(dG)–poly(dC) DNA molecules *Phys. Rev. Lett.* **87** 198102
[21] Zhu J-X, Rasmussen K Ø, Balatsky A V and Bishop A R 2007 Local electronic structure in the Peyrard–Bishop–Holstein model *J. Phys.: Condens. Matter.* **19** 136203
[22] Senthilkumar K, Grozema F C, Guerra C F, Bickelhaupt F M, Lewis F D, Berlin Y A, Ratner M A and Siebbeles L D A 2005 Absolute rates of hole transfer in DNA *J. Am. Chem. Soc.* **127** 14894–903
[23] Mehrez H and Anantram M P 2005 Interbase electronic coupling for transport through DNA *Phys. Rev.* B **71** 115405
[24] Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H and Gentleman R 2009 ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data *Bioinformatics* **25** 2607–8
[25] Pagès H, Aboyoun P, Gentleman R and DebRoy S 2015 Biostrings: string objects representing biological sequences, and matching algorithms *R Package Version 2.34.1*
[26] Dreos R, Ambrosini G, Périer R C and Bucher P 2013 EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era *Nucleic Acids Res.* **41** D157
[27] Miller J, McLachlan A D and Klug A 1985 Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes *EMBO J.* **4** 1609–14
[28] Nolte R T, Conlin R M, Harrison S C and Brown R S 1998 Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex *Proc. Natl Acad. Sci. USA* **95** 2938–43
[29] Liu L and Heermann D W 2015 The interaction of DNA with multi-Cys_2His_2 zinc finger proteins *J. Phys.: Condens. Matter.* **27** 064107
[30] Liu L, Wade R C and Heermann D W 2015 A multiscale approach to simulating the conformational properties of unbound multi-C_2H_2 zinc fingers proteins *Proteins* **83** 1604–15
[31] Schneider T D, Stormo G D, Gold L and Ehrenfeucht A 1986 Information content of binding sites on nucleotide sequences *J. Mol. Biol.* **188** 415–31
[32] Schneider T D and Stephens R M 1990 Sequence logos: a new way to display consensus sequences *Nucleic Acids Res.* **18** 6079–100
[33] Bailey T L and Elkan C 1994 Fitting a mixture model by expectation maximization to discover motifs in biopolymers *Proc. 2nd Int. Conf. on Intelligent Systems for Molecular Biology* pp 28–36
[34] D'haeseleer P 2006 What are DNA sequences motifs? *Nat. Biotechnol.* **24** 423–5
[35] Setubal J and Meidanis J 1997 *Introduction to Computational Molecular Biology* (Boston: PWS Publishing)
[36] Mount D W 2004 *Bioinformatics Sequence and Genome Analysis* 2nd edn (New York: Cold Spring Harbor Laboratory Press)
[37] Zhong J, Wasson T and Hartemink A J 2014 Learning protein-DNA interaction landscapes by integrating experimental data through computational models *Bioinformatics* **30** 2868–74
[38] Gerstein M B *et al* 2007 What is a gene, post-ENCODE? History and updated definition *Genome Res.* **17** 669–81
[39] Wing R, Drew H, Takano T, Broka C, Tanaka S, Itakura K and Dickerson R E 1980 Crystal structure analysis of a complete turn of B-DNA *Nature* **287** 755–8
[40] Fu Y, Sinha M, Peterson C L and Weng Z 2008 The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome *PLoS Genet.* **4** e1000138
[41] Frankish A *et al* 2019 GENCODE reference annotation for the human and mouse genomes *Nucleic Acids Res.* **47** D766–73
[42] Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M and Haussler D 2002 The human genome browser at UCSC *Genome Res.* **12** 996–1006
[43] Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore I K, Wang J-P Z and Widom J 2006 A genomic code for nucleosome positioning *Nature* **442** 772–8
[44] Clarkson C T, Deeks E A, Samarista R, Mamayusupova H, Zhurkin V B and Teif V B 2019 CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length *Nucleic Acids Res.* **47** 11181–96
[45] Park P J 2009 ChIP-seq: advantages and challenges of a maturing technology *Nat. Rev. Genet.* **10** 669–80
[46] Heger P, Marin B, Bartkuhn M, Schierenberg E and Wiehe T 2012 The chromatin insulator CTCF and the emergence of metazoan diversity *Proc. Natl Acad. Sci. USA* **109** 17507–12
[47] Heger P, Marin B and Schierenberg E 2009 Loss of the insulator protein CTCF during nematode evolution *BMC Mol. Biol.* **10** 84
[48] Cai L, Tabata H and Kawai T 2000 Self-assembled DNA networks and their electrical conductivity *Appl. Phys. Lett.* **77** 3105
[49] Taniguchi M and Kawai T 2006 DNA electronics *Physica* E **33** 1–12
[50] de Pablo P J, Moreno-Herrero F, Colchero J, Gómez Herrero J, Herrero P, Baró A M, Ordejón P, Soler J M and Artacho E 2000 Absence of dc-conductivity in λ-DNA *Phys. Rev. Lett.* **85** 4992
[51] Durst R A and Staples B R 1972 Tris/Tris·HCl: a standard buffer for use in the physiologic pH range *Clin. Chem.* **18** 206–8
[52] Llopis J, McCaffery J M, Miyawaki A, Farquhar M G and Tsien R Y 1998 Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins *Proc. Natl Acad. Sci. USA* **95** 6803–8
[53] Filippova G N, Fagerlie S, Klenova E M, Myers C, Dehner Y, Goodwin G, Neiman P E, Collins S J and Lobanenkov V V 1996 An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes *Mol. Cell. Biol.* **16** 2802–13
[54] Lozzio C and Lozzio B 1975 Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome *Blood* **45** 321–34
[55] Ohlsson R, Renkawitz R and Lobanenkov V 2001 CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease *Trends Genet.* **17** 520–7

