Predictive models for chromatin folding: connecting sequence to structure

by

Pau Farré

M.Sc., Simon Fraser University, 2014
B.Sc., University of Barcelona, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
in the
Department of Physics
Faculty of Science

© Pau Farré 2018
SIMON FRASER UNIVERSITY
Spring 2018

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Approval

Name: Pau Farré
Degree: Doctor of Philosophy (Physics)
Title: Predictive models for chromatin folding: connecting sequence to structure

Examining Committee: Chair: Dr. Malcolm Kennett
Associate Professor

Dr. Eldon Emberly
Senior Supervisor
Professor

Dr. John Bechhoefer
Supervisor
Professor

Dr. David Sivak
Supervisor
Assistant Professor

Dr. Maxwell Libbrecht
Internal Examiner
Assistant Professor
School of Computing Science

Dr. Gautam Menon
External Examiner
Professor
The Institute of Mathematical Sciences, Chennai, India

Date Defended: April 13, 2018
Abstract

The DNA packaged inside a nucleus shows complex structures stabilized by a host of DNA-bound factors. This combination of DNA and bound factors is known as chromatin. Both the distribution of bound factors and the contacts between different locations of the DNA can be now measured on a genome-wide scale. Nevertheless, to what extent is the likelihood of contact between sites in the genome encoded by the spatial sequence of bound factors? Current approaches at addressing this question primarily use simulations of heterogeneous polymers to generate structures using the locations of bound factors. In contrast, here we develop novel predictive models for connecting chromatin sequence to structure using statistical physics, information theory and machine learning. Since our methods do not require costly polymer simulations they can quickly predict the effect on structure due to changes in the distribution of bound factors. In addition, our methods are formulated in a manner that allows us to solve the inverse problem: namely, given just structural data, predict the likely sequence of bound factors. We show that the models developed can make biologically meaningful predictions, highlighting key features of the mechanisms through which the three-dimensional conformation of DNA is coordinated by the interactions between DNA-bound factors.

Keywords: Hi-C; ChIP-seq; Bayesian statistics; Maximum-entropy modelling; Artificial neural networks
For Gabi
I got war and peace inside my DNA.
– Kendrick Lamar
Acknowledgements

First and foremost, I would like to thank Eldon Emberly for his guidance, support and patience on my journey through graduate school. His unique approach to biophysical research has been a true inspiration for my work. As my Senior Supervisor, Eldon has always cared for my career advancement and opportunities and I will be forever grateful to him for welcoming me into his team.

I also thank David Sivak and John Bechhoefer for their enthusiasm for biophysics and knowledgeable advice. Their feedback during supervisory committees, courses and the continuous interactions we had during my years in the department have laid key foundations for my research.

For the mentorship and hospitality during my stay in France, I thank Olivier Cuvier and Raphael Mourad at the CNRS Toulouse.

I value the companionship of Lavisha Jindal, Saeed Saberi, Rachel Edgar, Sumaiya Islam, Sarah Goodman, Meaghan Jones, Alexandre Heurteau and Tiphaine Fabre. I’m thankful for their time, collaborations and for enduring hardships together.

I’m grateful for the friendship of Matthew Fitzpatrick and Cristina Cordoba, who have contributed immensely to my well-being throughout these years. I will miss those lunch breaks chatting about gardening and all the laughs we had.

I acknowledge the Musqueam, Squamish and Tsleil-Waututh First Nations on whose unceded territories we are privileged to live, work and play.

And last but not least, I thank my family for their love and encouragement. My parents, Jordi and Lole, and my brother Martí have always been highly supportive and ready to cheer me up from a distance. Most of all I cherish the help of my wife Gabi, whose optimism and endless support has been immeasurable. I could not have done this without her.
# Table of Contents

Approval ii

Abstract iii

Dedication iv

Quotation v

Acknowledgements vi

Table of Contents vii

List of Tables x

List of Figures xi

1 Introduction 1
  1.1 DNA, RNA and proteins: The central dogma 2
  1.2 Storing the DNA inside cells: the packing problem 4
  1.3 Hierarchical organization of DNA 6
    1.3.1 Higher-order chromatin structure 9
  1.4 Function follows form 10
    1.4.1 Promoter-enhancer loops 12
    1.4.2 Insulator loops 14
    1.4.3 Other mechanisms 15
  1.5 Measuring chromatin 16
    1.5.1 Chromatin Immuno-Precipitation (ChIP) 16
    1.5.2 Fluorescence in situ hybridization (FISH) 18
    1.5.3 Chromosome conformation capture 18
  1.6 Modelling chromatin data 22
  1.7 Thesis outline 24

