THE UNIVERSITY
of EDINBURGH

Computer simulations of chromatin structures at nucleosome resolution

Oliver S. Wiese

Doctor of Philosophy
The University of Edinburgh
August 2019

Abstract

In this thesis, I present the results from my research into the properties and organisation of chromatin structures at a nucleosome resolution. Nucleosomes, a secondary structure of DNA, are essential to the compaction and protection of DNA. However, they also play a role in the regulation of the expression of genes through changes in the 3D conformation of the chromatin fibre.

The initial work, described in chapter 3, was carried out by looking at the three dimensional conformation of the chromatin structure in Saccharomyces cerevisiae (brewers yeast). Data from a recently developed technique called "Micro-C" (published by Hsieh et. al. [1]) is used to build a contact map, detailing the interactions between nucleosomes in 3D. This raw contact data is translated to a nucleosome resolution by pairing it with nucleosome occupancy data (published by Dang et. al.[2]) to produce a nucleosome resolution contact map. A finding of the Micro-C experiments were small chromosomally interacting domains not previously observed in yeast. These "micro-domains" are at a much smaller length scale than previously observed domains in eukaryotes, typically only containing a few yeast genes per micro-domain.

The nucleosome occupancy data used to generate the nucleosome resolution maps can also be used to feed a simple "beads-on-a-string" computer simulation model discussed in chapter 4. The simulation model can be used to generate chromatin conformations. The output from the simulations can then be compared to the experimental data allowing us to deduce that just the spacing of the nucleosomes along the DNA has a significant effect on the position of domains in yeast chromosomes. The simulation output can also replicate domain boundaries to a high degree of accuracy compared when to the experimental data. Surprisingly a more detailed model does not improve the performance of feature replication.

One of the primary factors driving the formation of these micro-domains seems

to be the highly irregular nucleosome spacing found in yeast seldom discussed in the literature. When compared to the average nucleosome spacing in yeast, micro-domain boundaries have a significantly larger spacing.

Finally, in chapter 5 the findings from yeast were then taken and the same model was applied to data for the human genome in order to make predictions about the chromatin structure. The preliminary and speculative results suggest that micro-domains are also found in humans at a sub gene level and boundaries of these micro-domains are again preferentially found at long linkers.

Lay summary

In this thesis I have studied the three-dimensional organisation of chromatin, first in budding yeast *Saccharomyces cerevisiae* and then in humans. I did this by combining a bioinformatic analysis of experimental data with molecular dynamics simulations.

Chromatin is a compacted form of DNA which is rarely found in its pure state. Instead it forms complex structures with proteins found within a cell nucleus. One such structure is the nucleosome, a combination of a length of DNA wrapped around a protein core. Recent Micro-C experiments gave insights into the structure of chromatin by measuring the contact probability of different nucleosomes along the chromatin fibre. In general the data conforms with previous findings of experiments in eukaryotes, however, a novel feature is revealed. Namely, small "chromosomal interaction domains", which are much smaller than previously observed domains and which I refer to as micro-domains. Any nucleosomes within a micro-domain are much more likely to interact with nucleosomes within the same micro-domain as opposed to others.

Taking an analysis of nucleosome positioning and chromatin structure data in yeast, I created a simple computational model for chromatin, comprised of nucleosomes and linker DNA. The model fibre is generated purely from the positions of nucleosomes found experimentally and contact maps generated from the simulation are in good agreement with experimental findings. Quantitative comparison can be done by calling the locations of micro-domain "boundaries" and comparing the results between simulation and experiment. The model developed had sufficient detail to correctly determine the positions of 84% of these domain boundaries.

The data used as input to the simulations only consisted of the most-likely nucleosome positions as determined from MNase-seq data. The initial simulations

carried out with yeast studied the effects of irregular nucleosome positioning on the three-dimensional conformation of a chromatin fibre. An expected finding was that fibres with uniform nucleosome spacing in fact do not produce any domains. However, irregular spacing of nucleosomes does produce interaction domains in contact maps. The implication here is then, that the formation of micro-domain patterns is, at least in part, down to the positioning of the nucleosomes. Compared to regular spacing of nucleosomes on a chromatin fibre, the irregular spacing leads to an overall reduction in the size of the polymer. The local compaction of a region of chromatin is closely linked to the number of nucleosome found within that region. Compared to a region of linker DNA with no nucleosomes, the size is reduced with a low number of nucleosomes, but increased with a higher number of nucleosomes.

Yeast is a unicellular organism, but is often considered a model organism for higher eukaryotes. As such many of the findings from experiments and simulations in yeast are known to transfer to higher eukaryotes, but not without caveats. The results of the nucleosome interaction model for the human DNA follow this principle as it was designed as a preliminary foray into working with human nucleosome positioning data in order to draw some speculative conclusions. In fact, it proved surprisingly successful in recreating results gained from the work in yeast. The chromatin fibre model with irregularly spaced nucleosomes can be successfully used to predict elements of the structure of human chromatin. By using nucleosome positioning data alone, it shows that micro-domains are likely to exist at a nucleosome resolution level within humans. In fact new research using Micro-C experiments with human DNA confirm the predictions of my model.

Acknowledgements

I would like to thank my supervisors Davide and Chris for all the help and guidance they have given me over the past four years.

I'd like to thank my wife Laura for the amazing support she has given me in my studies and encouraging me through tough time. And also the rest of my family for believing in me and supporting me.

SDG

Declaration

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

Parts of this work have been published in:
O. Wiese, D. Marenduzzo, and C. A. Brackley, "Nucleosome positions alone can be used to predict domains in yeast chromosomes," *Proceedings of the National Academy of Sciences*, 2019 .

(*Oliver S. Wiese, August 2019*)

Contents

Abstract i

Lay summary iii

Acknowledgements v

Declaration vi

Contents vii

List of Figures xi

List of Tables xv

1 Introduction 1

1.1 Deoxyribonucleic acid and DNA compaction 1

1.2 The nucleosome 5

1.3 Further levels of compaction 7

1.4 Nucleosome occupancy and positioning 8

1.5 "3C" methods and the 3D organisation of the nucleus 12

1.5.1 Microscopy 12

1.5.2 3C - one versus one 13

1.5.3 4C - one versus many 14

1.5.4 HiC - all versus all 15
1.5.5 Compartments and domains 15
1.6 Saccharomyces cerevisiae 16
1.7 Polymer physics 18
1.7.1 Worm-like chain and persistence length 19
1.7.2 Radius of gyration 20
1.8 Aims of the thesis 22

2 Methods 24

2.1 Next generation sequencing (NGS) 24
2.2 Chromatin immunoprecipitation (ChIP) 26
2.3 MNase-seq 27
2.4 Micro-C and Micro-C XL 28
2.4.1 Micro-C and Micro-C XL process 30
2.5 Molecular simulations 31
2.5.1 Langevin equations 32
2.5.2 A simple DNA model 33

3 Data analysis: nucleosome positions and interactions in yeast 35

3.1 Nucleosome positions from MNase-seq data 35
3.1.1 Nucleosome finding algorithms 36
3.1.2 NucPosSimulator 37
3.1.3 Nucleosome positions 39
3.2 Linker lengths 40

3.3 Generating contact maps from Micro-C data 43
3.3.1 Binned contact map generation 43
3.3.2 Nucleosome resolution contact maps 44
3.4 Micro-C vs Micro-C XL 47
3.5 Gene data and Pol II 49
3.6 Nucleosome interactions around genes 50
3.7 Boundary finding 53
3.7.1 Comparing experiment and simulation 56
3.8 Representative regions 58

4 Computer simulations of yeast chromatin 59

4.1 Nucleosome model 61
4.1.1 Beads on a string 61
4.1.2 Simulation parameters 62
4.1.3 Simulation maps 64
4.1.4 Micro-domain boundaries 67
4.1.5 Radius of gyration 75
4.2 Simulation version 2 – Disk-shaped cylinders on a constrained string 81
4.2.1 Results from the more detailed model 83
4.2.2 Simulation version 3 – Complex fibre 86

5 Modelling nucleosome positioning in the human genome 89

5.1 Regions of interest and the genes within 90

5.2 Nucleosome positioning 91

5.2.1 Nucleosome repeat lengths and NucPosSimulator energy variation 94

5.3 Nucleosome position variation after TNFα treatment 95

5.4 Simulation contact maps 100

5.4.1 Contact maps 101

5.4.2 Changes in the gene body 111

6 Conclusion 116

Bibliography 120

List of Figures

(1.1) The fundamental structure and building blocks of the DNA double helix. 2

(1.2) The basic structure of DNA chromatin. 4

(1.3) The structure of a nucleosome. 6

(1.4) The crystal structure of a nucleosome. 6

(1.5) Electron micrographs of chromatin. 7

(1.6) The distribution of nucleosome positions around a TSS. 10

(1.7) The yeast cells. 18

(1.8) Schematic of the structure of a yeast cell. 18

(2.1) An example of a conventional HiC map. 27

(2.2) Simplified summary of the Micro-C process. 28

(2.3) A nucleosome resolution contact map. 29

(3.1) Nucleosome coverage map. 36

(3.2) NucPosSimulator process summary. 38

(3.3) Detail section from nucleosome coverage map. 39

(3.4) Mnase-seq data and nucleosome positions. 40

(3.5) Linker length distribution. 41

(3.6) Linker length distribution in gene-related regions. 42

(3.7) Linker length distribution for different gene types. 42

(3.8) Binned contact map comparison 44

(3.9) Upper triangular part of a symmetric contact map. 45

(3.10) Comparison of a Micro-C vs Micro-C XL interaction map. 46

(3.11) Combined figure of Micro-C and Micro-C XL interaction map. . 47

(3.12) The number of interaction reads between nucleosomes scales with their genomic separation. 48

(3.13) Plot of Pol II distribution for all genes of >1000bp length. 50

(3.14) Nucleosome occupancy and interactions around gene TSS comparison. 51

(3.15) Directionality index method. 54

(3.16) Sliding box method. 56

(3.17) Boundary comparison. 57

(4.1) Schematic representation of the simple nucleosome model. 61

(4.2) Snapshot render of the simulated model fibre. 61

(4.3) Nucleosome resolution contact map for region 9. 66

(4.4) Nucleosome resolution contact map comparison for region 3. . . . 67

(4.5) Boundary comparison. 68

(4.6) Venn diagram showing the boundary finding results of all 8 regions. 68

(4.7) Comparison contact maps of the simulation data against Micro-C data for regions 0, 1, 3 and 9. 69

(4.8) Comparison contact maps of the simulation data against Micro-C data for regions 17, 18, 19 and 24. 70

(4.9) Comparison of the insulation score between the Micro-C data and the simulation data. 71

(4.10) Plot showing the boundary signal linker lengths. 72

(4.11) Contact map comparison between Micro-C XL data and simulated long range cross-linker data. 73

(4.12) All contact maps created using Simulated XL data and Micro-C XL data. 74

(4.13) The radius of gyration as a measure of equilibrium for the simulation. 75

(4.14) Different measures for the geometric shape of a fibre. 77

(4.15) Radius of gyration found by sliding window. 78

(4.16) Render comparison of realistic fibre with artificial fibre. 78

(4.17) Plot of the radius of gyration as a function of the polymer length. 79

(4.18) Radius of gyration of a sliding window for realistic and artificial fibre. 80

(4.19) Schematic showing the more detailed model 81

(4.20) Contact map comparison between the more detailed simulation model and the Micro-C data 83

(4.21) Venn diagram summarising the boundaries found between the Micro-C data and the more detailed simulation model. 83

(4.22) All eight regions in contact map comparisons of Micro-C and more detailed simulation data. 85

(4.23) Contact map comparison of the more detailed model with inclusion of selectively switched off angles at acytelation marks. . 87

(5.1) MNase-seq data for the t=0min data set of the three regions of interest. 92

(5.2) MNase-seq data for the t=30min data set of the three regions of interest. 93

(5.3) Variation of the binding energy in NucPosSimulator 94

(5.4) Nucleosome intersection for *EDN1* 96

(5.5) Nucleosome intersection for *SAMD4A* TSS 98

(5.6) Nucleosome intersection for *SAMD4A* 98

(5.7) Nucleosome intersection for heterochromatin region 100

(5.8) Contact map for *EDN1* at different cut-off 102

(5.9) Contact map comparison between human and yeast 103

(5.10) Linker length probability for all regions and at boundaries 104

(5.11) Comparison of distance map with cut-off interaction map for *SAMD4A* 105

(5.12) Comparison of distance map with cut-off interaction map for HECHRO 106

(5.13) Comparison for *EDN1* at -10 +40 around TSS 107
(5.14) Comparison for *SAMD4A* at -10 +40 around TSS 108
(5.15) Comparison for HECHRO at -10 +40 around TSS 109
(5.16) Changes to nucleosome positions at TSS of *EDN1* 110
(5.17) Changes to nucleosome positions at TSS of *SAMD4A* 111
(5.18) Contact maps showing potential changes within gene body . . . 112
(5.19) Renders of regions at t=0min. 114
(5.20) Renders of regions at t=30min. 115

List of Tables

(3.1) Comparison of mapped interactions between the Micro-C and Micro-C XL. 45

(3.2) Regions used to represent the entire genome. 58

(4.1) Regions used for simulations and contact map creations 60

(5.1) Gene regions used for simulations and contact map creations. . . 90

(5.2) Nucleosome intersection for *EDN1* 96

(5.3) Nucleosome intersection for *SAMD4A* 97

(5.4) Nucleosome intersection for heterochromatin region 99

(5.5) Simulation regions 101

Chapter 1

Introduction

In this chapter I will give a brief introduction as well as some essential background information concerning DNA and nucleosomes. I will discuss a number of experimental methods that form the basis of much of modern bioinformatics, before giving a bit more detail on the model organism Saccharomyces cerevisiae, or brewers yeast, used for the main part of my work. Finally I'll talk a bit about polymer physics and simple DNA models, before giving a brief overview of the motivation and aims of this thesis as a whole.

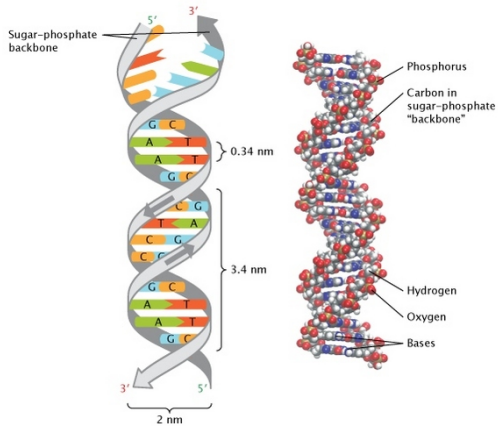
The main focus of this thesis is to explore a simple computational model for chromatin at a nucleosome resolution. The simulations for the model are generated from nucleosome position information that can be found from MNase-seq data. To measure the accuracy of the model, the simulation output can be compared to experimental nucleosome interaction data available from Micro-C experiments. Finally I explore in brief if the simple model can be extended to larger genomes such as the human genome.

1.1 Deoxyribonucleic acid and DNA compaction

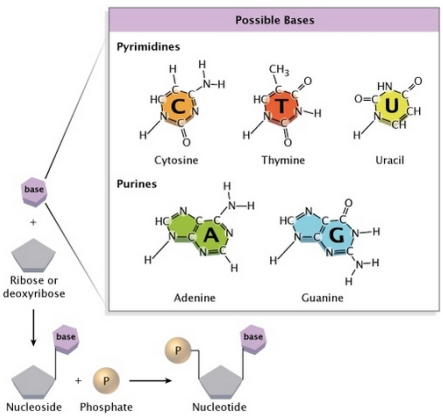
The core molecule of life is deoxyribonucleic acid or as it is more commonly known, DNA. It is one of the most complex macromolecules in the biological world despite being composed out of a simple and elegant double helix structure. This deceptive simplicity stems from its repeating and well defined molecular structure, whereas its complexity stems from all the functionally relevant interactions which the

DNA can have with surrounding proteins and biomolecules.

DNA has been continuously and extensively researched since it was first isolated by Friedrich Miescher in 1869. Its molecular structure was first identified by Francis Crick and James Watson in 1953, by using data from Rosalind Franklin, to be the now well-known and iconic double helix. This double helix is made up of two DNA chains that coil around each other, each chain being made up of a sugar phosphate backbone connecting nucleotide bases together. Each DNA strand is referred to as a polynucleotide as they are a construction of nucleotide monomers [4]. Each nucleotide has three main components: a nucleobase, a deoxyribose sugar and a phosphate group. The sugar and phosphate group form the backbone of the entire structure, using covalent bonds between a nucleotides sugar group and the next nucleotides phosphate group. The bonds between the nucleotides are formed by the phosphate groups bonding between the fifth and third carbon atoms of the sugar rings of nucleotides. This gives rise to a directionality of the strands, the five prime (5') and three prime (3') ends. The prime ending of the strand is often used to distinguish the strands. The strands encode the same information and are connected through the bases.



(a) Schematical and 3D model structure of the DNA double helix.



(b) basic chemical structure of the nucleotides making up the DNA.

Fig. 1.1 The fundamental structure and building blocks of the DNA double helix. Figures taken from Ref.[5] used under CC

Each of the four nucleotide bases pairs (bp) with its counterpart via hydrogen bonds, cytosine with guanine and adenine with thymine, linking the two DNA strands to form the double helix. The main function of DNA is the storing and transcription of a '*biological code*' which carries the instructions for the development, growth, and reproduction of all living organisms. This information

encoding is achieved through the sequence found in the nucleotide bases. Through transcription, the data stored in the DNA can be copied to RNA strands that then serve as code for the translation into the amino acid sequences which make up proteins. The mechanism of copying the DNA into RNA is by way of a copying protein (or enzyme) called RNA polymerase. The polymerase will attach itself to the double helix and split it apart locally. It then moves along the strand while pairing nucleotides, drawn out from the solution, with one the two strands. DNA to DNA copying on the other hand is done through DNA polymerase - an enzyme that works in pairs to synthesise two new strands from the original strand by combining deoxyribonucleotides into DNA.

A specific sequence of 'code' which can be transcribed to produce proteins and the like is called a gene. In humans, only about 3 % of the total DNA is gene coding with the remaining 97% appearing to have a generally organisational and regulatory role. In some cases this extra DNA has been referred to as 'junk DNA' [6], but the growing consensus is that it remains essential, though not well understood. Eukaryotes (animals, plants and fungi), store their DNA inside a cell nucleus (with some extra DNA in mitochondria and chloroplasts). This gives rise to the problem of cramming the long DNA double helix structure into the confined space of the nucleus, which is achieved by DNA compaction.

Since every cell within an organism contains a full copy of the entire DNA, one research area of interest is the problem of DNA compaction, the question of how it all fits into a cell. To illustrate, the entire human genome has an approximate combined length of 2 meters. These 2m of DNA have to fit into a cell with a diameter of the order of 10^{-5}m [7]. This means the DNA has to be compacted by a factor of as much as 10,000 to fit into the nucleus of the cell. It is of course very narrow (2×10^{-9} m wide), but despite this, the thread still requires serious compaction. Balance is key within the compaction, since DNA packaging can create both problems and opportunities. On the one hand, the packing of the DNA potentially limits and obstructs access to vital section of the DNA, such as gene coding regions. On the other hand, this obstruction can also be used to direct enzymes to particular sections of DNA, such as the RNA polymerase II (Pol II), which will initialise transcription of genes at the beginning of genes and not just randomly in the middle or near the end. DNA polymerase will make use of replication origins and DNA repair enzymes are focused towards damaged DNA. Although the DNA may get damaged, the packing also serves the function of protecting the DNA and enhancing its mechanical stability. These and other

significant advantages seem to outweigh the possible downsides of nucleosome packing.

The compaction proceeds through a number of stages via the combination of the DNA with various proteins to form chromatin [8]. DNA inside eukaryotic cells is rarely, if ever, present in its pure form, usually it is found in a complex of macromolecules together with proteins and RNA (ribonucleic acid) typically referred to as chromatin. The function of chromatin is, in part, to package DNA in such a way that its volume is reduced and to control gene-expression. The expression is regulated by packaging the DNA into more 'open' or more 'closed' regions and thus controlling accessibility for proteins.

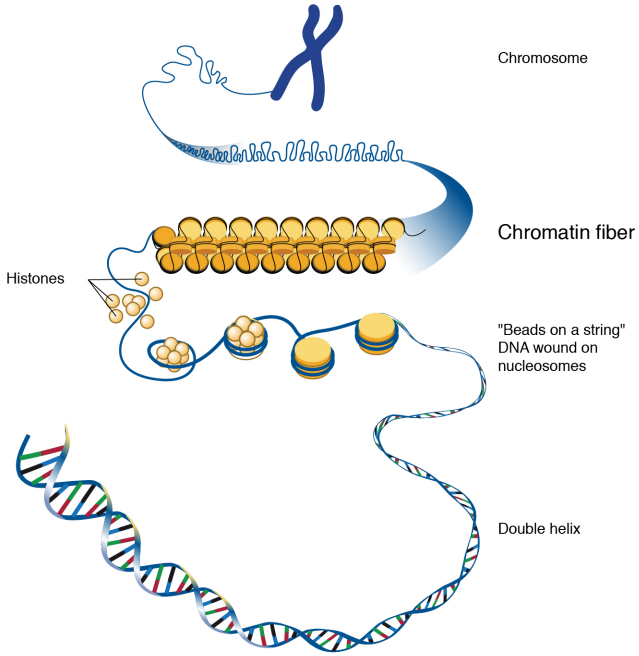


Fig. 1.2 The basic structure of DNA chromatin in its various forms of compaction from the double helix, through the nucleosome "beads-on-a-string" to chromatin fibres and ultimately full chromosome. Courtesy: National Human Genome Research Institute [1]

The packing of DNA is achieved by the formations of structures composed of DNA and proteins and can be categorised into three levels. The first level is the wrapping of the DNA double helix around histone proteins forming nucleosomes and a "beads-on-a-string" structure. The second level is the so-called 30-nm fibre which is still an active area of research and though it has been observed in some

[1] `genome.gov`

cell types (particularly in some higher eukaryotes it as been reconstituted in vitro) its presence is still much debated in yeast. The third level includes higher-level packaging of DNA and proteins into structures such as active chromosomes and metaphase chromosomes, also still an area of intensive research.

1.2 The nucleosome

In eukaryotic chromatin, the nucleosome [9] makes up the basic unit for the organisational packing of DNA, in which a histone core is wrapped with DNA to form DNA-protein structures. Each histone core contains 8 histone proteins (usually referred to as an octamer), with two of each out of four possible types (H2A, H2B, H3 and H4). The DNA has approximately 147 bp coiled around the histone octamer in a left handed toroid [10] in 1.65-1.67 turns [11]. Histones are proteins that serve essentially as a scaffolding for the DNA. They hold together by making use of electrostatic charges, where the histones are positively charged and as such attract the negatively charged DNA.

This structure is reminiscent of a spool, as the DNA wraps around the core histone octamer. The spool-like structure or squat disc-like structure is about 5.5nm in height and 11nm in diameter [12]. These spools are then connected by flexible/free sections of DNA giving rise to a "beads on a string" model in a linear arrangement of nucleosomes along the DNA polymer.

Histone proteins share a common structure with a globular core section and a flexible tail. This amino-terminal histone tail protrudes from the core past the DNA allowing for chemical alterations of the tails, such as acetylation (addition of an acetyl group CH_3CO) and methylation (addition of a methyl group CH_3). These acetylation and methylation marks are often associated with repression or activation, for example, the methylations H3K9me3 or H3K27me3 are associated with repression, whereas the acetylations H3K4me1, H3K4me3 and H3K27ac are classical activation marks. Though in some cases, such as in genes which are active or poised for activation, the nucleosomes will generally have their H2A and H3 histones replaced by their variants H2A.Z and H3.3 [13, 14]). Histone tails are considered "unstructured" portions of the histone protein with the tail regions adopting random coil conformations in free solution [15, 16]. The accessibility of the tails to enzymes allows for posttranslational modifications (histone modifications) which play an important role in epigenetic signalling.

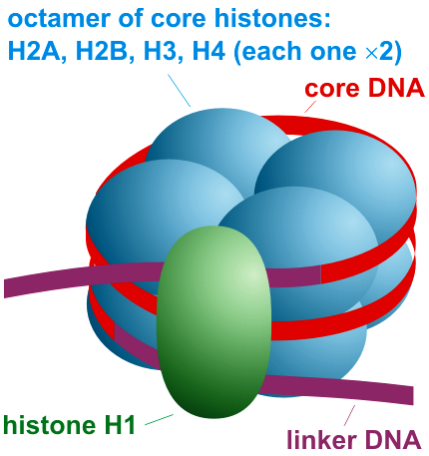


Fig. 1.3 The structure of a nucleosome with the histone octamer wrapped in DNA making up the core and linker DNA connecting different nucleosomes. The H1 histone sits at the entry/exit point of the linker DNA. taken from wikimedia.org commons

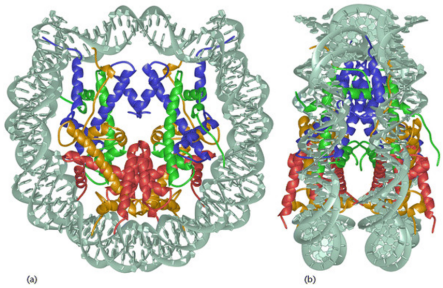


Fig. 1.4 The crystal structure of a nucleosome as produced by Ref: [10]. The DNA wraps around the histone proteins in two turns. Used with permission from Springer Nature

A further functionality is the regulation of gene expression by providing landmarks for gene regulatory proteins [17]. Histone modifications in particular are likely to control the chromatin compaction. This can be expressed to a level, that in addition to the genetic information encoded in the DNA basepairs, a further level of information is encoded in the nucleosome positioning. As such the study of nucleosome positioning is essential to understanding of genetic information stored in chromatin. The tails also play a role in inter-nucleosomal interactions when the chromatin is packaged into higher order structures and the tails appear to play a vital role in their organisation [18–20].

In higher eukaryotes an additional "gate-keeper" histone H1 binds at the entry/exit point of the DNA and acts as a linker histone which is connected to approx. 15 bp of DNA (∼50bp in metazoans) [21]. The H1 linker histones are primary component of all nucleosomes in higher eukaryotes, but is missing in yeast. Although it has the H1-like or analogue protein Hho1p, it seems to play a significantly different role and does not associate with most of the genome. Functions of the H1 histone include stabilizing the wrapping of the DNA around the nucleosome core and hence the assembly of the chromatin into higher order chromatin structures [22, 23]. Further, H1 also plays a role in the spacing of the nucleosome on the DNA strand [24, 25], gene expression regulation [26, 27] and transcription repression [28]. These linker histones also assist in chromatin

compaction and play a part in producing the 30nm fibre [7]. When talking about nucleosomes one generally refers to the combination of the core as well as the linker section, so an approximate total of 200bp.

The initial obvious advantage of nucleosomes is that they alone can produce an initial 7-fold compaction of the DNA and the DNA associated with the nucleosome is protected from nuclease digestion. A downside is that nucleosome formation contorts the DNA, bending the DNA around an octamer costs a lot of energy, so a strong affinity between the negatively charged DNA and positively charge nucleosomes is needed. The usual persistence length of DNA is about 150bp whereas in nucleosomes the DNA is bend to around 80bp/turn.[10, 29, 30]. The bending cost does depend on sequence which introduces some importance of DNA sequence to nucleosome positioning.

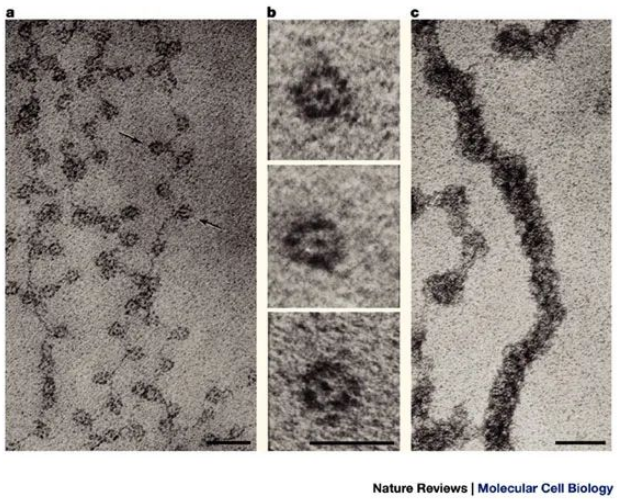


Fig. 1.5 Electron micrographs of chromatin showing the "beads-on-a-string" nucleosomes on the left, the nucleosomes themselves in the middle and a chromatin fibre to the right. Taken from Ref: [31] Used with permission from Springer Nature

1.3 Further levels of compaction

The next organisational packing level is the so-called chromatin fibre or 30nm fibre, since observations *in vitro* show nucleosome arrays as compact fibres of approx 30nm diameter. All that is currently known about the 30nm fibre is still based on either electron microscopy or theoretical models, as such there is still much debate surrounding the 30nm fibre. Although the structure of

the nucleosome core is well known, the structure of this 30nm fibre is still not completely solved despite continuous research and the fact that various models have been proposed for the structure *in situ*.

There are two main models which are proposed and explored in current research, though it has to be mentioned that the chromosome fibre has not yet been observed *in vivo* and as such its existence is still under debate. Nevertheless, it can be visualised when chromatin fragments are isolated or released from nuclei.

Most research agrees on the theory that the basic building block of the 30nm chromatin fibre is a so-called tetranucleosome formed by four nucleosome cores. The crystal structure of the tetranucleosome has been solved by Schalch et. al.[11] and supports the zigzag model, with the other common model being the solenoid model. The nucleosome-nucleosome interactions within tetranucleosomal units and between tetranucleosomal units appear to have a considerable role in the folding and formation of chromatin fibres but this is still a much debated and active are of research.

1.4 Nucleosome occupancy and positioning

There are a number of experimental approaches to measure the position of nucleosomes within the genome. Largely these consist of fragmenting the DNA to isolate nucleosomes and then sequencing the nucleosomal DNA and referencing it to known genomic features. The most common method is to use micrococcal nuclease (MNase) to cut and digest the linker DNA between the nucleosomes [32, 33] leaving fragments which consist of DNA which was wrapped around the histones [34]. The main downside is that all protected DNA (whether in a nucleosome or other protein complex) is left over and as such the sequencing (MNase-seq) can give false positives. This can of course be remedied with further steps such as treating the fragments with an antibody which recognises histone proteins. As an example, there are approximately 60,000 nucleosome in haploid yeast genomes.

Nucleosomes are found along the entire genome with a mostly uniform distribution. However, there are two significant deviations in this distribution, these are the nucleosome-free regions (NFR) or nucleosome-depleted regions (NDR). Although they appear to be the same, there are some subtle differences. A NFR is

characterised as an approximately 140bp size region that is devoid of nucleosomes, where a NDR will contain nucleosomes but generally at a much lower density than normal. The important difference between NDRs and NFRs is that the latter generally contain no nucleosomes, whereas the former can contain nucleosomes which when necessary will be removed. These regions are commonly found at the beginning and end of genes and it is now understood that misregulation of nucleosome positions can potentially lead to cancer and developmental defects [35–37]. So what brings about these NFRs and how are they regulated? This is an ongoing area of research and I will endeavour to outline some of the factors that control the nucleosome positioning.

Recent research and technological developments have allowed genome wide mapping of nucleosome positioning. This has shed some light on the questions of nucleosome positions. One big question was whether the nucleosome positions are random or whether nucleosomes are individually and specifically positioned. In the case of the former, a lack of positional cues makes the histone proteins solely a DNA packaging protein and in the case of the latter it would allow for specific physiological functions dependent on their position in the genome. There is evidence for both cases and different regions of the genome have displayed the two cases.

An important point to remember is that nucleosomes are not statically bound to specific locations on the genome, instead they are highly dynamic and can slide along the DNA [38, 39]. The interplay between nucleosome occupancy and nucleosome positioning gives rise to the concept of nucleosome dynamics [40–44]. Furthermore, they can also fully or partially disassociate from the DNA fibre [45] and are subject to post-translational modifications [46, 47]. In addition to that, the histones that make up the nucleosome core can be replaced by their sequence variants [48]. The underlying sequence of the DNA can have major influence on the positions of the nucleosomes and the location of nucleosomes can be changed by the cell using molecular machines which will position the nucleosomes. The dynamics of nucleosomes are now understood to be directly linked to genome regulation.

The accessibility of gene promoters (a region of DNA upstream of a gene where transcription begins) controls whether a gene gets transcribed or not and to this end the promoter needs to be accessible to chromatin regulators and transcription machinery proteins [40, 49]. Hence these promoters are generally characterised by nucleosome-free or nucleosome depleted regions, meaning that in the core

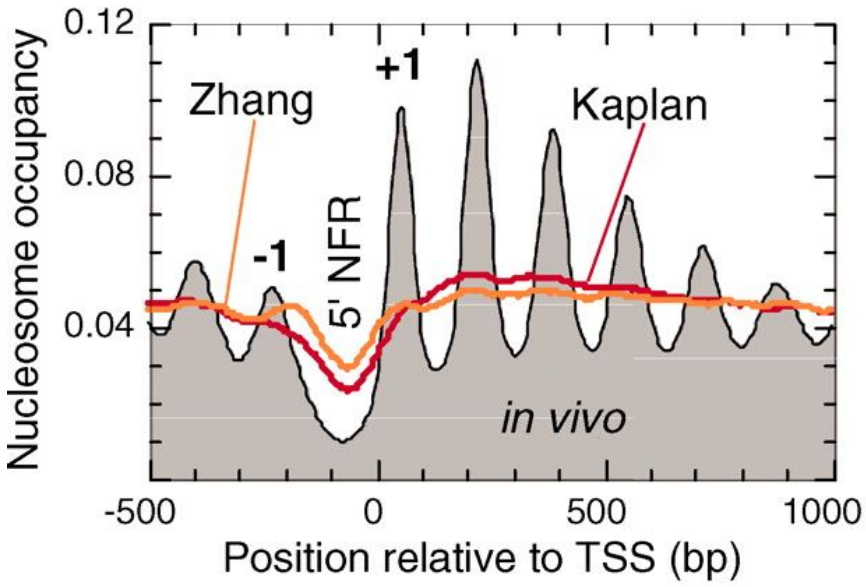


Fig. 1.6 The distribution of nucleosome positions around a TSS where the nucleosome occupancy varies to form likely ordered positions for the nucleosome up- and downstream from the TSS. A NFR is found just upstream of the TSS. Figure from Ref. [61] Reprinted with permission from AAAS

of the promoter region, none or very few nucleosomes are found. Environmental conditions can override this and cause a NFR to contain nucleosomes (for example gene repression). In this region a nucleosome could also missing.

Depletion of nucleosomes in intergenic regions where promoters are found [50–52] and activation of genes will result in a further nucleosome depletion [50, 52–56]. Further, the nucleosomes in promoter regions of highly transcribed genes are enriched with acetylated and methylated histones [57–59]. When a gene is no longer transcriptionally active, nucleosomes might bind in this region, blocking immediate access to the promoter. Studies [60] have shown that most genes will have nucleosomes organised around the beginning of the gene in very similar ways.

Typically the NFR is bordered by two specifically located nucleosomes, the downstream nucleosome (denoted the +1 nucleosome) sits a fixed distance from the TSS (transcription start site) and an upstream nucleosome (-1 nucleosome) sits at the other end of the NFR, followed by nucleosomes packaging the gene. When plotting the nucleosome density around a TSS (transcription start site), such as in Fig. 1.6, the data reveals that further downstream from the +1 nucleosome the following nucleosomes follow a canonical structure, meaning they

are well-positioned at defined intervals (in S. cerevisiae this is approximately 165bp from nucleosome centre to nucleosome centre) [62, 63]. The nucleosome centre, also know as the nucleosome dyad, is the point or rather base pair that sits at the centre of the nucleosome structure, creating a mirrored pseudo-symmetric structure. This basepair position is often given as the coordinate of the nucleosome.

The further along the gene, the more "fuzzy" the nucleosome positions get, i.e. they are less and less rigidly spaced and vary more from cell to cell [60, 64], possibly because the functional constraints of nucleosome are not required as much as near the beginning of a gene. Nucleosomes seem to adopt "canonical" positions around promoter regions. The +1 nucleosome is important for transcription of the gene as its position and structure (histone composition etc) affect RNA polymerase binding as well as transcription factors [65–68]. Nucleosome presence or absence can directly affect the function of the chromatin and will often regulate promoters. Most promoters will always contain a NFR to allow constant access, but some genes will contain a "placed" nucleosome in their NDR. This nucleosome might then be removed due to a cue such as stress signals. These NFRs can also be found in the regions of active enhancers.

Linker DNA, the DNA found between two nucleosomes, is usually short in length, but if long enough, can be considered as a NFR, although there is no limit in length for the two and an actual NFR is usually a binding site for polymerases. The lengths for linker DNA commonly quoted in the literature are measured between nucleosome midpoints and are approximately 165bp in S. cerevisiae, resulting in a 18bp linker on average [64, 65, 69] and approximately 185-200bp with 38-53bp linker in humans [70, 71] since 147bp are wrapped around the histones.

Nucleosome occupancy is a measure of the average number of nucleosomes within a genomic region. It is highly dependent on the cellular population and should therefore be understood as a probability measure of the likelihood of a nucleosome being present at a genomic location. Nucleosome occupancy directly influences chromatin functions due to the fact that the accessibility of the underlying DNA is affected by the presence or absence of a nucleosome. Nucleosome positioning on the other hand is related to the probability of a nucleosome dyad sitting at a specific genome coordinate as opposed to nearby. It is related to nucleosome phasing, a measure of a most likely position of a nucleosome at a genome coordinate. Further, even when no process requires the removal of DNA

from a nucleosome, the nucleosome may disassociate and re-associate from the DNA, locally changing nucleosome occupancy values. Some DNA sequences, specifically AT-rich sequences, are less likely to bind nucleosomes in a stable way [72, 73] passively influencing nucleosome dynamics. During the course of DNA transcription, recombination or repair, the DNA will be unwrapped from the nucleosome [61, 65, 74]. Although a nucleosome will bind again at the same site after the process is finished, it is unlikely to be the exact same nucleosome - it will likely have different histone and chemical modifications.[75] Active regions of the genome (promoters, enhancers and origins of replication) tend to have the highest turnovers of nucleosomes [76–80].

Instead of simply statically packing the chromatin, nucleosomes are a key regulatory component of the genome, dynamically adjusting and moving along the DNA stand. As such DNA regulation and nucleosome dynamics are closely interlinked.

1.5 "3C" methods and the 3D organisation of the nucleus

The packing of DNA inside the nucleus is a organisational challenge and the shape of the genome strongly relates to genome functioning. Hence understanding the shape and organisation of the genome are of vital importance. To this end a number of breakthrough technologies have been developed over the years. Here I will focus on the "3C technologies" starting with the name-giver, chromosome-conformation-capture developed by Dekker et. al. [81]. In short, it is a biochemical method to study and analyse the contact frequencies between genomic locations in a population of cells. The 3C-derived methods have allowed for a systematic and high-resolution study of the genome topology.

1.5.1 Microscopy

Before the modern "3C technologies" for genome conformation research had been developed much of the research and insight came and still comes from microscopy. Studies into the structure of the nucleus have long preceded studies into genome topology, but with the advances in technologies, the depth of

microscopy studies has increased. Microscopy has delivered numerous initial findings such as that chromosomes divide the nucleus into territories which they occupy and fill preferential radial positions within it [82–84]. This means that larger chromosomes tend to prefer the outer areas of the nucleus whereas smaller chromosomes stick to the interior. Studies also found that even within the chromosome territories a separation of gene-rich and gene-poor regions exists. This separation of active and inactive chromatin is especially interesting as it suggests there might be a link between gene activity and nuclear positioning. Modern microscopy techniques have supported these findings with showing that genes can leave their preferred regions if they change state [85, 86]. An example of that would DNA fluorescence in situ hybridisation (FISH), a technique that binds fluorescent probes to specific parts of DNA and which can then be observed using fluorescence microscopy. Although much has been found through microscopy, it is ultimately limited by throughput and resolution and the findings cannot conclusively tell if these findings are general principles of nuclear organisation.