[56] Phillips J E and Corces V G 2009 CTCF: master weaver of the genome *Cell* **137** 1194–211
[57] Zlatanova J and Caiafa P 2009 CTCF and its protein partners: divide and rule? *J. Cell Sci.* **122** 1275–84
[58] Tark-Dame M, Jerabek H, Manders E M M, Heermann D W and van Driel R 2014 Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling *PLoS Comput. Biol.* **10** e1003877

Supplementary material: Prediction and Comparative Analysis of CTCF Binding Sites based on a First Principle Approach

Nestor Norio Oiwa[1,2], Kunhe Li[2], Claudette Elísea Cordeiro[3], Dieter W. Heermann[2,*],

1 Department of Basic Science, Universidade Federal Fluminense, Rua Doutor Sílvio Henrique Braune 22, Centro, 28625-650 Nova Friburgo, Brazil
2 Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, D-69120 Heidelberg, Germany
3 Department of Physics, Universidade Federal Fluminense, Avenida Atlântica s/n, Gragoatá, 24210-346 Niterói, Brazil

* heermann@tphys.uni-heidelberg.de

Keywords: CTCF binding sites, Cys2His2 zinc finger, extended ladder model.

S1: Selection of Well-known CTCF binding sites for electronic nucleotide alignment

The first sequence in Table 1 is a file 2,139 bp long covering between genes DMPK and SIX5, related with Myotonic Dystrophy (DM) [1]. We can easily localize DM1 and DM2 CTCF binding sites, because they flank the repeated sequence $(CTG)_n$. The authors apply gel mobility shift assay for CTCF-biding site identification. The next file is the 1,161 bp long around the beginning of the first exon of the gene telomerase reverse transcriptase (TERT) [2]. This CTCF binding site is identified by ChIP, electrophoretic mobility shift (EMSA) and transient transfection assays We study the CTCF binding site at human retinoblastoma gene promoter [3] using a fasta 700bp long file. The existence of this particular binding site is confirmed by EMSA in HeLa. EMSA in HeLa cells are also used for the binding site confirmation in the promoter of amyloid β-protein precursor (APP) gene [4]. Here, we select a 2,149 bp nucleotide sequence for APP. The CTCF binding sites of the v-myc avian myelocytomatosis viral oncogene homolog (MYC) are identified by ChIP assay [5]. We use a fasta file with a length of 1,366bp around the CTCF binding sites a and n. The existence of CTCF sites in the breast cancer 1 (BRCA1) gene are confirmed using EMSA and ChIP [6]. We target the same region form the reference map using a sequence 2,310 bp long surround CTCF1 and CTCF2 binding sites. We also take the 2,870 bp long maternally expressed imprinted gene 3 (MEG3) between DLK1 and GTL2 genes. This is a putative CTCF binding sites, similar to H19 and Igf2 domains, validated by methylation assay [7]. In the β-globin (HBE) CTCF binding site, validated by EMSA, we consider a sequence with 1330 bp flanking the folate receptor 1 gene [8]. Finally, we have the H19/Insulin-like grown factor 2 gene (Igf2) CTCF binding site clusters for mouse and human. In the case of Mus musculus, we are using h1 to h5 cluster [9]. We select a 3,430 bp long nucleotide sequence around 3kbp upstream of H19. The h4 binding site of this cluster is particularly interesting because this putative binding site, spotted using traditional nucleotide alignment, is not confirmed experimentally. We take a fasta file with 550 bp with h1 to h7 for human. The methylation of this cluster has already studied experimentally using EMSA [10]. We are not using the original sequences in our work, but regions around the mentioned CTCF binding sites in the reference map. We apply BLAST for finding the sequences of the interest [11].