2 A pairwise model for chromatin contact free energies 26
  2.1 Methods 28
2.1.1 Chromatin contact maps .......................... 28
2.1.2 DNA-bound factors .............................. 29
2.1.3 Principal Component Analysis ..................... 29
2.1.4 Free energy matrix ............................... 30
2.1.5 Free energy decomposition: PCA-based normalization 30
2.1.6 Calculating chromatin coupling energies .......... 31
2.2 Results ............................................. 32
   2.2.1 Distance-dependent free energy .................. 32
   2.2.2 PCA of free energy profiles .................... 34
   2.2.3 Interaction profiles and chromatin structure .... 34
   2.2.4 Coupling energies between chromatin factors .... 36
2.3 Conclusion ......................................... 41

3 Mutual information between chromatin sequence and structure: quantifying the role of non-local effects 43
   3.1 Methods .......................................... 44
      3.1.1 Chromatin sequence states .................... 44
      3.1.2 Chromatin contact affinities ................. 45
      3.1.3 Information-theoretic quantities ............. 47
      3.1.4 Maximum-entropy method ...................... 50
   3.2 Results .......................................... 52
      3.2.1 The importance of neighbours in forming chromatin contacts 52
      3.2.2 Non-local chromatin information ............. 56
   3.3 Conclusion ....................................... 69