1.5.2 3C - one versus one

This is where the introduction of the 3C method [81] has sparked a veritable arms-race in the development of genomics methods. The underlying principle of these methods is to build up a representation of the DNAs 3D organisation. The first steps in the process are more or less the same no matter the technique, with the chromatin being "frozen" in place using a fixation agent such as formaldehyde. In the next step restriction enzymes cut the DNA at specific locations (these enzymes are able to recognise specific sequences of 4bp or larger). Multiple cutter enzymes are available but once the DNA is cut it essentially creates cross-linked DNA fragments, of which two of the four ends are "sticky". This essentially means that where the enzymes cut the double helix, the cut is not clean and there might be uneven strands with overhangs of unpaired bases, making them more likely to bind to other fragments, nucleotides or otherwise. These ends can be religated to each other, essentially forming a single fragment with two sections. These sections when "frozen" were close together in 3D space, but might have been distant on the chromatin chain. This process essentially creates a 1D representation of the 3D structure. To measure the 3D conformation from this, the number of ligations between sites that don't neighbour each other, has to be measured. In the standard 3C method, this is done using semiquantitative [81] or quantitative [87, 88] PCR amplification of selected ligation junctions. In polymerase chain

reaction (PCR) a specific DNA segment can be replicated exponentially in order to generate up to millions of copies.

This can be achieved by using primers at the restriction fragment ends and building a matrix of ligation frequencies. The main drawback of this method is that the method is highly specific, i.e. the primers needed have to be specifically designed to match a certain region. This in turn means that you can only investigate regions if you know where to look. One of the primary issues with 3C data is the simple fact that sequences that are close together on the linear chromosome will always ligate more often as they are more close in 3D space, hence, to find actual long distance 3D interactions it is imperative to rely on quantitative measurements to produce adequate data. However, the further the linear distance between crosslinked sites, the less likely the interactions are quantifiable above the background noise in the data, making results less clear. However, advances in genome scale methods (microarrays and high-throughput sequencing) have helped in developing less range-biased methods and in studying the long range contacts. One of the first discoveries from this method was that the chromosome III in yeast forms a contorted ring and once adapted to work for mammalian systems, it showed that *in vivo* chromatin loops exist between genes and their regulatory DNA elements.

1.5.3 4C - one versus many

The chromosome conformation capture-on-chip (4C) technology was an advancement on the 3C method wherein the 3C method was combined with microarrays to produce a "one versus all" analysis of a genomic site with all fragments on the array [89]. An alternative to normal 4C is 4C-seq which uses next-generation sequencing instead of microarrays. The basic principle of 4C is to take the ligated 3C template and process it with a further treatement of DNA digestion and ligation to produce DNA circles. View-point-specific primers are added and inverse PCR amplifies the relevant sequences. The result can then be analysed using microarrays or NGS. Next generation sequencing (NGS) or high-throughput sequencing is a modern technique that parallelises the sequencing process and produces a high number of sequences at the same time. The primary advantage of 4C, when compared to 3C, is that for a given viewpoint all interactions genome-wide can be captured. So one no longer needs to know where to look apart from choosing the viewpoint. A major drawback of this method is its lack of resolution with local interactions not being resolved.

1.5.4 HiC - all versus all

From the "one versus one" (3C) method to the 4C "one versus all" the next step is the "many versus many" 5C (chromosome conformation capture carbon copy technology) and the "all versus all" HiC methods. The HiC method was only made possible through advances in NGS methods bringing the cost down and resolution up for chromatin analysis. The HiC method [90] is a variant on the 3C methods where the initial procedure is slightly altered and, before ligation takes place, the ends cut by the restriction enzymes are filled in with biotin-labeled nucleotides. These ends are then ligated before the DNA is purified and then sheared into pieces. By using a biotin pull-down the ligation products can be isolated for sequencing. The resulting sequences can be compared with a reference genome, identifying where in that genome the ligated fragments originated. Pairs that are found on different fragments are counted as interactions between the fragments. This can then be used to build up a matrix of ligation frequencies. Major findings from this method were that the genomes of many organisms are spearated into compartments and domains.

1.5.5 Compartments and domains

When considering the genome as a whole, it can be split into two compartments (type "A" and "B"), within which regions tend to preferentially interact or associate with each other [90]. These regions are on the multi-Mb scale and the "A"- type regions are associated with open and expression-active chromatin, whereas "B"- type regions are associated with more closed and expression-inactive chromatin. Found in the interior of the nucleus, the A compartments are gene-rich and contain histone markers for active transcription. The B compartments are the opposite, with gene-poor and histone markers for gene silencing while lying on the outer areas of the nucleus.

Within these compartments a further level of organisation can be found, the domains. In the active parts of the genome these are generally referred to as Topologically Associating Domains (TADs). The TADs, or also known as self-interacting domains, range in size from the 1-2 mb scale in eukaryotes [91] to 10s of kb in single celled organisms [92]. A key feature of these domains and the one that makes them self-interacting is that they have a higher ratio of interchromosomal contacts within itself, i.e. the interactions between components

of DNA within this domain are increased and the distances between them are decreased. Chromatin loops are also a common feature of these TADs where in higher eukaryotes some TADs contain loops between the edges of the domains. These TADs are generally formed actively by proteins. Self-interacting domains are also thought to influence regulation of gene expression where specific domains are associated with transcription activation or repression. This might be the case if a domain boundary isolates a promoter from an enhancer or conversely it might make a promoter more likely to interact with an enhancer.

1.6 Saccharomyces cerevisiae

Saccharomyces cerevisiae, also know as brewers yeast, is a budding yeast and a robust and versatile model system for eukaryotic genetics. It has been cultivated by humans for millennia in order to produce beer, bread and wine, becoming the most commonly used industrial yeast as a domesticated microorganism and sexual eukaryote. Yeasts are fungi and share a common cellular architecture and life cycle with multicellular eukaryotes (plants and animals). Being non-pathogenic and nonmotile makes them particularly suitable for propagation and manipulation in laboratories. Research on yeast has led to much of what has been learned in cell and molecular biology, as the yeast genome is a suitable analogue for higher eukaryotes. Although yeast as a single celled eukaryote is much simpler than multicellular organism, it shares many features with them, and many crucial cellular mechanisms are variations of those in higher eukaryotes.

The life cycle of the yeast cell can be summarised as follows. Similar to other sexual eukaryotes, the yeast cell alternates between haploid (single chromosomal complement) and diploid (two chromosomal complements). A “mother” cell will start the cycle by budding to produce a genetically identical daughter cell through cell division and via mitosis a copy of each chromosome gets transferred to the daughter cell. When two haploid (containing half the usual number of chromosomes) cells mate, they fuse and produce a diploid (2 sets of chromosomes) cell, which contains two of each chromosome. The diploid can either grow by budding, such as the mother cell, or undergo meiosis. During meiosis, the cell divides into four haploid cells or spores. These are initially held together in an ascus or tetrad but can either split into individual haploid clones or end up mated

with each other to produce new diploids. This allows for rapid growth of the cells especially with a generation time of 1.25-2 hours at 30 °C.

The structure of S. cerevisiae is common to all eukaryotes with membrane enclosed organelles. The most important of these organelles is the nucleus, in which the budding yeast carries its genome. The genome is split into 16 linear chromosomes, with 12 megabase pairs[2] containing over 6000 genes[3][93].

A special feature of the yeast chromatin is that it appears to maintain its Rabl conformation even when not in telophase (final stage of mitosis - where the cell splits in two). The Rabl conformation was discovered by Carl Rabl in 1885 [94]. It is a conformation of the chromosomes where they are folded over at the centromere and the telomeres are found close together (similar to holding a wet spaghetti near the middle). The centromeres themselves are held close to one side of the nucleus and the chromosomes extend out from there.

Brewers yeast was the first eukaryotic genome to be fully sequenced and released to the public in 1996. The yeast genome is very compact for a eukaryotic genome, with a small number and size of genes, but high in density. This makes it very suitable for studies as well as the fact that yeast genes have few introns and short intergenic regions. Introns are DNA sequences within a gene that are removed during RNA transcription, i.e. essentially non coding DNA within a gene.

It was a promising candidate for HiC experiments and initial experiments were able to confirm the Rabl configuration of the chromatin ([95–97]) This facilitates the analysis of gene functions, aided by a low genetic redundancy. Overall most of the yeast genome is active, giving rise to not much compartmentalisation but domains can still be found at the Mbp range. However recent studies such as the Micro-C paper (Ref. [1]) have found that smaller domains can also be observed within yeast chromatin. This investigation forms much of the foundation for the work in this thesis.

For more information about yeast please refer to Ref: [98]

[2] 12,156,677 bp
[3] 6275 genes

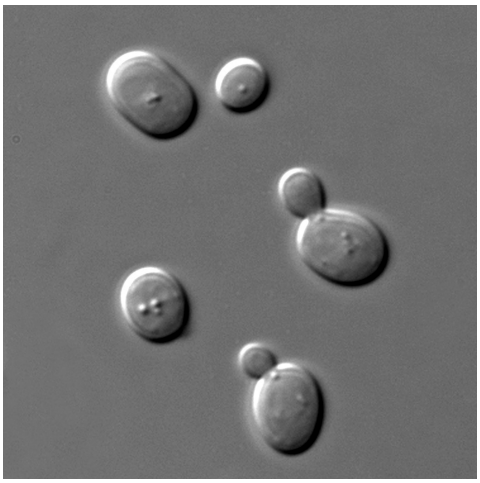


Fig. 1.7 The yeast cells are round to ovoid and approximately 5–10 μm in diameter. Image from Public Domain

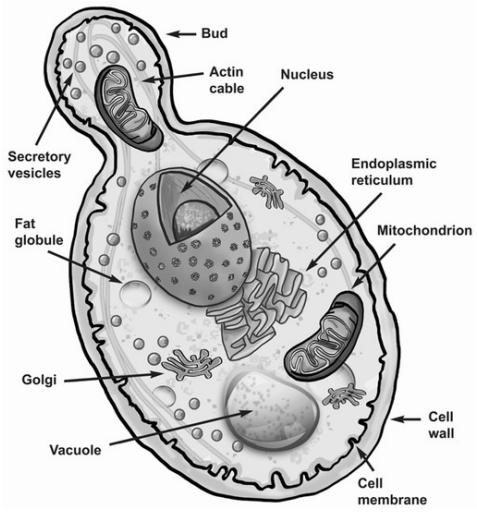


Fig. 1.8 Schematic of the structure of a yeast cell showing the key features. Figure taken from Ref: [99]

1.7 Polymer physics

Polymers or in particular biopolymers (such as DNA) are biomacromolecules and can be classified into a number of categories. The category of particular interest here is the nucleic acids. Like all polymers, biopolymers are made up of strings of monomeric units. These monomers will usually form linear chains, but they can also form other shapes, such as closed and circular chains.

Fundamentally all biopolymers exhibit a common characteristic in that they form hierarchical structures at successive length scales. The primary structure of all polymers are the monomeric units which when organised into a local molecular structure form the secondary structure. Following on from this, the polymer can also adopt a 3D conformation which gives the tertiary structure. And finally the biopolymers can move on to interact with other biopolymers to form macromolecular components. An interesting feature of biopolymers are the emergent features, features that the monomers or lower tier structures do not exhibit but in combinations can emerge. DNA is an obvious biopolymer and its basic features have been discussed above.

When it comes to modelling polymers, there are a number of approaches, the simplest of them all being the **ideal chain**[4]. In this model, the monomers are a fixed separation from one to the next, but can turn in any direction. If we have

[4]Though there are technically a number of ideal chain models.

$N+1$ monomers in the chain with each monomer's position described by a centre of mass position vector $\vec{R}_i$, then the step vector between subsequent monomers is given by $\vec{l_i} = \vec{R}_i - \vec{R}_{i-1}$. This will then simply describe a random walk with step length $l = |\vec{l_i}|$. In this model there are no interactions between the monomers and the orientation of each link is uncorrelated to another. This model can also be referred to as the random flight chain and is the simplest of the ideal chains.

There are a number of interesting properties exhibited by an ideal chain, with a particularly relevant one being the radius of gyration (elaborated in section 1.7.2). For the ideal chain it is given as:

$$\langle R_g^2 \rangle = \frac{1}{6} N l^2$$

where N is the total number of monomer links and l the step length. For a derivation of the results please refer to Ref. [100].

1.7.1 Worm-like chain and persistence length

A model which increases in complexity slightly but models DNA better is the **worm-like chain** (WLC) as it models semi-flexible polymers. In the WLC, the segments are taken to point in roughly the same direction and the joints are stiffened up. The model can be likened to a thin elastic filament which obeys Hooke's law for small deformations. In this model the step length l goes to zero by increasing the number of segments to infinity, while maintaining the total contour length at a constant $L = Nl$. An important feature which comes into play here is the persistence length L_p, a measure of the typical length scale over which orientation correlation is lost or rather it quantifies the stiffness of a polymer. The formal definition is given by:

$$\langle \vec{r_i} \cdot \vec{r}_{i+s} \rangle = e^{-\frac{s}{L_p}}$$

where s is the contour distance between two monomer units. A good informal explanation of the persistence length is that for a polymer, section shorter than the L_p, it behaves like a flexible elastic rod, while for sections longer, it behaves more like an ideal chain and random walk (coil).

The radius of gyration can again be calculated for the WLC (and again I would like to direct the reader to Ref. [100] for a full derivation), but in this case it is related to the persistence length.

$$\left\langle R_g^2 \right\rangle = \frac{1}{12} L^2 \quad , \quad L \ll L_p(\text{rod})$$
$$\left\langle R_g^2 \right\rangle = \frac{1}{3} L L_p \quad , \quad L \gg L_p(\text{coil})$$

1.7.2 Radius of gyration

The Radius of Gyration (R_g) is a good measure to check the dimensions of the polymer. It also allows for comparing the persistence length of the model polymer with literature values as the R_g is linked to the persistence length as mentioned above.

The definition of the R_g is slightly dependant on the field of interest, in polymer physics it is usually used to describe the dimension of a polymer chain. More formally it is known as the second moment of the mass distribution or in other words - as the distance from the axis of rotation to a point where the total mass of the body is supposed to be concentrated, in order that the moment of inertia about the axis remains the same.

The general mathematical description of the R_g is as follows:

$$R_g^2 = \frac{1}{N} \sum_{i=1}^{N} (\vec{r}_i - \vec{r}_{mean})^2$$

where r_{CM} is the centre of mass of the components given as:

$$\vec{r}_{CM} = \frac{1}{N} \sum_{i=1}^{N} \vec{r}_i$$

with N being the number of particles and $\vec{r}_i$ being the i$^{\text{th}}$ particles position vector.

The R_g is not the only measure which can be used to describe a polymer. Other properties include the asphericity, acylindricity and relative shape anisotropy and a simple way to calculate all four is to make use of the gyration tensor. By summing the principal moments of the gyration tensor, R_g can be calculated.

Similarly to the definition of the R_g above, the coordinate system of the particle is rescaled to sit at the centre of mass and each coordinate has the r_{mean} subtracted from its value.

By definition the gyration tensor then is a symmetric 3x3 matrix where each component S_{mn} is given by:

$$S_{mn} = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (r_m^{(i)} - r_m^{(j)})(r_n^{(i)} - r_n^{(j)}) \tag{1.1}$$

where m and n are all combinations of the Cartesian coordinates x, y, z

The gyration tensor can then be diagonalised by finding its eigenvalues to give:

$$S = \begin{bmatrix} \lambda_x^2 & 0 & 0 \\ 0 & \lambda_y^2 & 0 \\ 0 & 0 & \lambda_z^2 \end{bmatrix} \tag{1.2}$$

The diagonal elements are referred to as the "principal moments" of the gyration tensor and can be used to find quantities such as the radius of gyration by summing them.

$$R_g^2 = \lambda_x^2 + \lambda_y^2 + \lambda_z^2$$

The other quantities that can be calculated from the principal moments are the asphericity,

$$b = \lambda_z^2 - \frac{1}{2}(\lambda_x^2 + \lambda_y^2)$$

a measure of how spherically symmetric the distribution of particles is or how symmetric with respect to the three coordinate axes. It is always non-negative and the zero condition only occurs when all moments are equal, this only happens when the particles are spherically symmetric.

The acylindricity is given by:

$$c = \lambda_y^2 - \lambda_x^2$$

and is a measure of how cylindrically symmetric a distribution of particles is, with a similar conditions of being non-negative and zero only if the symmetry condition holds.

And finally the relative shape anisotropy is given by:

$$\kappa^2 = \frac{3}{2}\frac{\lambda_x^4 + \lambda_y^4 + \lambda_z^4}{(\lambda_x^2 + \lambda_y^2 + \lambda_z^2)^2} - \frac{1}{2}$$

This is a value bounded between zero and one. The former only occurs when all particles are spherically symmetric and the latter when all points lie on a line.

1.8 Aims of the thesis

As mentioned at the start of this chapter, the focus of this thesis is to explore a simple computational model for chromatin with particular focus on the nucleosome resolution. The primary inspiration for this work came from the results published by Hsieh et. al. [1], where they presented a new HiC based method which allowed them to study the 3D conformation of yeast chromatin at the nucleosome level. Their work presented a new type of nucleosome level interaction map which showed domains at a much smaller level than previously observed in HiC maps. These "micro-domains" seemed to exist at gene-level scale with preferential interaction of the nucleosomes within the domains. The aim for the work in this thesis then, was to reproduce the experimental results, but in a simulation, to see if a simple model could predict these micro-domains in yeast.

Nucleosome positioning data was used to serve as initial input to these simulations and further refinements could be done to the model as required. Surprisingly the simple model was able to reproduce the micro-domains and gave insight into how nucleosome positioning is in part responsible for the formation of these domains. A final step was to move the model over to human genome data. This presented a challenge as no Micro-C data is as of yet publicly available, so the model is limited to nucleosome occupancy data with limited results.

The thesis is divided into six chapters, with this first chapter forming the introduction and the final chapter drawing the thesis together with a conclusion. The second chapter gives an outline of the main experimental methods used

to produce sequencing data as well as give an introduction into Molecular simulations that form the basis of the model developed in this thesis. The third chapters focuses on the bioinformatics work carried out to analyse and process the experimental data used. Some initial results drawn from the analysis are presented, some of which influenced further decision in the development of the model. The forth chapter presents the main bulk of the simulation work for the yeast chromatin giving details of the simple nucleosome model. It then continues into a discussion of the results and conclusions drawn from the model before ending with work carried out to develop a more detailed model, which ultimately failed to produce better results. The fifth chapter then gives an outline of the work done to apply the simple nucleosome model to human nucleosome occupancy data and gives some speculative results and conclusions.

Chapter 2

Methods

In this chapter I will give an overview of the different methods used or relied on throughout the rest of the thesis. It will primarily focus on already established methods whereas methods developed specifically for the present work will be detailed in later chapters. The methods fall into the two separate categories of analysing the experimental data and the analysis of computer simulations and associate data. For the experimental side I will outline the process on how it was obtained and how it is generally processed.

2.1 Next generation sequencing (NGS)

Most modern bioinformatics and DNA analysis relies on next generation sequencing (NGS) to provide the necessary DNA sequencing data for genomic research. Also known as massively parallel sequencing it is used to refer to any of a number of high-throughput approaches to DNA sequencing. It ultimately achieves its speed in processing by using the concept of massively parallel processing to perform sequencing of millions of small fragments of DNA. Several bioinformatics techniques can then be used to piece together these fragments by using a reference genome to locate the reads. As the sequences can vary in length and overlap as well as be sequenced a number of times, a high resolution in data is achieved to provide accurate sequencing results.

The general process of NGS (the precise process depends on which technology is used), follows a three-step format: library preparation, sequencing and analysis.

In the preparation phase, the DNA or RNA samples are processed to be compatible with the sequencing process, typically by fragmenting the DNA and ligating special adapters to both ends of the fragments. The fragments are then amplified, i.e. they are copied many times to produce millions of single-stranded DNA fragments. A process used often to produce large numbers of fragments from an initial batch is polymerase chain reaction (PCR). In PCR a specific DNA segment can be replicated exponentially in order to generate up to millions of copies. This is done most often by thermally cycling, where different processes happen at hot and cold intervals in the cycle. The first step of PCR is usually DNA melting, where the high temperature causes the double strand to separate into two single strands. Primers, or short single strand DNA fragments can then bind at the exposed complimentary sites on the DNA sequence. DNA polymerase is then used to take these templates and build new DNA strands from nucleotides. As any newly created DNA can be used as templates for further replication, this process exponentially amplifies the original sequence.

The second step of NGS is the actual sequencing where artificial nucleotides bind to the DNA template strand. These nucleotides are chemically modified to contain a fluorescent tag which indicates which nucleotide it is. The sequence can then be read by the sequencer and the process can, if required, continue to repeat for the reverse strand. In the case that sequencing occurs from both ends of the strand, it is referred to as paired-end sequencing, otherwise it is called single-end sequencing. Paired-end sequencing is a process that provides more accurate read alignment by providing twice the data for a single fragment.

The third step of NGS is the base-calling done by the machine software where the nucleotides are identified along with an accuracy measure and any further analysis is done to the data by the user. This analysis will again vary depending on the experiment sequenced and the processes involved. In general, quality control is a mandatory step which can be done with a variety of softwares, a popular one being the fastqc software[1], which I used for my analysis. The quality control can analyse the raw sequence data for any preliminary problems as well as detect any adapter contamination (read-through in short fragments), of which the latter issue can be resolved with further post-processing. Although necessary for sequencing, the adapters are not part of the original genome sequence and contaminated reads won't map correctly to the reference genome. For uncontaminated reads, the sequencing adapters are found at both ends where they were purposefully ligated

[1] `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

to before sequencing. After sequencing, these were trimmed from the paired-end data which is simply sequence reads from both ends of a hybridized fragment.

The next major step in any sequencing data analysis is always the aligning of the reads to a reference genome, this can be achieved with the the bowtie2 software [101][2]. This tool allows for the efficient aligning of sequencing reads to long reference sequences. The samtools suite [3] allows for manipulation of the large data files. In particular it can be used to sort and filter the data as well as remove any duplicates found (most likely PCR duplicates). Further processing can be done as required, but will depend on the usage and experiment.

2.2 Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation (ChIP) is an experimental technique used to study interactions between proteins and DNA. It is particularly useful in finding protein associations within the genome, in particular, it can be used to find transcription factors on promoters and histone modifications. In the process, the cell are 'fixed' where the DNA and proteins are cross-linked and then sheared into fragments. A specially selected antibody can then be used to single out specific proteins with associated DNA from the whole, with the remaining solution being discarded.

The precipitated result can then be processed and analysed by sequencing, giving rise to the process known as ChIP-sequencing (ChIP-seq). The sequencing happens as described in the previous section and using this technique binding sites for most proteins in the genome can be mapped. An older method used microarrays for the analysis process, termed ChIP-on-chip, it made use of known sequence probes for the entire genome on a DNA chip. The binding sites of the ChIP solution on the microarray could then be used to identify the genomic locations of the proteins in question.

[2] `http://bowtie-bio.sourceforge.net/bowtie2/index.shtml`
[3] `http://www.htslib.org/`

2.3 MNase-seq

MNase-seq data uses micrococcal nuclease to digest the DNA not protected by nucleosomes. A key difference between the MNase-seq and ChIP-seq is missing out the step of linking nucleosomes together. Hence rather than interactions, the data provides a genome-wide map showing nucleosome coverage within a population of cells (Fig. 2.1). So crucially MNase-seq is simply a process to find the 1D locations of nucleosomes along the genome.

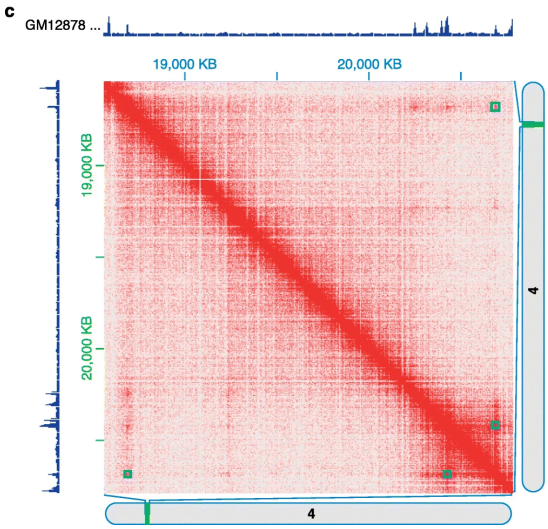


Fig. 2.1 An example of a conventional HiC map for a locus on chromosome 4 of human kidney cells. Here the intensity of the colour gives the strength of the interactions measured. Figure taken from Ref. [102] used under CC

The principle behind MNase-seq is to use the enzyme to digest the "free" DNA between nucleosomes as DNA "protected" by nucleosomes and proteins restricts access to it. The chromatin is then broken into millions of fragments by shearing and once the DNA is isolated it can be sequenced using NGS to provide MNase-seq data. The data set used can be accessed at GEO:GSM53721 and a similar procedure as with most sequencing data is followed. The data is extracted from the SRA (Short Read Archive), fastqc quality control is run on it, sequencing adapters are removed and the reads are aligned to the S. Cerevisiae reference genome (SacCer3) using bowtie2 in paired end mode. Quality control was done by removing low quality reads. This then provides a large data set of individual reads that map to certain places in the genome, indicating the possible presence of a nucleosome. To find the position of the nucleosome from this population data, an algorithm must be used and this explained in section 3.1.

2.4 Micro-C and Micro-C XL

Recently a new method based of the HiC protocol was presented which allowed sequencing and mapping resolutions to reach mononucleosome level at around 200bp. The method developed by Hsieh et al., dubbed Micro-C (micrococcal nuclease chromosome conformation assay) follows much the same process as other 3C based methods. The organism in question, here the yeast S. Cerevisiae, is grown in lab conditions. The DNA in the cell nucleus is fixed in place with formaldehyde so that the 3D structure is preserved and the nucleosomes are crosslinked. Instead of restriction enzymes in order to fragment the chromatin it makes use of micrococcal nuclease (MNase) to yield $> 95\%$ mononucleosomes. This step is followed by mononucleosomal end repair, ligation of the fragments and a two step process to purify the ligation products. Illumina paired end deep sequencing can then be used to characterise the ligation products, i.e. the sequencing data can be processed to reveal where in the genome the fragment originated – see section 2.4.1 below. Paired end deep sequencing is different from single end sequencing in that the ligated fragment is sequenced from both ends as opposed to just one of the ends.

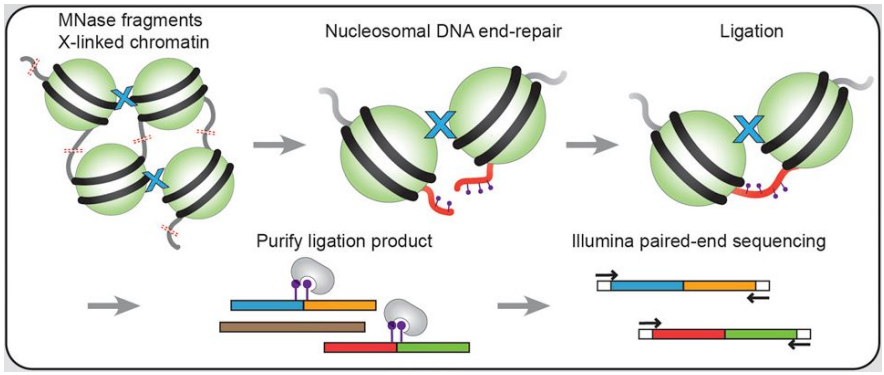


Fig. 2.2 Simplified summary of the Micro-C process in which the nucleosomes are initially stuck together and cross-linked. MNAse digestions fragments the chromatin before DNA end-repair and ligation link the DNA in nucleosomes together. Once the nucleosomes are removed, the ligation products can be sequenced. Figure taken from Ref. [1] used with permission from Elsevier

This then allows for the creation of nucleosome level chromosome interaction maps (referred to as contact maps such as in Fig. 2.3). One of the primary finding of Hsieh et al. is that there is an abundance of self-associating domains in yeast, similar to other species, albeit at a much smaller scale. Typically they

only encompass between one to five genes. The boundaries of these domains tend occur at the promoters of highly transcribed genes or regions with a large histone turnover, also known as nucleosome free regions (NFRs). A further major observation was a distinct lack of evidence for any regular organization of the chromatin fibre above the nucleosomal scale. If found, this would have suggested the formation of 30nm chromatin fibres in yeast, which so far have only been found in higher eukaryotes.

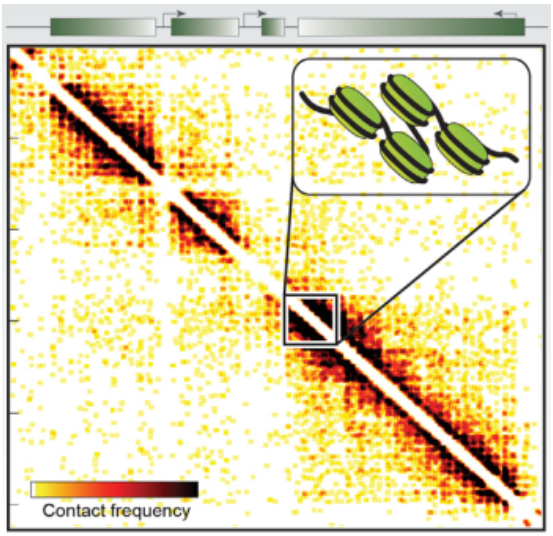


Fig. 2.3 A nucleosome resolution contact map from the Hsieh et al. paper [1]. Inset a stylisation of nucleosome interaction contacts found near the diagonal. Figure taken from Ref. [1] used with permission from Elsevier

A major limitation on the Micro-C process is the missing out on long range interactions between chromatin sections such as centromere-centromere interactions which are present in conventional HiC [81, 103]. To remedy this, a follow-up method termed Micro-C XL (Micrococcal nuclease-based analysis of Chromosome folding using long X-linkers, X for cross) was developed by the same team [104]. Much of the same method remains but the crosslinking on the nucleosomes is supplemented with a longer range agent, which allows for the observation of long range interactions. Instead of just using the "zero-length" cross-linker formaldehyde, it is supplemented with disuccinimidyl glutarate (DSG, a 7.7-Å crosslinker) and ethylene glycol bis(succinimidyl succinate) (EGS, a 16.1-Å crosslinker).

In the Micro-C XL data, chromosomally interacting domains (CIDs) were again observed like before, but with the longer crosslinkers, visualisation of the structures was improved and increased the assay's overall signal-to-noise ratio.

The major complaint of the previous method was also resolved with the longer range crosslinkers giving clear centromere–centromere and telomere–telomere interactions characteristic of the Rabl configuration found in yeast.

2.4.1 Micro-C and Micro-C XL process

The data related to the paper by Hsieh et. al. [1] is provided via the NCBI Gene Expression Omnibus under the accession number GSE68016 while Micro-C XL data are obtained from Ref. [104] (available at GEO:GSE85220, specifically samples GSM2262329, GSM2262330 and GSM2262331). The most abundant data was available for the *Saccharomyces cerevisiae* strain Yeast_BY4741, of which there were 20 replicates in SRA format, therefore my work focused on this data set. The processing procedure is very similar to the one mentioned in the reference and began with extraction from the archive and splitting into fastq files. A quality control check using the fastqc software analyses the reads for sequencing adapter contamination. The reads were then aligned to the *S. Cerevisiae* reference genome (SacCer3 build) with bowtie2 [101]. The data were treated as single-end reads, with each of the pairs being aligned independently as ligation fragments will not form proper pairs when aligned. After sorting, duplicates can be removed as they are likely an artefact of the PCR process and it is unlikely that identical fragments are found. The data is then consolidated into a single file. For quality control, all pairs with a mapping quality score of less than 30 are sorted out. The mapping quality is a measure of how many places in the reference genome a fragment maps to, in this case a value of 30 is roughly equivalent to mapping to a single location as unique fragments are required.

To avoid issues when potentially including reads resulting from runs of undigested nuclesomes, the reads are further filtered according to the strand each read in the pair maps to. The data can now be binned to create contact maps or can be mapped onto specific nucleosomes to obtain a nucleosome-nucleosome interaction map (see section 3.3 for details).

2.5 Molecular simulations

Computer Simulations of complex biological systems have only been possible in recent years due to advances in computing power, but have already proven invaluable in furthering our understanding in the field. In many cases it is possible and more convenient to use a simple model to investigate something before investing in complex experimental procedures. It is unlikely that computer models will ever replace experimental processes, but they are proving to be a good companion to the experiments.

At its core molecular dynamics (MD) is a process where the motions of particles over time is simulated within a computational framework. MD simulations, traditionally aim to solve Newtons Second Law for all the particles found in the system. To this end a 'force field', a set of quantum mechanically determined interaction potentials, is used to calculate the forces between the atoms. This leads to one particular issue that arises from the complexity of the system and which the simulations have not yet overcome, the problem of scale. Consider how many atoms there are in a small piece of DNA; and always when trying to simulate this, the answer will be "too many". Increased resolution and complexity always comes at a computational cost which in modern systems translates to time. The more detailed a simulated system, the more time it will take to process. A method of getting around this drawback is to make use of **coarse grained modelling**. It basically means that complex systems are simplified and "coarse-grained" to reduce the computational complexity. For example, instead of resolving every single atom in a molecule, instead resolving groups of atoms or similar. This is particularly useful for molecular dynamics simulations as degrees of freedom can be reduced and thus simulation times are reduced.

A further adaptation to the method is to remove any solvents present in the system and dealing with it implicitly by including its impact on the motion of the particles. This way the solvent itself does not have to be simulated, but its effects are still felt. A common method is to use Brownian or Langevin Dynamics and the simplest model includes only viscous drag "thermal jostling" the water molecules would impart on the particles.

2.5.1 Langevin equations

In Langevin dynamics the effects of any solvent in the system are approximated by adding two new forces into the equations of motions. Hence the time dynamics of the beads within this coarse grained molecular dynamics simulation are given by the Langevin equation, where changes in the positions of the ith bead $\vec{r}_i$ are given by:

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\nabla U_i - \gamma_i \frac{d\vec{r}_i}{dt} + \sqrt{2k_B T \gamma_i}\vec{\eta}_i(t), \tag{2.1}$$

where m_i is the mass of bead i. So if the right hand side of the equation is simply mass times acceleration, then the left hand side is a sum of all the forces the particle experiences. The sum of all interactions for a bead i and all other beads is represented by the potential U_i, the first term. This term basically represents the "force field" experienced by a particle. The second term is the first of two terms for the effects of the implicit aqueous solvent, representing viscous drag. This drag is proportional to the velocity and dependant on the friction γ_i. The third term relates to the "thermal jostling" mentioned above, i.e random uncorrelated noise that gives the particles "nudges" in random directions. The term $\vec{\eta}_i$ obeys the following relations

$$\langle \eta_\alpha(t) \rangle = 0 \quad \text{and} \quad \langle \eta_\alpha(t)\eta_\beta(t') \rangle = \delta_{\alpha\beta}\delta(t - t'). \tag{2.2}$$

The final term of Eq. (2.1), related to the noise variance, is scaled by the thermal energy of the system which in turn is given by the Boltzmann factor k_B multiplied by the system temperature T (set at 310 K for a cell). To keep things simple, all beads in the system are assumed to have the same mass and friction $m_i \equiv m$, and $\gamma_i \equiv \gamma$.

2.5.2 A simple DNA model

The DNA model in Refs. [105–107] was used as inspiration in the work of this thesis. Here a DNA molecule is represented by a simple connected chain of beads. Ultimately this is a coarse grained model in which DNA is represented as a bead-and-spring polymer. The 2.5nm beads in the model represent 7.35 bp of DNA and are connected to each other via finitely extensible non-linear elastic (FENE) spring, where the ith bead in the sequence with position $\mathbf{r}_i$ is connected to the $i+1$th bead with the FENE spring. The potential associated with this is given by:

$$U_{\mathrm{FENE}}(r_{i,i+1}) = \\ U_{\mathrm{WCA}}(r_{i,i+1}) - \frac{K_{\mathrm{FENE}}R_0^2}{2}\log\left[1-\left(\frac{r_{i,i+1}}{R_0}\right)^2\right], \tag{2.3}$$

where $r_{i,i+1} = |\mathbf{r}_i - \mathbf{r}_{i+1}|$ is the separation of the beads, and the first term is the Weeks-Chandler-Andersen (WCA) potential

$$\frac{U_{\mathrm{WCA}}(r_{ij})}{k_BT} = \begin{cases} 4\left[\left(\frac{d_{ij}}{r_{ij}}\right)^{12} - \left(\frac{d_{ij}}{r_{ij}}\right)^{6}\right] + 1, & r_{ij} < 2^{1/6}d_{ij} \\ 0, & \text{otherwise,} \end{cases} \tag{2.4}$$

The WCA potential represents a steric interaction to prevent the adjacent beads from overlapping. The mean of the diameter of beads i and j is given by d_{ij} in Eq. (2.4).

By setting the bead size to 2.5nm the diameter presents a natural length scale to parametrize the system, denoted σ. This then becomes a reference to all other length scales. The second term in Eq. (2.3) gives the maximum extension of the bond, R_0, which is set to $R_0 = 1.6\ \sigma$ and the bond energy is set to $K_{\mathrm{FENE}} = 30\ k_BT$ for linker DNA beads.

The bends within the polymer are dictated by the rigidity between every three adjacent DNA beads. This is governed by a Kratky-Porod potential and given by:

$$U_{\mathrm{BEND}}(\theta) = K_{\mathrm{BEND}}\left[1 - \cos(\theta)\right], \tag{2.5}$$

where θ is the angle between the three beads as given by

$$\cos(\theta) = [\mathbf{r}_i - \mathbf{r}_{i-1}] \cdot [\mathbf{r}_{i+1} - \mathbf{r}_i], \tag{2.6}$$

and K_{BEND} is the bending energy. The persistence length in units of σ is given by $l_p = K_{\mathrm{BEND}}/k_B T$.

Finally, steric interactions between non-adjacent DNA beads are also given by the WCA potential [Eq. (2.4)] The masses of the DNA beads and nucleosome beads are set to 1 in simulation units for the the simple model.

To solve the dynamics of the system and Eq. (2.1), a MD software code can be used. In this case the LAMMPS software [108] is used which applies a standard velocity-Verlet algorithm. LAMMPS or Large-scale Atomic/Molecular Massively Parallel Simulator is a molecular dynamics code developed by Sandia National Laboratories. It is freely available open-source software used to model ensembles of particles in a liquid, solid or gaseous state and can model various systems (such as atomic or polymeric) through the use of various interatomic potentials and boundary conditions.

Chapter 3

Data analysis: nucleosome positions and interactions in yeast

In this section I will go into a bit more detail on the work done with the experimental data. Most of this work was done with the simulations in mind and as such were a preparation for things to come or to generate later data for the simulation models. Most sequencing data is not available in a format which allows it to be used immediately, as every user tends to have their own requirements for the data, it is generally left in an unprocessed state. The purpose of this chapter now, is to outline the work I have done with the raw data, in order to bring it into a state in which it can be used for my purposes. I also discuss a number of preliminary observations that can be drawn from this data and how it used for the simulations to come.