S2: Extended ladder model [12]

Since a $L \times L$ matrix with billion size L is not possible for the eigenvalue and eigenvector computation, we slipt the complete genome with L nucleotides in windows with length n=200bp. Then we compute the local density of states from the eigenvalues E_k and eigenvectors $\phi_i^k, k = 1, ..., n_e$, of the nucleotide in position i using the extended ladder model. Here n_e is the number of electrons in the double helix (dDNA) and we have $2n$ nucleotides with n base pairs.

The model consider one double DNA chain with n base pairs, totaling $2n$ nucleotides. Actually our model does not consider nucleotides, but **nucleosides**, i.e. the **nucleotide** with the phosphate group. But, we simplify the nomenclature calling nucleosides by nucleotides. The spinless free electron of the nucleotide π-orbital is described by [12, 13],

$$H = H_e + H_{eb} + H_b. \tag{1}$$

Here H_e is the electronic degree of freedom without nucleotide coupling,

$$\begin{aligned} H_e &= \sum_{i=1}^{2n} \epsilon_i C_i^\dagger C_i + (\sum_{i=1}^{n-1} t_{2i-1,2i+1} C_{2i-1}^\dagger C_{2i+1} + \sum_{i=1}^{n-1} t_{2i,2i+2} C_{2i}^\dagger C_{2i+2} \\ &+ \sum_{i=1}^{n-1} t_{2i-1,2i} C_{2i-1}^\dagger C_{2i} + \sum_{i=1}^{n-1} t_{2i-2,2i+1} C_{2i-2}^\dagger C_{2i+1}) + H.c. \end{aligned} \tag{2}$$

where $C_i^\dagger$ and C_i are the electron creation and annihilation operators at site i, ϵ_i is the on-site ionization energy, n is the number of nucleotides and t_{ij} is the electron hopping rate between nucleotides i and j. The lattice considered In Eq. 3 is the extended ladder and the electronic hopping rates in H_e are the same in the literature [12, 14, 15, 16, 17]. Moreover, H_{eb} represents the coupling between the free electron and the nucleotide displacement field,

$$H_{eb} = \alpha_v \sum_{i=1}^{2n} y_i C_i^\dagger C_i \tag{3}$$

where y_i is the displacement of the electronic cloud from the equilibrium in the nucleotide. H_{eb} controls the gap size between HOMO and LUMO and we fix $\alpha_v = 1.0$. In this way, the gap in our spectra will be in accordance with those reported in literature [14, 15, 16, 17, 18, 19]. Finally, H_b is the interaction of the electron with the nucleotide:

$$H_b = \sum_{i=1}^{2n} [D_i (e^{-a_i y_i} - 1)^2 + \frac{k_v}{2} (y_i - y_{i-1})^2], \tag{4}$$

where D_i and a_i are parameters of the Morse potential, k_v is the spring constant of the anharmonic interaction between two contiguous base-pairs. Concerning the parameters for the Morse potential, we are using those extensively suggested in the density functional literature: D_A, D_T, D_C and D_G are respectively 0.25eV, 0.44eV, 0.33eV and 0.45eV [20, 21]; a_A, a_T, a_C and a_G are correspondingly 3.0Å^{-1}, 3.0Å^{-1}, 3.0Å^{-1} and 2.5Å^{-1} [22, 23]; and $k_v = 0.0125$eV [12].