4 A maximum-entropy model for predicting chromatin contacts 70
   4.1 Results .......................................... 72
      4.1.1 Maximum-entropy model ...................... 72
      4.1.2 Model parameters ............................. 74
      4.1.3 Contact map prediction ....................... 76
      4.1.4 Structural changes due to sequence mutation 77
      4.1.5 Predicting sequence from structure .......... 78
   4.2 Conclusion ....................................... 80
   4.3 Methods .......................................... 81
      4.3.1 DNA contact maps ............................ 81
      4.3.2 Binary chromatin sequence classification .... 81
      4.3.3 Training and test data sets ................... 81
      4.3.4 Extracting sequence neighbourhoods .......... 82
      4.3.5 Calculating sequence-neighbourhood statistics 83
      4.3.6 Fitting the maximum-entropy distributions 83
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.7 Fitting first-order maximum-entropy distributions</td>
<td>84</td>
</tr>
<tr>
<td>4.3.8 Inspection of model parameters</td>
<td>85</td>
</tr>
<tr>
<td>4.3.9 Structural changes due to sequence mutation</td>
<td>88</td>
</tr>
<tr>
<td>4.3.10 Predicting sequence from structure</td>
<td>88</td>
</tr>
<tr>
<td>5 Neural networks for chromatin conformation</td>
<td>91</td>
</tr>
<tr>
<td>5.1 Dense neural networks for connecting conformation to sequence</td>
<td>92</td>
</tr>
<tr>
<td>5.2 Results</td>
<td>96</td>
</tr>
<tr>
<td>5.2.1 Model predictions</td>
<td>96</td>
</tr>
<tr>
<td>5.2.2 Spatial analysis of conformational effects</td>
<td>100</td>
</tr>
<tr>
<td>5.3 Conclusion</td>
<td>105</td>
</tr>
<tr>
<td>5.4 Methods</td>
<td>106</td>
</tr>
<tr>
<td>5.4.1 ChIP-seq data</td>
<td>106</td>
</tr>
<tr>
<td>5.4.2 Hi-C data</td>
<td>106</td>
</tr>
<tr>
<td>5.4.3 Datasets for training, validating and testing</td>
<td>107</td>
</tr>
<tr>
<td>5.4.4 Neural network model details</td>
<td>107</td>
</tr>
<tr>
<td>5.4.5 Gradient analysis of DNNs</td>
<td>108</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>110</td>
</tr>
<tr>
<td>Bibliography</td>
<td>113</td>
</tr>
</tbody>
</table>
List of Tables
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Schematic of DNA, Central Dogma, Genome Sizes, Lysed E. coli.</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Hierarchical organization of DNA</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>The main chromatin ensembles</td>
<td>11</td>
</tr>
<tr>
<td>1.4</td>
<td>Protein-mediated DNA loops</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>ChIP-seq</td>
<td>17</td>
</tr>
<tr>
<td>1.6</td>
<td>fluorescence in situ hybridization</td>
<td>19</td>
</tr>
<tr>
<td>1.7</td>
<td>Chromosome conformation capture</td>
<td>21</td>
</tr>
<tr>
<td>2.1</td>
<td>Average free energy of interaction</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Free energy principal components and chromatin-binding profiles.</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>Specific energies of interaction and associated chromatin factor contacts.</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>PCA filtering can improve fit to interaction model</td>
<td>39</td>
</tr>
<tr>
<td>2.5</td>
<td>Chromatin factor coupling energies from fitting</td>
<td>40</td>
</tr>
<tr>
<td>3.1</td>
<td>HMM of contact counts</td>
<td>46</td>
</tr>
<tr>
<td>3.2</td>
<td>Venn diagram of info-theory quantities</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Autocorrelation of binding factor occupation</td>
<td>53</td>
</tr>
<tr>
<td>3.4</td>
<td>Non-local effects of the BEAF insulator</td>
<td>55</td>
</tr>
<tr>
<td>3.5</td>
<td>Local contact and chromatin specificities</td>
<td>57</td>
</tr>
<tr>
<td>3.6</td>
<td>Chromatin specificity of relative distances</td>
<td>59</td>
</tr>
<tr>
<td>3.7</td>
<td>Detailed composition</td>
<td>62</td>
</tr>
<tr>
<td>3.8</td>
<td>Binary classification chromatin specificity</td>
<td>65</td>
</tr>
<tr>
<td>3.9</td>
<td>Chromatin specificity of max-entropy distribution</td>
<td>66</td>
</tr>
<tr>
<td>3.10</td>
<td>Chromatin specificity for ( w ) flanking neighbours</td>
<td>68</td>
</tr>
<tr>
<td>4.1</td>
<td>Schematic of spin-model</td>
<td>72</td>
</tr>
<tr>
<td>4.2</td>
<td>Spin statistics of model versus experiment</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>Contact prediction in test set.</td>
<td>76</td>
</tr>
<tr>
<td>4.4</td>
<td>Structural changes by chromatin mutation</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>Prediction of bound factors from contact map</td>
<td>79</td>
</tr>
<tr>
<td>4.6</td>
<td>Schematic of the construction of spin-neighbourhoods</td>
<td>82</td>
</tr>
<tr>
<td>4.7</td>
<td>Spin statistics of first-order maximum-entropy model</td>
<td>85</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The idea that living systems inherit biological traits has been part of human knowledge for millennia. The observation that ancestors pass traits to descendants not only underlies the sophisticated breeding of species performed by humans since the Neolithic [1], but it has also been key in many forms of human organization [2]. Perhaps surprisingly, very few laws governing these mechanisms of inheritance were written down until the 19th century, when Gregor Mendel found that the inheritance of various biological traits could be explained by simple mathematical rules [3]. This crucial finding only gained popularity some time later after Charles Darwin developed the theory of evolutionary adaptation through natural selection that elegantly explained the diversity of species [4]. The concept was revolutionary and yet so simple that it sounded almost tautological: what best ensures its survival will survive.

Many decades later, in the 1930s, Max Delbrück [5] and Erwin Schrödinger [6], among others, hypothesized that the inheritable trait units that had been given the name “genes” must be coded into long polymers. Nevertheless, much of the nature of these polymers remained unknown for decades until Watson and Crick finally showed that genes were encoded in the molecular structure of deoxyribonucleic acid (DNA). The influential 1953 paper where they presented its double helical structure [7] mentions “this structure has novel features which are of considerable biological interest”. The newly found “molecule of life” looked like a computer tape, having a long double sequence of elements that even hinted at a copying mechanism. At this point, genetics was transformed into information technology.

This DNA polymer that contains all the information for how to make an organism is typically many orders of magnitude longer than the width of its confining space; in humans, two meters of DNA are packed in a micron-sized cell nucleus. It thus needs to be packed and confined into a small space, which in principle could pose a challenge for getting information out of it if some regions of the polymer are inaccessible. The recent advance of techniques for characterizing the conformation of the DNA inside a cell has allowed us to discover that this polymer is not simply randomly folded; instead it is organized into spatial domains and
this organization is important for the proper execution of genetic information. While some
regions of the DNA contain the genetic instructions for constructing cell machinery, other
DNA regions regulate when and how those instructions are executed by folding the DNA into
particular 3D configurations and attracting a host of gene-regulatory molecules. This thesis
intends to improve the understanding of the mechanisms that determine the conformation of
dNA inside the nucleus by using theoretical methods from physics and information theory.
In particular, we focus on developing models that can predict the folding of DNA inside a
cell based on the genomic distribution of regulatory factors that interact with it.