3.1 Nucleosome positions from MNase-seq data

Although it is possible to infer the nucleosome positions from Micro-C data, a better approach is to use data from MNase-seq as it provides an independent measurement and at better resolution. A high read-depth MNase data set for the same yeast strain used in the Micro-C experiment is available from Dang et al. [2] and was used in further analysis. Once processed, MNase-seq data can be used to generate nucleosome coverage maps. This is done by piling up the centre points of each read. This allows for an approximate idea of where nucleosomes might be

located, but since MNase-seq data is obtained from a population of cells, there will be be a high cell-to-cell variability in the nucleosome positions. As can be seen in Fig. 3.1, a number of peaks can be identified in the data, but there is a significant variation between the peaks. The peaks all have very different heights, signifying different numbers of reads found for that location and the peaks are differently well-defined, with some being much narrower or wider than others. Some peaks are also much too close together, which would mean overlapping nucleosome, something that is not possible if discrete nucleosome positions are sought after.

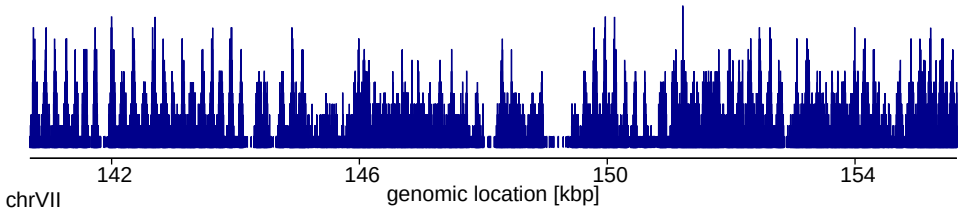


Fig. 3.1 Example nucleosome coverage map for a sample genomic region on *S. Cer* chromosome VII, generated from a pileup of reads, i.e. the height of the bars represents the number of reads found centred at that location. Nucleosome positions can be inferred from this data, but their most likely position is obscured.

3.1.1 Nucleosome finding algorithms

There are a number of different software packages and approaches available for extracting the nucleosome positioning from MNase-seq data. In principle they all deal with the problem of analysing millions of small and fuzzy enrichment peaks which might or might not correspond to individual nucleosomes.

In general there are three main approaches to dealing with MNase-seq data to extract nucleosome positions, the first being a detection of enriched peaks with width of around 147bp from the experimental data with a precision of one to a few bp. An example for this approach would be PuFFIN (Positioning for Fuzzy and FIxed Nucleosomes) [109], a software tool developed to build nucleosome maps for the entire genome. The benefit of being free of user-tuned parameters made it initially an interesting candidate but it was later replaced by a different approach because ultimately it is only reasonable to call nucleosome peaks when they are well defined. In yeast this tends to be the case, but in higher eukaryotes the peaks become too blurred for this approach.

In this case a second style of approach lends itself better to the nucleosome position analysis. Instead of finding specific nucleosome positions, one works with a continuous occupancy profile which can be used to define and find regions of interest and of differential nucleosome occupancy.

A further complication of working with MNase-seq data is that it is a representation of a whole ensemble of cells and not of an individual cell. Therefore only an average or approximate nucleosome landscape can be found. The third style of approach aims to deal with this issue by using an approach of Monte Carlo simulations. The nucleosome occupancy data is used to generate an effective nucleosome interaction potential in order to build up a representation of "most probable non-overlapping" nucleosome positions in the cell. This approach uses ideas common in physics and is employed by the NucPosSimulator software. As this is the software used in this work, it will be detailed further below.

3.1.2 NucPosSimulator

The NucPos Simulator [110] [1] software implements a method for finding non-overlapping nucleosome positions from a set of MNase-seq data. This is done by combining a binary-variable analysis and a Metropolis Monte Carlo (MMC) approach with a simulated annealing process. It was developed to deal with the intrinsic biological variability of nucleosome positions found in MNase-seq data. As each cell's nucleosome conformation can be different from the others, an average created from these conformations will contain overlapping and ambiguous positions for the nucleosomes. The method was developed as a response to other methods which cannot deal with overlapping nucleosomes. The method creates dynamic populations of non-overlapping nucleosomes which will quench to a single population and conformation.

The basics of the software is that it will take the centre point for each paired end read and build a frequency count profile. After smoothing with a Gaussian kernel and normalisation it is used to generate an effective potential landscape for nucleosome positions.

The MMC simulation will then run recursively to add, remove and move nucleosome in this potential landscape to bring the nucleosomes into "favourable" positions. Specifically, moves of nucleosomes are chosen at random and a move

[1] software available from: `http://bioinformatics.fh-stralsund.de/nucpos/index.html`

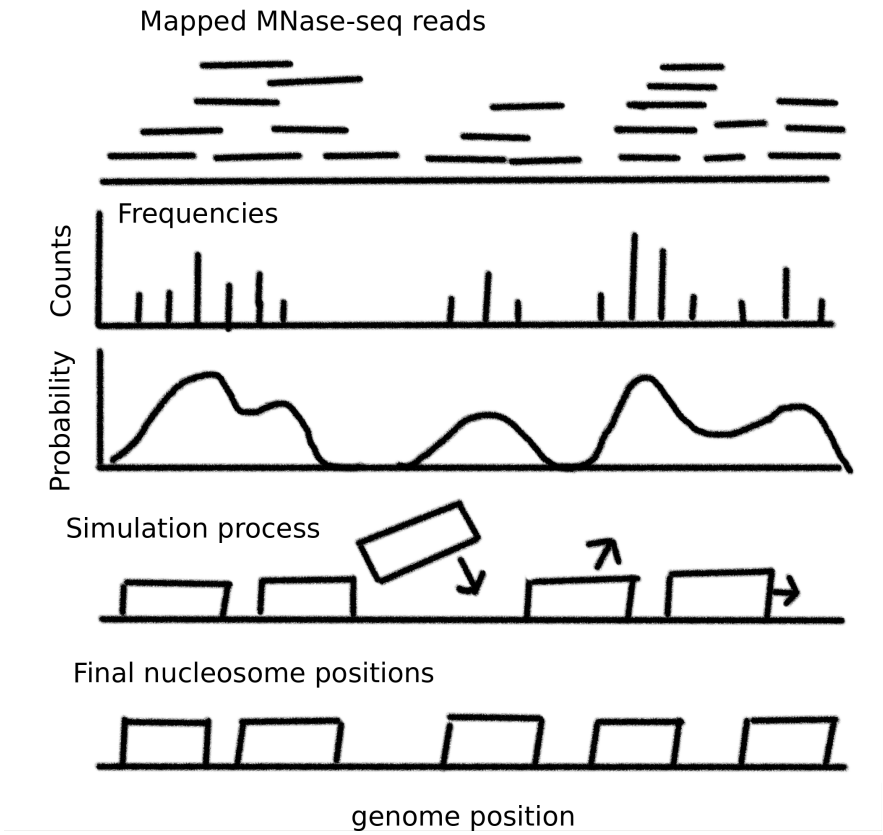


Fig. 3.2 Summary of the nucleosome position finding process of the NucPosSimulator software. Initially the mapped MNase-seq reads are piled up and converted into frequencies of counts for the nucleosome occupancy found. From there a probability "potential" is generated which can then be used in the simulation process to find a most-favourable positioning of the nucleosomes by minimising the energy. Finally a set of "most-likely" nucleosome positions is generated as output. (Figure inspired by the summary on the NucPosSimulator website.)

is proposed. The Metropolis algorithm will then accept or reject an attempted move based on the resulting change in the energy of the nucleosome landscape. Ultimately the aim is to minimise the energy of the landscape to produce a most-favourable positioning of the nucleosomes. When used in the "simulated annealing" mode, the system starts at a high temperature, allowing for a highly dynamic system, before being slowly cooled where nucleosomes will settle into more likely positions. The final output is then a list of nucleosomes found with their most likely positions. Since MNase-seq is a "population-level" method, the simulated annealing mode will produce a single most likely positioning for the nucleosomes. The other option is to use an alternative fixed T mode (system temperature) in which a whole ensemble of likely configurations can be generated.

3.1.3 Nucleosome positions

When using the software in the "simulated annealing" mode, the system starts at a high temperature and the nucleosomes will be highly dynamic in their positions. As the system is slowly cooled the nucleosomes settle into their most likely positions. Using the software, all the chromosomes for yeast can be analysed and the most likely nucleosome positions determined.

In Fig. 3.3 a small section of the genome is plotted with the pile-up data above and the nucleosome positions below. Each major peak in this case corresponds to a nucleosome, but the centring can vary depending on the "fuzzyness" of the peak.

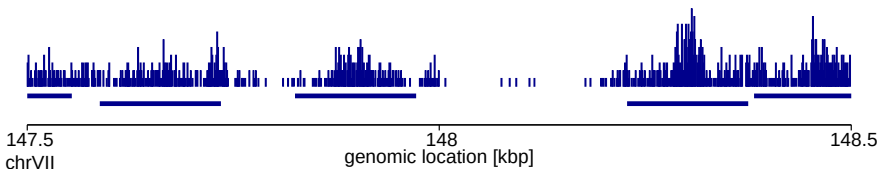
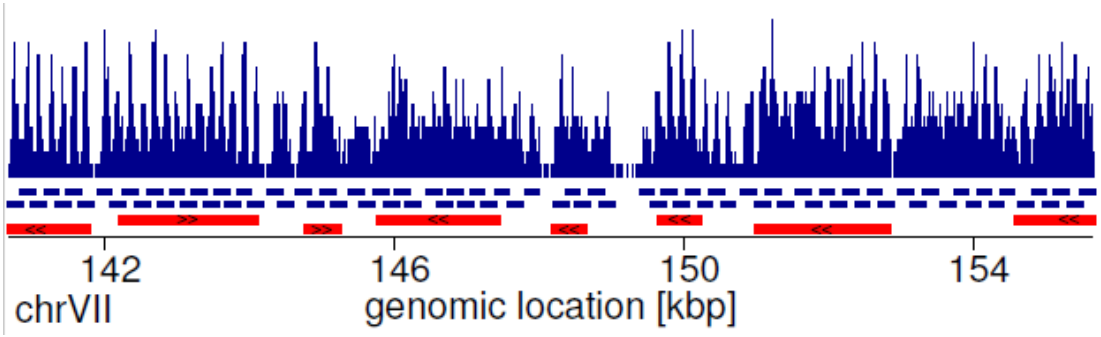


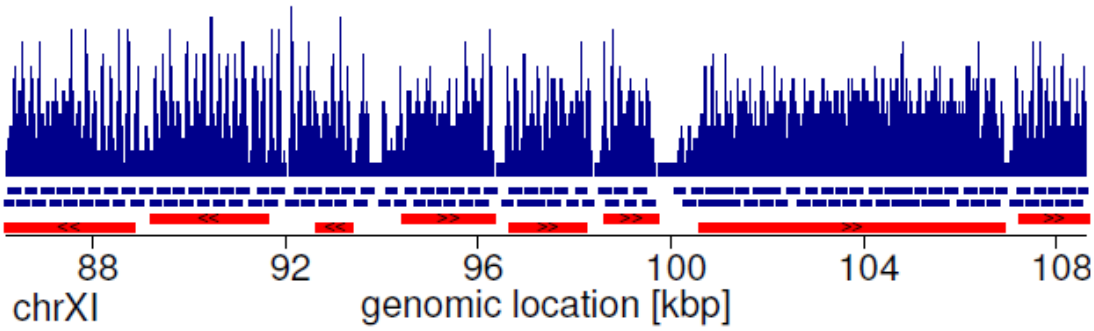
Fig. 3.3 Small detail section from the plot in Fig. 3.1 showing the fuzzyness of the peaks and the possible ambiguity of placing nucleosomes. The centre peak is quite well defined and would warrant a placement, but the peak to the far right is more contested visually. The software did place a nucleosome there and it is non-overlapping with neighbouring nucleosomes.

When comparing larger regions, such as in Fig. 3.4, the results of the nucleosome placement become more obvious. In some cases the nucleosome positions are a bit more debatable, but in other cases clear gaps can be seen in the MNase-seq data.

To ensure that the nucleosome positions for the relatively small representative regions are not skewed by them being so small, the entire chromosome in which each region resides was used when finding nucleosome positions with NucPosSimulator. Since the chromosomes in yeast are relatively small (compared to human chromosomes for example) the software manages to process the entirety of the data. Any effect that the ends of a region might have on the nucleosome positioning are mitigated in this way.



(a) chrVII:140680 - 155644



(b) chrXI:86225 - 108599

Fig. 3.4 Two full representative regions are shown with the MNase-seq data above and the placed nucleosomes in blue below. The red bars represent genes found within the region.

3.2 Linker lengths

One of the aspects that was interesting to consider, especially after some initial work with the bioinformatic analysis of nucleosome positions, was the linker length between the nucleosomes. The analytical inspection of the chromatin and nucleosome positions generated shows that nucleosome spacing is highly irregular, which leads to the formation of a heterogeneous fibre. Interestingly, although nucleosome positioning data has been available for some time now, the textbook picture usually given is that of regular spacing. So it might be that irregular nucleosome spacing is a generic feature of yeast chromatin *in vivo* and it appears as if it is just not often discussed in chromatin fibre formation studies.

To examine the nucleosome spacing, the distribution of linker lengths genome-wide can be plotted as in Fig. 3.5. The linker length distribution of the regions analysed has been added to show an agreement between the distributions, making the simulated regions a good representation of the genome. The nucleosome

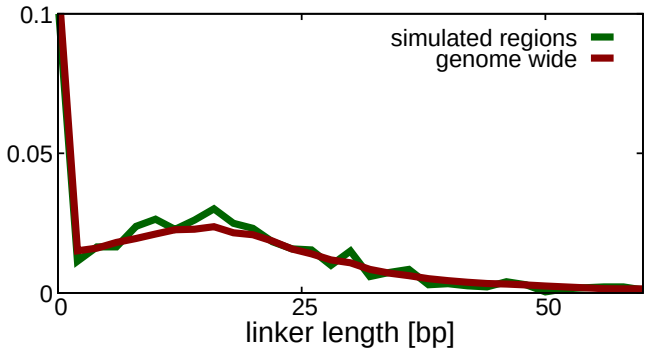


Fig. 3.5 The linker length distribution genome wide (red) and for the eight representative regions (green) based on the nucleosome positions generated from the NucPosSimulator software.

positions are all generated from the MNase-seq data with NucPosSimulator. An interesting feature of the distribution is its multi-modal shape as opposed to a Gaussian distribution, which would be expected by a regular distribution. If the distribution had been a random spacing, brought about by a Poisson process, an exponential distribution would be expected. Instead, a large number of very short linkers (about 25% of linkers genome wide have length 1-3 bp) is followed by a broad peak at around ∼16 bp. Furthermore, many linkers are longer than this with about 12% of linkers being between 50 and 200 bp in length. These are most likely part of NDRs around gene promoters. Much longer linker lengths are likely to be artefacts from un-mappable regions of the genome. The literature value for yeast nucleosome repeat length is often quoted as 165 bp [111, 112], which corresponds to a linker length of 18 bp. The distribution in Fig. 3.5 gives a mean linker length of ~ 28.7 bp. If only those linkers which are ≤ 100 bp are considered, this decreases to ~ 18 bp, i.e. NDRs are excluded.

The nucleosome linkers can be split up according to their location along the genome with regard to genes. Thus, in addition to the genome-wide distribution, we can consider first the linker lengths within genes, second the linker lengths within regions 500 bp upstream of genes (i.e. promoters) and third the linker lengths in non-genic regions. For the gene related regions, only genes of length ≥ 1 kbp are considered, therefore only linkers within the genes are taken into account. For the non-genic regions, all annotated genes are excluded to determine the correct linkers. When considering these different genomic regions in Fig. 3.7, it can clearly be seen that the genome average, linkers within genes and in non-genic regions have a similar length distribution although there are more short linkers (< 3 bp) within gene bodies (around $\sim 30\%$ as opposed to $\sim 25\%$ in other

regions). As expected in the upstream regions of genes, typically the location of promoters and NDRs, there is a higher proportion of long (50-200 bp) linkers ($\sim 40\%$), with a lower proportion of short and medium length linkers.

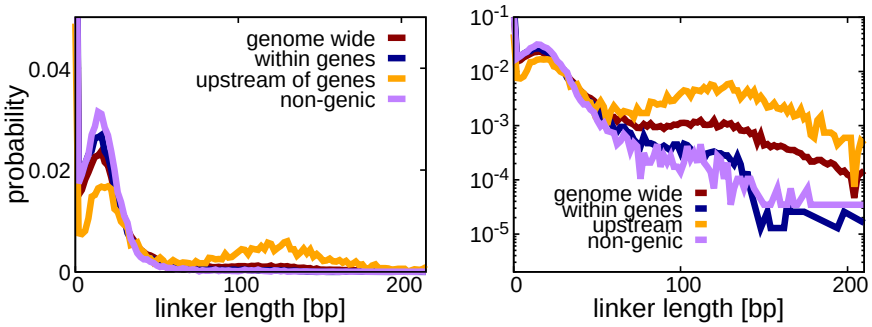


Fig. 3.6 The linker length distributions for within genes (blue) (annotated of length $\geq$ 1kbp), the distribution within the 500bp upstream of the TSS for genes (yellow) and for the non-genic regions (purple). These are all compared to the genome-wide distribution (red)

Fig. 3.7 plots the linker length distribution of genes and promoters in addition to the linker length results of the boundary finding algorithm for the entire genome. This confirms that, genome-wide, boundaries tend to be found at long linkers (with the adjacent linkers tending to be short or medium in length). The histogram shows that most linkers are found within genes and only a smaller proportion are found upstream and at boundaries. However the result also shows that the majority of long linkers are found within boundary regions and upstream of genes.

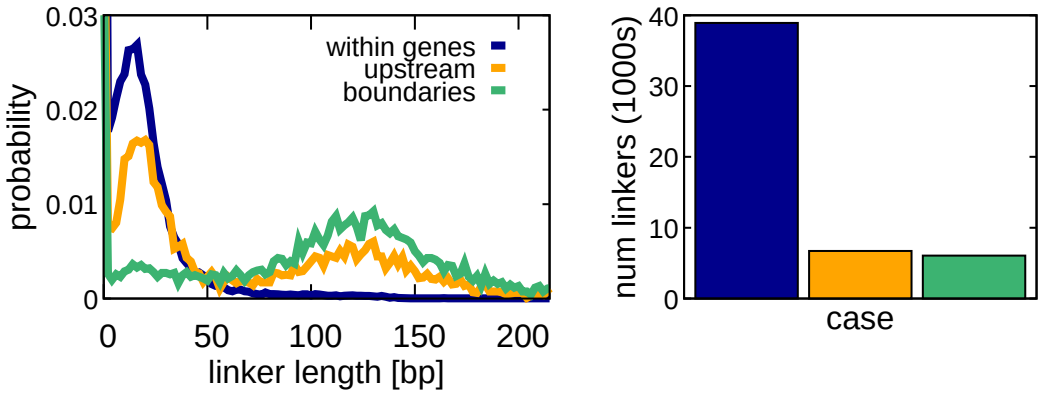


Fig. 3.7 The distribution of linker lengths for three different cases. In green the results for domain boundaries for the entire genome (found in the same way as other boundaries), in yellow the distribution upstream of genes and in blue the distribution within genes. To the right the distribution of linkers for the three cases is given in 1000s.

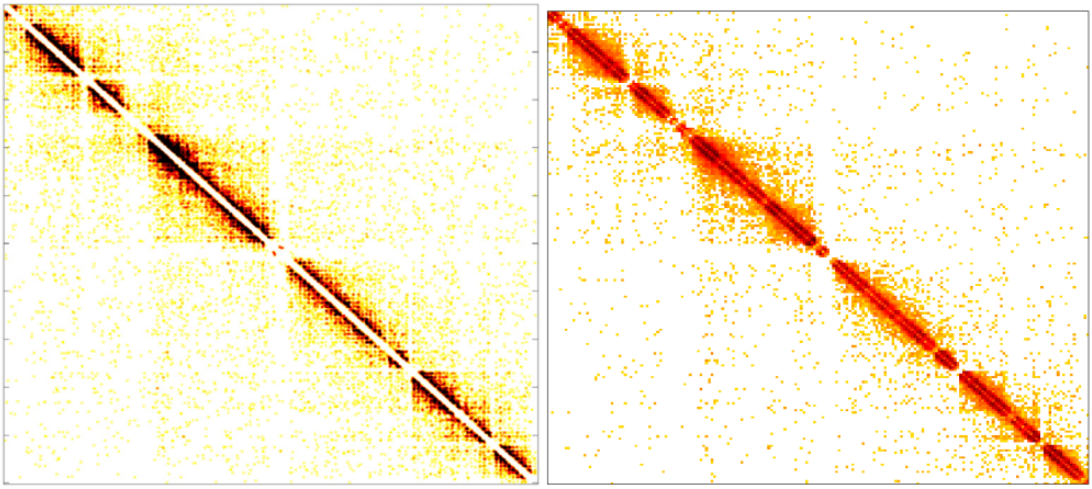
3.3 Generating contact maps from Micro-C data

In total three different types of Contact map had to be generated from different types of data sets. Initial maps were a reproduction of the maps found in Ref. [1], where the genome is bins of fixed width for the interaction maps. Once I had the nucleosome positioning data, the Micro-C data was translated to nucleosome resolution contact maps. Later, once simulation data was available, this also was used to create nucleosome resolution contact maps.

3.3.1 Binned contact map generation

The initial contact map generation focused on creating the contact maps found in the Hsieh et al. paper. For this the data sets from the paper were analysed as outlined in the previous sections to obtain a list of detected interactions. The binned maps were meant as a proof of concept that I could replicate the data, therefore, only a number of replicates were combined into a single contact map since the analysis of the contact maps took significant computing resources and time. While some of the later data sets were processing, the first sets were binned and combined to create the contact map.

The initial tests were made with bin sizes of 100, 200 and 1000bp and the 100bp bin size gave the best comparison to the figures from the paper. In Fig. 3.8 a comparison between an initial contact map created by me with a 100bp bin size is shown next to the same region from the paper. Just as in the Hsieh et al. paper the small nucleosome level domains observed were also visible in the contact maps I created. As the goal was to create simulations with the nucleosome positions, the next step was to convert the binned contact maps into nucleosome resolution maps.



(a) Binned contact map from Hsieh et al. Ref. [1] used with permission from Elsevier

(b) Binned contact map from my data analysis

Fig. 3.8 Contact map comparison for Chromosome 9 (360,001 – 380,000) using a binning method and 100bp bin sizes. The version on the right did not have the diagonal removed.

3.3.2 Nucleosome resolution contact maps

To generate nucleosome level interaction maps, I used the nucleosome positions generated by NucPosSimulator from the MNase-seq data as detailed above. Treating each of the pair from each Micro-C read separately, reads which overlap with a single nucleosome are unambiguous; reads which do not overlap with a nucleosome, but map to a position where their centre point is within 200 bp of the centre of one or more nucleosomes are assigned to their closest nucleosome. Reads which overlap with more than one nucleosome are assigned to the nucleosome with which they have the largest overlap. Reads which do not map within 200 bp of a nucleosome or which overlap with two nucleosomes by the same amount are discarded. Only read pairs where both members of the pair are assigned to nucleosomes are retained as informative interactions.

For the Micro-C data, across 20 replicates, starting with 73,943,603 interactions, I was able to assign 73,803,602 of these unambiguously to pairs of nucleosomes; i.e. less than 1% of read pairs were discarded as it was ambiguous as to which nucleosomes they represented. For the Micro-C XL data, across 3 replicates, starting with 130,958,525 interactions, 114,652,179 of these could be assigned unambiguously to pairs of nucleosomes, this time discarding less than 0.05% of read pairs (summarised in table 3.1). Although the Micro-C XL data actually

has more interactions in total, these interactions are spread over a much larger scale. For the focus on the micro-domains, I required the majority of interactions to lie close to the diagonal and not far away from it, even if there are more total interactions.

Experiment	Replicates	Interactions		
		total	mapped	% discarded
Micro-C	20	73,943,603	73,803,602	>1%
Micro-C XL	3	130,958,525	114,652,179	>0.05%

Table 3.1 Comparison of mapped interactions between the Micro-C and Micro-C XL data set.

To create a single map from the twenty replicates, the individual maps are combined. This is achieved by summing all corresponding interaction values between the different replicates. As the entire chromosomes are too large to be plotted into maps at adequate resolution, the work from here on continued with representative regions which were extracted from the primary map. These regions are meant to represent the genome across several chromosomes excluding the telomeres and centromeres (more detail in section 3.8). For region 0, the smallest region at chrVII:140,380-155,644 one half of the symmetric contact map is given below in Fig.3.9. The red bars show the genes found in the region. From this simple contact map, a certain grouping of nucleosome is obvious and indicative of the "micro-domains" found at the nucleosome level in yeast.

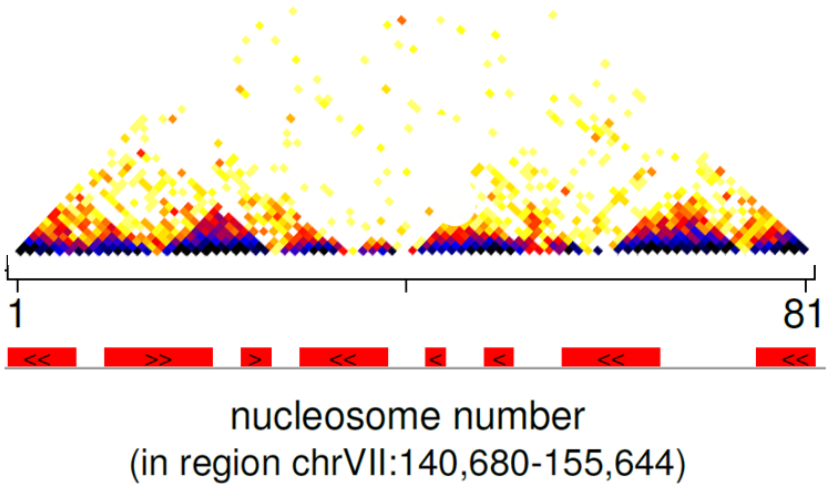
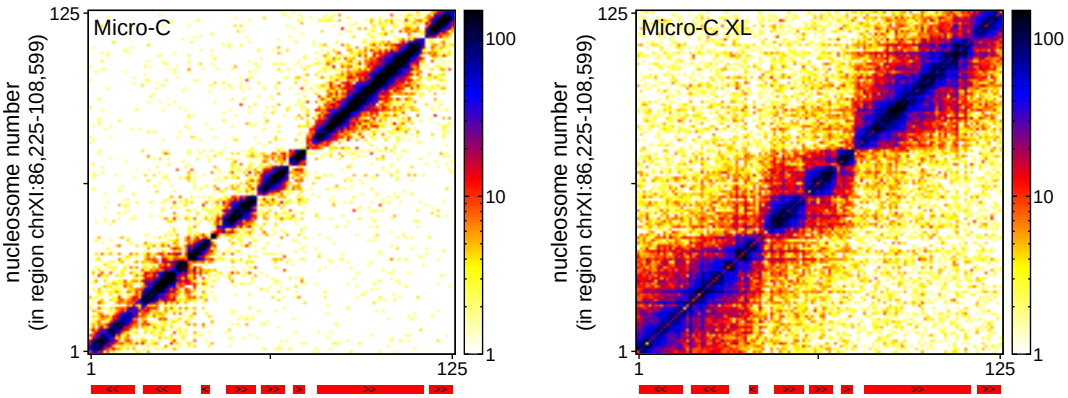


Fig. 3.9 Upper triangular part of a symmetric contact map for a region at chrVII:140,380-155,644 , with very long range interactions cropped out. Below the genes found in the region are given by the red bars with transcription direction indicated by the arrow.

Using this method, contact maps were created for all regions with the experimental data from the Micro-C experiment and the Micro-C XL experiment. Two representative figures (Fig. 3.10) are shown below for region 9 found in chrXI:86,225-108,599.



(a) Micro-C **(b)** Micro-C XL

Fig. 3.10 Comparison of a Micro-C vs Micro-C XL interaction map for the same region in chrXI:86,225-108,599. Similar micro-domains are clearly visible, but there is a stark difference in signal intensity further away from the diagonal in line with expectations of more longer range interactions in the Micro-C XL data.

3.4 Micro-C vs Micro-C XL

Comparing the two contact maps in Fig. 3.10, it is obvious that although they do represent very similar data, the signal is very different between the maps, with the Micro-C XL map (b) being much busier away from the diagonal. With a close comparison of the two in Fig. 3.11, the micro-domains are distinguishable and even boundaries found in both coincide.

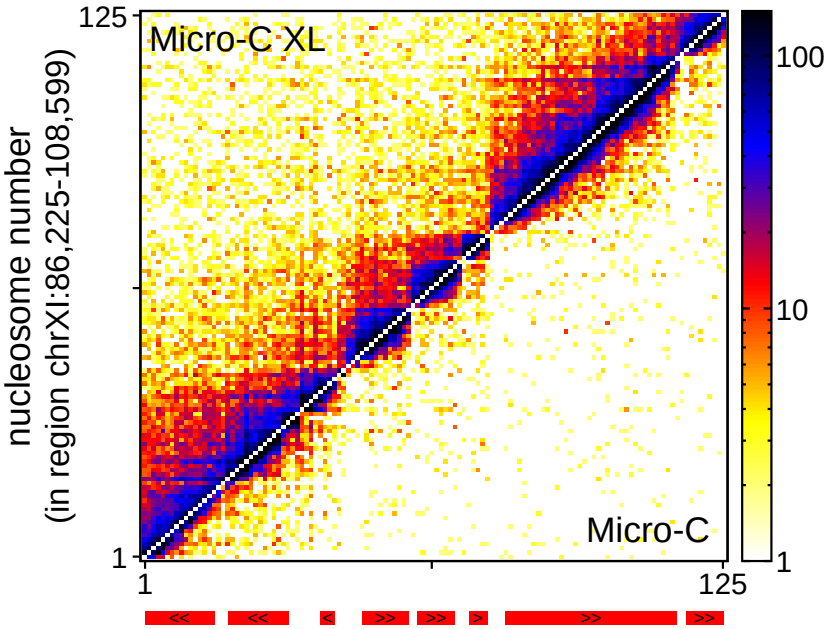


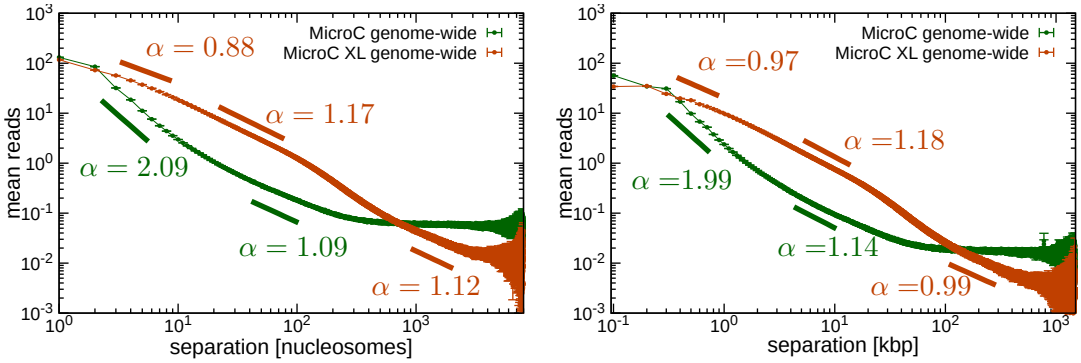
Fig. 3.11 Combined figure of the two contact maps from Fig. 3.10 with the upper triangular showing the Micro-C XL data and the lower triangular showing Micro-C. Genes are given again below the map in red. Found boundaries in both maps are shown above, with agreement between the two data-sets.

The data from the Micro-C XL experiments also reproduced higher-order chromatin interactions not found in the Micro-C data, but a striking difference between the data is in how the number of interactions depends on the genomic separation, s. In the contact map this presents itself as a higher long-range interaction count, but can also be shown explicitly via a log-log plot of the mean number of interactions for a given (nucleosomal) separation (Fig. 3.12), where a linear relationship on the plot shows a power law relationship (mean number of reads $\sim s^{-\alpha}$) often found in other experimental methods. Here both the Micro-C and Micro-C XL data show linear regions, though interestingly, there seem to be different power law regimes for short range (separation $s \approx$ 2–10 nucleosomes) and long range ($s \approx$ 10–100 nucleosomes) interactions. The difference between

the regimes is much stronger for the Micro-C than the Micro-C XL data, and both data sets show a similar slope in the 10–100 nucleosome range.

In Fig. 3.12a the number of reads for the Micro-C data initially decreases steeply with genomic separation (with an exponent close to 2), easing off to a shallower gradient before reaching another linear regime with exponent close to 1. The plateau at large separations can be explained with the fact that the original Micro-C method fails to capture long range interactions. The Micro-C XL data initially shows a much shallower slope for small separation, than Micro-C, with an exponent less than 1. It then becomes steeper with a linear regime and a slope close to 1, before reaching a small steeper region until at long ranges (separation $s > 500$) there is again a slope close to 1.

It is unclear what the origin of these different regimes is, though HiC data typically shows an exponent close to 1 for large separations. Fig. 3.12b shows a plot with separations as DNA length as opposed to nucleosomes, but shows a very similar picture with similar exponents.



(a) **(b)**

Fig. 3.12 Plots showing how, on average, the number of interaction reads between nucleosomes scales with their genomic separation. A linear relationship on a log-log plot implies a power law behaviour (reads $\sim s^{-\alpha}$) and exponents α are approximated using linear fits to different ranges of the log-data. (a) Genome wide data from Micro-C (green; obtained from Ref. [1]) and Micro-C XL (orange; obtained from Ref. [104]) experiments are shown on the same plot. Separations are measured in nucleosomes, so a value of 1 means adjacent nucleosomes. (b) A similar plot is shown but here separations are measured in kbp.

Both the Micro-C and Micro-C XL data give very similar results with the latter producing more interactions at longer length scales (as expected), but from a analysis standpoint, the twenty replicates of data combined for the Micro-C data give a cleaner data-set than the three replicates for Micro-C XL, at least for the focus on domains in this thesis. The close agreement in boundaries, domains and general structure allows for conclusions drawn from the one to be applied to the other. Since the focus of this study was on the micro-domains and not the longer-range interactions of Micro-C XL, the simulations were primarily compared to the Micro-C data.

3.5 Gene data and Pol II

In order to relate the nucleosome positions to the positions of genes along the genome, a data set was needed for this. The Saccharomyces genome project [2] provides a list of all verified open reading frames, i.e. gene bodies. To give an indication to the activity of the genes, it can be combined with Pol II ChIP-on-chip data from Kim et al [113]. This type of experiment uses probes on the genes to measure RNA polymerase II (Pol II) levels, which give an indication of mRNA transcription activity which in turn is correlated to gene transcription activity.

After removing all genes found on the Chromosome M (as this was not used) the data was cross-checked against the Pol II data to find the Pol II enrichment over each gene promoter. This was done using the bedtools software suit[3] .

With these data sets a list of genes ranked by activity could be constructed and the data could be divided into groups of high and low transcription rates. For specific investigations into the correlation between gene activity and nucleosome positioning, genes of lengths of less than 1kb were filtered out. Fig. 3.13 shows a distribution of all genes of >1000bp length with Pol II enrichment data. The Pol II data is scaled to a range of 0 to 10 and binned into 200 bins. The number of genes per bin can then be plotted and the 100 genes with lowest Pol II signal shown with green highlighting and the 100 genes with the highest Pol II signal with the blue highlighting.

[2] www.yeastgenome.org

[3] software available from: `http://bedtools.readthedocs.io`

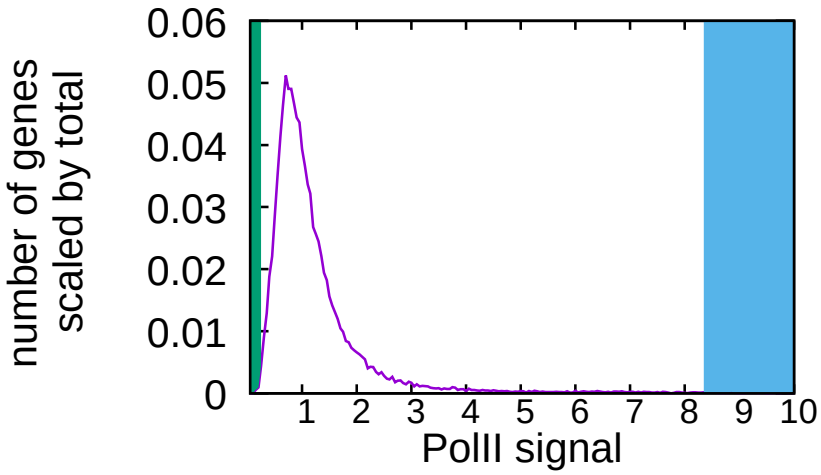


Fig. 3.13 Plot of Pol II distribution for all genes of >1000bp length. The Pol II signal is scaled to a range of 0 to 10 then binned (200 bins) and given on the x-axis. The number of genes found in that bin is given on the y-axis scaled by the total number of genes analysed. The green colouring to the left indicates the bins containing the 100 genes with lowest Pol II signal and the blue colouring indicates the 100 genes with the highest Pol II signal.

3.6 Nucleosome interactions around genes

For most eukaryotic genes, near the 5' end, the nucleosomes arrange themselves in highly ordered positions at fixed distances from the TSS. Alluded to in section:1.4 this has been studied in more detail in Ref. [61], but here I thought it to be meaningful to consider the nucleosome *interactions* around gene start sites. To this end, the gene activity data was combined with the nucleosome occupancy gained from MNase-seq data.

To prepare the gene activity data, some assumptions were made about the TSS of the genes, specifically that the TSS is at the start of the gene and the promoter region is directly upstream from this. The +1 nucleosome is taken to be immediately downstream or containing the TSS. Genes are divided into three categories, the 100 most active genes, 100 most inactive genes and 100 randomly chosen genes with a typical level of activity where the activity is given by Pol II $\log_2$(observed/expected) levels between -0.5 and 0.5.

Nucleosome occupancy is calculated by taking the MNase-seq reads from within a gene region. The occupancy around each gene is aligned and oriented at the TSS and the profile is averaged. By averaging over a certain gene region, cell-to-

cell variability can be taken into account as nucleosome occupancy data coverage might differ between regions. If the NucPosSimulator output had been taken, then the most likely conformation would only be considered as it only shows where the nucleosomes are and not the occupancy.

In the Fig. 3.14, the three regions are shown in the columns. The first row shows a plot of the mean nucleosome occupancy data for the regions as a function of distance from the TSS. The middle and bottom rows show the same interaction map data in separate ways. In the middle row, the interaction maps for nucleosomes around the TSS for each group are shown. The data is averaged over all genes within the group and the number of reads is given by the logarithmic colour scale. The bottom plots normalise the interaction maps by the expected genome wide average interaction strength for the given separation (colour scale is set to $\log_2$(observed/expected)). As such genome average interaction levels are displayed as white, whereas increased interactions are red and decreased interactions blue.