We study the electronic part H_e and H_{eb} of the Hamiltonian in Eq. 1 computing the eigenvalue E_k and eigenvectors ϕ_i^k, $i, k = 1, ..., 2n$, of the $2n \times 2n$ Hermitian matrix $H_e + H_{eb}$ [12]. Given an initial $\{y_i\}$, we diagonalize $H_e + H_{eb}$ calculating the electronic occupation in each site $< n_i >$, where $n_i = \sum_{k=1}^{n_e} |\phi_i^k|^2$ and n_e is the number of electrons in the system. This set of $< n_i >$ will be used for the y_i estimate in the Langevin equation, given by

$$< \frac{\partial H_b}{\partial y_i} + \frac{\partial H_{eb}}{\partial y_i} >= 0, \tag{5}$$

where $< ... >$ is the average over the free electrons in the system. We update the values of $\{y_i\}$, using fourth-order Runger-Kutta method in the Langevin equation. The new $\{y_i\}$ set is inserted again in the matrix $H_e + H_{eb}$. We repeat the iteration until we achieve the minimum local adiabatic electronic and structural configuration. Since we wish to analyze massive amount of data, we rewrite the code in R used in [12] to C++, increasing the performance over the original program by factor of a thousand. The iteration method for solving Eqs. $H_e + H_{eb}$ and the self consistent Eq. 5 have already been described in [12, 13].

Using the results for SP1 and EGR1 in our previous work [12], we define electrons with a maximum 8.02eV of the energy as bottom of the molecular orbital. In this work we call them ground states in order to simplify their understand in the context of the paper, since they include the ground states. We call lowest unoccupied molecular orbital (LUMO) those electrons with $9.1 \leq E_k \leq 9.4$eV, and highest occupied molecular orbital (HOMO) are electrons with $8.52 \leq E_k \leq 8.60$eV. We show a typical result for the H19

mouse CTCF 5 in Fig. 1. The local density of states (LDOS) of the ground states is in black lines in Fig. 1(a), and HOMO are in orange and the LUMO electrons are in red. Once we estimate the shape of the electronic cloud along the DNA chain, the nucleotides with at least 10% of probability in finding ground state, HOMO or LUMO electrons are marked respectively in yellow, orange or red bordered boxes (c.f. Fig. 1(b)). Assuming that the valence band is completed filled and the conduction band is empty, $n_e = n$, we usually have 100% of probability n_i in finding electron in cytosine and thymine (pyrimidines), Fig. 1(c), as we reported in our previous article [12]. Finally, we may distinguish the different nucleotides too [12], because guanine and cytosine have a displacement field y_i around -0.11Å, while adenine and thymine have -0.12Å in Fig. 1(d). The displacement field y_i is the rearrangement of the π-orbital of nucleotide i in function of electron-base interaction [12, 13].

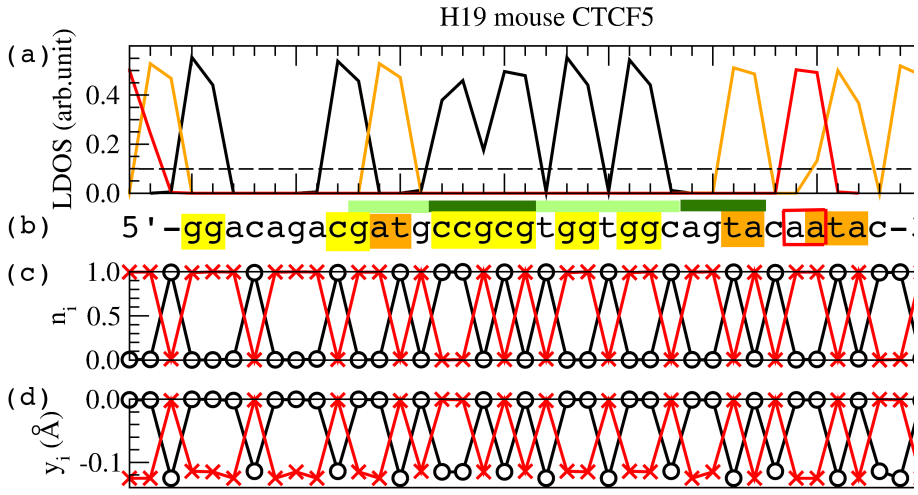


Figure 1: (a) is the ground state (black), HOMO (orange) and LUMO (red) of the local electronic density of states (LDOS) for H19 Mouse CTCF5. The quota of 10% used in (b) is in dashed line. (b) is nucleotide sequence of the CTCF5-DNA binding site in reverse complementary strand. Nucleotides with at least 10% of probability in finding one ground state, HOMO and LUMO electrons are indicated respectively in yellow, orange and red bordered boxes. We have in light and dark green the four modulus in [27]. (c) is the probability for finding one electron in the direct strand (black) and the complementary strand (red), when the valence band is completely filled, $n_e = n$. (d) is the field displacements y_i in the Morse potential with $n_e = n$ for the direct strand (black) and for the complementary strand (red).