In this chapter, we summarize key discoveries about the packaging of DNA inside cells
that will motivate the theoretical methods and their corresponding application to the DNA
folding problem presented in the following chapters.

1.1 DNA, RNA and proteins: The central dogma

DNA is the polymer that encodes the instructions for building all the molecular machinery
that living systems are made of. This information is encoded as a polymeric sequence made
of an alphabet of four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T).
The order of this sequence is what distinguishes one organism from another. Each nucleotide
is formed of a phosphate group, a sugar and a nitrogenous base. Nucleotides on the same
strand are linked to each other by covalent bonds (Fig. 1.1A).

DNA is normally found as two complementary-sequence strands that coil onto each
other forming a structure known as the double helix. The double helix is a right-handed
helical structure where the nucleotides of one strand complement the nucleotides of the
other through specific hydrogen bond pairs (typically A with T and G with C), forming
what are known as Watson-Crick base-pairs (bp). Therefore, if one strand contains the
sequence “CGTAT” the other one will have “GCATA”. The double helix makes a complete
turn every 10.5 bp, measures 2 nm in diameter, and 0.3 nm/bp in contour length.

In addition to DNA, cells contain two other main biopolymers that carry information
(Fig. 1.1B). One is RNA, or ribonucleic acid, which also consists of a sequence of nu-
cleotides, but containing uracil (U) instead of T. One of the functions of RNA is to carry
a transcript of a DNA section (“CGATAT” → “CGAUUA”) as a single strand called mes-
senger RNA (mRNA) to a molecular machine called the ribosome where the message is
translated into proteins. In addition, RNA can fold onto itself to perform many of functions
including genetic regulation (by interfering with other mRNAs or the DNA itself) as well
as the catalysis of biological reactions. These properties form the basis of the “RNA world
hypothesis,” which speculates that RNA was a precursor of DNA in the origin of life [8].
The third type of information-carrying biopolymers are proteins. They are polymers made
out of 20 possible monomers called amino acids, which fold into specific three-dimensional
conformations. Proteins can bind specifically to other molecules and perform essential func-
Figure 1.1: (A) Schematic of the DNA double helix. *Credit: Encyclopaedia Britannica, Inc.; https://www.britannica.com/science/human-genome; December 2017.* (B) Central dogma of biology. *Credit: Clancy, S. and Brown, W. (2008); Translation: DNA to mRNA to Protein; Nature Education 1(1):101.* (C) Genome size of various organisms. Bars indicate size range. *Credit: Latorre and Silva, Metode (2013).* (D) DNA from a lysed E. coli cell *Credit: Dr. Ruth Kavenoff (1944-1999).* — All images are reproduced with permission from the owners.
tions such as molecular transportation inside cells, gene regulation, transcription of DNA into RNA, catalysis of biochemical reactions, cell communication and the formation of fibres. Importantly for this thesis, certain proteins can bind to DNA and interact with each other to help the DNA to form loops and fold.

Despite notable exceptions, the flow of biological information is largely in one direction (Fig. 1.1B). DNA is transcribed into mRNA by a protein called RNA polymerase, and RNA is translated into protein by a complex of protein and RNA called the ribosome. This directionality is often referred as the “central dogma of molecular biology” [9], although the word dogma must not be taken literally.

1.2 Storing the DNA inside cells: the packing problem

DNA can be thought as the “hard drive” of a cell, where all of its information is initially stored. This information is called the genome, which includes regions that code for particular proteins, known as genes, as well as non-coding regions which include sequences that regulate gene expression. Prokaryotes and archaea are simpler (and smaller) types of organisms whose genome is a circular polymer of double-stranded DNA called a chromosome. Prokaryotes lack cellular compartments such as the nucleus. They are often thought to be a well-mixed “bag of molecules,” though this is far from the truth. In contrast, the more complex eukaryotic cells have multiple linear chromosomes, and their DNA is contained inside a nucleus. In addition, about 88% of the prokaryotic and archaean genomes are occupied by protein-coding genes, whereas eukaryotic genomes are largely non-coding, with only about 2% of the sequence coding for genes.