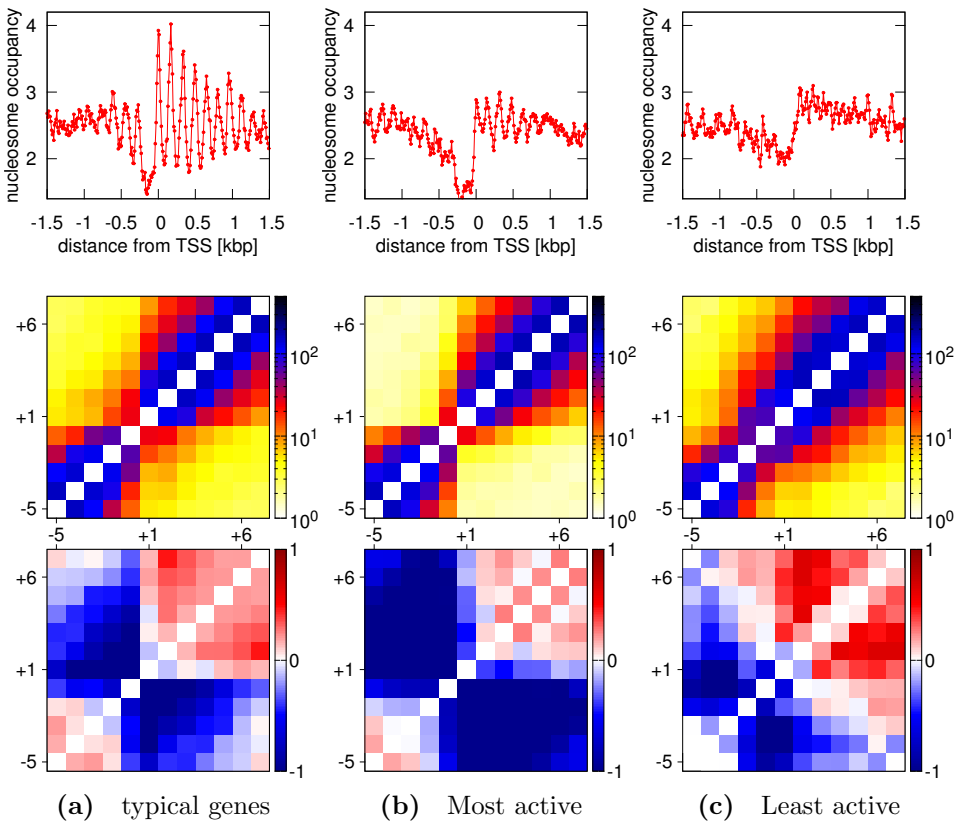


Fig. 3.14 Nucleosome occupancy and interactions around gene TSS comparison. Genes are divided based on their Pol II activity level scores (a) Typical genes - 100 randomly chosen genes with a typical level of Pol II activity. (b) the 100 most active genes and (c) the 100 least active genes. The first row in the figure shows the nucleosome occupancy value around the TSS whereas the second and third row show an interaction map for the nucleosome around the TSS.

From the mean nucleosome occupancy plots we can see that there is a clear pattern of positioned nucleosomes around the TSS, just as found by other studies. However, we can also see that the pattern is most clearly defined for the typical genes, with clear nucleosome positions downstream of the TSS and at least partially upstream as well. There also seems to be a clearly defined NDR before the TSS where at least one nucleosome is missing in the pattern. However a sample size of 100 genes might be too small to draw full conclusions. When considering the 100 most active genes, there is a clear indication of a larger NDR with only weak signals upstream of the gene. The signal within the gene is visible, but not as strong as with the typical genes. It could be because the genes are active and as such the DNA might be disassociated from the nucleosome for transcription. It is known that in more active regions there is a higher "turn-over" rate for the nucleosomes [76].

The least active genes are the least clear as there is only a small NDR, but only a weak signal of defined nucleosome positions upstream and downstream of the TSS. It appears as if the occupancy in the 500bp upstream of the TSS, is higher than in the other regions, indicating that there are more nucleosomes in the regions, i.e. it is not a NFR as such. This indicates that there is no strong alignment of nucleosomes around the TSS.

The average occupancy (not shown in the figure) within the genes is approximately the same for each set which suggests that it is the positioning of the TSS rather than overall nucleosome occupancy that is responsible for differences in signal.

The interaction maps in the middle and bottom rows of Fig. 3.14, show a clear distinction between the most active and least active genes, with there being less of a difference between the active and typical genes. There is a clear boundary between interactions upstream of the TSS and the gene body. In the most active genes the interactions are higher than average within the gene (top right quadrant), but comparable to average within the region, upstream of the TSS (bottom left quadrant). The interactions between the gene body and the upstream region are strikingly repressed (blue area in the off-diagonal quadrant, i.e. there are very few interactions between the gene body and the upstream region). It seems that the +1 nucleosome interacts less with the gene body than in the case of the typical genes and similarly the +2 seems to have less than gene average interactions with gene body.

For the least active genes, there is a clear above average level of interactions within the gene body. However, there is again a significant dip in interactions upstream of the TSS. There seems to be a genome average level of interactions right around the TSS, this might be in line with the thought of a nucleosome sitting at the TSS to "switch-off" the gene.

3.7 Boundary finding

As some of the subsequent work focuses on finding boundaries between small domains within the yeast nucleosome structure, a brief introduction into boundary finding algorithms is in order. Here I will outline two methods used in the literature, which were both initially used before settling on one of the two.

The principle behind finding boundaries in contact maps is to come up with some measure which can quantify when the interactions in one domain end and begin in another. The first method often used in the literature is the **directionality index** [91] where the interactions from a nucleosome forward and backwards are summed. The idea is that at a boundary, the last nucleosome in a domain will primarily have interactions with previous nucleosomes. Conversely, the first nucleosome in the next domain will have most interactions forward with further nucleosomes. For any given nucleosome, the interactions with previous nucleosomes are denoted by r_k and interactions with coming nucleosomes by f_k. For each nucleosome $x_k k$ the directionality index d_k is then found by taking the difference between interactions forward and backwards:

$$d_k = f_k - r_k,$$

where

$$f_k = \sum_{j=k+1}^{k+l+1} x_{kj} \qquad r_k = \sum_{i=k-l}^{k-1} x_{ik}.$$

The window of interactions l can be chosen so that it is big enough to clear small gaps between domains, but small enough to also detect smaller domains. A boundary can then be called when between two neighboring nucleosomes the directionality index changes from negative (mostly backwards interactions) to positive (mostly forward interactions).

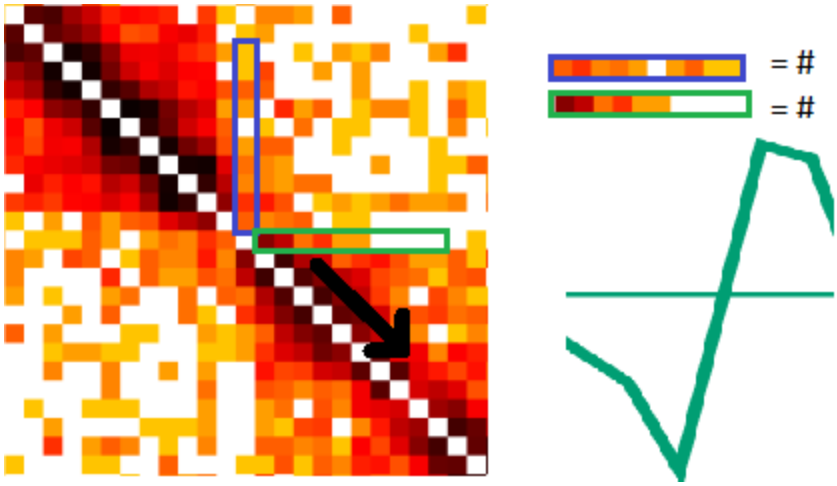


Fig. 3.15 Directionality index method - interactions between nucleosomes backwards (blue box) and forwards (green box) are summed up. The difference can then be plotted as seen on the right and whenever a change from negative (backwards) interactions to positive (forward) interactions is observed a boundary can be called.

This method was not used in the end since it gave good predictions with most boundaries, but it struggled with weak boundaries and boundaries that were close together. The issue with very weak and close together boundaries was that the signal would oscillate around zero, so to call a boundary, a minimum change from negative to positive would have to be defined as a cut-off. The alternative method allowed for more refinement work and for parameters to be set for boundaries that are close together.

The second method from the literature is the **sliding box** algorithm. Essentially a square box is placed off the diagonal of the interaction map (with its corner on $i, i+1$), and the values for nucleosome interactions falling within the box are summed. Then the box is slid along the diagonal, nucleosome-by-nucleosome to obtain a boundary signal as a function of box position. Symbolically the signal at nucleosome k (used to determine if there is a boundary between nucleosomes

k and $k+1$) is given by

$$s_k = \frac{1}{2l} \sum_{i=k-l+1}^{k} \sum_{j=k+1}^{k+l} x_{ij} \quad \text{for } l < k < N - l,$$

where x_{ij} is the number of interactions between the ith and jth nucleosome, N is the number of nucleosomes in the region and l is the window of interactions (10 in this case). At the edges of a region (i.e. $k < l$ or $k > N - l$) the same function is used but l is reduced, e.g. for $k = 4$ the window becomes $l = 3$. This is essentially the same as the algorithm used in Ref. [1], where for each nucleosome the number of upstream to downstream interactions (within some range) is counted.

This signal gives a measure the number of interactions between regions either side of a given nucleosome (the lower the value the fewer crossing interactions), thus minima in s_k are potential boundaries.

To call a boundary at a minima it is required that value of s_k is smaller than its local average by at least some threshold factor; hence it is defined to be:

$$\bar{s}_k = \frac{1}{9} \sum_{i=k-4}^{k+4} s_i,$$

and then if $s_k < \gamma \bar{s}_k$ a minima can be called a boundary. The value of the factor γ can be tuned by visual inspection of the called boundaries and the interaction map.

Due to noise in the data, occasionally the boundaries found by the algorithm are not the same as would be expected from visual inspection of the interaction maps: this highlights the difficulty in unambiguously defining domains and boundaries within 3C based interaction maps. To remove these we take the distribution of the strengths of putative boundaries across the simulated regions and discard any boundaries with a score s_k above the 90th percentile of that distribution (higher value means weaker boundary).

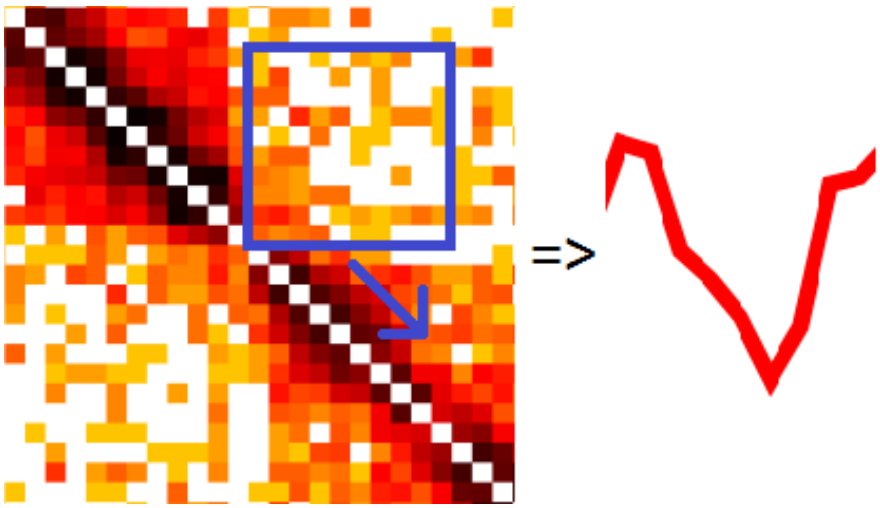


Fig. 3.16 Sliding box method, interactions within the box are summed and the box is slid along the diagonal giving rise to the curve on the right. This can then be inspected for local minima which are boundary candidates.

3.7.1 Comparing experiment and simulation

To compare the found boundaries between the experiment data and simulation data, I developed an algorithm which would analyse the two boundary data sets and attempt to align them in order to 'call' matching boundaries. An overlap value can be used to allow for stricter or more lenient matching of boundaries, i.e. a value that dictates by how many nucleosomes found boundaries are allowed to be apart to still be considered the same boundary. In the strictest case, the boundary would only be considered the same if it was found between the same neighbouring nucleosomes.

The algorithm makes use of a recursive method to compare the two boundary lists. Two values from the list are taken and compared, if two values are the same or within the threshold, then they are stored in the final results.

"Correct predictions" are all boundaries that lie within the threshold of each other. "Missing boundaries" are all boundaries that are found in the first file (generally the Micro-C experiment data) but not in the second, and conversely all boundaries found in the second (simulation file) but not the first are counted as "extra boundaries".

In cases where two simulated boundaries are within one nucleosome of a "file one" boundary this is counted as one correct prediction and one extra boundary; likewise if there is one simulation boundary within one nucleosome of two Micro-C boundaries, this is counted as one correct prediction and one missing boundary.

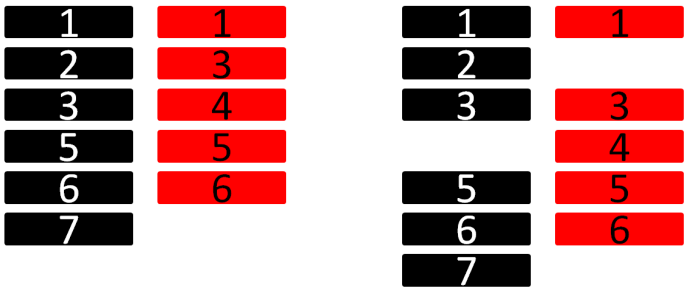


Fig. 3.17 When two lists of boundaries are compared, they are shifted so the correct nucleosomes match with each other. This then gives "correct" predictions (4 in this example), "missing" boundaries (red 2 and red 7) and "extra" boundaries (red 4).

3.8 Representative regions

For the simulations that follow, a number of regions were required for analysis as an entire chromosome would be too large to simulate efficiently. Initial tests were done with a region referenced in a figure of the Micro-C paper and later named region 0.

The regions on the final list all came from a much larger list and as such their names are not sequential. A total of seven regions were defined in addition to the test region in order to increase the number of simulated regions. The choice here was based on picking medium to long regions with a good mix of Pol II active and inactive regions and a high number of genes per region. Therefore, the overall aim of the regions was to give a good representation of the genome. These regions come from six different chromosomes and are between 15-43 kbp long to cover around ~ 240 kbp.

region	chrom	coords (bp)	bp	nucs	genes	counts
Region 0	chrVII	140680 - 155644	14964	82	8	18326
Region 1	chrI	39259 - 81951	42692	243	20	112973
Region 3	chrIV	1254937 - 1287938	33001	193	15	55690
Region 9	chrXI	86225 - 108599	22374	126	8	50177
Region 17	chrIV	267698 - 300003	32305	188	19	58220
Region 18	chrIV	832859 - 870563	37704	218	19	74086
Region 19	chrVIII	325598 - 352453	26855	158	15	59378
Region 24	chrXV	194970 - 222902	27932	161	16	74106

Table 3.2 Regions used to represent the entire genome.

Chapter 4

Computer simulations of yeast chromatin

In this chapter I will discuss my main computer simulation work on yeast chromatin fibres. The primary objective of this work was to make use of nucleosome positioning data to feed a simulation model which would allow to reproduce the contact maps generated from the Micro-C data in order to study the mechanism of formation of nucleosome level domains (micro-domains) in yeast.

The aim is to investigate how these domains might be formed and what determines the boundaries of the domains. Specifically what could allow a model to reproduce the domain patterns found in the Micro-C data from Ref. [1]. The data from this reference did not show the larger domains that had been observed in previous low-resolution studies [114], in particular the longer ranged centromere-centromere and telomere-telomere interactions. An improvement to the Micro-C method has been published in Ref. [104] and that data covers both short range and longer range interactions. Although both data sets are available, the primary focus of these simulations is based on the original data set [1]. This data set contained more replicates and as such was much larger in scope. Since the focus of these simulations is on short range interactions of the micro-domain level, the long-range interactions from the newer data were not required.

As a starting point of these simulations, the idea was to start with the simplest possible model to study the 3D organisation and more detail could be added into the simple model framework later on if required. However the simple model was already able to predict a number of features of the 3D organisation. Specifically

the locations of boundaries between the micro-domains can be inferred from only nucleosome positions as input to the simulation. This suggests that the irregular spacing of nucleosomes within the yeast chromatin may lead directly to the boundary formation.

Although there are similarities in how the Micro-C data and MNase-seq data is obtained, they focus on different key insights and differ in data contained. The MNase-seq data focuses on finding the positioning of the nucleosomes, whereas the Micro-C data focuses on the distance between nucleosomes. The nucleosome positions can be deduced from the Micro-C data and could be compared for the nucleosome positioning, however by using the nucleosome positioning to produce the model, the 3D distance relation can be related back from the nucleosome positions to the Micro-C contact map. As such the purpose of the structural model is to relate the positions of the nucleosomes to the interactions and the data from the Micro-C experiment serves to verify the results of the model.

The regions defined in Section: 3.8 are used for the simulations with the information most relevant for the simulations summarised in the table below. The coordinates in nucleosomes are derived from the first and last nucleosome found in the regions as determined from the MNase-seq data. The number of beads is the total number of DNA beads and nucleosome beads simulated.

The regions simulated are listed in Table 4.1 with the number of nucleosome found in the regions and the number of beads used in the simulation.

region	chrom - coords (bp)	coords (nuc)	beads
Region 0	chrVII: 140,680 – 155,644	802 – 883	467
Region 1	chrI: 39,259 – 81,951	211 – 453	1194
Region 3	chrIV: 1,254,937 – 1,287,938	6692 – 6884	845
Region 9	chrXI: 86,225 – 108,599	493 – 618	656
Region 17	chrIV: 267,698 – 300,003	1427 – 1614	813
Region 18	chrIV: 832,859 – 870,563	4496 – 4713	995
Region 19	chrVIII: 325,598 – 352,453	1766 – 1923	665
Region 24	chrXV: 194,970 – 222,902	1062 – 1222	731

Table 4.1 Regions used for simulations and contact map creations

4.1 Nucleosome model

The nucleosome model is an extension of the simple DNA model outlined in section: 2.5.2 with the inclusion of "nucleosome beads" joined together by "linker DNA beads". To keep things simple, the histone DNA complex known as a nucleosome is represented as a singular bead of 10nm diameter (equivalent to 4 σ in simulation units). Each nucleosome bead contains the equivalent of 147bp of DNA. In reality a nucleosome shape is more disk-like and further refinements to the model do include this.

4.1.1 Beads on a string

The addition of the nucleosome beads leads to a simple "beads-on-a-string" model, such as is depicted in figures 4.1 and 4.2. For this model there are **no interactions** between nucleosomes themselves and the DNA beads apart from excluded volume interactions. These steric interactions are given by the WCA potential Eq. (2.4). Bonds between DNA and nucleosome beads and between nucleosome beads are given by FENE bonds as before according to the Eq. (2.3). The value of R_0 changes depending on the bond, hence between two nucleosomes it is set to $R_0 = 5.6\ \sigma$ and between a DNA bead and a nucleosome it becomes $R_0 = 3.6\ \sigma$.

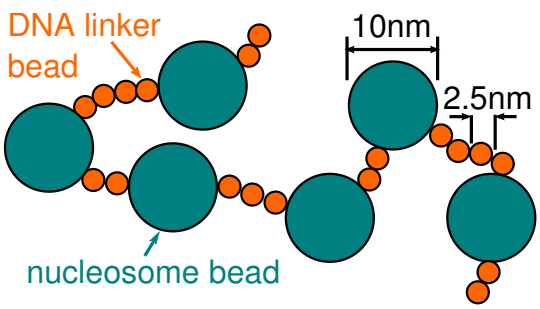


Fig. 4.1 Schematic representation of the simple nucleosome model as "beads on a string", with DNA beads as 2.5nm beads and nucleosomes as 10nm beads. Beads are connected by springs with bending rigidity.

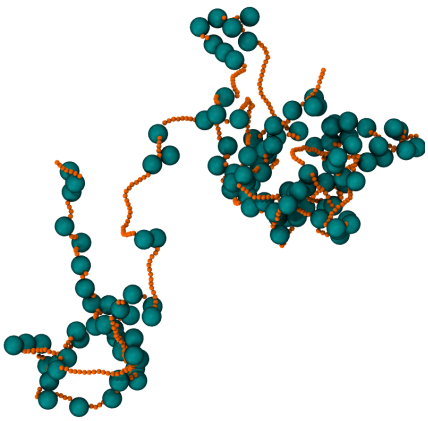


Fig. 4.2 Snapshot render of the simulated model fibre.

The model is set up to not include any orientational or bending constraints between DNA linker and nucleosomes. In practice this means that when there is a bead succession of DNA-nucleosome-DNA, this acts as a freely rotating joint. Similarly nucleosome-nucleosome joints are freely rotating as well.

A number of different coarse grained models for chromatin have been used in the past to study the yeast genome. Some have resolved individual nucleosomes as simple spheres without linker DNA [115] while more detailed models of chromatin at a nucleosome level included a complex nucleosome geometry with surface charges and histone tails [116–119]. There are some models in which nuclear organisation was studied with much lower resolution representations of the chromatin fibre, allowing whole nuclei to be simulated [120, 121]. The model used in this work is a simpler and less computationally expensive model.

The aim of the work in this thesis, is to be at a level of detail which sits between that of these two previous approaches. With this approach, large chromosome regions spanning multiple domains can be simulated as well as their internal structure resolved. Importantly, and unlike most previous work, this model also includes realistic nucleosome spacing, which plays a key role in determining domain features. More detail on this is found in section 3.2.

4.1.2 Simulation parameters

The simulation is run by using the LAMMPS software [108] which uses a standard velocity-Verlet algorithm to solve the Langevin Equation Eq. (2.1) describing the system.

In order to relate the simulation units to real physical units, a natural simulation time unit can be defined from the length units of σ=2.5 nm, and energy units of k_BT. The masses are given in units of the mass of a DNA bead, approximately 8×10^{-24} kg, and a choice of $K_{\mathrm{BEND}} = 20\ k_BT$ therefore gives a realistic DNA persistence length of $l_p = 20\ \sigma = 50$ nm. The natural time unit then becomes $\tau = \sqrt{m\sigma^2/k_BT}$. A time step of $\Delta t = 0.005\ \tau$ is used in the simulations.

Another important time scale is the Brownian time $\tau_{\mathrm{B}} = \sigma^2/D_i$, which is the time scale over which a DNA bead diffuses across its own diameter σ. Here D_i is the diffusion constant for bead i, given through the Einstein relation by $D_i = k_BT/\gamma_i$ where γ_i is the friction or drag of the system. With the choice of $\gamma_i = 1$ this

means $\tau_B = \tau$. To map to real times the mean squared displacement (MSD) for all beads can be measured. The experimental results from Ref. [122], who measured the MSD for various chromatin loci in live yeast cells, can then be used to find a best fit of the value of τ_B. This results in $\tau_B = 80$ μs, meaning that each 10^7 time step simulation run represents approximately 4 s of real time.

The nucleosome positioning data from the NucPosSimulator software is used to create a sequence of nucleosome beads interspersed with DNA beads for each of the simulated regions. The simulations themselves are then initialized in a way that the nucleosome and DNA bead positions follow a random walk. The visualisation of this random walk configuration led in part to the investigations of linker lengths between nucleosomes covered in section: 3.2. The random walk was chosen for its simplicity in creating a starting configuration and any overlap in the fibre as well as unlikely conformations could be removed by the equilibration phase. Equilibration is possible as the yeast chromosome is considerably shorter than most eukaryotic chromosomes. Although long eukaryotic chromosomes do not have the properties of an equilibrated polymer melt, the sections simulated in my work are even shorter still. In the paper by Rosa and Everarers[123], the author show that simulations of yeast chromosomes do in fact equilibrate over time. This result allows my simulations to be run to an equilibrated state for analysis. And although for eukaryotic chromosome polymers the same cannot be said, it is however a reasonable approximation for short polymers as simulated here, that they will equilibrate.

An initial run of the system for 122 τ evolves the dynamics to obtain an equilibrium polymer conformation. During this equilibration run of the system, the potential between beads is replaced by a soft potential to remove any overlap that might be introduced by the random walk starting conformation. The main simulation run is then a further 50×10^3 τ and the system configuration is saved every 250 τ. For each region this was repeated 20 times with a different initial starting configuration and a set of random numbers to generate the noise $\boldsymbol{\eta}_i(t)$. Periodic boundary conditions are used, with a simulation box size of $400\times400\times400$ σ, meaning that the system is dilute.

The simulations were run on multi-core computer clusters and each individual simulation gave 200 output files with the configurations of the beads. Of these I took the latter set of 100 out of 200 and used these to generate interaction maps as detailed below. This gave a total of 2000 conformations across all simulated regions.

4.1.3 Simulation maps

Initially the contact maps created from the simulations used the standard approach of taking the 3-dimensional data from the beads, calculating distances between them and using a cut-off to determine a contact between two nucleosome beads. The different conformations where combined and averaged to create a single output contact map. This gave promising contact maps, but in comparison to the experimental maps, the signal in the simulation maps was much stronger making comparison difficult, i.e. unless a perfect cut-off distance was chosen the number of contacts was much larger and on a different scale than for the experimental map. It might have been possible to mitigate this by scaling the result of the simulation by the experimental map, but instead the program was altered to mimic the experimental process and give simulation interaction results on the same scale as the experiment.

Although the standard distance threshold method would give us the true interaction map from the simulation data, this is not necessarily the data that is wanted and not necessarily very close to the data provided by the experiment. Instead the novel method was used to mimic the Micro-C method in how the interaction data is produced. The new stochastic method allows for a variation in "cross-linking length" as the cells are fixed using formaldehyde and there is no fixed length over which the formaldehyde would bind proteins and nucleic acids to each other. It binds those which are "close together" where this is not a fixed distance. It has been suggested that formaldehyde molecules can polymerise and as such increase the interaction distance[124]. As such a fixed interaction threshold is not a accurate representation for the experiment.

In the experimental procedure of Micro-C experiments nucleosomes are cross-linked and unprotected DNA is digested and protected DNA fragments are ligated. Hence the expectation is that the probability of DNA fragments being ligated is linked to their 3D separation, i.e. the process is stochastic as ligated fragments are pulled down with some probability. Random ligation events are not excluded from occurring during the protocol and other processes in the experiment are also stochastic in nature.

Another reason for using a stochastic method is that it has the potential to replicate the pairwise interactions of Micro-C. With the threshold method, the pairwise interactions can break down for cases where nucleosomes can be close to more than one nucleosomes. For example if a nucleosome A is always close to B and C, D and E are close to each other, then pairwise interactions between A-B would show up correctly in both methods, however interactions between C-D, C-E and D-E would occur at only 1/3 the rate of A-B for each cell. With the stochastic method, this behaviour should be reproduced. A similar method was also used by Buckle et. al. in [125].

Therefore in order to mimic this experimental process *in silico* two random nucleosomes are picked from a configuration and with a probability $P(r)$ this is accepted as an interaction or rejected otherwise. The probability is a Gaussian shaped function $P(r) = e^{-r/l_c}$ of the nucleosome separation r. The interaction length scale is set at $l_c = 15$ nm (6 σ) or 1.5 times the nucleosome bead centre-to-centre distance. This operation is repeated N^2 times for a simulation of N nucleosomes. For ease of comparison to the experimental maps, this process can be repeated to give a simulated number of interactions that is close to the number of Micro-C reads in a given region. The final step is then to take the simulated number of interactions and apply a scaling factor γ to bring the interaction number exactly in line with the total number of reads for the interaction maps of the experiment.

This produces contact maps such as the representative map below in Fig. 4.3.

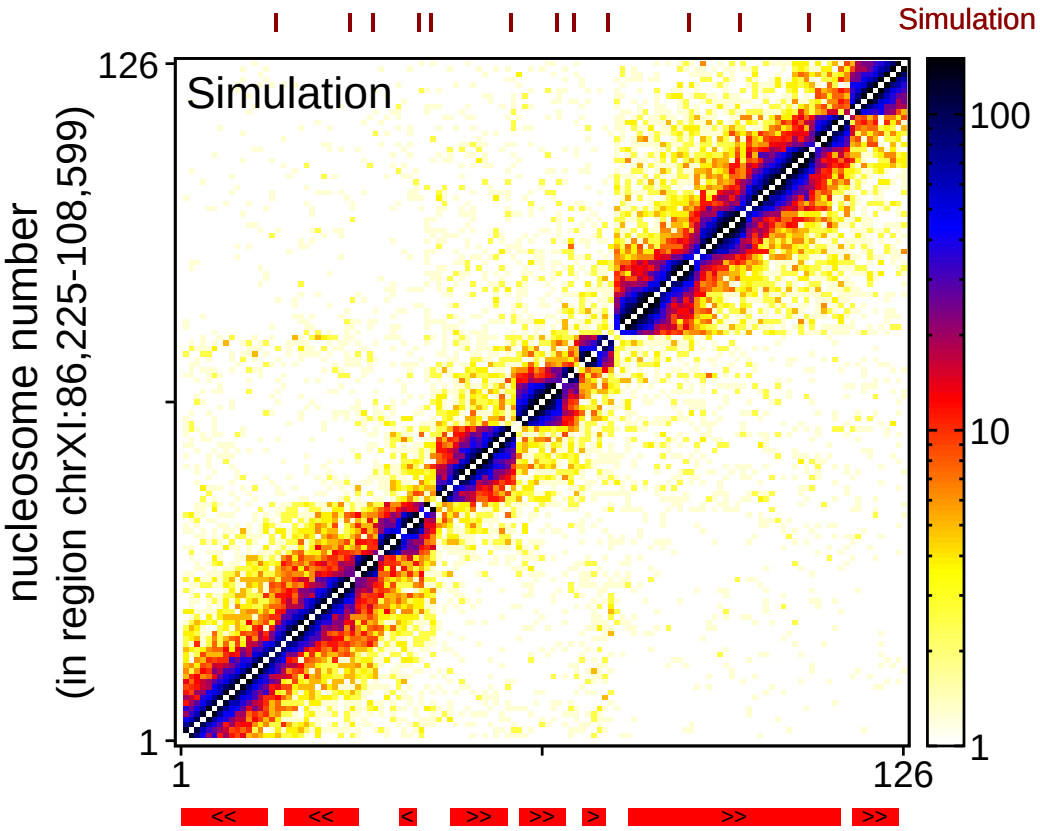
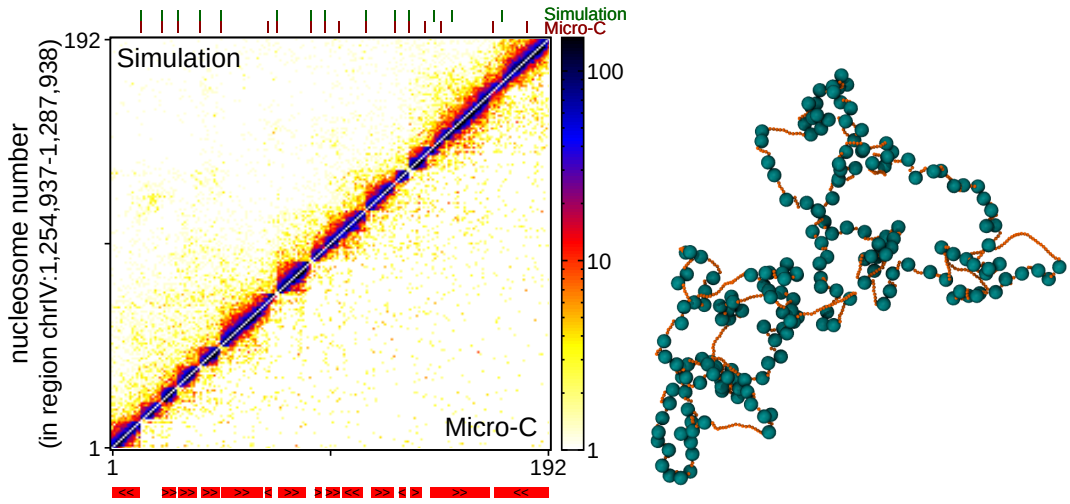


Fig. 4.3 Nucleosome resolution contact map for a region at chrXI: 86,225 – 108,599 produced from simulation data only (render or fibre given in Fig. 4.2). The nucleosome number is given on the axis as this is a nucleosome resolution contact map (nucleosomes start at 1). Below the map the genes found in the region are given mapped to the closest nucleosome and above the figure the called boundaries in this region are shown.

Though this model is simple, as it treats nucleosomes as spheres, rather than a more realistic disk-like shape, and it ignores the complex inter-nucleosome interactions mediated by histone tails, surprisingly it captures sufficient detail to correctly predict many features of short-range nucleosome contacts in 3-D. Inspection of the map clearly shows a reproduction of the micro-domains found in the experimental maps and a clear definition of boundaries. Both features are compared to the Micro-C map below.



(a) Contactmap

(b) Render of the fibre conformation

Fig. 4.4 (a) Contact map comparison for the region at chrIV: 1,254,937 – 1,287,938, where the upper triangular part is generated purely from simulation data and the lower triangular is generated from Micro-C data. Boundaries are shown for both data sets above the contact map. (b) A representative render of a conformation of the region shown in the contact map.

To compare the experimental data with the simulation results, for each region of simulated chromatin conformations a map of *nucleosome-nucleosome* interactions is generated, which is then combined with a Micro-C map as in Fig. 4.4. The figure shows a contact map for a 33 kbp region of the yeast genome (chrIV: 1,254,937 – 1,287,938). A sample configuration is given to the right. The figure shows the simulation results in the upper left half of the contact map and the Micro-C results are in the lower right half.

As mentioned it is clear that close to the diagonal, the two maps are extremely similar, only further from the diagonal do minor differences become apparent. To quantify the similarities and differences between the maps, the domain boundaries can be called using the algorithm described in section 3.7. Fig. 4.5 considers the boundary finding in more detail and comparative maps for all regions follow after.

4.1.4 Micro-domain boundaries

Fig. 4.5 shows the same region as Fig. 4.4, but in this case the contact maps have been rotated and longer range interactions were cut off. The top shows the simulation map and the bottom the Micro-C experiment. In between the boundary calls are shown with corresponding to each map. Using the boundary

finding algorithm, a total of 17 boundaries are found in the Micro-C data and 14 are found for the simulation, of these 11 match the Micro-C boundaries giving a correct prediction. Although 64% of the boundaries are correctly identified, an extra 3 boundaries are predicted and 6 boundaries are not found. Visual inspection throws some doubt on some of the extra boundaries and their validity, but computationally they are valid given the criteria set for the algorithm.

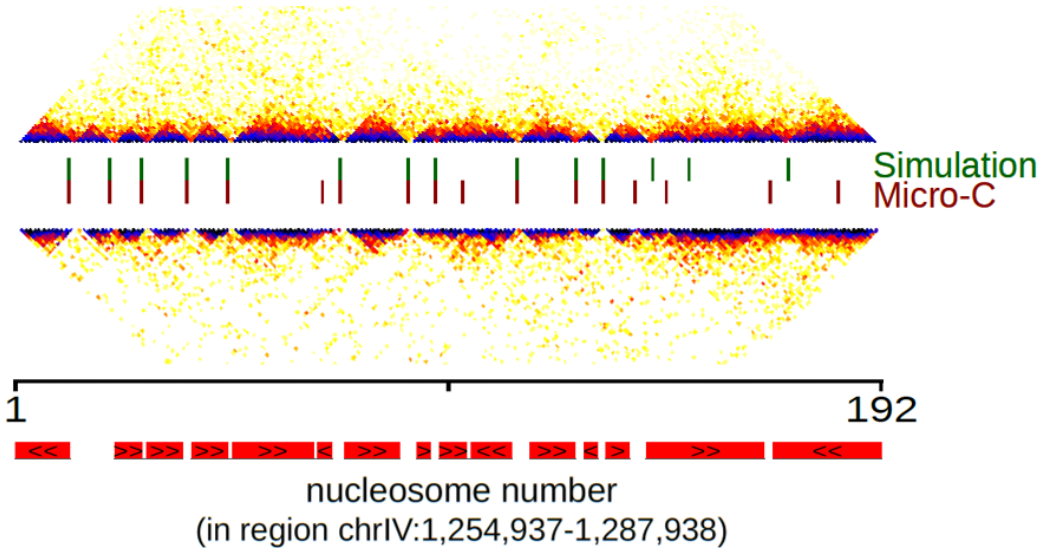


Fig. 4.5 Same contact map as in Fig.4.4, but rotated and cropped to focus on the diagonal and the found boundaries. The top half shows the simulation and bottom half the Micro-C data with the respective boundaries given in between as called by the algorithm discussed in section 3.7. The same algorithm with the same criteria is used for both cases.

When considering all simulated regions (Fig. 4.7 & 4.8), which cover a total of 240 kbp, the simulations are able to predict 99 out of 119 boundaries (83.2%). An additional 31 boundaries were found in the simulation contact maps bringing the total percentage of correct simulation boundaries to 76.0%. The Venn diagram in Fig. 4.6 shows a summary of the boundaries found between the simulations and experiment.

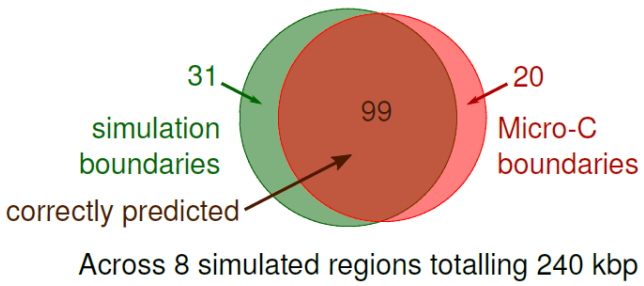


Fig. 4.6 Venn diagram showing the boundary finding results of all 8 regions. Of the found boundaries, 83.2% are correctly identified, with 31 boundaries being found by the simulation but not matched to the Micro-C data.

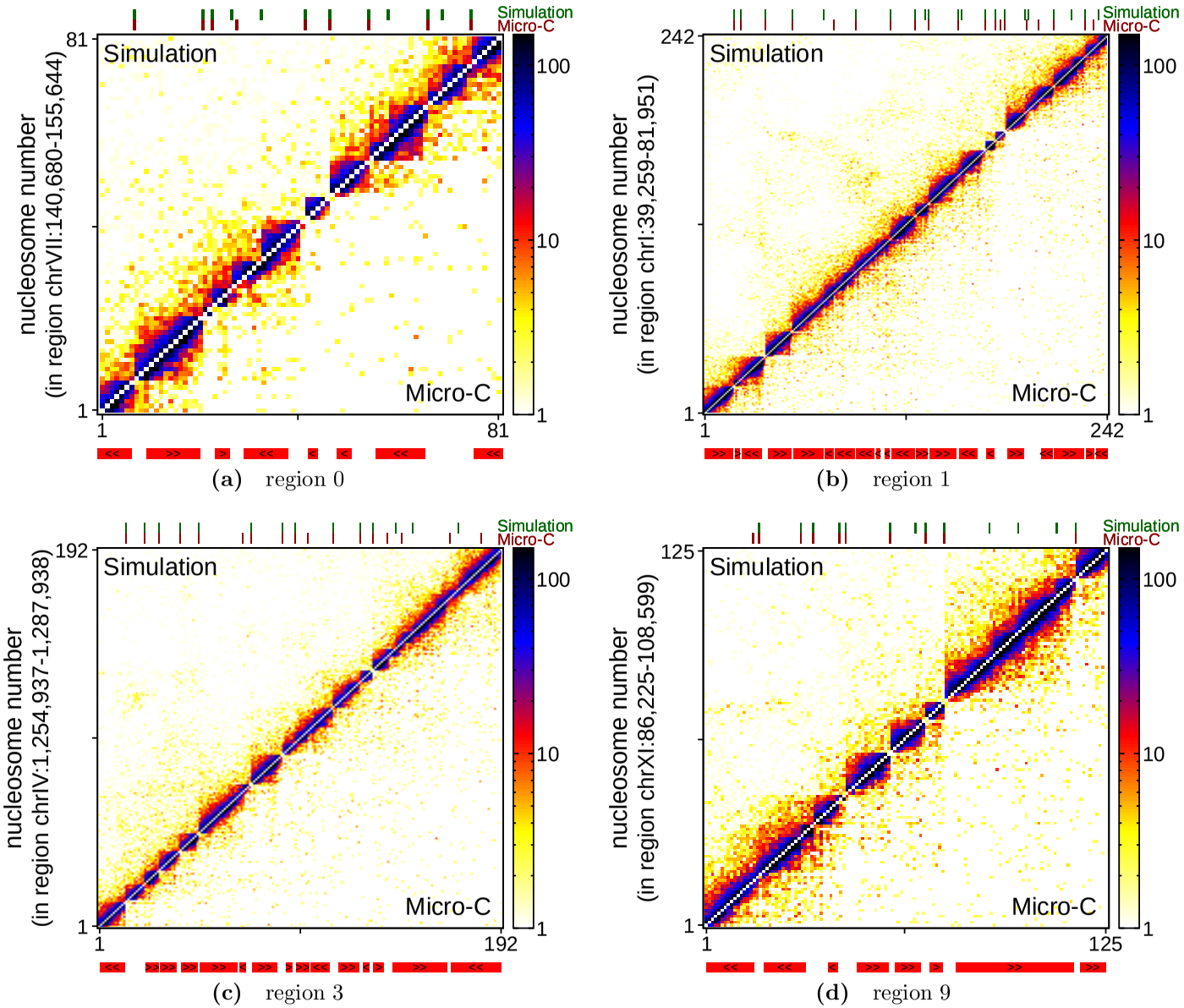


Fig. 4.7 Comparison contact maps of the simulation data against Micro-C data for regions 0, 1, 3 and 9.