S3: Selected Genomes

We apply the method for the 24 *Homo sapiens* chromosomes (GRCh38/hg38) [24]. Although the human genome was drafted in 2001 [25], the numerous gaps remains due to repetitive domains. The assemble hg38 still has 303 contiguous sequences (contigs), instead of 24 assembled molecules. The statistics of the fragmentation of human genome is N50=56,413,054bp and L50=19, where contigs with length N50 or longer include half of the bases of the molecular chromosomal assembly and L50 is the number of contigs that contains half of base pairs [24]. Since most of 303 contigs are small and restricted in particular regions, we consider only those bigger than 1 million of base pairs (1 Mbp), diminishing the amount in 96. The contigs have also small gaps with few base pairs of length, filled with N or another letter. Although the statistics of N50 and L50 is provided by [24], the real genomic fragmentation must be checked before, since these small gaps are frequently neglected in L50 and N50. We admit a maximum of 10 small gaps per 1Mbp and the sum of the small gaps should be smaller than 1kbp as the acceptable contiguous sequence. Despite these exclusion criteria, we still cover around 91.8% of the 3,088,269,837 bp long complete genome (column L in Tab. 1). Our genome length account is smaller than [24], because we consider the assembly molecule,

excluding unlocalized scaffolds.

The 21 chromosomes of *Mus musculus* (mouse, build GRCm38.p6) are also studied. This genome is 2,725,521,371 long and the contigs cover 96.1%. This genome has N50=32,273,079bp and L50=26 [24]. We reduce the 353 contigs to 159 using the same criteria for human contigs.

Although GenBank holds genetic information of thousands of species [24], most of reference genomes are still very fragmented as we will discuss later. However, in 2016 pig and goat became available with acceptable N50 and L50 statistics, *i.e.* N50 bigger than 1Mbp and L50 smaller than 100. Beyond this values, the genome is too fragmented and not practical.

The 19 chromosomes of *Sus scrofa* (pig, breed Duroc, build Sscrofa11.1) have N50= 48,231,277bp and L50=15 [24]. This is a 2,435,262,063bp genome with 98.1% of coverage. We do not study the chromosome Y, because it is too fragmented.

The 29 chromosomes of *Capra hircus* (goat, build ARS1, breed San Clemente, N50= 26,244,591bp, L50=32) are 2,466,191,353bp long and cover 84.3% of genome. Our length is shorter than those reported by [24], because we do not consider the chromosome X due to its excessive fragmentation and there is not data about chromosome Y in ARS1.

We do not restrict our CTCF analysis to mammals. [26] reports CTCF in insects too. We consider the 6 chromosomes of *Drosophila melanogaster* (fruit fly, Release 6, N50=19,478,218bp and L50=3) [24]. We exclude the chromosome Y, since it is divided in too many segments. The 133,880,608bp long genome has 96.0% of covering. In the case of the 3 chromosomes of *Aedes aegypti* (build AaegL5.0, N50=11,758,062bp and L50=30), the genome is 1,195,030,408 bp long, covering entire genome [24]. This is the mosquito responsible for the transmission of yellow fever, denge, zika and chinguya. Again the problems of fragmentation of the genomes do not allow us to advance beyond the mentioned insect genomes.

At the end, we apply the extended ladder model for some negative tests.

Caenohabditis elegans is a worm with 6 chromosomes that lost its CTCF gene along the evolution [26, 27]. Since there are no gaps in the sequence, the evaluation of N50 and L50 is meaningless for this genome. We use the build WS262, a 100,272,607bp long assembly [24].

We study the genome of *Plasmodium falciparum*, build ASM276v2. This is the protozoan with 14 chromosomes, which causes malaria. The 23,264,338 long genome has not gene for CTCF.

We analyze the genome of *Arabidopsis thaliana* (buid TAIR10, N50=11,194,537bp and L50=5) [24]. The 5 chromosomes are 119,146,138bp long and have 97.4% of coverage. Although [24] announce *A. thaliana* as complete, there are many gaps filled by Y (pyrimidine) or other letters. Since plants have no CTCF, we do not expect them in this genome.

After working with the many GenBank files enlisted above, we conclude that only genomes with N50 bigger than 1Mbp and L50 smaller than 100 are viable for CTCFbs analysis proposed in this paper.

S4: Example of $P_{\mathbf{core}}$ and $P_{\mathbf{flank}}$ estimate

As an illustration of the pattern identification method, consider 5'-aa cc gg ccg cg agg ttg cag tg ca-3'.