Genomes display a large variety of sizes. Simpler organisms tend to have smaller genomes than more complex ones (Fig. 1.1C), with important outliers (the genome of the *Polychaos dubium* amoeba is 200 times longer than the human genome). Next are some examples or model organisms sorted by their genome size.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome length</th>
<th>Number of genes</th>
<th>Container volume</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em> (Bacterium)</td>
<td>4.6 Mbp</td>
<td>4288</td>
<td>0.5 μm³ cell</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em> (Yeast)</td>
<td>12.1 Mbp</td>
<td>6294</td>
<td>3.3 μm³ nucleus</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (Fly)</td>
<td>175 Mbp</td>
<td>13,600</td>
<td>80 μm³ nucleus</td>
</tr>
<tr>
<td><em>Homo sapiens</em> (Human)</td>
<td>3.3 Gbp</td>
<td>20,000</td>
<td>100 μm³ nucleus</td>
</tr>
</tbody>
</table>

Differences in genome size, container volume and number of genes of these organisms are suggestive of the necessity of diverse packaging principles. For instance, *Drosophila* and humans have a similar-sized nucleus and a similar number of genes but very different genome lengths.

As a comparative example, an *E. coli* bacterium fits a 1.5 mm-long genome into a 0.5 × 0.5 × 2 μm cell while humans have 2 m of DNA contained in a 5 μm-diameter nucleus.
Treating the DNA as a long flexible cylinder that is 2 nm wide, we estimate that there is plenty of room to fit these DNAs inside their respective receptacles: E. Coli DNA occupies 1% of the volume and human DNA occupies 2%. However, given that these genomes are 3 to 6 orders of magnitude longer than their containers, we ask whether those genomes can be simply randomly accommodated, or whether they need to be packed in a more complex manner.

As a simple first approximation to answering this question, we can view the DNA double helix as a continuous polymer with a certain bending rigidity and no volume exclusion. If we use $\mathbf{r}(s)$ to indicate the position of each genomic location $s$, and also define the unit-vector tangent to the polymer as $\hat{t}(s) = d\mathbf{r}(s)/ds$, the correlation of the polymer orientation as a function of genomic separation $\Delta s$ must follow an exponential decay, $\langle \hat{t}(s) \hat{t}(s+\Delta s) \rangle = e^{-\Delta s/l_p}$, where $l_p$ is called the persistence length [10]. For the DNA double helix, this quantity is measured to be approximately 50 nm, or 150 bp.

For a distance $\Delta s$ twice the persistence length, the orientation of the polymer will be practically uncorrelated $\langle \hat{t}(s)\hat{t}(s+2l_p) \rangle \approx 0.1$. This allows us to move from the continuous model to a discrete random-chain model made of 100 nm-long freely jointed rods, known as Kuhn segments. This discrete polymer model predicts that the root-mean-squared end-to-end distance of an $N$-segment polymer scales as the square root of its length, $\sqrt{\langle \mathbf{r}^2 \rangle} = 2l_p \sqrt{N}$ [11]. This distance happens to be 6 times greater than the length of the receptacle for the E. Coli genome (Fig. 1.1D) and 20 times for the human chromosome 1 (250 Mbp long).

An alternative estimate of the volume occupied by a random polymer can be obtained by calculating the radius of gyration $R_g$, defined as the average distance of each segment $\mathbf{r}_i$ with respect to the polymer centre of mass $\mathbf{r}_{CM}$. This quantity offers the advantage of being easier to measure experimentally than the end-to-end distance. In the limit of a long chain $N \to \infty$, $R_g$ can be related to the root-mean-squared end-to-end distance $R_g^2 = \frac{1}{6} \langle \mathbf{r}^2 \rangle$, which gives $R_g = 2l_p \sqrt{N/6}$ [11]. This result implies that if the DNA was not confined by any boundaries, a randomly coiled DNA polymer would be 5 times greater than the length of the receptacle for E. Coli and 16 times greater for the human chromosome 1. Furthermore, the calculation of the bacterial DNA size is an overestimate since the chromosome is circular and therefore forced to walk back to the origin. In contrast, the human genome contains 46 chromosomes, occupying a larger volume than just the estimated chromosome 1.