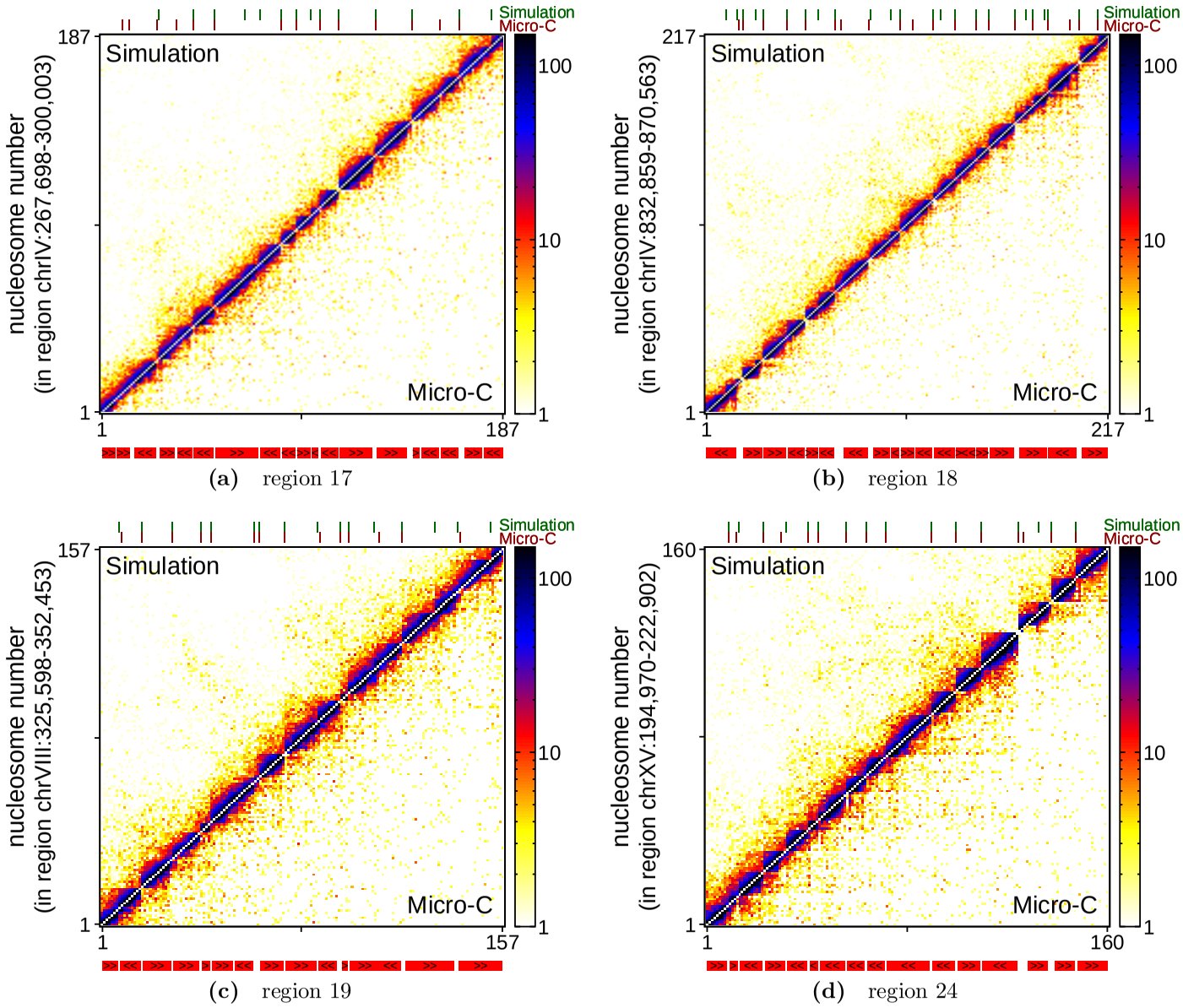


Fig. 4.8 Comparison contact maps of the simulation data against Micro-C data for regions 17, 18, 19 and 24.

Another measure that can be used to measure the level of chromatin interactions is the "insulation signal", a measure of the interactions between regions on opposite sides of a nucleosome. This value can be calculated from the value s_k used in the boundary finding algorithm (see section 3.7) by scaling its mean value across all regions and taking a negative log:

$$u_k = -\log\left(\frac{s_k}{\langle s_k \rangle}\right),$$

Hence the insulation signal is lower if more interactions are observed between the regions either side of a nucleosome. In Fig. 4.9 the insulation signal is plotted in a simulation vs Micro-C data relation. In the left figure all scores are plotted, whereas in the right figure, only insulation signals at correctly predicted boundaries are shown. To quantify the correlation between the data sets, the Spearman rank correlation coefficient can be calculated and it is given in each figure. Comparing the simulated and Micro-C insulation signals, a correlation coefficient of $r = 0.62$ ($p < 10^{-10}$) is found and this increases to $r = 0.76$ ($p < 10^{-10}$), if only boundaries are considered.

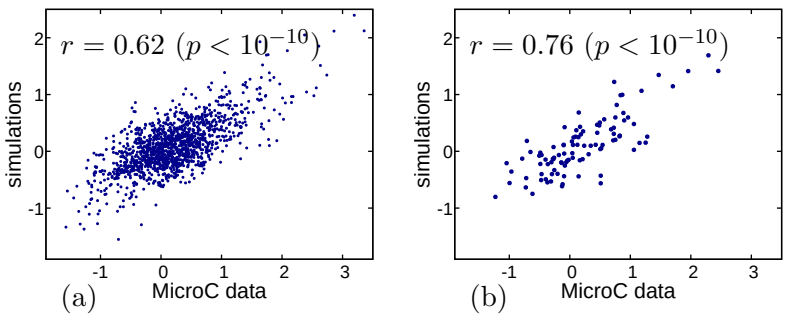


Fig. 4.9 Comparison of the insulation score between the Micro-C data and the simulation data. (a) Shows a scatter plot of the insulation signal for each nucleosome in all the simulated regions. (b) Shows a similar plot, but only nucleosomes at which a correct boundary was identified are included.

The predictions of the model concerning the boundaries are surprisingly accurate considering how simple it is. Additionally, there is a definite similarity between the contact maps. Since the only input data for the model are the nucleosome positions, it is an obvious assumption that these are a primary driver for chromatin interactions at this scale. The boundaries seem to, at least in part, come about through wider spacing of the nucleosomes when compared to the

genome average. The linker lengths for different boundaries are plotted in Fig. 4.10 in addition to all linkers in the regions. Compared to the average overall linker length within the regions of $\sim$ 28 bp, the linker length at the boundaries found in the Micro-C data are about $\sim$ 117 bp. Conversely, the "missing" boundaries, i.e. the boundaries found in Micro-C but not by the simulations, are situated at shorter linkers of on average $\sim$ 25 bp and the "extra "boundaries found by the simulation but not in the Micro-C data are at longer linkers of $\sim$ 96 bp. The overall average linker length for the boundaries found in the simulations is $\sim$ 130 bp, meaning that the boundaries found in the simulation are primarily a feature of the nucleosome spacing, whereas micro-domain boundaries in the Micro-C data are a feature of the nucleosome spacing as well as other factors.

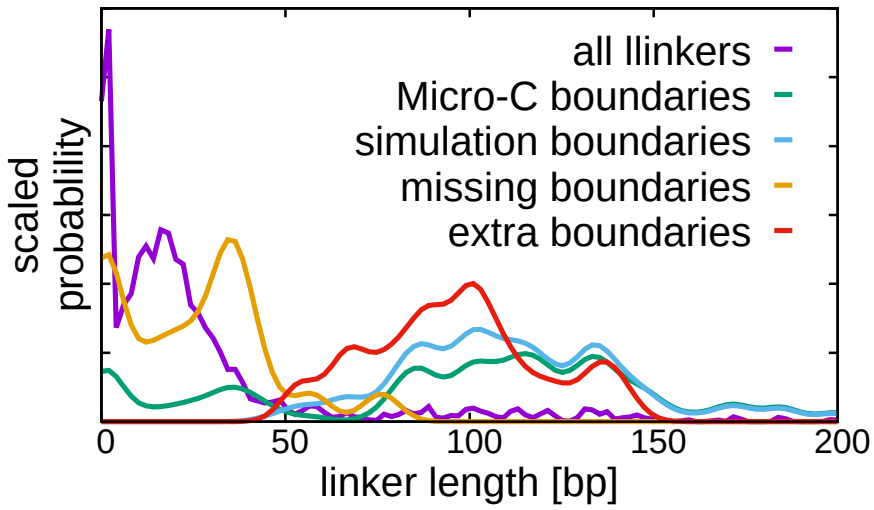


Fig. 4.10 Plot showing the boundary signal linker lengths. Here all the linkers are given (purple), all Micro-C boundaries (green), all simulation boundaries (blue), all missing boundaries, i.e those found by Micro-C but not the simulations (orange) and all the extra boundaries found by the simulation, but not the Micro-C data (red).

Contact maps can also be created from the simulation data for comparison with the Micro-C XL data. By altering the "cross-linker length scale" to $l_c = 26.25$nm (10.5σ) the longer range interactions visible in the Micro-C XL data are captured. A comparison figure is shown in Fig.4.11 with all other maps given in Fig.4.12 on page 74. An important point to make is that the domain structure found in the Micro-C XL data is very similar if not identical to the Micro-C data. A similar boundary analysis can be done to this data and the results are very similar to that of Micro-C, with 92 out of 110 Micro-C XL boundaries (84%) correctly identified. However a larger number of 42 "extra" boundaries are found by the simulations

and only 18 “missing” boundaries are found by Micro-C only, summarised in Fig.4.12h.

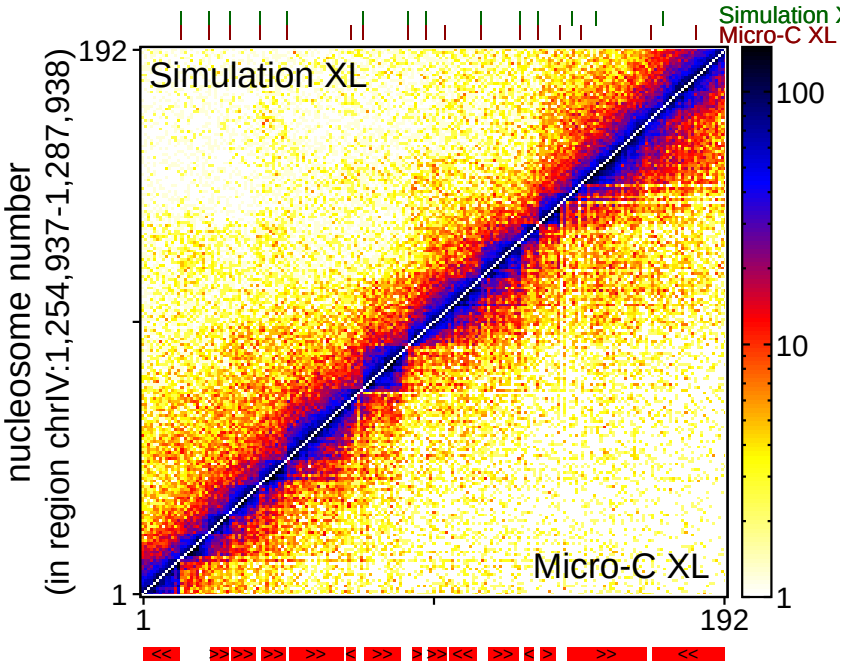
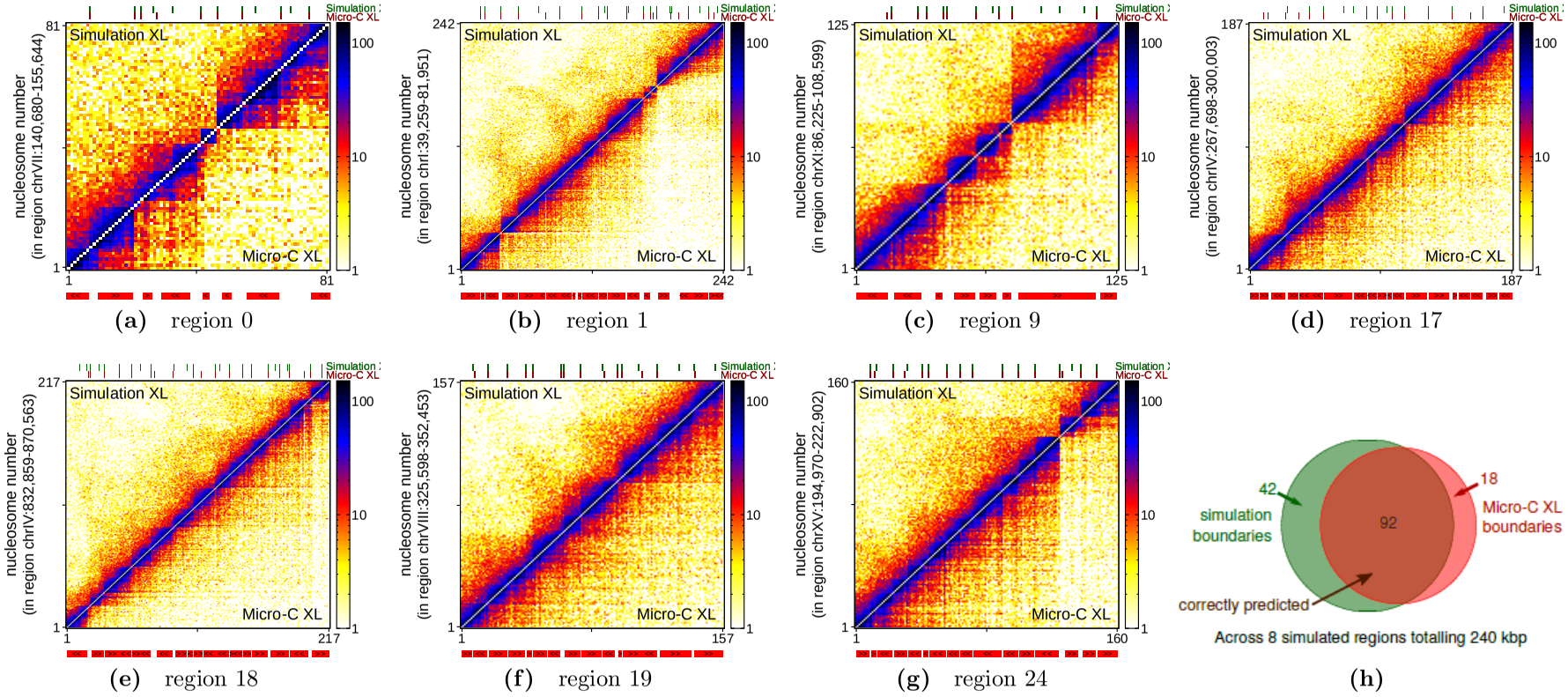


Fig. 4.11 Contact map comparison between Micro-C XL data and simulated long range cross-linker data.



(a) region 0 **(b)** region 1 **(c)** region 9 **(d)** region 17

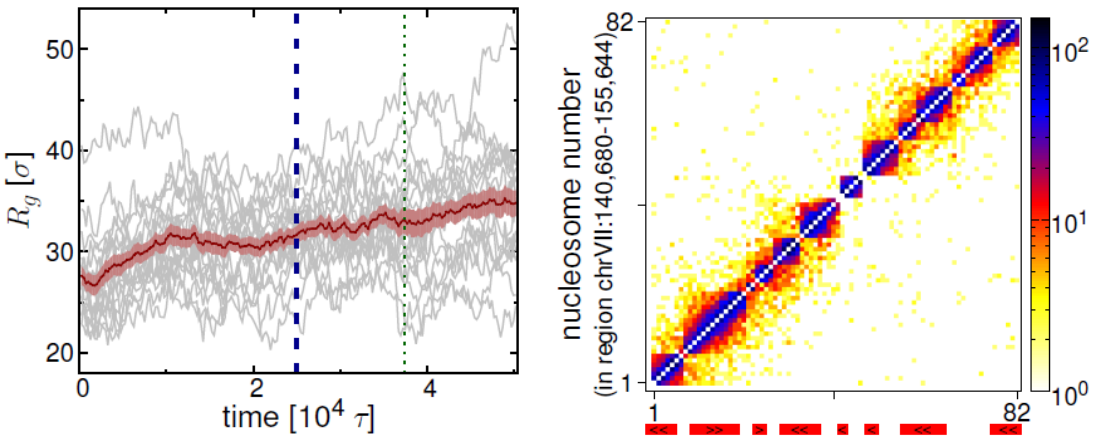
(e) region 18 **(f)** region 19 **(g)** region 24 **(h)**

Fig. 4.12 All contact maps created using Simulated XL data and Micro-C XL data.

4.1.5 Radius of gyration

A downside of the data from the experiments is that it only presents a picture of interactions at a population level. The simulations however allow for a much deeper analysis as a full configuration of a model chromatin fibre is available and more detailed measurements can be made.

As mentioned before in section: 1.7.2, the radius of gyration R_g is a good measure for the dimensions of a polymer. By using the output from the simulations, the positions of the particles can be used to calculate the radius of gyration for that time frame. By tracking the radius of gyration as a function of time (Fig.4.13a), changes in the conformations of the modelled system can be tracked. Once the system stops systematically changing (right of blue line), an equilibrium configuration is reached. When plotting a contact map (Fig. 4.13b) from the conformations between the blue and green line (upper triangle) and comparing it to conformations taken from the green line to the end (lower triangle), there is no appreciable difference in structure. This measure was one of the factors that informed the decision as to from what time-point in the simulation the data could be used for all the plots shown so far.



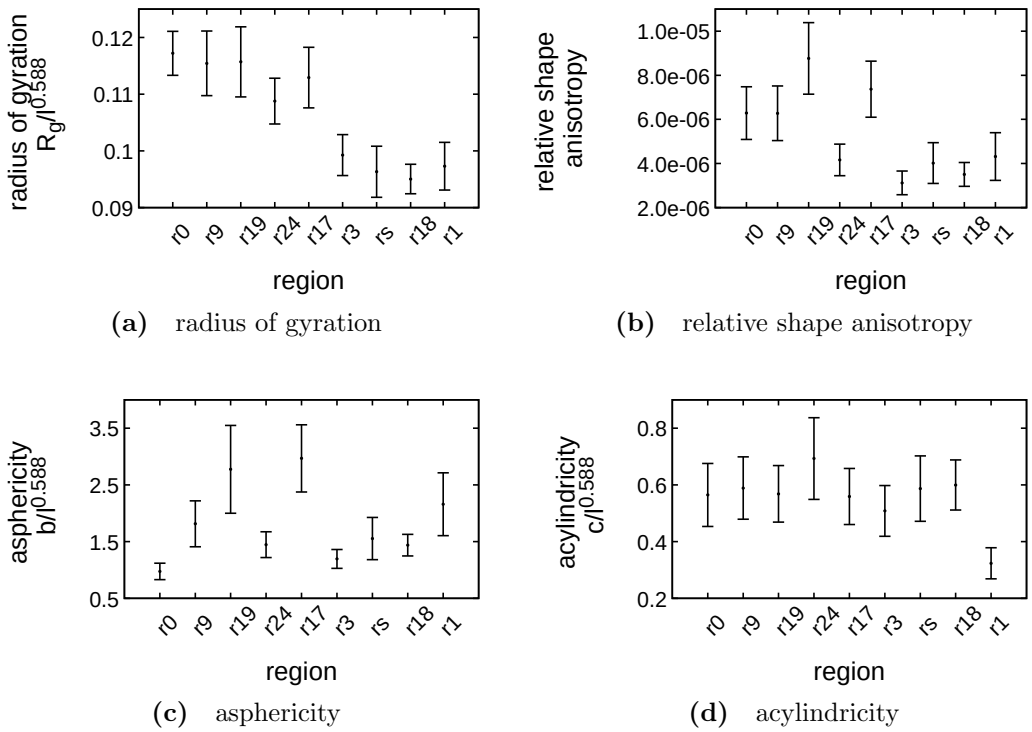
(a) Radius of gyration over time.

(b) Comparative contact map.

Fig. 4.13 The radius of gyration as a measure of equilibrium for the simulation, where in (a) the radius of gyration for simulation of region 0 (chrVII: 140,680 – 155,644) is shown as it varies with time. Grey lines represent 20 independent simulations with red giving the mean and error as shaded region. Configurations for analysis are taken from the right side of the blue line. In (b) a contact map is given where the conformations between the blue and green line form the upper triangle and conformations taken from the green line to the end form the lower triangle.

The conformation data from the simulations can be used to calculate the R_g for the fibre in question. As well as calculating the R_g for the entire polymer, the software can also calculate the R_g for a specific subsection of the fibre. This subsection can either be a specific gene of interest or particular section (eg. around the TSS) within the simulated region or a sliding window which calculates the R_g for a fixed window size as it moves along the polymer chain. All values for the R_g are calculated from a sample of time steps of the latter half of the simulation output. Calculations are done by making use of the Gyration tensor as outlined in section 1.7.2, as this also allows for the calculation of the asphericity, acylindricity and relative shape anisotropy.

Fig. 4.14a shows the average R_g for all the simulated fibres, scaled by the $l^{0.588}$ in bp of each region with to allow for comparison. The exponent comes from the fact that the characteristic size of a polymer chain scales according to a power law and the exponent for a polymer in good solvent is given as 0.588. Scaling by this factor should cause the R_g of a self-avoiding polymer to remain constant and any deviations are due to the irregular spacing of nucleosomes in the fibre. The ordering of the regions is from shortest to longest and this shows a potentially interesting correlation between fibre length and R_g. It appears as if the R_g decreases as the fibre length increases. This is discussed further when considering the comparison of a realistic fibre with an artificial regularly spaced fibre. However, it would make sense for the R_g to decrease as the length of the polymer increases as the local irregular nucleosome spacing begins to average out over the entire length of the polymer. This would also explain why the two regions longer than the regularly spaced fibre (rs) have values of R_g very similar to the regular spaced fibre. In Fig. 4.14b the relative shape anisotropy for each fibre is plotted, as the value is independent of length by definition it is not scaled by the length as the other parameters. The anisotropy is a measure of how spherical (0) or linear (1) the fibre is, here most fibres are much more spherically symmetric than linear. The irregular spacing does seem to affect the shape of the fibre as the spacing for regions 19 and 17 cause them to more aspherical/linear than regions 0, 24 and 3 for example. This is confirmed by the asphericity in Fig. 4.14c, a measure of how spherically symmetric a fibre is. In both cases regions 19 and 17 seem to be less spherical than the other regions. The final measure, acylindricity in Fig. 4.14d is the measure of how cylindrically symmetric a fibre is and the longest region 1 seems to be the strongest in this category.



(a) radius of gyration **(b)** relative shape anisotropy

(c) asphericity **(d)** acylindricity

Fig. 4.14 Different measures for the geometric shape of a fibre with the radius of gyration (a), the relative shape anisotropy (b), the asphericity (c) and acylindricity (d). Each regions value is plotted where the region names are abbreviated (rs for regularly spaced) and in all cases except the anisotropy, the measure is scaled by the length of the region in bp. Errors are given by the std. deviation.

The variations of the R_g within a fibre can be better understood when considering a window sliding along the fibre, for each bead it encounters, the R_g is calculated within this window. As the window slides along, variations in the fibre can be observed. This is shown in Fig. 4.15top, where for region 1 (chrI: 39,259 – 81,951) the window R_g is plotted. There are a number of level areas which are much better understood when considering the number of nucleosomes found within the window. In fact, it appears as if the R_g increases with a high number of nucleosomes in the window and decreases with a low number of nucleosomes, but if no nucleosomes are in the window, the R_g levels off at an average. This also is considered in more detail below.

An interesting point to investigate using the R_g is how the "realistic" fibres with irregularly spaced nucleosomes compare to fibres with regularly spaced nucleosomes. For this, one of the regions with irregularly spaced nucleosomes (chrIV: 1,254,937 – 1,287,938) was taken and compared to a artificial fibre with regular nucleosome spacing and linker length of 22 bp and similar total length. In

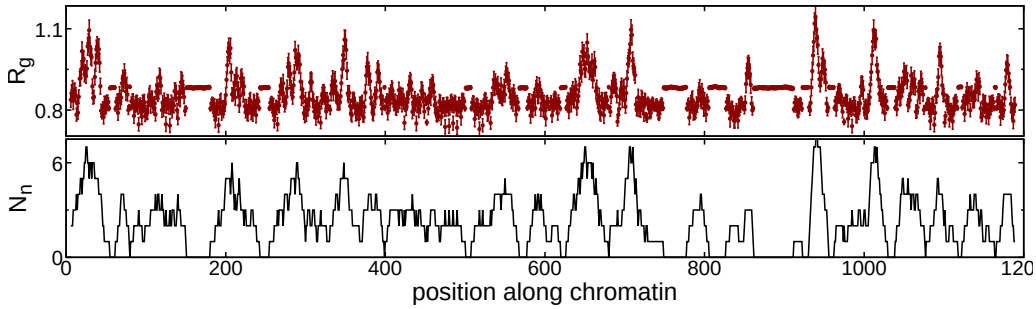


Fig. 4.15 (Top) The R_g for region 1 in chrI: 39,259 – 81,951 as found by a sliding window of size 11 beads along the length of the fibre. (Bottom) The number of nucleosomes found within each window.

Fig. 4.16, the configurations are visualised and a few differences (apart from the nucleosome spacing and sections with less nucleosomes) are visible. It appears as if the regularly spaced fibre is overall a bit more extended, whereas the realistic fibre has more bends that cause it to go back on itself to form a larger complex structure instead of a worm-like structure. What causes these differences can hopefully be understood when considering the R_g in more detail.

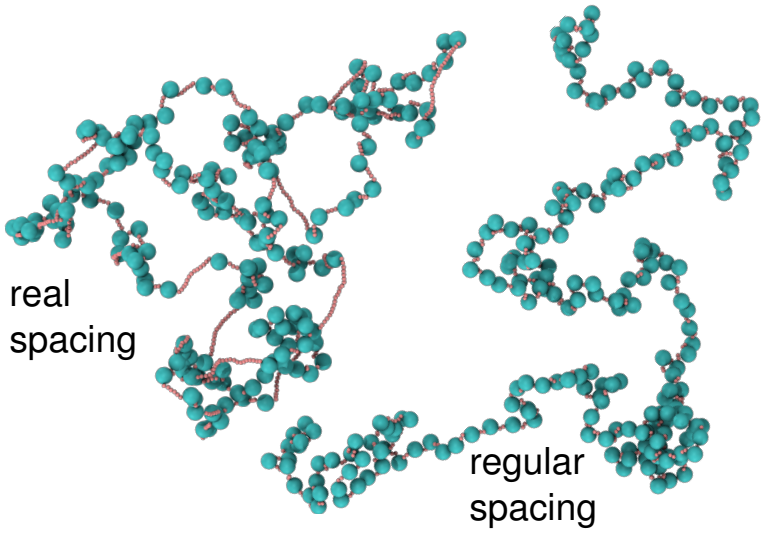


Fig. 4.16 Comparison of a render between a "realistic" fibre with irregularly spaced nucleosomes (left) and a fibre with regularly spaced nucleosomes (right)

In Fig. 4.17 the radius of gyration (as a measure of the size of the fibre) is plotted for both cases as a function of fibre length. This is calculated by sliding a window of size L beads along the entire length of the fibre and the bead types (Nucleosome and DNA beads) are treated as the same. So the R_g for the first L beads of the fibre is found, then the window moves to beads 2 to $L+1$, then beads 3 to $L+2$, and so on. The value of R_g for the fibre is then found by averaging over all the windows of length L and over the interval snapshots from the simulations. The value of L can then be varied to create the plot.

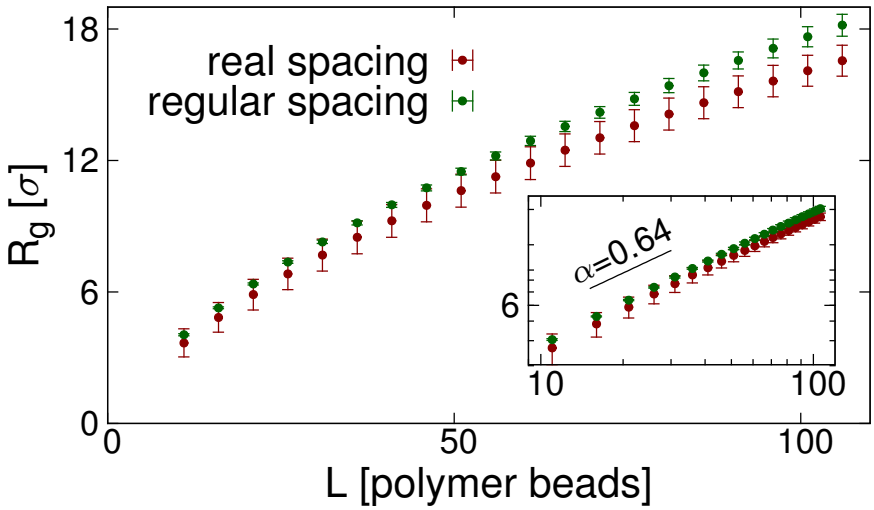


Fig. 4.17 Plot of the radius of gyration, R_g, as a function of the polymer length L, with the error bars giving the std. deviation. With an irregular spacing, the size of the polymer is reduced. The inset shows the same data as a log-log plot with the value giving the exponent of a power-law relationship exponent for the real spacing.

So when relating the R_g of the fibre as a measure of its volume, we find that the irregularly spaced fibre is smaller than the artificial fibre. In fact, the R_g reduces by about 10% between the cases. An interpretation of this would be to assume a decrease in the effective persistence length or stiffness of the fibre. A power law can be fitted to the data and a similar exponent is found for both cases. For the irregularly spaced nucleosomes $\alpha \approx 0.64 \pm 0.0057$ and for the regularly spaced nucleosomes it is $\alpha \approx 0.67 \pm 0.0034$. It seems as if these might be finite N crossovers to the literature value expected for large N for a polymer in a good solvent ($\alpha \approx 0.588$), i.e. with the polymer having two types of beads with flexible joints, at these low values of N, these disrupt the fibre and at larger values of N, the literature value would be recovered.

The local fibre compaction can also be measured using the radius of gyrations. For this, a similar window is slid along the fibre, but the size ($L = 11$ beads) is kept constant and the R_g is calculated for each window and averaged over the different snapshots from all the simulations. Since each window, as it slides along, will contain a different number of nucleosome and DNA beads, the R_g is scaled by a factor $\lambda = \sqrt{N_d + 4N_n}$, where N_d and N_n are the numbers of DNA and nucleosome beads within the window respectively. This is equivalent to the square root of the contour length, since nucleosome beads are four times larger than the DNA beads. The results from these calculations are given in Fig.4.18 below.

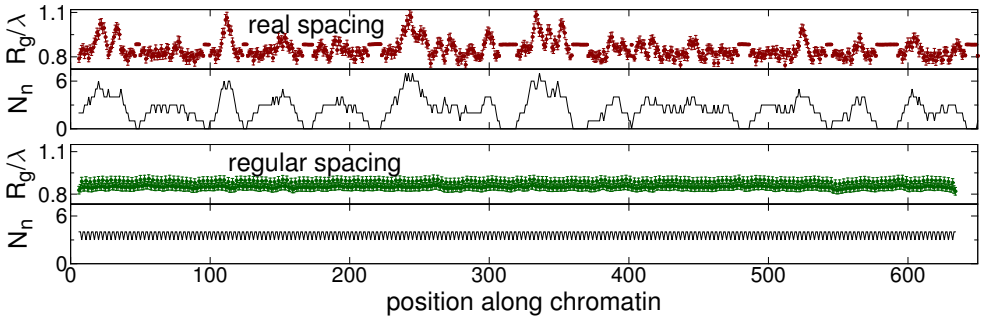


Fig. 4.18 The top plot of each pair shows the average R_g/λ of an $L = 11$ bead region, as a function of position along the fibre. The top set shows data for the region chrIV:1,254,937-1,287,938 and the bottom set for the artificial regularly spaced fibre. The bottom plot in each pair gives the number of nucleosomes within each $L = 11$ bead region.

In the top panel of Fig. 4.18, the value of R_g/λ varies significantly with position along the fibre. Below this, the number of nucleosomes N_n is given for each window and the variation of the value for R_g/λ coincides with the variation for the nucleosomes. So the variation in R_g is directly linked to the number of nucleosomes found within the window. When considering a window with no nucleosomes, with an approximate persistence length of 20 beads, the value $R_g/\lambda \approx 0.88$, which is in agreement with the expected value from a worm like chain model in the rigid rod limit. In this model, the expected upper bound value is $R_g/\lambda = (L/\sqrt{12})/\sqrt{L} \approx 0.96$.

However, the addition of nucleosome into a region adds turning points into the polymer (since nucleosome beads effectively act as freely rotating joints). With only a few nucleosomes, the turning points reduce R_g/λ, but when many nucleosomes are added, then although more joints are added, the steric interactions between the nucleosome beads limit the possible rotation of the fibre leading to a stiffening of the chain. This in turn then increases R_g/λ, in line with the observations from the figure.

In contrast, the artificial regularly spaced fibre has only minor variations in nucleosome number per window and as such the R_g profile is pretty much flat in line with the nucleosome number plot. Unsurprisingly, a regular fibre also yields a nucleosome interaction map which contains no domains.

A major conclusion that can be drawn from these simulations is that the different spacing of nucleosomes leads to relatively small, yet significant, differences in the global and local 3-D organisation of chromatin.

4.2 Simulation version 2 – Disk-shaped cylinders on a constrained string

The first revision of the simple nucleosome model aimed to improve the structural representation of the nucleosome geometry. From the results of crystallography [126, 127] a number of features could be included, but for now a more realistic "disk-like" shape for the nucleosomes and a DNA entry-exit angle constraint are added.

The strategy was that this addition of "more realistic" features might increase the agreement of the simulation results with the chromatin interaction data from the experiments. This however was not the case.

The more detailed model is shown in Fig. 4.19. Here the nucleosome, instead of being a single sphere is represented as a rigid body made up of five beads, to give a more cylinder-like shape. Four 5 nm (2 σ) beads make up the nucleosome core and a 2.5 nm (σ) bead functions as a "connector bead". The masses of these beads remain at 1 in simulation Units.

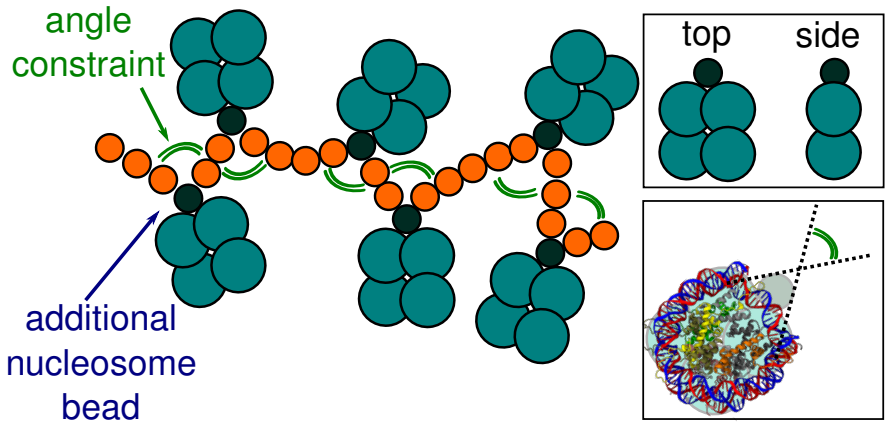


Fig. 4.19 Schematic showing the more detailed model in which nucleosomes are made of up from five beads acting as single body. The four larger beads are arranged for form a more disk-like shape of the nucleosome, with diameter roughly 10 nm and height 5 nm. DNA beads attach to the nucleosome via the "connector" bead on the side. This allows for control of the entry/exit angle by which the DNA connects. Top inset: top and side views show a more disk-like nucleosome shape. Bottom inset: the nucleosome schematic is overlaid on an image of the nucleosome crystal structure (obtained from Ref. [128]) to show the preferred linker exit/entry angle.

The four core beads are arranged with their centres on the corners of a square of size 4.2 nm (1.68 σ); the connector bead is positioned 5.75 nm (2.3 σ) from the centre of the square. In the more complex model the mass of the DNA beads remain at 1 mass unit, the linker bead is set to 1 mass unit, but the nucleosome beads and central bead are set to 11.2 mass unit. However since the central bead is excluded from any interactions the mass of the nucleosome is set at 44.8 mass units. The reasoning for this mass came from a "back of the envelope" calculation for the mass of nucleosomes vs mass of DNA base-pairs. If the mass of a DNA bp is taken as 650 daltons [4] then a DNA bead with approximately 7.35bp has a mass of 4777.5 daltons. Each histone has an approximate weight of 11-15kDa[129]. As such the approximate weight of a nucleosome bead is $8 * 15$kDa$+147 * 650$Da$=$ 215550Da$=$ 45.12DNA beads. For simplicity this was set to 11.2 mass units per nucleosome core bead ($4 * 11.2 = 44.8$). The normal linker DNA beads from the previous model are connected to the nucleosome connector beads using harmonic springs with the associated potential:

$$U_{\mathrm{HARM}}(r_{i,i+1}) = K_{\mathrm{HARM}}(r_{i,i+1} - R_0)^2, \tag{4.1}$$

where $r_{i,i+1} = |\mathbf{r}_i - \mathbf{r}_{i+1}|$ is the separation of the beads, and $R_0 = 1.1\ \sigma$ is the equilibrium separation.

Instead of allowing the nucleosome bead to act as freely rotating joint, with no limitations on the entry and exit angle for the DNA beads, the DNA beads now connect to the small "connector bead". This allows me to constrain the entry-exit angle for DNA linkers connecting to a nucleosome. The potential governing the bending interaction between three connected DNA-connector-DNA beads is given by

$$U_{\mathrm{NUC-BEND}}(\theta) = K_{\mathrm{BEND}}\left[1 - \cos(\theta - \theta_0)\right], \tag{4.2}$$

where θ is the angle between the three beads, and θ_0 is the desired equilibrium angle measured from the straight angle of 180°. So θ_0 is set at θ_0 =72°, so as to match the entry-exit angle (108°) measured from the canonical nucleosome crystal structure [126]. The interaction energy is set to be the same as that used for the linker DNA beads.

4.2.1 Results from the more detailed model

Despite the refinements added to this more detailed model, the results did not in fact increase the agreement of the simulation results with the Micro-C data. Visual inspection (Fig. 4.20) of the contact maps shows that the output is very similar to that of the simpler first model. The only difference I can see by eye would be a slight increase in longer range interactions and, at least for this region, the number of boundaries is the same.