The subsequence 5'-agg tgg cag-3' belongs to the core zinc fingers $\{ZF5, ZF4, ZF3\}$. In the case of first triplet agg, we have 9 nucleotide sequences for the motif agg in the column ZF5 along the 23 selected CTCFbs in Fig. 1(c). We mark this motif as agg(9) in Fig. 1(f). Then, we associate a probability of $P(S_{\mathrm{ZF5}}) = \frac{9}{23}$, when the sequence agg appear along the genome. Same procedure is made for $P(S_{\mathrm{ZF3}}) = \frac{6}{23}$ and $P(S_{\mathrm{ZF4}}) = \frac{12}{23}$, resulting $P_{\mathrm{core}} = 37,62\%$. This probability indicates a valid CTCFbs, because $P_{\mathrm{core}} \geq 9.0\%$.

The subsequences 5'-aa cc gg ccg cg-3' and 5'-tg ca-3' are the flanking sequences. They are associated with $\{ZF9, ZF8a, ZF8b, ZF7, ZF6, ZF2a, ZF2b\}$. When we consider only the motifs associated, we have $P(S_{\mathrm{ZF9}}) = \frac{6}{23}$, $P(S_{\mathrm{ZF8a}}) = \frac{3}{23}$, $P(S_{\mathrm{ZF8b}}) = \frac{8}{23}$, ... , $P(S_{\mathrm{ZF2b}}) = \frac{7}{23}$, resulting in $[\prod_k P(S_k)]^{1/7} = 33,19\%$ in P_{flank}. We also compute $P(a_{\text{-}11}) = \frac{11}{23}$, $P(a_{\text{-}10}) = \frac{9}{23}$, $P(c_{\text{-}9}) = \frac{7}{23}$, ... , $P(g_{\text{-}1}) = \frac{13}{23}$, $P(t_{10}) = \frac{13}{23}$, $P(g_{11}) = \frac{14}{23}$, $P(c_{12}) = \frac{11}{23}$, $P(a_{13}) = \frac{11}{23}$ and calculate the geometric average of the nucleotide occurrence $[\prod_i P(S_i)]^{1/15} = 52.72\%$. So, $P_{\mathrm{flank}} = 42,96\%$, which is a putative CTCFbs, since it is bigger than 6.5%.

Since both P_{core} and P_{flank} are valid CTCFbs, the sequence 5'-aa cc gg ccg cg agg ttg cag tg ca-3' is a good candidate for the CTCFbs.

S5: K562 ChIP-seq data

We verify our electronic pattern using ChIP-seq data of the K562 cells, deposited at The Encyclopedia of DNA Elements (ENCODE). K562 is an immortal cell strain that come from a 53 year woman [28] and ENCODE is a databank seeking the integration of the many biological functions along the genomes [29, 30]. We use the following K562 files: ENCFF002CEL, ENCFF002CLS, ENCFF002CLT, ENCFF002CWL and ENCFF002DDJ with respectively 51,992, 45,603, 11,533, 54,387 and 43,247 CTCFbs each one. Since they are GRCCh37 build (hg19) and we consider GCCh38 assembly coverage (hg38), we apply the NCBI Remapping Service available in [31], converting hg19 to hg38 assemble. Only the ubiquitous binding sites in ChIP-seq data are used, because the ChIP-seq technology is not mature with possible false sites as we discuss along the paper. We localize 61,254 binding sites for K562 cells, of which 8,786 are ubiquitous. Since we are using updated ENCODE files, these values are different from [32], where they found 67,986 CTCFbs with 19,036 ubiquitous. 5,817 sites of the ChIP-seq data are in the negative G-bands, representing 66.2% of the total. The bands with the 25% and 50% of Giemsa stain responses have respectively 1122 and 987, resulting in 12.8% and 11.2% of the experimental data. The darker bands with quota 75% and 100% have 547 (6.2%) and 298 ChIP-seq binding sites in K562 cells (3.4%), respectively. And we report 15 binding sites (0,2%) in the heterochromatic domains.

S6: NucMap

In the usual nucleosome positioning method, the distribution profile of nucleosome positioning come from micrococal nuclease digestion with high-throughput sequencing data (MNase-seq) [33]. After denoising, inflection points are detected in this profile, using Laplacian of Gaussian Convolution. Then the nucleosome positions are estimated from the region delimited by theses inflection points as maximums or minimums. The improved nucleosome-positioning algorithm (iNPS) increase the number of detected nucleosomes, considering derivatives of Gaussian convolution too [34].