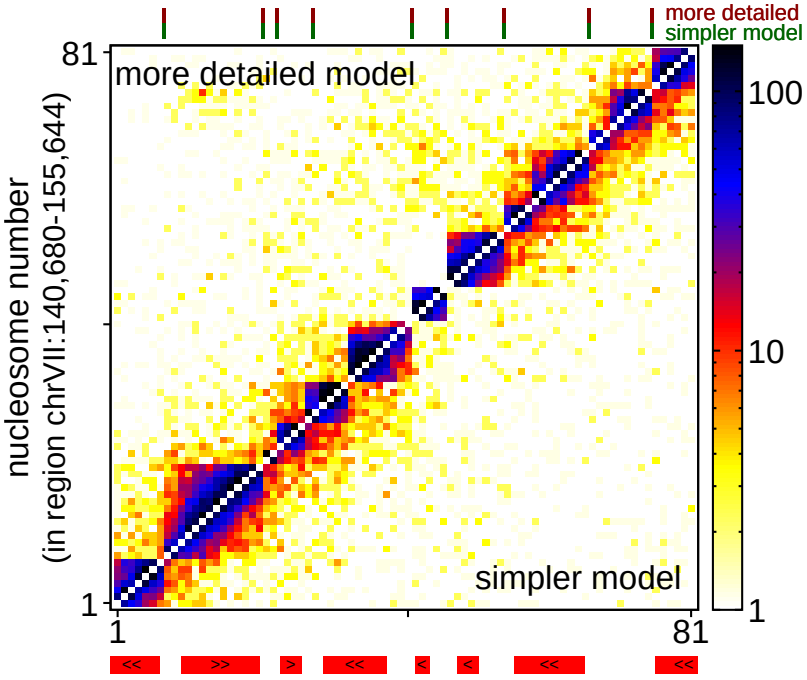


Fig. 4.20 Contact map comparison between the more detailed simulation model and the Micro-C data in the same arrangement as before. Domains are clearly identifiable, but there is no significant improvement of the agreement of data.

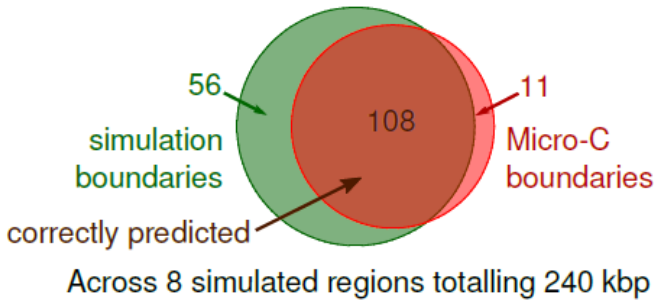
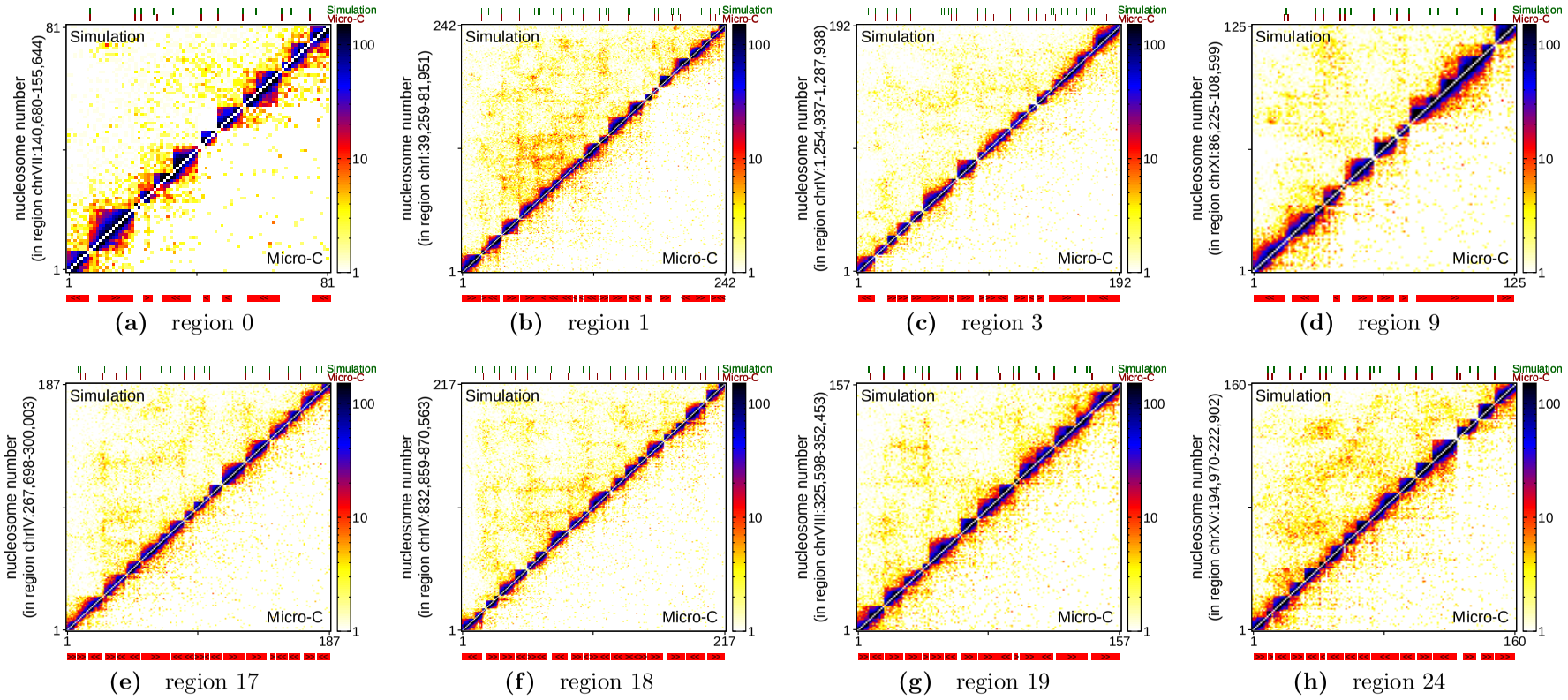


Fig. 4.21 Venn diagram summarising the boundaries found between the Micro-C data and the more detailed simulation model. Here 90.8% of experimental boundaries correctly found, but a much larger number of simulation boundaries were found in total.

Actually when considering boundaries, the more detailed model manages to identify more correct Micro-C boundaries. A total 90.8% of experimental boundaries correctly found, compared to 83.2% found by the simpler model. However, this might be down to the fact that it finds more boundaries in general with the same criteria as used in the simpler model. Compared to the simpler model, where 76.0% of the found boundaries were correct, only 65.9% of the boundaries are correct, giving a higher percentage of "extra" boundaries. The boundary results can be seen summarised in the Venn diagram in Fig. 4.21.

The insulation signal score can also be calculated for this model and the correlation to the Micro-C signal ends up as $r = 0.52$ ($p < 10^{-10}$), about 16% smaller than the previous model.



(a) region 0 **(b)** region 1 **(c)** region 3 **(d)** region 9

(e) region 17 **(f)** region 18 **(g)** region 19 **(h)** region 24

Fig. 4.22 All eight regions in contact map comparisons of Micro-C and more detailed simulation data.

4.2.2 Simulation version 3 – Complex fibre

Since the more detailed model did not give any appreciable improvements, two further refinements were added. The first, is an attractive interaction potential between nucleosomes. The reasoning for this would be that surface charges on the nucleosome core and/or the histone tails would bring about these interactions. The second refinement was to add in the effects of certain histone modifications. Although the model is too simple to add this in detail, a simplified way of including it would be to remove the angle constraint on the entry/exit DNA for nucleosomes that have an acetylation modification.

For the first refinement, the short range attractive interaction between nucleosome centres is added by making use of a centre bead placed at the core of the nucleosome structure. This bead does not interact with any of the other beads within the rigid body nucleosome structure or the DNA beads. Instead it interacts with other nucleosomes by using a Lennard-Jones interaction potential between any pair of nucleosome centre beads, given by

$$U_{\mathrm{LJcut}}(r) = \begin{cases} U_{\mathrm{LJ}}(r) - U_{\mathrm{LJ}}(r_{\mathrm{cut}}), & r < r_{\mathrm{cut}} \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where

$$U_{\mathrm{LJ}}(r) = 4\epsilon_n \left[\left(\frac{d_n}{r} \right)^{12} - \left(\frac{d_n}{r} \right)^{6} \right], \tag{4.4}$$

where r is the separation between the centre beads of the nucleosomes, ϵ_n is the interaction energy, $d_n = 2.0\ \sigma$, and $r_{\mathrm{cut}} = 3.0\ \sigma$ is the range of the interaction.

As would be expected, this refinement leads to an overall increase in nucleosome-nucleosome contacts. However, the interaction energy ϵ_n needs to be carefully tuned to give reasonable interactions as too large values can promote long range interactions in particular and cause nucleosomes to collapse into a globule.

The second refinement was implemented by annotating the nucleosome lists used to create the model fibre with histone modification data. During the creation of the LAMMPS input files, nucleosomes with acetylation marks were modified to not have an angle constraint on its entry/exit DNA. Acetylation has been shown to decrease the charge of a modified histone from positive to neutral, this in turn weakens the interaction between the histones which rely on the positive and negative charge interactions[130, 131]. Similarly the DNA also binds less strongly and we decided to model this "loosening" of the structure by opening up the entry/exit angle, giving the polymer increased freedom.

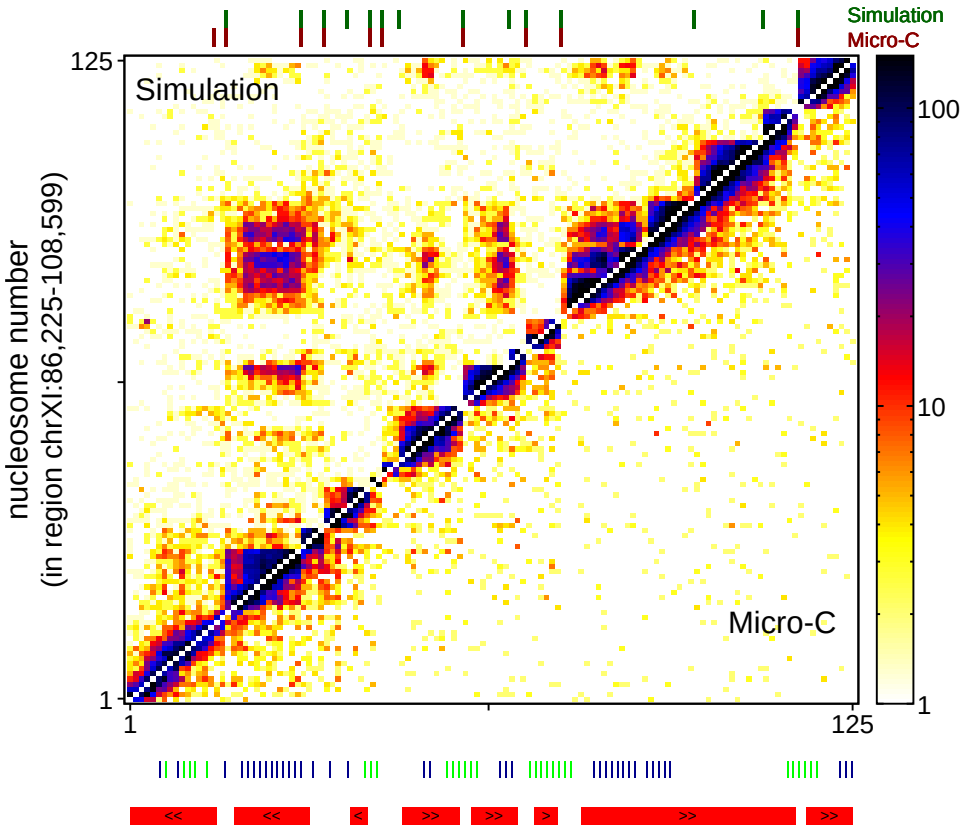


Fig. 4.23 Contact map comparison of the more detailed model with inclusion of selectively switched off angles at acytelation marks. Boundaries are given as before but in addition, the type of histone modification is given below through the green/blue bars, where green is acytelation and blue is methylation.

Ultimately, neither of the two refinements showed an improved agreement with either the Micro-C or Micro-C XL data compared to the simple model. Perhaps with even further refinements and careful tuning of the parameters a no longer simple model could be created which would give better agreement with the Micro-C data. However, the simple sphere model seems sufficient to reproduce the micro-domains and it its unlikely that the model can be improved further in order to give better results. To understand other mechanism at play in the regulation of micro-domains, such as the role of histone modifications, a more detailed model would likely be required to understand the fine details of chromatin fibre structures such as Ref. [116, 117, 132]

Chapter 5

Modelling nucleosome positioning in the human genome

Following on from the success of the simple model with reproducing micro-domains in yeast, the next step was to try and apply the model in a predictive way to some human nucleosome positioning data. Starting from nucleosome positioning data the aim was to make a prediction about the structure of mammalian chromatin and how it changes in response to external stimuli (e.g. inflammation). As there was and at the time of writing still is no Micro-C data available for humans a different approach had to be used to apply the model. Recently two preprints were published [133, 134] which discuss experiments in which Micro-C data was gathered for the mouse and human genome. The data associated to these papers has not been made publicly available yet and as such my work carried on in a slightly different direction.

The idea was to use the data from Diermeier et al. [135] which they used to study the changes in nucleosome positioning in primary human umbilical vein endothelial cells (HUVECs) when treated with tumour necrosis factor alpha (TNFα). After treatment, large parts of the genome experience a rearrangement of nucleosomes accompanied by a change in 3D structure and expression level of many genes.

The aim of my work then was to take the nucleosome occupancy data and feed it into the model to observe any major changes in domain structures through simulation contact maps.

5.1 Regions of interest and the genes within

In the paper by Diermeier et al. [135], the authors look at a number of different genes which elicit some form of response from the (TNFα) treatment. In my work I will use two genes which in their work gave two different responses. The first gene is *SAMD4A* (Sterile Alpha Motif Domain Containing 4A) which is a protein coding gene. Upon treatment it was found to be *responsive* by being up regulated and formed significantly more new contacts within the chromosome arm after 30min. It also lost a high number of the 0-min contacts with 131 contacts (as measured in HiC experiments) forming *de novo* out of the original 167 contacts [135]. The other gene considered is *EDN1* (Endothelin 1) another protein coding gene which was found to be unresponsive to treatment. Out of the contacts found initially, many remain after treatment though some change does occur in NFRs. The third region I considered was a heterochromatin region (abbreviated as HECHRO), containing no major genes. This region was chosen to give a comparison of nucleosome occupancy changes to non-genic regions.

Care has to be taken with these two genes as their length and/or position changes depending on the reference genome used. For this work, the slightly older Human reference genome **hg19** was used as this was also used in the paper. A major difference between the older hg19 and the newer hg38 is that *EDN1* changes in size from 6,899 bases to 35,981 bases.

Gene / region	chromosome	coordinates (bp)	size
SAMD4A	chr14	55,033,815 – 55,260,033	226,219 bases
EDN1	chr6	12,290,529 – 12,297,427	6,899 bases
heterochromatin (HECHRO)	chr14	49,200,000 – 49,400,000	200,000 bases

Table 5.1 Gene regions used for simulations and contact map creations.

The idea was to create simulations of the 3D conformation of these regions for the t=0min state and the t=30min state and compare the two contact maps in how they change.

5.2 Nucleosome positioning

As with the yeast work before, the simulations require a set of nucleosome positioning data to start from. The supplied data from Diermeier et al. [135] at GEO: GSE53343 contains all the MNase-seq data needed and the analysis followed a similar procedure as outline in the yeast work.

The data was aligned to the hg19 reference genome[1]. The data was filtered for mapping quality and unmapped reads were removed using the samtools suite. Any unpaired reads can be removed as well as any duplicates before sorting. The final step is converting the data into the bedpe format usable by NucPosSimulator.

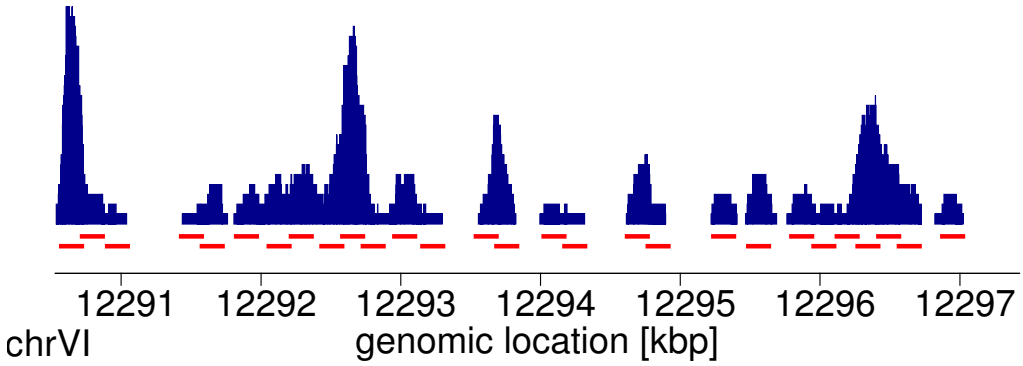
The region around *SAMD4A* and the heterochromatin region are much larger than any of the regions used with yeast, the *EDN1* region however is considerably smaller than the smallest yeast region. In order to adequately position any nucleosomes in the region, a 1Mbp region centred roughly around the midpoint of the gene is used by NucPosSimulator to convert the nucleosome occupancy data into most likely nucleosome positions. Once again, the process is set up to generate an effective potential and the simulated annealing mode is used to find the most likely nucleosome positions that do not overlap. This process took more work than with yeast and is elaborated on in section 5.2.1.

For the t=0min data set the MNase-seq data is shown along with the found likely nucleosome positions in Fig. 5.1, where the gene or region is extracted and plotted. The END1 gene is small enough that details are visible without magnification, but *SAMD4A* and the HECHRO regions are too large to make out detailed nucleosome positioning. Therefore an extraction of the *SAMD4A* region is given in Fig. 5.1c (although the data is the same and the region is correct, it looks different due a different binning of the data).

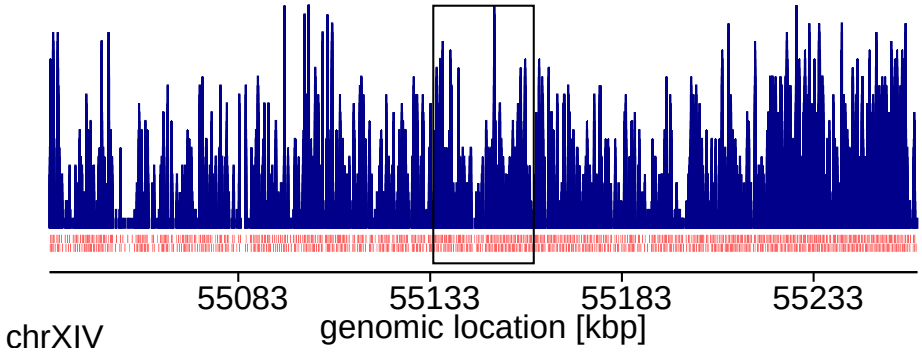
The same process was repeated for the t=30min data set and the nucleosome occupancy data for the regions of interest is given in Fig.5.2. The MNase signal for this data set was significantly weaker and sparser giving less defined peaks.

As the data sets for human DNA are much larger and also much sparser concerning nucleosome positioning compared to yeast, a couple of modifications had to be considered to the previously established method.

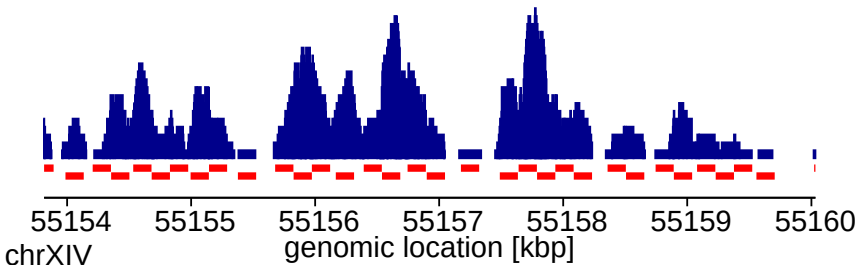
[1] available from the bowtie2 website, requires building from archive.



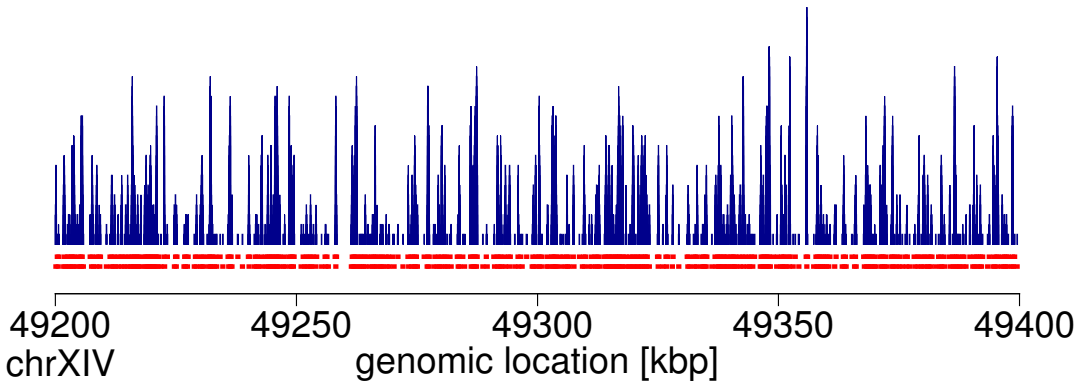
(a) *EDN1* gene chr6: 12,290,529 – 12,297,427



(b) *SAMD4A* gene chr14: 55,033,815 – 55,260,033, the contents of the box are given in the detail view below.

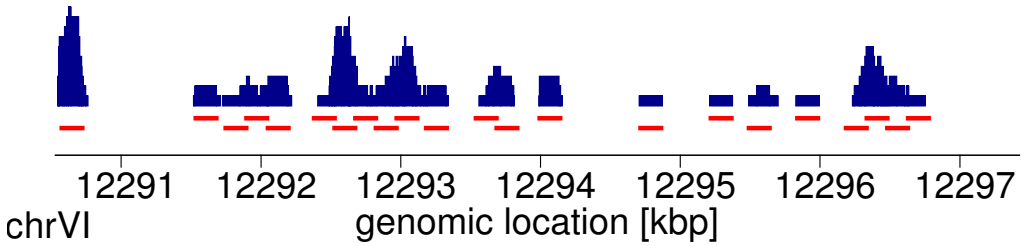


(c) *SAMD4A* gene detail view at chr14: 55133815 – 55160033

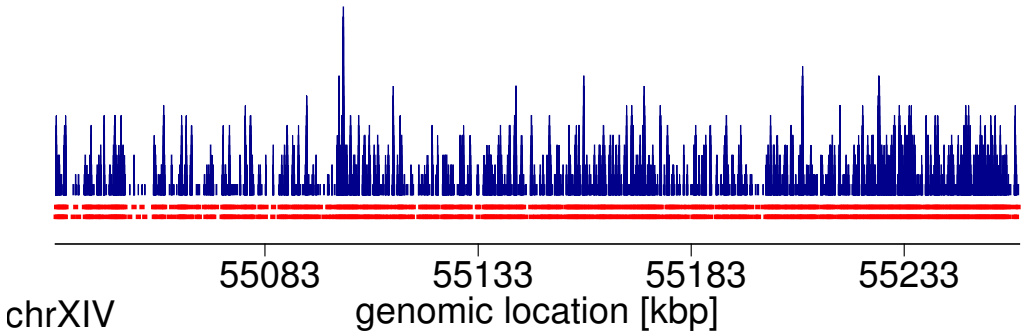


(d) heterochromatin region chr14: 49,200,000 – 49,400,000

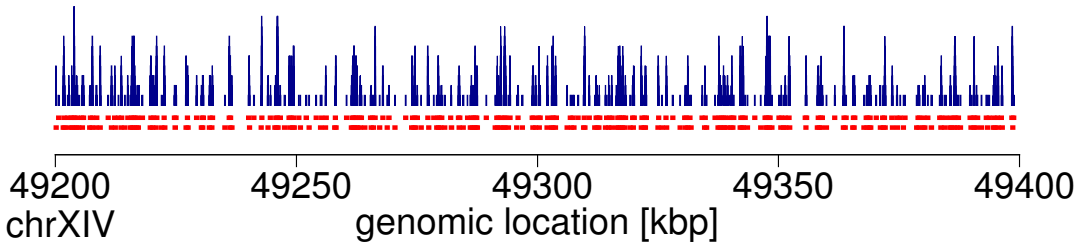
Fig. 5.1 MNase-seq data for the t=0min data set of the three regions of interest. Nucleosome occupancy pile-up data is given in blue with most likely nucleosome positions given below in red.



(a) *EDN1* gene chr6: 12,290,529 – 12,297,427



(b) *SAMD4A* gene chr14: 55,033,815 – 55,260,033



(c) heterochromatin region chr14: 49,200,000 – 49,400,000

Fig. 5.2 MNase-seq data for the t=30min data set of the three regions of interest. Again the nucleosome occupancy pile-up data is given in blue with most likely nucleosome positions given below in red.

5.2.1 Nucleosome repeat lengths and NucPosSimulator energy variation

The first issue encountered with the nucleosome positions was observed when looking at the nucleosome repeat length (NRL). The literature cites a number of values for human nucleosome repeat lengths which differ from cell to cell and depending on the activity level. Most estimates range around 190–205bp [136–138], giving an approximate linker length of 43–58bp when considering a 147bp nucleosome.

To find the NRL I used the *SAMD4A* nucleosome data set, it contains 1Mbp with approximately 3000–4000 nucleosomes. The linker length is calculated by going through the list of nucleosomes and finding the base pair number between the end of one nucleosome and the start of the next. To translate this to a NRL the number of base pairs in a nucleosome can be added. Simple statistics can then be run on this data to find the average linker length over the entire region.

The data set available for the human nucleosome occupancy has a much lower coverage (i.e. reads per bp) than the yeast data used for the previous work, as such it is much noisier. The default settings for NucPosSimulator were ideal for finding nucleosome positions in yeast, but with the higher noise and lower coverage, it gave too low of a nucleosome occupancy for the human data. This also include too large a value for the mean NRL compared to the literature values available.

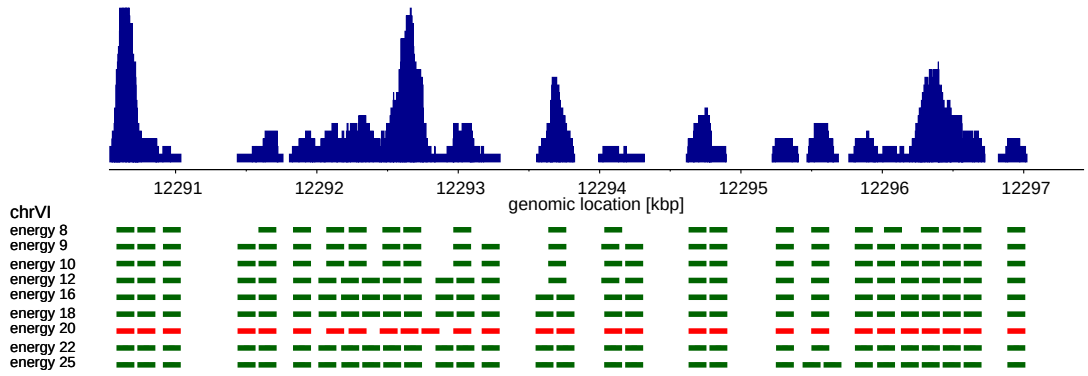


Fig. 5.3 Variation of the binding energy in NucPosSimulator. The MNase-seq data is given above and the nucleosome positions found by NucPosSimulator at the different binding energies are given below. The nucleosomes for the chosen energy of -20.0 are given in red.

In order to find nucleosome positions which agreed with an approximate nucleosome repeat length, the binding energy value parameter for NucPosSimulator can be varied. By decreasing the value from the default of -8.0, which increases the energy, the likelihood of nucleosomes being placed is increased. As the number of nucleosomes increases the values for occupancy and NRL improved, but an energy too large would place nucleosomes in NFRs. Therefore, a balance must be found. A number of simulations with different binding energies were run to find a value that gave a better agreement with the NRL, but still matched the MNase data as well. In the Fig. 5.3 the MNase data for *EDN1* is given at the top and the different nucleosome positions found at different energies are given at the bottom.

There is a clear variation between the lower and higher energies. The final decision was to use a binding energy of -20.0 to find the nucleosome positions as over all regions. This seemed to give an adequate value for the NRL without overcrowding the MNase data.

5.3 Nucleosome position variation after TNFα treatment

As mentioned at the start of the chapter, Diermeier et al. [135] observed a significant change in chromatin contacts and in the 3D structure. To investigate this further the bedtools suite [2] can be employed to get an "intersection" of the two nucleosome data sets. This allows for tracking of nucleosomes between two data sets where three types of nucleosomes can be isolated. First, the nucleosomes that potentially move between sets but still overlap by at least 60% (around 88bp) with a nucleosome from the other set are considered "unchanged". Second any nucleosomes that are found in the t=0min set but not the t=30min set are considered "removed". Third, any nucleosomes that appear in the t=30min but not in the t=0min set are considered "new". For the analysis (with results shown in tables: 5.2,5.3 and 5.4) the entirety of the NucPosSimulator (NPS) nucleosome dataset for each region is used in the left hand column and just the gene or heterochromatin region is considered in the right hand column. As the regions are too large to plot, the smaller genes or subsets of the genes are plotted for comparison in figures: 5.4, 5.6 and 5.7.

[2] available from `https://bedtools.readthedocs.io`

Nucleosome intersection *EDN1*		
	NPS region	*EDN1* gene
total nucs found at t=0min	3932	28
total nucs found at t=30min	2917	22
nucs at t=30min which match t=0min to within 7bp	342 (8.7 %)	5 (18 %)
nucs at t=30min which match t=0min to within 15bp	625 (16 %)	5 (18 %)
nucs at t=30min which match t=0min to within 88bp (unchanged)	2110 (54 %)	15 (54 %)
nucs found at t=0min but not at t=30min (removed)	782 (20 %)	4 (14 %)
nucs found at t=30min but not at t=0min (new)	109	0

Table 5.2 Nucleosome intersection for *EDN1* of the t=0min and t=30min data set. The nucleosome positions found by NucPosSimulator (NPS) are compared to just the nucleosome positions found within the gene.

The intersection results for the nucleosome around END1 are given in Table: 5.2, with a significant decrease in nucleosomes between the data sets. However, if a 60% overlap between nucleosomes is allowed to consider them the same nucleosome, then just over 50% of nucleosomes are affected by only small relocations. In total 782 (about 20%) nucleosomes disappear completely and 109 (about 3%) appear *de novo*.

Considering visual analysis of just the *EDN1* gene in Fig. 5.4, only 6 nucleosomes are lost between the data sets and no new nucleosomes are found. Overall, there is a smaller change between sets in contrast to the greater NucPosSimulator region.

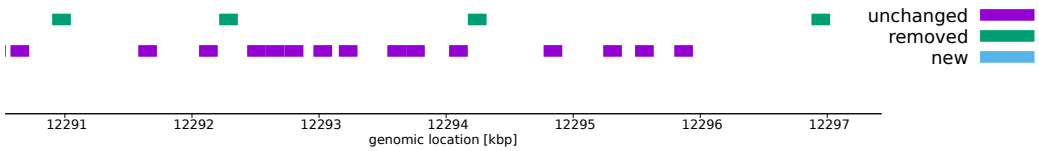


Fig. 5.4 Nucleosome intersection for the *EDN1* gene chr6:12,290,529-12,297,427 where nucleosomes whose position changed by less than 88bp, between the t=0min and t=30min data set, are considered 'unchanged' and given in purple. Nucleosome that are not found in the t=30min set are 'removed' and given in green and any 'new' nucleosome found only in t=30min are given in blue.

For *SAMD4A* the changes in nucleosome positioning (Table: 5.3) are quite similar in percentages, though there seems to have been a bit more movement as a lower percentage matches the stricter criteria of having a 95% overlap (or moved less than 7bp). Despite being a larger region, less new nucleosomes appeared overall.

Nucleosome intersection *SAMD4A*		
	NPS region	*SAMD4A* gene
total nucs found at t=0min	4401	1055
total nucs found at t=30min	3436	812
nucs at t=30min which match t=0min to within 7bp	433 (10 %)	96 (9.1 %)
nucs at t=30min which match t=0min to within 15bp	796 (18 %)	186 (18 %)
nucs at t=30min which match t=0min to within 88bp (unchanged)	2538 (58 %)	610 (58 %)
nucs found at t=0min but not at t=30min (removed)	680 (15 %)	169 (16 %)
nucs found at t=30min but not at t=0min (new)	75	11

Table 5.3 Nucleosome intersection for *SAMD4A* of the t=0min and t=30min data set. The nucleosome positions found by NucPosSimulator (NPS) are compared to just the nucleosome positions found within the gene.

Fig. 5.6a shows the intersection data for the gene *SAMD4A* and Fig. 5.6b shows a subset of this gene. It appears that there is a greater change to nucleosome positions in the responsive gene, however, in Fig. 5.5 there seems to be no loss of nucleosomes in particular around the promoter region (approx 1kbp upstream of TSS). It is likely then that the change in activation state for *SAMD4A* comes from a small positioning change in the nucleosomes or a different factor. Although no conclusions can be drawn from a single example gene, it does support the hypothesis that a change in expression is accompanied by a change in nucleosome positions even though the TSS has no changes in the nucleosome number. Potential changes in nucleosome positioning around the TSS are discussed later in section 5.4.1.3.

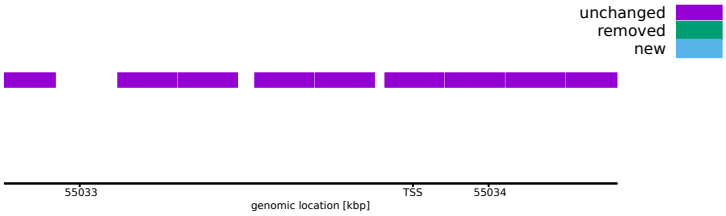
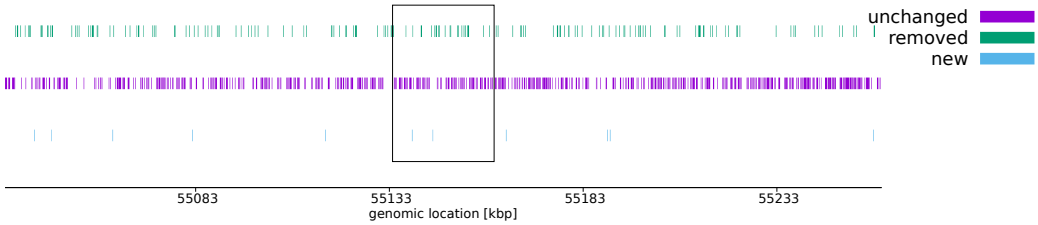
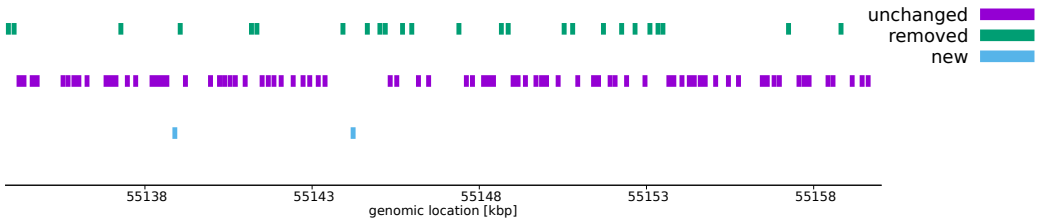


Fig. 5.5 Nucleosome intersection for the *SAMD4A* gene TSS at chr14:55,033,815, where nucleosomes whose position changed by less than 88bp, between the t=0min and t=30min data set, are considered 'unchanged' and given in purple. If any were found in this region around the TSS, then 'removed' nucleosome would be given in green and any 'new' nucleosomes in blue.



(a) chr14: 55,033,815 – 55,260,033



(b) subsection chr14: 55,133,815 – 55,160,033

Fig. 5.6 Nucleosome intersection for the *SAMD4A* gene, where nucleosomes whose position changed by less than 88bp, between the t=0min and t=30min data set, are considered 'unchanged' and given in purple. Nucleosome that are not found in the t=30min set are 'removed' and given in green and any 'new' nucleosome found only in t=30min are given in blue. The entire gene is given in (a) and a subsection for increased detail is given in (b).

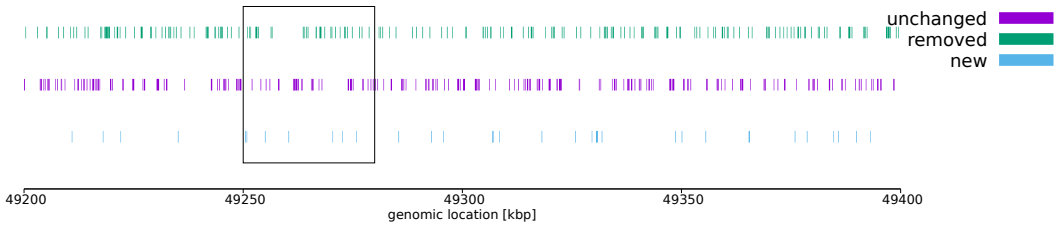
The heterochromatin intersection data is given in table: 5.4 and shows the highest variation of nucleosome positions with only 41% of nucleosomes overlapping between data sets by 88bp or more. With 34% it also has the highest percentage of nucleosomes lost entirely between sets, though it also has the highest number of new nucleosomes formed over the entire NPS region (1Mbp).

Nucleosome intersection HECHRO		
	NPS region	HECHRO region
total nucs found at t=0min	3077	595
total nucs found at t=30min	1932	376
nucs at t=30min which match t=0min to within 7bp	233 (7.5 %)	48 (8.1 %)
nucs at t=30min which match t=0min to within 15bp	432 (14 %)	85 (14 %)
nucs at t=30min which match t=0min to within 88bp (unchanged)	1268 (41 %)	235 (39 %)
nucs found at t=0min but not at t=30min (removed)	1054 (34 %)	205 (34 %)
nucs found at t=30min but not at t=0min (new)	172	35

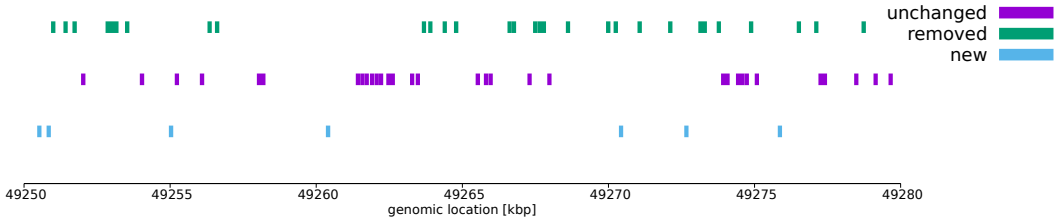
Table 5.4 Nucleosome intersection for heterochromatin region of the t=0min and t=30min data set. The nucleosome positions found by NucPosSimulator (NPS) are compared to just the nucleosome positions found within the gene.

It appears that the high variation in nucleosome positions in this non-genic and repressed region could be consistent with the hypothesis that in heterochromatin regions, nucleosomes are under less strict control and their positions are more free to vary.

Overall, this data suggests that there is a significant rearrangement of the nucleosome positions just from the treatment with (TNFα). The question now would be if these changes in positioning affect the 3D conformation enough that a simulation of the chromatin fibre can reflect these changes in contacts. However, with the much larger size of the human genome and also human genes, the nucleosome data coverage is much lower. Further, the data is not as good as the yeast data was, primarily due to the lower resolution, so care must be taken when drawing conclusions from these results.



(a) chr14: 49,200,000 – 49,400,000



(b) chr14: 49,250,000 – 49,280,000

Fig. 5.7 Nucleosome intersection for the heterochromatin region, where nucleosomes whose position changed by less than 88bp, between the t=0min and t=30min data set, are considered 'unchanged' and given in purple. Nucleosome that are not found in the t=30min set are 'removed' and given in green and any 'new' nucleosome found only in t=30min are given in blue. The entire region is given in (a) and a subsection for increased detail is given in (b).

5.4 Simulation contact maps

For the simulations carried out with the human nucleosome positioning data, the first simple model developed for yeast was used without any modifications (see sec 4.1). To make sure region boundary effects do not influence the nucleosome positioning, the regions of interest are placed within a 200,000bp region (i.e. with 100,000bp either side). A time step of $\Delta t = 0.005\ \tau$ is used in the simulations as before but the equilibration time was increased to $25{\times}10^3\ \tau$ with the main simulation running for a further $25{\times}10^3\ \tau$. The system configuration is saved every 250 τ and for each region this process was repeated 10 times with a different initial starting configuration and a set of random numbers to generate the noise $\boldsymbol{\eta}_i(t)$. Periodic boundary conditions, with a simulation box size of $1000{\times}1000{\times}1000\ \sigma$ are used.

The simulated regions are given in table 5.5 showing the number of nucleosomes and beads found in each of the six data sets and renders of the three regions generated from the nucleosome positions before and after treatment are given at the end of the chapter (Fig. 5.19 and Fig. 5.20).

region	sim coords (bp)	nucleosomes		beads	
		t=0	t=30	t=0	t=30
EDN1	chr6: 12,190,529 – 12,397,427	918	691	9909	13869
SAMD4A	chr14: 54,933,815 – 55,360,033	1941	1512	19544	27007
HECHRO	chr14: 49,100,000 – 49,500,000	1160	686	29850	38104

Table 5.5 Simulation regions for each region of interest, where the gene is given a 100kbp buffer on either side. The number of nucleosomes for each region and time set is given together with the number of beads simulated.

5.4.1 Contact maps

The process for creating contact maps from the simulation data is very similar to the process outlined in section 4.1.3. However, as there is no Micro-C data available, the stochastic process to find a representative number of interactions within the map does not work as no target number can be reached nor can any final number be scaled appropriately. Therefore, the program used to generate the contact maps was modified to use a cut-off value on the maximum distance that two beads can be to still count as interacting. Although this value would benefit from further fine tuning, it was initially defined to be 15σ or 37.5nm. Further, to reduce the load the large fibres put on the program, the maximum distance that any nucleosome interaction would be considered was set to 150 nucleosomes. A cut-off is needed to define the maximum distance where an interaction between nucleosomes is considered a contact. In order to see if micro-domains are visible in the data, a distance map can be used as no cut-off is needed and the distance between them is plotted directly. The colour-scheme in the distance based map is inverted to the contact maps, but the same features are visible. Distance maps also consider only nucleosome interactions between nucleosomes up to 150 away. As micro-domains are indeed visible within the distance maps (Fig. 5.8a), this can be used to make an informed decision on what cut-off is needed to also show the micro-domains in the interaction maps. The cut-off of 15σ was chosen as it gave the best visual match to a simple distance map (see Fig. 5.8a and Fig. 5.8b). To generate a single contact map, the conformations of the latter half of ten simulations are sampled and combined, bringing the contact map to be an

average of 50 different conformations.

In the following, a number of different figures which were generated from the simulations of the three regions, at both t=0min and t=30min, are presented. Further work is required to refine the analysis process.

5.4.1.1 t=0min contact maps

For the first data set of t=0min, a figure is given below showing the entire gene of *EDN1* (Fig. 5.8), as its small size and number of nucleosomes make it easy to show detail. The same section is shown in each of the four plots with the distance map given in the first and only the cut-off distance varied between the other three. Visually all four maps reproduce similar features, but a good agreement is given between the 15σ cut-off and the distance map. A similar process of analysis can be repeated for the second set of data at t=30min.

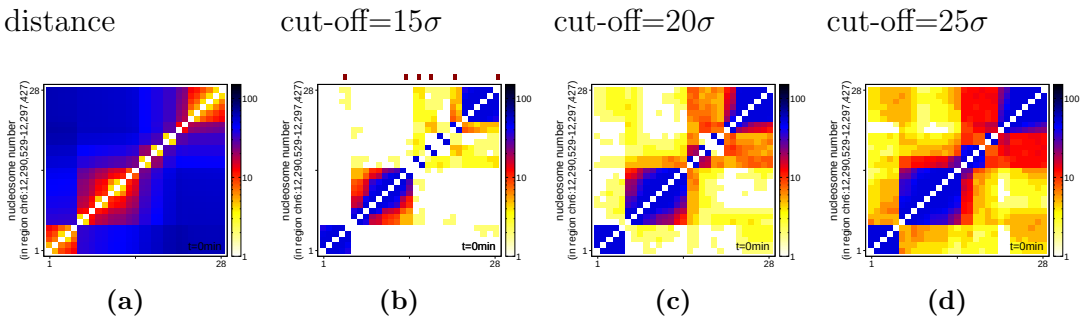
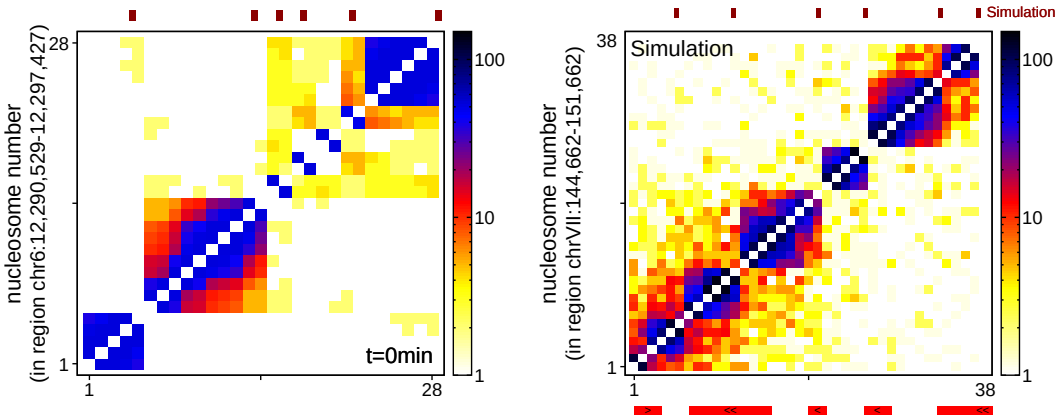


Fig. 5.8 A comparison of contact maps for the *EDN1* gene at t=0min. In (a) the distance based map is shown and micro-domains can be seen. In (b)–(d) contact maps at different cut-off values (shown above the maps) are given. Micro-domains remain visible, but interactions further from the diagonal increase as the cut-off increases.

In figures 5.11 (page 105) and 5.12 (page 106) the *SAMD4A* and *HECHRO* regions are given for completion. As these regions are so large and any detail is lost in a full region contact map, a detail view is given below the figure showing the entire region. It is interesting that in all cases a micro-domain pattern is visible, these are at sub-gene level and hence different from the common HiC domains. As such this could be considered a prediction that sub-gene domains exist in the human genome, whereas in yeast they seem to incorporate one or more genes.

To investigate this further, the region containing *EDN1* can be compared to an extract of region 0 from yeast (chrVII: 144,662 – 151,662), which has been trimmed to 7kbp. The comparison is shown in Fig. 5.9, where the entire *EDN1* gene is given in the left and the trimmed yeast region is given on the right. In both cases a number of micro-domains are visible, but as expected the yeast data shows more nucleosomes in a similar sized region. For both maps, boundary calling can be done using the same algorithm as described before (sec. 3.7) and found boundaries are shown above the plot. At least visually, it appears the micro-domains are of a similar size, on the order of 4-10 nucleosomes (at least in this region), but the overall larger nucleosome spacing seems to give rise to less interactions between nucleosomes. However comparison between the contact maps is hindered by the difference in how they were generated, with the yeast region having the interaction contacts scaled to Micro-C results.



(a) EDN1 **(b)** Subsection of region 0

Fig. 5.9 Contact map comparison between human (*EDN1* gene) and yeast (region 0). The map for *EDN1* in (a) gives the entire gene whereas the map in (b) for yeast is a subsection of region 0, trimmed to be comparative in length of base pairs to *EDN1*. Both maps show boundaries found by algorithm within the region and in (b) any genes found in the region are given below.

Domain boundaries are again more likely found at long linker lengths, in fact, at much longer linkers than in yeast. For the entire regions simulated in humans, the average nucleosome linker length is around 67bp compared to 28 for yeast. This is likely to be partially caused by the lower resolution of the MNase-seq data as well as a sparser nucleosome positioning in the human genome. Experiments in electrophoresis suggest that nucleosomes in humans have an average centre to centre distance of 200bp (147bp nucleosome + 53bp linker) and yeast only has a 167bp spacing (147bp nucleosome + 20bp linker). The linker length at

boundaries for yeast was almost four times longer than the average linker length for the Micro-C boundaries, but around four and a half times longer for average boundaries found in the simulations. In humans the linker lengths at boundaries in the simulation seem to be much larger, approximately 700bp. Currently it is unclear whether this is a genuine feature of human chromatin or an artefact due to the relatively low coverage of MNase data used as an input. Once Micro-C data becomes available for the human genome, this could be tested further. Fig. 5.10 shows the scaled probability for certain linker lengths in all simulated regions as well as for the boundaries found at t=0min and t=30min. Compared to the linker lengths found throughout the entire simulated regions, the linker lengths at boundaries are much less likely, but much longer. Thus, it is safe to say that boundaries, at least in this data, are still primarily found at long linkers.

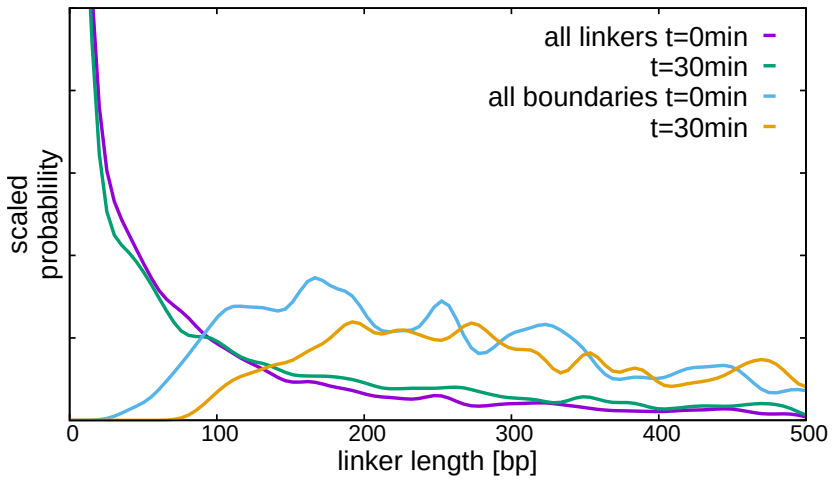


Fig. 5.10 Linker length probability for all regions and at boundaries for both the t=0min and t=30min data sets. The y-axis range is shortened to allow the boundary linkers to be more visible. A kernel density estimation method with bandwidth of 10bp is is used, where curves are normalised to enclose unit area.

distance cut-off=15σ

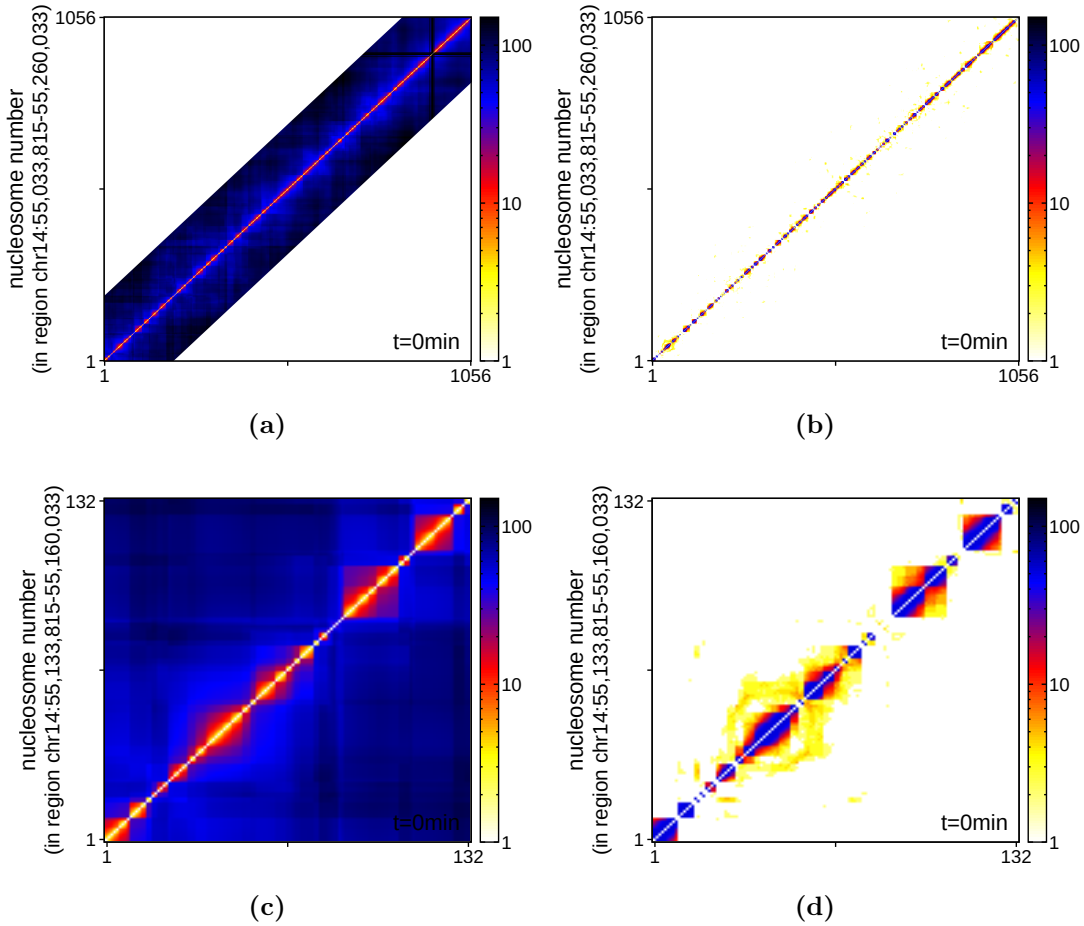


Fig. 5.11 Comparison of distance map with cut-off (15) interaction map for *SAMD4A*, where in (a) and (b) the entire gene is shown and in (c) and (d) a subsection is given for detail

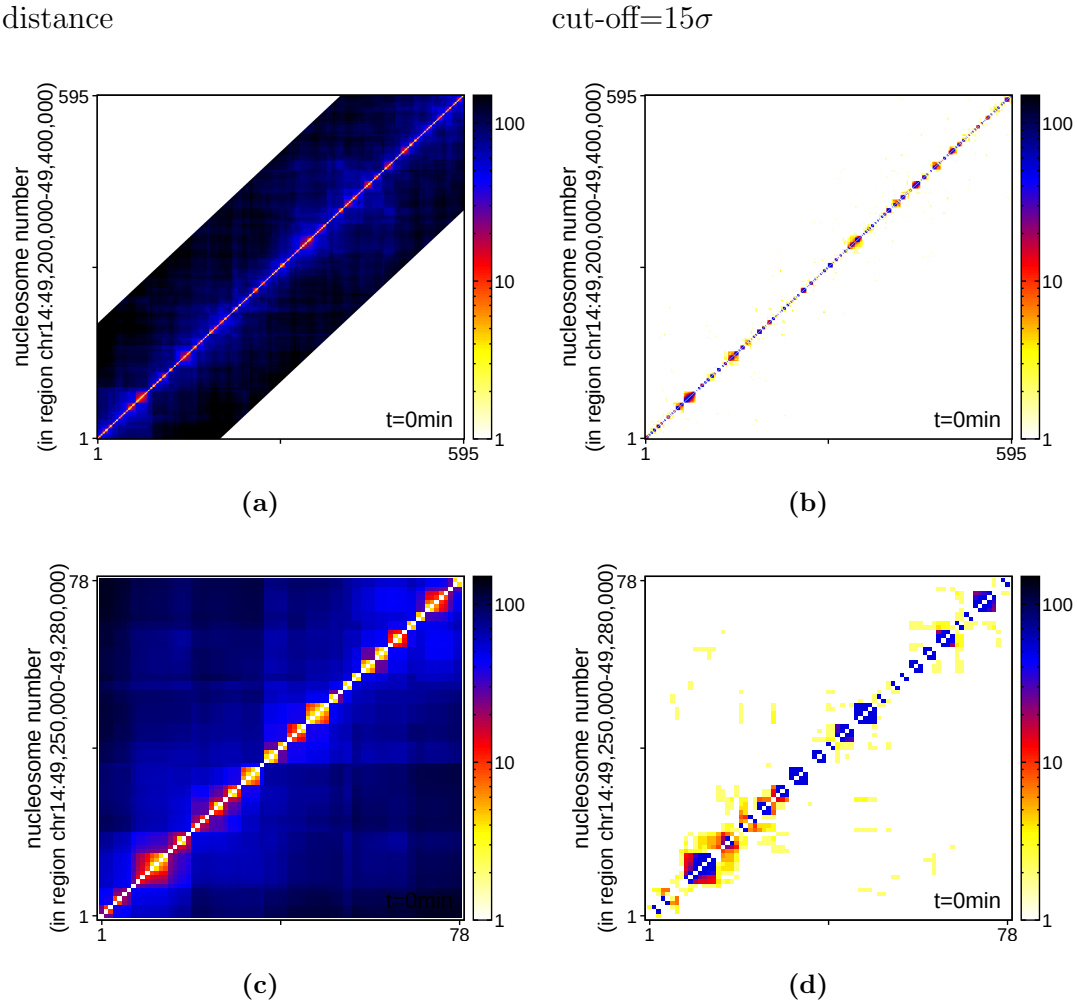


Fig. 5.12 Comparison of distance map with cut-off (15) interaction map for the heterochromatin region, where in (a) and (b) the entire region is shown and in (c) and (d) a subsection is given for detail

5.4.1.2 TNFα treatment comparison

An interesting observation to make is that the changes in the number of nucleosomes between the two data sets, before and after TNFα treatment, do produce significantly different contact maps. This can best be observed when the data sets are placed next to each other. This is however complicated by the fact that the regions differ in length, even if they may be the same length in base pairs, the number of nucleosomes differ and hence they will not line up correctly. A method of working around this is to use nucleosome contact maps that are located by a single coordinate and not a range. The simplest way of doing this is to centre the region on the nucleosome closest to the TSS of the gene. Although the heterochromatin region does not contain a gene, the start of the region will be referred to as TSS in order to simplify discussions. The following figures exploring the comparison between the data sets are all regions of 41 nucleosomes where the 11th nucleosome is the TSS nucleosome of the gene.

For the region around the TSS of *EDN1* (Fig. 5.13), some significant changes have occurred with the contacts between nucleosomes having shifted significantly. However most interaction regions appear to have merely changed in size or shifted in position, which would be explained by any missing nucleosomes. In the studies by Diermeier et al. this gene was unresponsive to the treatment and remained active. This is surprising as similarly large changes in *SAMD4A* caused a change in activation levels, whereas *EDN1* was unresponsive. It may well be that some other factors apart from nucleosome positioning control the activity of *EDN1*.

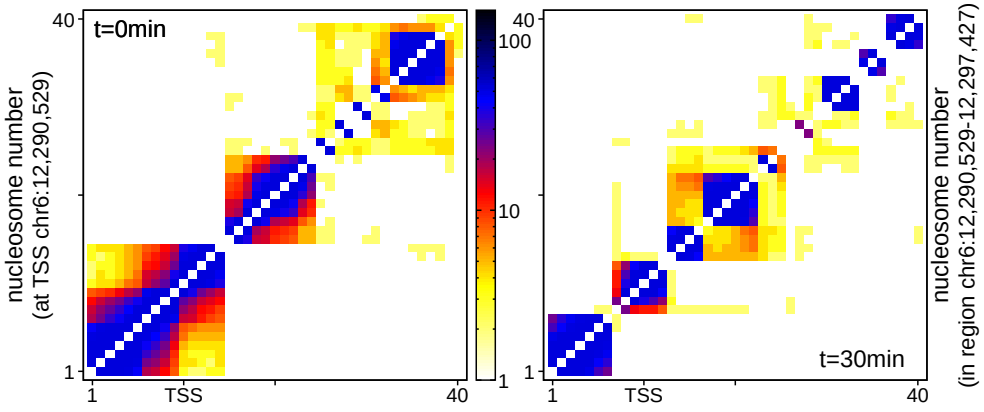


Fig. 5.13 Comparison of *EDN1* map between t=0min (left) and t=30min (right) sets for region of -10 +30 nucleosomes around the TSS. Although the same number of nucleosome is plotted, the number of base pairs differs.

Similarly to the region around the TSS of *EDN1*, the TSS of *SAMD4A* seems to have undergone a few changes as well. Other than changes in position of interaction regions, interactions near the TSS have gone down to open up an interaction boundary. However, downstream a boundary disappeared and two smaller interaction regions appear to have merged into a single larger one. In the studies by Diermeier et al. [135] this gene was a responsive gene, meaning it changed its state due to the treatment. At t=0min it is inactive and after treatment at t=30min it has become active. This is consistent with the yeast results, that more active genes tend to have stronger domain boundaries at or near the TSS.

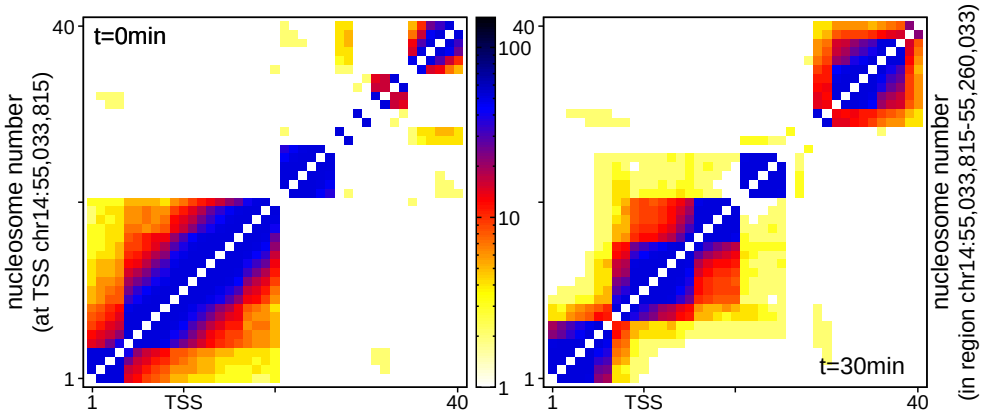


Fig. 5.14 Comparison of *SAMD4A* map between t=0min (left) and t=30min (right) sets for region of -10 +30 nucleosomes around the TSS. Although the same number of nucleosome is plotted, the number of base pairs differs.

For the HECHRO region the contact map has the most significant changes with most interaction domains shifting and most larger regions disappearing. At first sight this is surprising as the region which was expected to change the most was *SAMD4A*, however large changes are also seen in *EDN1* and the heterochromatin region. A caveat to add then, would be that the model does not include any bridging proteins which are likely to play important roles in active regions to create loops between promoters and enhancers and as such add a further level of control to the expression levels.

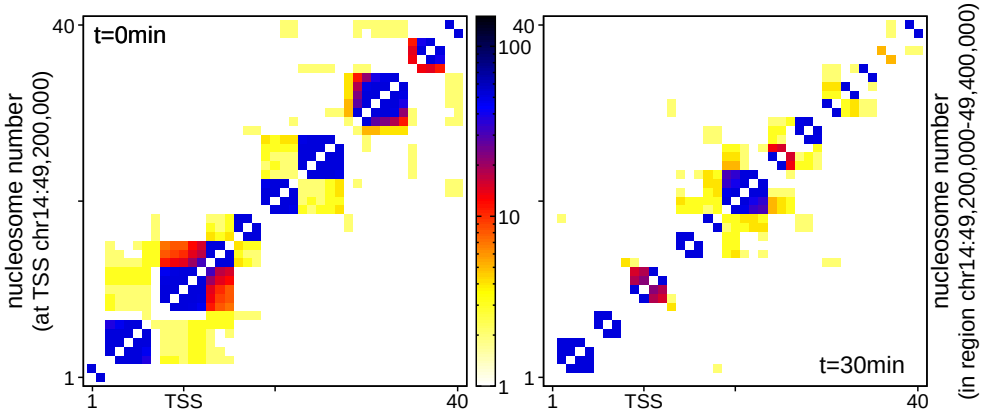
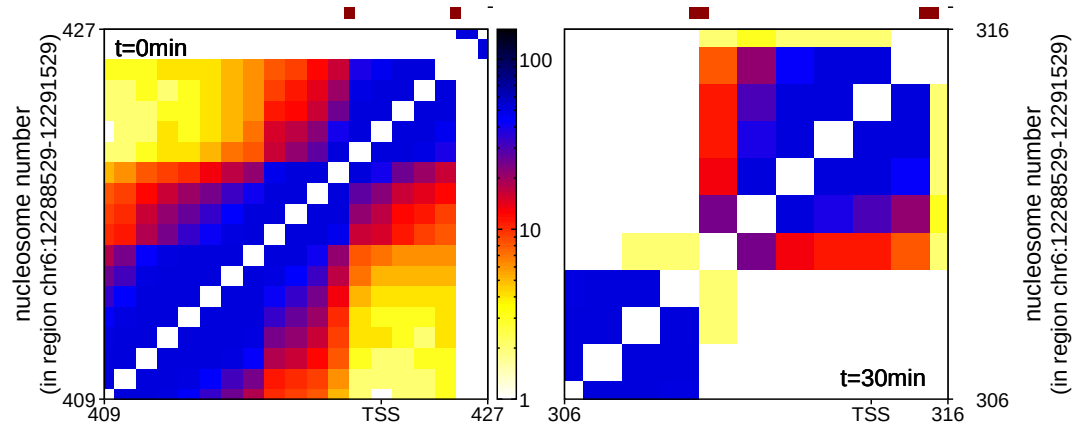


Fig. 5.15 Comparison of heterochromatin region map between t=0min (left) and t=30min (right) sets for region of -10 +30 nucleosomes around the TSS. Although the same number of nucleosome is plotted, the number of base pairs differs.

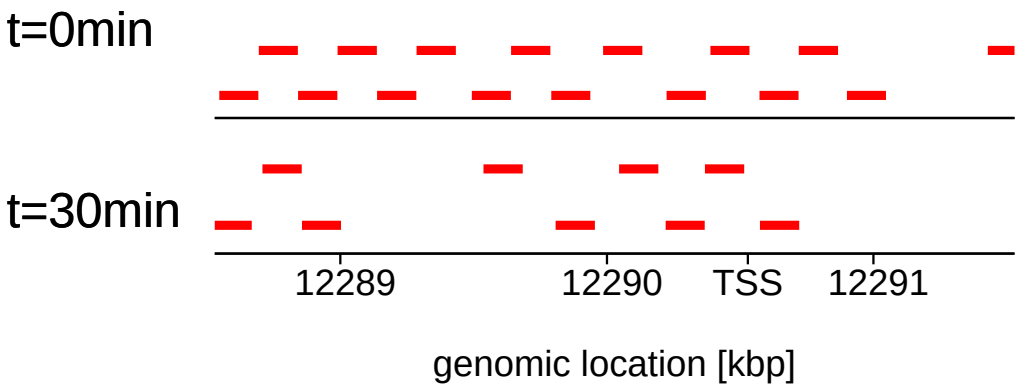
5.4.1.3 Changes at the TSS

An interesting point to consider, is how the nucleosome positioning and interaction of nucleosomes changes upstream and at the TSS. In figures 5.16 and 5.17 all the nucleosomes within a base pair region of -2000bp +1000bp around the TSS are given. This gives unequal numbers of nucleosomes but makes changes around the TSS more visible. The region is an extract from a larger region, to focus on the TSS. Further, the nucleosome positions in relation to the TSS are given below for the same region.

The results for *EDN1* are given in Fig. 5.16, apart from a large change in nucleosome numbers (18 to 10), a boundary has become more apparent in front of the TSS. From a look at the nucleosome positioning data, it is clear that a loss of nucleosomes in that area caused the strong domain boundary to appear, but if the gene was unresponsive to the treatment, it must mean that the changes did not affect the promoter.



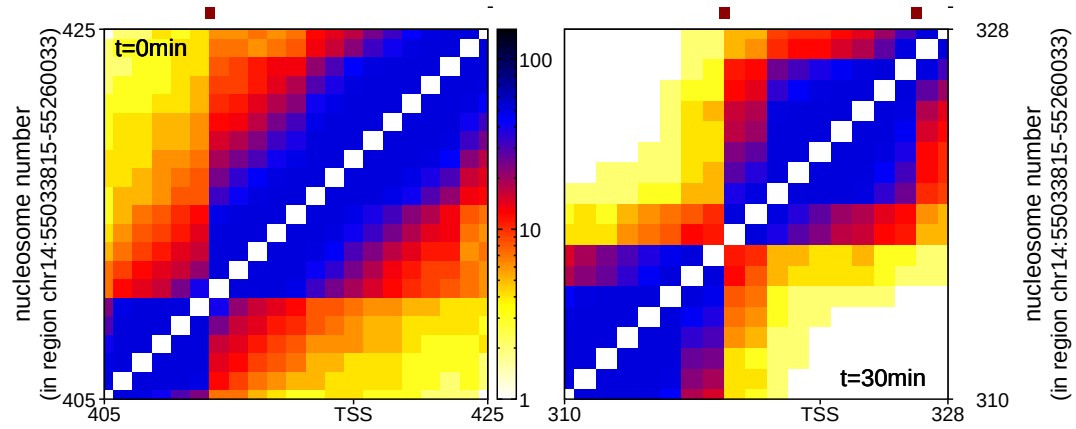
(a) Contact maps for a region of -2000bp +1000bp around the TSS, as the region is given in base pairs, the number of nucleosomes differs between the t=0min and t=30min maps. Boundaries found by the algorithm are shown above then contact map.



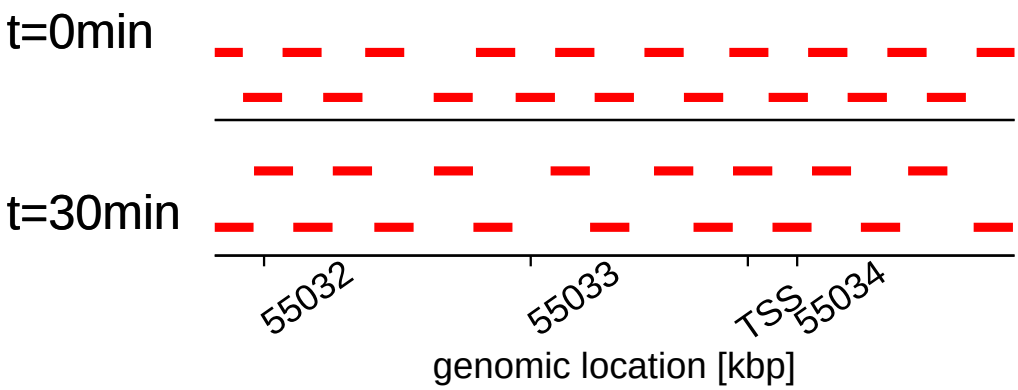
(b) Nucleosome positions along the genome for the same -2000bp +1000bp region around the TSS

Fig. 5.16 Changes to nucleosome positions at TSS of *EDN1* shown both in contact map (a) and actual nucleosome positions along the genome (b).

For *SAMD4A* the results are in Fig. 5.17. With a much lower change in nucleosome numbers (20 to 18), the appearance of a strong boundary seems to come from the loss of a key nucleosome which shifts the interactions between the nucleosomes. Since *SAMD4A* was found to be responsive to treatment, changing to an active state, this is likely to be a nucleosome near the promoter, which either shifts or detaches. This causes a shift in the interaction pattern.



(a) Contact maps for a region of -2000bp +1000bp around the TSS, as the region is given in base pairs, the number of nucleosomes differs between the t=0min and t=30min maps. Boundaries found by the algorithm are shown above then contact map.

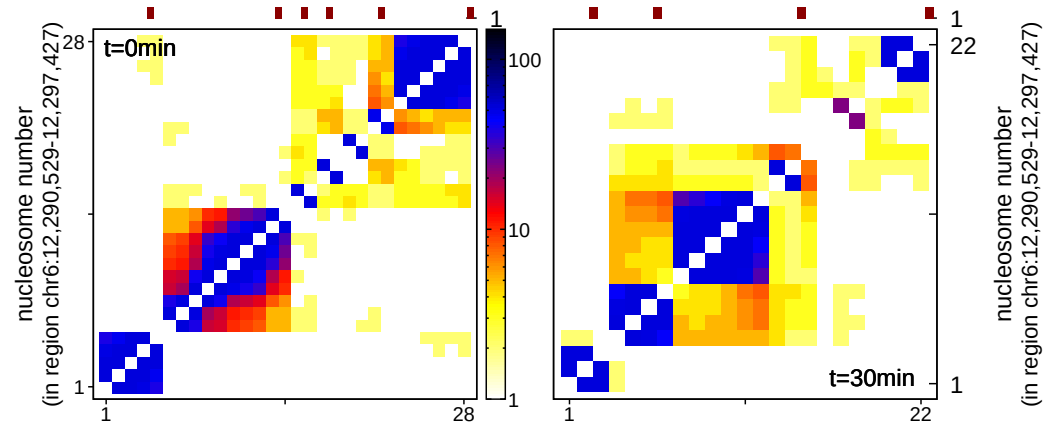


(b) Nucleosome positions along the genome for the same -2000bp +1000bp region around the TSS

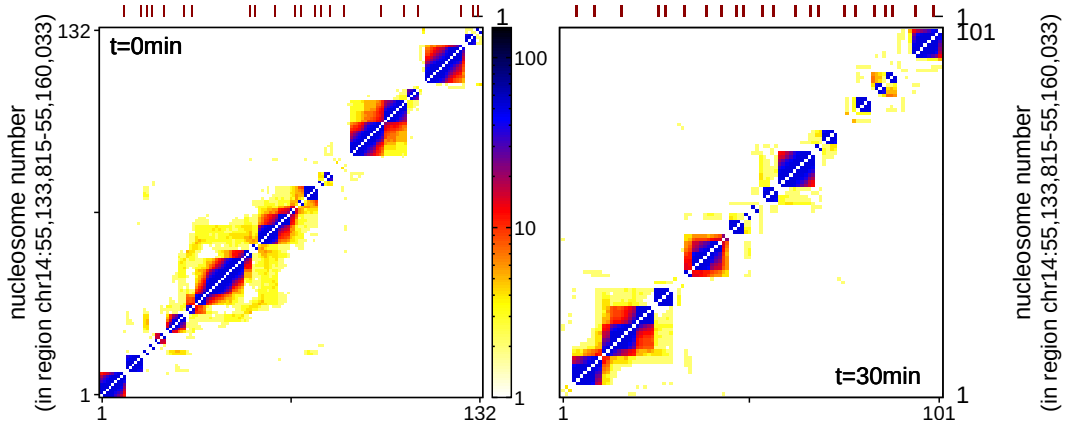
Fig. 5.17 Changes to nucleosome positions at TSS of *SAMD4A* shown both in contact map (a) and actual nucleosome positions along the genome (b).

5.4.2 Changes in the gene body

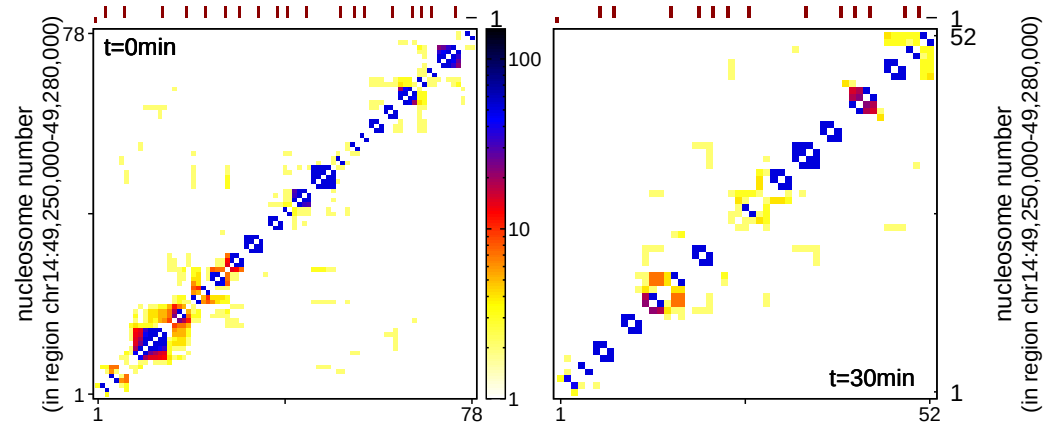
In Fig. 5.18, the entire gene is shown in the case of *EDN1* and a subsection of the region for *SAMD4A* and HECHRO. Since the number of nucleosomes varies in the regions, the t=0min and t=30min data sets are plotted on the left and right respectively. Each region has a number of changes to the interaction domains with the HECHRO region losing almost all strong interactions beyond 3-4 nucleosomes.