References

[1] Filippova GN, Thienes CP, Penn BH, Cho DH, Hu YJ, Moore JM, Klesert TR, Lobanenkov VV and Tapscott SJ 2001 CTCF-biniding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus *Nature Genetics* **28** 335-343

[2] Renaud S, Loukinov D, Abdullaev Z, Guilleret I, Bosman FT, Lobanenkov V and Benhattar J 2007 Dual role of DNA methylation inside and outside of CTCF-binding regions in the transcriptional regulation of the telomerase hTERT gene *Nucleic Acids Res.* **35** 1245-1256

[3] Rosa-Velázquez IA, Rincón-Arano H and Benítez-Bribiesca L 2007 Epigenetic Regulation of the Human Retinoblastoma Tumor Supressor Gene Promoter by CTCF *Cancer Res.* **67** 2577-2585

[4] Vostrov AA and Quischke WW 1997 The Zinc Finger Protein CTCF binds to the APBβ Domain of the Amyloid β-Protein Precursor Promoter *J. Bio. Chemistry* **272** 33353-33359

[5] Gombert WM and Krumm A 2009 Targeted Deletion of Multip-le CTCF-Binding Elements in the Human C-MYC Gene Revels a Requirement for CTCF in C-MYC Expression *PLoS One* **4** e109

[6] Butcher DT, Mancini-DiNardo DN, Archer TK and Rodenhiser DI 2004 DNA binding sites for putative methylation boundaries in the unmethylated region of the BRCA1 promoter *Int. J. Cancer* **111** 669-678

[7] Wylie AA, Murphy SK, Orton TC and Jirtle RL 2000 Novel Imprinted DLK1/GTL2 Domain on Human Chromosome 14 Contains Motifs that Mimic Those Implicated in IGF2/H19 Regulation *Genome Res.* **10** 1711-1718

[8] Saitoh N, Bell AC, Recillas-Targa F, West AG, Simpson M, Pikaart M and Felsenfeld G 2000 Structural and functional conservation at the boundaries of the chicken β-globin domain *The EMBO Journal* **19** 2315-2322

[9] Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM and Tilghman SM 2000 CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus *Nature* **405** 486

[10] Bell AC and Feisenfeld G 2000 methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene *Nature* **405** 482

[11] NCBI Resource Coordinator 2017 Database Resources of the National Center for Biotechnology Information *Nucleic Acids Res.* **45** D12-D17

[12] Oiwa NN, Cordeiro CE and Heermann DW 2016 The Electronic Behavior of Zinc-Finger Protein Binding Sites in the Context of the DNA Extended Ladder Model *Frontiers in Physics* **4** 13/1-10

[13] Zhu JX, Rasmussen KO, Balatsky AV and Bishop AR 2007 Local electronic structure in the Peyrard-Bishop-Holstein model *J. Phys.: Condens. Matter* **19** 136203

[14] Senthilkumar K, Grozema FC, Guerra CF, Bickelhaupt FM, Lewis FD, Berlin YA, Ratner MA and Siebbeles LDA 2005 Absolute Rates of Hole Transfer in DNA *J. Am. Chem. Soc.* **12** 14894-14903

[15] Mehrez H and Anantram MP 2005 Interbase electronic coupling for transport through DNA *Physical Review* **B71** 115405.

[16] Sarmento RG, Albuquerque EL, Sesion Jr. PD, Fulco UL and Oliveira BPW 2009 Electronic transport in double-strand poly(dG)-poly(dC) DNA segments *Physics Letters* **A373** 1486-1491

[17] Zilly M, Ujsaghy O and Wolf DE 2010 Conductance of DNA molecules: Effects of decoherence and bonding *Physical Review* **B82** 125125

[18] Shapir E, Cohen H, Calzolari A, Cavazzoni C, Ryndyk DA, Cuniberti G, Koltlyar A, di Felipe R and Porath D 2008 Electronic structure of single DNA molecules resolved by transverse scanning tunneling spectroscopy *Nature Materials* **7** 68-74

[19] Wang H, Lewis JP and Sankey OF 2004 Band-Gap Tunneling States in DNA *Phys. Rev. Lett.* **93** 016401

[20] Richardson NA, Gu J, Wang S, Xie Y and Schaefer III HF 2004 DNA Nucleosides and Their Radical Anions: Molecular Structures and Electron Affinities *J. Am. Chem. Soc.* **126** 4404-4411

[21] Gu J, Leszczynski J and Schaefer III HF 2012 Interactions of Electrons with Bare and Hydrated Biomolecules: From Nucleic Acid Bases to DNA Segments *Chemical Reviews* **112** 5603