(a) *EDN1* gene



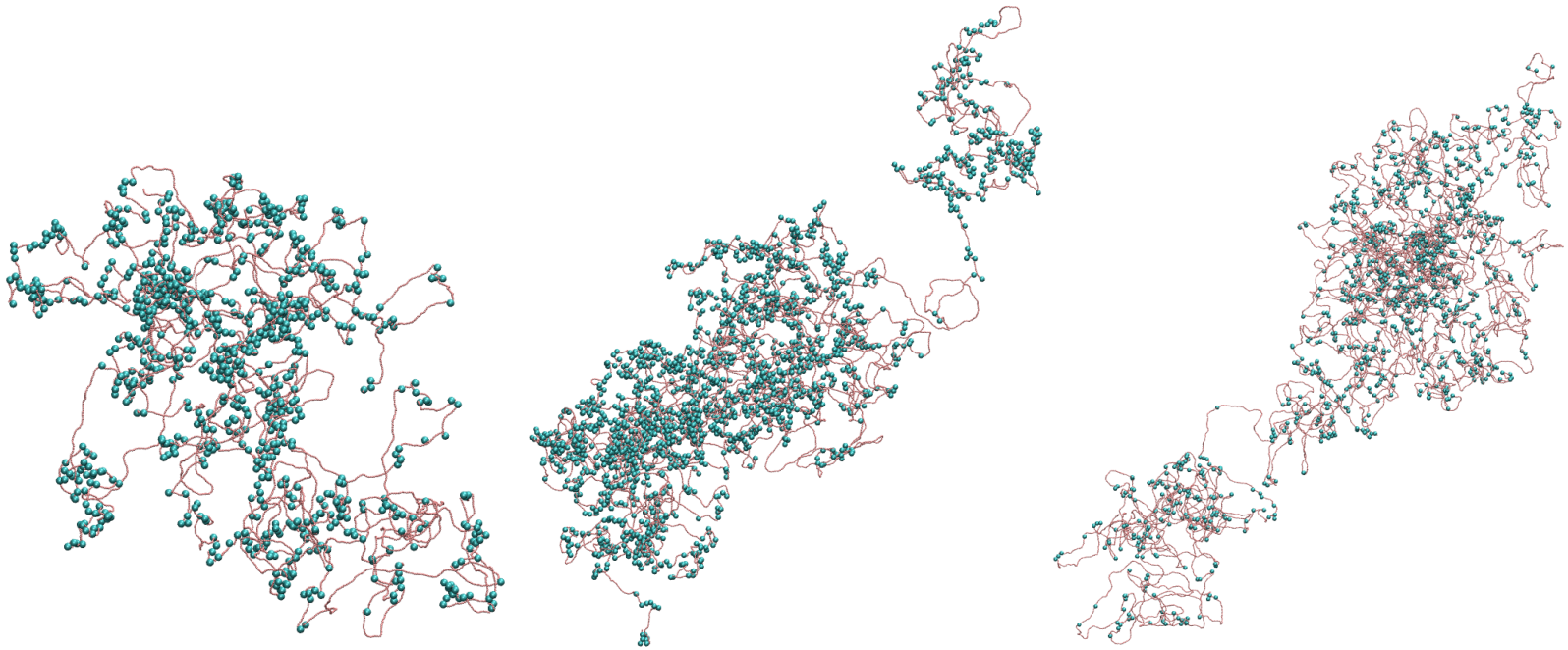
(b) *SAMD4A* gene subsection



(c) heterochromatin region subsection

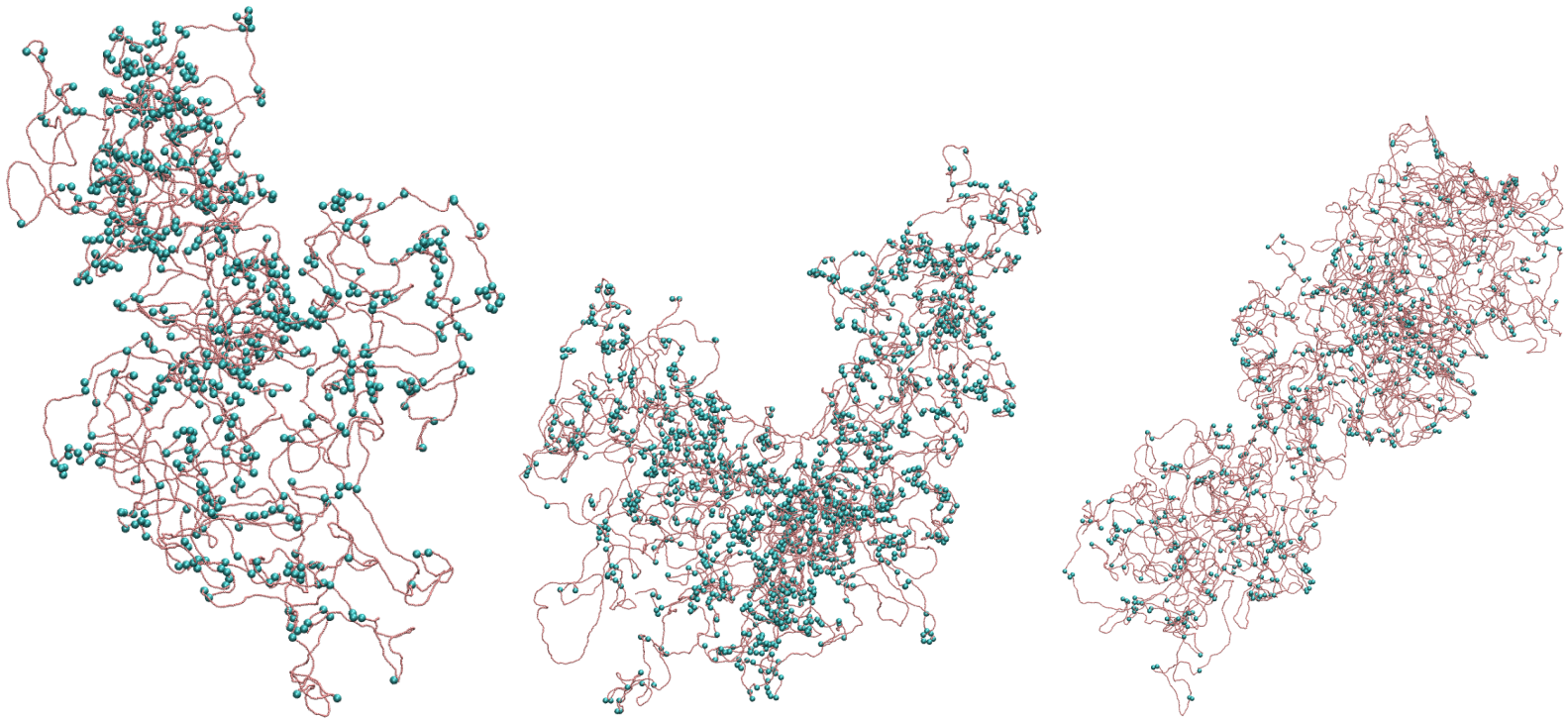
Fig. 5.18 Contact maps showing potential changes within gene body for *EDN1* (a), *SAMD4A* (b) and the heterochromatin region (c). As the latter two regions are too large to show detail, a subsection is given for both. Boundaries found within the regions are given above each contact map.

Overall it appears that, similarly to yeast, micro-domains are also found within the human genome. However, the much lower resolution of the MNase-seq data makes these predictions very speculative and no firm conclusions can be drawn yet. Despite this, the simulations were able to predict boundaries between micro-domains and changes in the interaction structure of the nucleosomes could be observed. The next step would now be to wait for the release of the Micro-C data for humans from Ref. [133, 134], which would allow for these simulation predictions to be verified. In fact the preprint papers do confirm the presence of micro-domains in both humans and mice. Once the data is publicly available, the simulations could be refined and there might be a potential in the simulation approach.



(a) *EDN1* **(b)** *SAMD4A* **(c)** heterochromatin region

Fig. 5.19 Renders of regions at t=0min generated from a single conformation. Contact maps are generated from a combination of 10 simulations.



(a) *EDN1* **(b)** *SAMD4A* **(c)** heterochromatin region

Fig. 5.20 Renders of regions at t=30min generated from a single conformation. Contact maps are generated from a combination of 10 simulations.

Chapter 6

Conclusion

In this thesis I have studied the three-dimensional organisation of chromatin, first in yeast and then in humans. I did this by combining a bioinformatic analysis of experimental data with molecular dynamics simulations. For this I presented a simple computational model for chromatin as a heteromorphic polymer, comprised of nucleosomes and linker DNA. I used the model to study nucleosome interactions within the chromosomes of the budding yeast *Saccharomyces cerevisiae* and make speculative observations of nucleosome interactions in humans.

This seemingly simple model, which represents nucleosomes as 10 nm spheres connected through linker DNA beads, is consistent with microscopy results but the inhomogeneity of the fibre is not usually included in computer simulations. This inhomogeneity is vital for the model to correctly predict the nucleosome interaction patterns observed in recent Micro-C data [1, 104].

The model is build from a bioinformatic analysis of nucleosome positioning and chromatin structure data in yeast, outlined in chapter 3. From the analysing of MNasa-seq data, a nucleosome position distribution can be established and used to find the most-likely positions of nucleosome in the chromatin fibre. The positioning data could then be used to analyse the linker lengths of DNA joining the nucleosomes and the distribution of these linkers is much larger than anticipated from the literature. In fact, the variety of linker lengths gives reason to chromatin being best viewed as a heteromorphic fibre, instead of a homogenous fibre.

In chapter 3 I also look at the structure of chromatin by analysing Micro-C data from Refs. [1, 104] which measures the contact probability of different nucleosomes along the chromatin fibre. In general the data conforms with previous findings of HiC experiments in eukaryotes, however, a novel feature is revealed. Namely, "chromosomal interaction domains" of typical length ∼1-2 kbp, which are much smaller than previously observed domains and which I refer to as micro-domains. The nucleosomes throughout the chromatin fibre have a higher likelihood of interacting with nucleosomes within the same micro-domain as opposed to other micro-domains.

In chapter 4 I described the newly developed model and the simulations carried out with it of a heterogeneous fibre in yeast. The fibre is generated purely from the positions of nucleosomes found experimentally and contact maps generated from the simulation are in good agreement with experimental findings. Quantitative comparison can be done by calling the locations of micro-domain "boundaries" and comparing the results between simulation and experiment. The model developed had sufficient detail to correctly determine the positions of these domain boundaries. For the eight simulated regions, 84% of domain boundaries were correctly identified as boundaries also present in Micro-C data. The initially good agreement between simulation and experiment warranted an attempt at improving the model further in order to increase the agreement. Inclusion of additional features found in nucleosomes however, such as disk-like nucleosome shapes and constraints on the entry/exit angle of linker DNA, did not show any significant improvement in the agreement.

The data used as input to the simulations only consisted of the most-likely nucleosome positions as determined from MNase-seq data. The implication here is then, that the formation of micro-domain patterns is, at least in part, down to the positioning of the nucleosomes. Previous work [1] found that the domain boundaries had increased binding of some proteins and nucleosomes either side of a boundary were enriched for transcriptional activating histone modifications. The work here suggests that these may not be directly responsible and the binding of proteins might directly or indirectly maintain nucleosome depleted regions or cause them to occur, which in turn cause micro-domain boundaries to appear.

These results suggest that the nucleosome positions are a cause of the micro-domains observed in the experimental Micro-C data, as such they are a "signature" of regulatory mechanisms (the domains follow a function). This is in contrast to much of the current understanding on the formation of larger domains

in eukaryotes, which appear to regulate the chromatin interactions and as such control expression (the function is driven by domains).

The initial simulations carried out with yeast studied the effects of irregular nucleosome positioning on the three-dimensional conformation of a chromatin fibre. This is also discussed in chapter 4 and was done by using the specific conformations gained from the simulations and calculating the radius of gyration as a measure of the physical properties of the fibre. In particular it allows the probing of relative shape, density and rigidity via the persistence length. An expected finding was that fibres with uniform nucleosome spacing in fact do not produce any domains. However, irregular spacing of nucleosomes does produce interaction domains in contact maps, as such there must be a close link between the nucleosome positioning and the chromatin interactions at the nucleosome level. Compared to regular spacing of nucleosomes on a chromatin fibre, the irregular spacing leads to an overall reduction in the size of the polymer. The local compaction of a region of chromatin is closely linked to the number of nucleosome found within that region. Within the model, the 3D size is in a non-linear relationship to the number of nucleosome. The size would be reduced with a low number of nucleosomes compared to a region of linker DNA with no nucleosomes, but increased with a higher number of nucleosomes. It appears that this phenomenon is closely linked to how the persistence length within the model changes depending on the number of nucleosomes. Although the simple model presented does not include a constraint on the entry/exit angle of linker DNA, the introduction of this would certainly affect these results. The more detailed version of the model presented did include the angle constraint, but further work is needed to investigate the effect on the size of the polymer.

Although genome-wide data on nucleosome positions have been available for several years, the striking irregularity in nucleosome spacing is often overlooked. An interesting aspect to study in future would be how the irregular spacing affects the 3D structure at longer length scales [139]. A future model could include effects of torsional rigidity, limiting the rotation of DNA and nucleosomes and as such would control the relative orientation of the nucleosome as currently they are free to rotate. Another challenge this model faces is that it is generated from a set of "most-likely" nucleosome positions for each region. The MNase-seq data provides data from a population of cells with continuous variation and not from a single cell [140]. Further, the model only considers static positions of nucleosomes, which in reality are likely to be much more dynamic. Multiple factors such as

DNA sequence, the action remodelling complexes and histone chaperones, as well as transcription and replication, play important roles in defining the positions of nucleosomes.

Yeast is a unicellular organism, but is often considered a model organism for higher eukaryotes as it itself is a simple type of eukaryote to study. However, many of the findings from experiments and simulations are known to transfer to higher eukaryotes [**?**], but not without caveats. The results in chapter 5 follow this principle and as such when the nucleosome interaction model is used for the human DNA, it is primarily designed as a preliminary foray into working with human nucleosome positioning data in order to draw some speculative conclusions. In fact it proved surprisingly successful in recreating results gained from the work in yeast. The inhomogeneous chromatin fibre model can be successfully used to predict elements of the structure of human chromatin. By using nucleosome positioning data alone, it shows that micro-domains are likely to exist at a nucleosome resolution level within humans. As the nucleosome positioning is the primary factor in predicting chromatin structure, changes in the positioning, due to stimuli for example, will cause the structure to be prone to change accordingly.

There are however a number of limitations to the work carried out so far. The primary issue would be the lack of Micro-C data for humans to verify any predictions from the model as well as the much lower resolution of the MNase-seq data. Thus, the nucleosome position data is based on limited confidence, as such, predictions are made without firm conclusions. Also simulations for models in higher eukaryotes would also benefit from features such as active protein bridges to paint a clearer picture of the structure. This is especially true in cases such as *SAMD4A* where changes in the structure caused a change in activation, but ultimately the structural changes were not considerably different from the other regions. However, the preprint papers in Refs. [133, 134] are good news in this regard since upon publication Micro-C data should become available for both mice and humans. In fact, the simulation model's prediction of micro-domains in humans is confirmed by Ref. [134] and as Micro-C data is essentially nucleosome positioning data, a higher resolution MNase-seq data set could be gained as well. The preprints give high hopes that the model developed in this Thesis could be extended further to cover nucleosome interactions in more detail and any further developments on the model itself might further the studies of chromatin conformation.

Bibliography

[1] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, "Mapping nucleosome resolution chromosome folding in yeast by micro-c," *Cell*, vol. 162, no. 1, pp. 108–119, 2015.

[2] W. Dang, G. L. Sutphin, J. A. Dorsey, G. L. Otte, K. Cao, R. M. Perry, J. J. Wanat, D. Saviolaki, C. J. Murakami, S. Tsuchiyama, *et al.*, "Inactivation of yeast isw2 chromatin remodeling enzyme mimics longevity effect of calorie restriction via induction of genotoxic stress response," *Cell metabolism*, vol. 19, no. 6, pp. 952–966, 2014.

[3] O. Wiese, D. Marenduzzo, and C. A. Brackley, "Nucleosome positions alone can be used to predict domains in yeast chromosomes," *Proceedings of the National Academy of Sciences*, 2019.

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

[5] L. Pray, "Discovery of dna structure and function: Watson and crick," *Nature Education*, vol. 1, no. 1, p. 100, 2008.

[6] A. F. Palazzo and T. R. Gregory, "The case for junk dna," *PLoS genetics*, vol. 10, no. 5, p. e1004351, 2014.

[7] C. R. Calladine, H. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA: the molecule and how it works (3rd edn)*. Elsevier Academic press, 2004.

[8] H. Schiessel, "The physics of chromatin," *Journal of Physics: Condensed Matter*, vol. 15, no. 19, p. R699, 2003.

[9] R. D. Kornberg and A. Klug, "The nucleosome," *Scientific American*, vol. 244, no. 2, pp. 52–65, 1981.

[10] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[11] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond, "X-ray structure of a tetranucleosome and its implications for the chromatin fibre," *Nature*, vol. 436, no. 7047, pp. 138–141, 2005.

[12] T. Richmond, J. Finch, B. Rushton, D. Rhodes, and A. Klug, "Structure of the nucleosome core particle at 7 å resolution," *Nature*, vol. 311, no. 5986, p. 532, 1984.

[13] R. T. Kamakaka and S. Biggins, "Histone variants: deviants?," *Genes & development*, vol. 19, no. 3, pp. 295–316, 2005.

[14] K. Sarma and D. Reinberg, "Histone variants meet their match," *Nature reviews Molecular cell biology*, vol. 6, no. 2, p. 139, 2005.

[15] I. Walker, "Differential dissociation of histone tails from core chromatin," *Biochemistry*, vol. 23, no. 23, pp. 5622–5628, 1984.

[16] R. Smith and R. Rill, "Mobile histone tails in nucleosomes. assignments of mobile segments and investigations of their role in chromatin folding.," *Journal of Biological Chemistry*, vol. 264, no. 18, pp. 10574–10581, 1989.

[17] Z. Zhang and B. F. Pugh, "High-resolution genome-wide mapping of the primary structure of chromatin," *Cell*, vol. 144, no. 2, pp. 175–186, 2011.

[18] J. Allan, N. Harborne, D. C. Rau, and H. Gould, "Participation of core histone" tails" in the stabilization of the chromatin solenoid.," *The Journal of cell biology*, vol. 93, no. 2, pp. 285–297, 1982.

[19] P. M. Schwarz and J. C. Hansen, "Formation and stability of higher order chromatin structures. contributions of the histone octamer.," *Journal of Biological Chemistry*, vol. 269, no. 23, pp. 16284–16289, 1994.

[20] M. Garcia-Ramirez, C. Rocchini, and J. Ausio, "Modulation of chromatin folding by histone acetylation," *Journal of Biological Chemistry*, vol. 270, no. 30, pp. 17923–17928, 1995.

[21] C. L. White, R. K. Suto, and K. Luger, "Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions," *The EMBO journal*, vol. 20, no. 18, pp. 5207–5218, 2001.

[22] J. C. Hansen, "Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions," *Annual review of biophysics and biomolecular structure*, vol. 31, no. 1, pp. 361–392, 2002.

[23] J. Allan, T. Mitchell, N. Harborne, L. Bohm, and C. Crane-Robinson, "Roles of h1 domains in determining higher order chromatin structure and h1 location," *Journal of molecular biology*, vol. 187, no. 4, pp. 591–601, 1986.

[24] T. A. Blank and P. B. Becker, "Electrostatic mechanism of nucleosome spacing," *Journal of molecular biology*, vol. 252, no. 3, pp. 305–313, 1995.

[25] Y. Fan, T. Nikitina, E. M. Morin-Kensicki, J. Zhao, T. R. Magnuson, C. L. Woodcock, and A. I. Skoultchi, "H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo," *Molecular and cellular biology*, vol. 23, no. 13, pp. 4559–4572, 2003.

[26] X. Shen and M. A. Gorovsky, "Linker histone h1 regulates specific gene expression but not global transcription in vivo," *Cell*, vol. 86, no. 3, pp. 475–483, 1996.

[27] Y. Fan, T. Nikitina, J. Zhao, T. J. Fleury, R. Bhattacharyya, E. E. Bouhassira, A. Stein, C. L. Woodcock, and A. I. Skoultchi, "Histone h1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation," *Cell*, vol. 123, no. 7, pp. 1199–1212, 2005.

[28] X. Lu, S. N. Wontakal, H. Kavi, B. J. Kim, P. M. Guzzardo, A. V. Emelyanov, N. Xu, G. J. Hannon, J. Zavadil, D. V. Fyodorov, *et al.*, "Drosophila h1 regulates the genetic activity of heterochromatin by recruitment of su (var) 3-9," *Science*, vol. 340, no. 6128, pp. 78–81, 2013.

[29] T. J. Richmond and C. A. Davey, "The structure of dna in the nucleosome core," *Nature*, vol. 423, no. 6936, p. 145, 2003.

[30] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, "A novel roll-and-slide mechanism of dna folding in chromatin: implications for nucleosome positioning," *Journal of molecular biology*, vol. 371, no. 3, pp. 725–738, 2007.

[31] D. E. Olins and A. L. Olins, "Chromatin history: our view from the bridge," *Nature reviews Molecular cell biology*, vol. 4, no. 10, p. 809, 2003.

[32] C. A. Meyer and X. S. Liu, "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology," *Nature Reviews Genetics*, vol. 15, no. 11, p. 709, 2014.

[33] M. Tsompana and M. J. Buck, "Chromatin accessibility: a window into the genome," *Epigenetics & chromatin*, vol. 7, no. 1, p. 33, 2014.

[34] X. Zhou, A. W. Blocker, E. M. Airoldi, and E. K. O'Shea, "A computational approach to map nucleosome positions and alternative chromatin states with base pair resolution," *Elife*, vol. 5, p. e16970, 2016.

[35] P. Bu, Y. A. Evrard, G. Lozano, and S. Y. Dent, "Loss of gcn5 acetyltransferase activity leads to neural tube closure defects and exencephaly in mouse embryos," *Molecular and cellular biology*, vol. 27, no. 9, pp. 3405–3416, 2007.

[36] A. Shilatifard, "Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression," *Annu. Rev. Biochem.*, vol. 75, pp. 243–269, 2006.

[37] C. M. Whittle, K. N. McClinic, S. Ercan, X. Zhang, R. D. Green, W. G. Kelly, and J. D. Lieb, "The genomic distribution and function of histone variant htz-1 during c. elegans embryogenesis," *PLoS genetics*, vol. 4, no. 9, p. e1000187, 2008.

[38] G. J. Narlikar, R. Sundaramoorthy, and T. Owen-Hughes, "Mechanisms and functions of atp-dependent chromatin-remodeling enzymes," *Cell*, vol. 154, no. 3, pp. 490–503, 2013.

[39] C. Y. Zhou, S. L. Johnson, N. I. Gamarra, and G. J. Narlikar, "Mechanisms of atp-dependent chromatin remodeling motors," *Annual review of biophysics*, vol. 45, pp. 153–181, 2016.

[40] C. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nature Reviews Genetics*, vol. 10, no. 3, p. 161, 2009.

[41] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, p. 772, 2006.

[42] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "Nucleosome sequence preferences influence in vivo nucleosome organization," *Nature structural & molecular biology*, vol. 17, no. 8, p. 918, 2010.

[43] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nature structural & molecular biology*, vol. 20, no. 3, pp. 267–273, 2013.

[44] A. L. Hughes and O. J. Rando, "Mechanisms underlying nucleosome positioning in vivo," *Annual review of biophysics*, vol. 43, pp. 41–63, 2014.

[45] H. S. Rhee, A. R. Bataille, L. Zhang, and B. F. Pugh, "Subnucleosomal structures and nucleosome asymmetry across a genome," *Cell*, vol. 159, no. 6, pp. 1377–1388, 2014.

[46] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature*, vol. 403, no. 6765, p. 41, 2000.

[47] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, *et al.*, "Genome-wide map of nucleosome acetylation and methylation in yeast," *Cell*, vol. 122, no. 4, pp. 517–527, 2005.

[48] C. M. Weber and S. Henikoff, "Histone variants: dynamic punctuation in transcription," *Genes & development*, vol. 28, no. 7, pp. 672–682, 2014.

[49] O. J. Rando and K. Ahmad, "Rules and regulation in the primary structure of chromatin," *Current opinion in cell biology*, vol. 19, no. 3, pp. 250–256, 2007.

[50] C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb, "Evidence for nucleosome depletion at active regulatory regions genome-wide," *Nature genetics*, vol. 36, no. 8, p. 900, 2004.

[51] E. A. Sekinger, Z. Moqtaderi, and K. Struhl, "Intrinsic histone-dna interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast," *Molecular cell*, vol. 18, no. 6, pp. 735–748, 2005.

[52] B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, and S. L. Schreiber, "Global nucleosome occupancy in yeast," *Genome biology*, vol. 5, no. 9, p. R62, 2004.

[53] B. Guillemette, A. R. Bataille, N. Gévry, M. Adam, M. Blanchette, F. Robert, and L. Gaudreau, "Variant histone h2a. z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning," *PLoS biology*, vol. 3, no. 12, p. e384, 2005.

[54] M. A. Schwabish and K. Struhl, "Evidence for eviction and rapid deposition of histones upon transcriptional elongation by rna polymerase ii," *Molecular and cellular biology*, vol. 24, no. 23, pp. 10111–10117, 2004.

[55] S. J. Zanton and B. F. Pugh, "Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock," *Genes & development*, vol. 20, no. 16, pp. 2250–2265, 2006.

[56] H. Zhang, D. N. Roberts, and B. R. Cairns, "Genome-wide dynamics of htz1, a histone h2a variant that poises repressed/basal promoters for activation through histone loss," *Cell*, vol. 123, no. 2, pp. 219–231, 2005.

[57] S. K. Kurdistani, S. Tavazoie, and M. Grunstein, "Mapping global histone acetylation patterns to gene expression," *Cell*, vol. 117, no. 6, pp. 721–733, 2004.

[58] M. Vogelauer, J. Wu, N. Suka, and M. Grunstein, "Global histone acetylation and deacetylation in yeast," *Nature*, vol. 408, no. 6811, p. 495, 2000.

[59] B. E. Bernstein, E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides, and S. L. Schreiber, "Methylation of histone h3 lys 4 in coding regions of active genes," *Proceedings of the National Academy of Sciences*, vol. 99, no. 13, pp. 8695–8700, 2002.

[60] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, "Genome-scale identification of nucleosome positions in s. cerevisiae," *Science*, vol. 309, no. 5734, pp. 626–630, 2005.

[61] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh, "A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome," *Science*, vol. 332, no. 6032, pp. 977–980, 2011.

[62] C. Jiang and B. F. Pugh, "A compiled and systematic reference map of nucleosome positions across the saccharomyces cerevisiae genomes," *Genome biology*, vol. 10, no. 10, p. R109, 2009.

[63] F. Cui, H. A. Cole, D. J. Clark, and V. B. Zhurkin, "Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing," *Nucleic acids research*, vol. 40, no. 21, pp. 10753–10764, 2012.

[64] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh, "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome research*, vol. 18, no. 7, pp. 1073–1083, 2008.

[65] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V. R. Iyer, "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation," *PLoS biology*, vol. 6, no. 3, p. e65, 2008.

[66] O. I. Kulaeva, F.-K. Hsieh, H.-W. Chang, D. S. Luse, and V. M. Studitsky, "Mechanism of transcription through a nucleosome by rna polymerase ii," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1829, no. 1, pp. 76–83, 2013.

[67] C. M. Weber, S. Ramachandran, and S. Henikoff, "Nucleosomes are context-specific, h2a. z-modulated barriers to rna polymerase," *Molecular cell*, vol. 53, no. 5, pp. 819–830, 2014.

[68] R. Reja, V. Vinayachandran, S. Ghosh, and B. F. Pugh, "Molecular mechanisms of ribosomal protein gene coregulation," *Genes & development*, vol. 29, no. 18, pp. 1942–1954, 2015.

[69] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow, "A high-resolution atlas of nucleosome occupancy in yeast," *Nature genetics*, vol. 39, no. 10, p. 1235, 2007.

[70] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[71] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, no. 5, pp. 887–898, 2008.

[72] I. P. Ioshikhes, I. Albert, S. J. Zanton, and B. F. Pugh, "Nucleosome positions predicted through comparative genomics," *Nature genetics*, vol. 38, no. 10, p. 1210, 2006.

[73] E. Segal and J. Widom, "From dna sequence to transcriptional behaviour: a quantitative approach," *Nature Reviews Genetics*, vol. 10, no. 7, p. 443, 2009.

[74] A. T. Annunziato, "Split decision: what happens to nucleosomes during dna replication?," *Journal of Biological Chemistry*, vol. 280, no. 13, pp. 12065–12068, 2005.

[75] S. Venkatesh and J. L. Workman, "Histone exchange, chromatin structure and the regulation of transcription," *Nature reviews Molecular cell biology*, vol. 16, no. 3, p. 178, 2015.

[76] R. B. Deal, J. G. Henikoff, and S. Henikoff, "Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones," *Science*, vol. 328, no. 5982, pp. 1161–1164, 2010.

[77] D. Ray-Gallet, A. Woolfe, I. Vassias, C. Pellentz, N. Lacoste, A. Puri, D. C. Schultz, N. A. Pchelintsev, P. D. Adams, L. E. Jansen, *et al.*, "Dynamics of histone h3 deposition in vivo reveal a nucleosome gap-filling mechanism for h3. 3 to maintain chromatin integrity," *Molecular cell*, vol. 44, no. 6, pp. 928–941, 2011.

[78] D. C. Kraushaar, W. Jin, A. Maunakea, B. Abraham, M. Ha, and K. Zhao, "Genome-wide incorporation dynamics reveal distinct categories of turnover for the histone variant h3. 3," *Genome biology*, vol. 14, no. 10, p. R121, 2013.

[79] O. Yildirim, J.-H. Hung, R. J. Cedeno, Z. Weng, C. J. Lengner, and O. J. Rando, "A system for genome-wide histone variant dynamics in es cells reveals dynamic macroh2a2 replacement at promoters," *PLoS genetics*, vol. 10, no. 8, p. e1004515, 2014.

[80] J. P. Svensson, M. Shukla, V. Menendez-Benito, U. Norman-Axelsson, P. Audergon, I. Sinha, J. C. Tanny, R. C. Allshire, and K. Ekwall, "A nucleosome turnover map reveals that the stability of histone h4 lys20 methylation depends on histone recycling in transcribed chromatin," *Genome research*, vol. 25, no. 6, pp. 872–883, 2015.

[81] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *science*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[82] T. Cremer, C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders, "Analysis of chromosome positions in the interphase nucleus of chinese hamster cells by laser-uv-microirradiation experiments," *Human genetics*, vol. 62, no. 3, pp. 201–209, 1982.

[83] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, *et al.*, "Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes," *PLoS biology*, vol. 3, no. 5, p. e157, 2005.

[84] T. Cremer, M. Cremer, S. Dietzel, S. Müller, I. Solovei, and S. Fakan, "Chromosome territories–a functional nuclear landscape," *Current opinion in cell biology*, vol. 18, no. 3, pp. 307–316, 2006.

[85] S. Chambeyron and W. A. Bickmore, "Chromatin decondensation and nuclear reorganization of the hoxb locus upon induction of transcription," *Genes & development*, vol. 18, no. 10, pp. 1119–1130, 2004.

[86] C. Ferrai, I. J. de Castro, L. Lavitas, M. Chotalia, and A. Pombo, "Gene positioning," *Cold Spring Harbor perspectives in biology*, vol. 2, no. 6, p. a000588, 2010.

[87] E. Splinter, H. Heath, J. Kooren, R.-J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. de Laat, "Ctcf mediates long-range chromatin looping and local histone modification in the β-globin locus," *Genes & development*, vol. 20, no. 17, pp. 2349–2354, 2006.

[88] H. Würtele and P. Chartrand, "Genome-wide scanning of hoxb1-associated loci in mouse es cells using an open-ended chromosome conformation capture methodology," *Chromosome Research*, vol. 14, no. 5, pp. 477–495, 2006.

[89] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat, "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c)," *Nature genetics*, vol. 38, no. 11, p. 1348, 2006.

[90] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *science*, vol. 326, no. 5950, pp. 289–293, 2009.

[91] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, p. 376, 2012.

[92] T. B. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-resolution mapping of the spatial organization of a bacterial chromosome," *Science*, vol. 342, no. 6159, pp. 731–734, 2013.

[93] A. Goffeau, B. G. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, *et al.*, "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[94] C. Rabl, "Uber zelltheilung," *Morphol. Jahrb.*, vol. 10, pp. 214–330, 1885.

[95] H. Jin, H. Zhou, X. Cheng, R. Tang, M. Munoz, and N. Nguyen, "Recombinant respiratory syncytial viruses with deletions in the ns1, ns2, sh, and m2-2 genes are attenuated in vitro and in vivo," *Virology*, vol. 273, no. 1, pp. 210–218, 2000.

[96] K. Bystricky, P. Heun, L. Gehlen, J. Langowski, and S. M. Gasser, "Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques," *Proceedings of the National Academy of Sciences*, vol. 101, no. 47, pp. 16495–16500, 2004.

[97] C. Zimmer and E. Fabre, "Principles of chromosomal organization: lessons from yeast," *The Journal of cell biology*, vol. 192, no. 5, pp. 723–733, 2011.

[98] J. C. Mell and S. M. Burgess, "Yeast as a model genetic organism," *e LS*, 2001.

[99] G. Walker and G. Stewart, "Saccharomyces cerevisiae in the production of fermented beverages," *Beverages*, vol. 2, no. 4, p. 30, 2016.

[100] J. R. Van der Maarel, *Introduction to biopolymer physics*. World Scientific Publishing Company, 2007.

[101] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods*, vol. 9, p. 357, 2012.

[102] G. G. Yardımcı and W. S. Noble, "Software tools for visualizing hi-c data," *Genome biology*, vol. 18, no. 1, p. 26, 2017.

[103] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, no. 7296, p. 363, 2010.

[104] T.-H. Hsieh, G. Fudenberg, A. Goloborodko, and O. Rando, "Micro-c xl: assaying chromosome conformation from the nucleosome to the entire genome," *Nature Methods*, vol. 13, p. 1009, 2016.

[105] C. A. Brackley, M. E. Cates, and D. Marenduzzo, "Facilitated diffusion on mobile dna: Configurational traps and sequence heterogeneity," *Physics Review Letters*, vol. 109, p. 168103, Oct 2012.

[106] C. A. Brackley, S. Taylor, A. Papantonisc, P. R. Cook, and D. Marenduzzo, "Nonspecific bridging-induced attraction drives clustering of dna-binding proteins and genome organization," *Proceedings of the National Academy of Sciences USA*, vol. 110, pp. E3605–E3611, 2013.

[107] C. A. Brackley, A. N. Morozov, and D. Marenduzzo, "Models for twistable elastic polymers in brownian dynamics, and their implementation for lammps," *Journal of Chemical Physics*, vol. 140, no. 13, p. 135103, 2014.

[108] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," *Journal of Computational Physics*, vol. 117, no. 1, pp. 1–19, 1995.

[109] A. Polishko, E. M. Bunnik, K. G. Le Roch, and S. Lonardi, "Puffin-a parameter-free method to build nucleosome maps from paired-end reads," *BMC bioinformatics*, vol. 15, no. 9, p. S11, 2014.

[110] R. Schöpflin, V. B. Teif, O. Müller, C. Weinberg, K. Rippe, and G. Wedemann, "Modeling nucleosome position distributions from experimental nucleosome positioning maps," *Bioinformatics*, vol. 29, no. 19, pp. 2380–2386, 2013.

[111] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, "Genome-scale identification of nucleosome positions in s. cerevisiae," *Science*, vol. 309, no. 5734, pp. 626–630, 2005.

[112] H. J. Szerlong and J. C. Hansen, "Nucleosome distribution and linker dna: Connecting nuclear function to dynamic chromatin structure.," *Biochemistry and Cell Biology*, vol. 89, p. 24, 2011.

[113] T. S. Kim, C. L. Liu, M. Yassour, J. Holik, N. Friedman, S. Buratowski, and O. J. Rando, "Rna polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast," *Genome biology*, vol. 11, no. 7, p. R75, 2010.

[114] Z. Duan, M. Andronescu, K. Schutz, S. Mcllwain, Y. J. Kim, J. Lee, C. Shendure, S. Fields, A. Blau, and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, pp. 363–367, 2010.

[115] T. M. Cheng, S. Heeger, R. A. Chaleil, N. Matthews, A. Stewart, J. Wright, C. Lim, P. A. Bates, and F. Uhlmann, "A simple biophysical model emulates budding yeast chromosome condensation," *eLife*, vol. 4, 2015.

[116] G. Arya and T. Schlick, "A tale of tails: How histone tails mediate chromatin compaction in different salt and linker histone environments," *Journal of Physical Chemistry A*, vol. 113, pp. 4045–4059, 2009.

[117] O. Perišić, R. Collepardo-Guevara, and T. Schlick, "Modeling studies of chromatin fiber structure as a function of dna linker length," *Journal of Molecular Biology*, vol. 403, pp. 777–802, 2010.

[118] O. Müller, N. Kepper, R. Schöpflin, R. Ettig, K. Rippe, and G. Wedemann, "Changing chromatin fiber conformation by nucleosome repositioning," *Biophysical Journal*, vol. 107, no. 9, pp. 2141–2150, 2014.

[119] C. Brackley, J. Allan, D. Keszenman-Pereyra, and D. Marenduzzo, "Topological constraints strongly affect chromatin reconstitution in silico," *Nucleic Acids Research*, vol. 43, pp. 63–73, 2015.

[120] H. Tjong, K. Gong, L. Chen, and F. Alber, "Physical tethering and volume exclusion determine higher-order genome organization in budding yeast," *Genome Research*, vol. 22, 2012.

[121] H. Wong, H. Marie-Nelly, S. Herbert, P. Carrivain, H. Blanc, R. Koszul, E. Fabre, and C. Zimmer, "A predictive computational model of the dynamic 3d interphase yeast nucleus," *Current Biology*, vol. 22, pp. 1881–1890, 2012.

[122] H. Hajjoul, J. Mathon, H. Ranchon, I. Goiffon, J. Mozziconacci, B. Albert, P. Carrivain, J.-M. Victor, O. Gadal, K. Bystricky, and A. Bancaud, "High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome," *Genome Research*, vol. 23, pp. 1829–1838, 2013.

[123] A. Rosa and R. Everaers, "Structure and dynamics of interphase chromosomes," *PLoS computational biology*, vol. 4, no. 8, 2008.

[124] E. A. Hoffman, B. L. Frey, L. M. Smith, and D. T. Auble, "Formaldehyde crosslinking: a tool for the study of chromatin complexes," *Journal of Biological Chemistry*, vol. 290, no. 44, pp. 26404–26411, 2015.

[125] A. Buckle, C. A. Brackley, S. Boyle, D. Marenduzzo, and N. Gilbert, "Polymer simulations of heteromorphic chromatin predict the 3d folding of complex genomic loci," *Molecular cell*, vol. 72, no. 4, pp. 786–797, 2018.

[126] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, p. 251, 1997.

[127] A. R. Cutter and J. J. Hayes, "A brief review of nucleosome structure," *FEBS Letters*, vol. 589, pp. 2914 – 2922, 2015. 3D Genome structure.

[128] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9Å resolution††we dedicate this paper to the memory of max perutz who was particularly inspirational and supportive to t.j.r. in the early stages of this study.," *Journal of Molecular Biology*, vol. 319, no. 5, pp. 1097 – 1113, 2002.

[129] A. R. Cutter and J. J. Hayes, "A brief review of nucleosome structure," *FEBS letters*, vol. 589, no. 20, pp. 2914–2922, 2015.

[130] J. D. Watson, *Molecular biology of the gene*, vol. 1. Pearson Education India, 2004.

[131] M. Grunstein, "Histone acetylation in chromatin structure and transcription," *Nature*, vol. 389, no. 6649, pp. 349–352, 1997.

[132] G. D. Bascom, C. G. Myers, and T. Schlick, "Mesoscale modeling reveals formation of an epigenetically driven hoxc gene hub," *Proceedings of the National Academy of Sciences*, 2019.

[133] T.-H. S. Hsieh, E. Slobodyanyuk, A. S. Hansen, C. Cattoglio, O. J. Rando, R. Tjian, and X. Darzacq, "Resolving the 3d landscape of transcription-linked mammalian chromatin folding," *bioRxiv*, p. 638775, 2019.

[134] N. Krietenstein, S. Abraham, S. Venev, N. Abdennur, J. Gibcus, T.-H. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. Mirny, *et al.*, "Ultrastructural details of mammalian chromosome architecture," *bioRxiv*, p. 639922, 2019.

[135] S. Diermeier, P. Kolovos, L. Heizinger, U. Schwartz, T. Georgomanolis, A. Zirkel, G. Wedemann, F. Grosveld, T. A. Knoch, R. Merkl, *et al.*, "Tnfα signalling primes chromatin for nf-κb binding and induces rapid and widespread nucleosome repositioning," *Genome biology*, vol. 15, no. 12, p. 536, 2014.

[136] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, no. 7352, p. 516, 2011.

[137] H. J. Szerlong and J. C. Hansen, "Nucleosome distribution and linker dna: connecting nuclear function to dynamic chromatin structure," *Biochemistry and Cell Biology*, vol. 89, no. 1, pp. 24–34, 2010.

[138] D. A. Beshnova, A. G. Cherstvy, Y. Vainshtein, and V. B. Teif, "Regulation of the nucleosome repeat length in vivo by the dna sequence, protein concentrations and long-range interactions," *PLoS computational biology*, vol. 10, no. 7, p. e1003698, 2014.

[139] J.-M. Arbona, S. Herbert, E. Fabre, and C. Zimmer, "Inferring the physical properties of yeast chromatin through bayesian analysis of whole nucleus simulations," *Genome Biology*, vol. 18, p. 81, 2017.

[140] E. C. Small, L. Xi, J.-P. Wang, J. Widom, and J. D. Licht, "Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity," *Proceedings of the National Academy of Sciences USA*, vol. 111, no. 24, pp. E2462–E2471, 2014.