[22] Chen ECM and Chen ES 2007 Thermal electrons and Watson Crick AT(-) *Chemical Physics Letters* **435** 331-335

[23] Chen ES and Chen ECM 2009 The Role of spin in biological processes: O2, NO, nucleobases, nucleosides, nucleotides and Watson-Crick base pairs *Molecular Simulation* **35** 719-724

[24] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J and Sayers EW 2013 GenBank *Nucleic Acids Res.* **41** D36-42

[25] Lander ES et al. 2001 Initial sequencing and analysis of the human genome *Nature* **409** 860921

[26] Heger P, Marin B, Bartkuhn M, Schierenberg E and Wiehe T 2012 The chromatin insulator CTCF and the emergence of methazoan diversity *PNAS USA* **109** 17507-17512

[27] Ong CT and Corces VG 2014 CTCF: an architectural protein bridging genome topology and function *Nature Review Genetics* **15** 234-246

[28] Lozzio CB and Lozzio BB 1975 Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome *Blood* **45** 321334

[29] The ENCODE Project Consortium 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project *Nature* **447** 799-816

[30] Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Jan O. Korbel JO, Emanuelsson O, Zhang ZD, Weissman S and Snyder M 2007 What is a gene, post-ENCODE? History and updated definition *Genome Res.* **17** 669-681

[31] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

[32] Chen H, Tian Y, Shu W, Bo X and Wang S 2012 Comprehensive Identification and Annotation of cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome *PLoS One* **7** e41374

[33] Zhang, Y., Shin, H., Song, J.S. et al. 2008 Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq *BMC Genomics* **9** 537

[34] Chen W, Liu Y, Zhu S, Green CD, Wei G and Han J-D J 2014 Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data *Nature Communications* **5** 4909

Chapter 7

Conclusion

Generalizable methods are developed to study the organizational principles of the genomic system, construct the 3D genome architecture, and unlock the mechanical code. Our methods are based on the analysis of physical properties, patterns, structures, mechanisms, and thermodynamical statistics.

After introducing the basic knowledge in chapter 2, the essential physical properties are discussed in chapter 3. As an indispensable parameter for chromosome conformation measurements, the contact probability for different polymer chains is carefully examined. We conclude that the asymptotic behavior of contact probability for the same type of chain is preserved even with different contact definitions. In addition to the contact probability, the persistence length, a characteristic parameter for bending rigidity, is computed for different polymer chains, and its behavior is inspected in both homogeneous and heterogeneous cases. Our results show that the existence of heterogeneity systematically decreases the persistence length, which demands an investigation of patterns for heterogeneous complex polymers.

We detect consistent patterns with experimental data of chromosome conformations and present them in chapter 4. By applying hierarchical clustering, a machine learning method, to the auto-correlation function of nucleosome positioning data, genome-wide clustering of chromatin regions is achieved. The clustering results display distinctive gene expression patterns corresponding to different nucleosome interactions and different gene densities. At the center stage of the procedure is a coarse-graining approach. It is observed that in the original length scale, the noise of the signals is overwhelming, and the pattern emerges only in the coarse-grained

scale. The method is tested on the *Candida albicans* genome, but it can be generalized to others. The result insists on a classification of nucleosome organization with more than two states, which expands the possibilities of future genomic research.

Having succeeded in utilizing machine learning to provide information on the organization pattern, we further examine the formation mechanism of nucleosome organization in chapter 5. We access the mechanism by establishing a method to extract the effective potentials for each section of the genome. Based on the parameters of the effective potentials, a genome-wide classification is accomplished. Furthermore, benefiting from the effective potentials, thermodynamic compressibilities can be computed for the whole genome. The genome-wide compressibility map serves as a quantitative characterization describing the mechanism of chromosome organization. Specifically, it is a quantitative parameter that measures the fluctuation and regularity of nucleosome organization in a region. By representing chromosome dynamics, both the compressibilities and the effective potentials facilitate further calculation of gene activities.

CCCTC transcription factor (CTCF), a primary factor reported to have an evident impact on the chromosomal structure, is inspected in chapter 6. We calculate the CTCF binding sites through a first principle approach. By examining the patterns of CTCF binding sites, cluster-like structures on a large scale are found, and a power law for two consecutive sites is noticed. Besides, the density curve of nucleosome positioning near the CTCF binding site is displayed, and the accurate averaged locations of individual nucleosomes in the vicinity are measured.