Despite a number of DNA-packaging mechanisms being discovered in bacteria [12], it is evident from the calculations above that prokaryotes can accommodate more randomly packed DNA given their smaller genomes. We therefore move to focus on describing the necessarily more complex packaging mechanisms of eukaryotic DNA.
1.3 Hierarchical organization of DNA

Eukaryotic DNA is typically found surrounded by crowds of protein complexes that bind to it and help the formation of higher-order structures. The combination of DNA and bound proteins is called chromatin. Eukaryotic chromosomes take different conformations depending on the stage of the cell cycle that they are in. Here we will focus on the interphase stage, where the cell spends most of its time growing, metabolizing nutrients, expressing genes and copying the genetic material in preparation for division.

One of the most abundant DNA-bound protein complexes are histones, whose total mass inside the nucleus is equal to the mass of DNA. Pairs of four types of histones (H2A, H2B, H3 and H4) assemble as an octameric protein core on which 147 bp of DNA wraps around. This complex is called a nucleosome and constitutes the most basic unit of DNA packaging (Fig. 1.2A). Interestingly, the DNA segment is able to wrap around the core 1.6 times while being of the same magnitude as the DNA persistence length. The average free energy cost of this bend \( \approx 70k_BT \) is compensated by the electrostatic attraction between the negatively charged DNA and the positively charged histones [13]. Nevertheless, part of the free-energy cost of nucleosome formation comes from the bending rigidity of DNA, which depends on sequence. Therefore, DNA sequences can vary in their nucleosome-binding affinity by over three orders of magnitude [14], and their positioning patterns can be largely predicted by a combination of sequence affinity and the presence of other proteins that impede nucleosome formation [15]. In-vitro experiments have observed that the 10-nm diameter nucleosomes further wrap into a compact tertiary structure called the 30-nm fibre [16–18] (Fig. 1.2B), and computer simulations agree with this picture [19]. This structure involves the addition of a fifth type of histone, H1, which neutralizes DNA charge facilitating its compaction. However, this tertiary structure has never been detected in living cells, casting many doubts about its existence [20–22].

Histones have protein tails that can be chemically modified at various positions by a host of different enzymes (Fig. 1.2C). These modifications do not affect the stability of the nucleosomes, but affect how nucleosomes interact with each other as well as with other proteins, ultimately regulating higher-order chromatin organization. Some of their known functions are the compaction of chromatin or the facilitation of the enzymatic access of the DNA, which in turn affects both gene expression and DNA replication. In fact, each tail can simultaneously accommodate multiple chemical modifications, allowing for a large number of possible histone states often referred to as the histone code. Chromatin remodelling enzymes can be attracted by already present histone modifications as well as by the presence of sequence-specific DNA-binding factors [25]. As a consequence, DNA conformation is reflective of the specific biological processes that a cell is undergoing. Interestingly, histone states are inheritable, allowing for the passing of gene-regulatory information to daughter cells. Inheritable mechanisms that alter gene expression with no underlying changes in
Figure 1.2: (A) Nucleosomes are made of eight histones and 146 basepairs of DNA. Credit: Francesca Felicia Caputi (2017) CC BY 3.0 [23] (B) Electron micrograph of the 30 nm chromatin fibre. Credit: Olins D. E. and Olins A. L.; Chromatin history: our view from the bridge; Copyright 2003, Nature Publishing Group. (C) Schematic representation of histone modifications. Credit: Mariuswalter, CC BY-SA 4.0 (D) Transmission electron micrograph of a cell. Heterochromatin and euchromatin DNA in the nucleus can be distinguished by image tone. Credit: Francastel et al. (2000) [24] (E) Schematic of a protein-mediated chromatin loop. (F) TADs are organized into A/B compartments. (G) Chromosomes occupy distinct territories in the nucleus. — All images are reproduced with permission from the owners.
the DNA sequence itself, such as histone modifications, DNA-binding proteins, or DNA methylation for example, are called epigenetic mechanisms [26].

Pioneering optical microscopy studies in the 1920s [27] distinguished between two types of chromatin in-vivo, a highly condensed type called heterochromatin and a less-condensed type called euchromatin (Fig. 1.2D). Euchromatin tends to reside in the inner region of the nucleus, whereas heterochromatin tends to localize closer to the nuclear envelope [28]. Both chromatin types contain genes, although heterochromatin has less of them and they tend to be silenced [29]. Heterochromatin is further divided into two types; constitutive heterochromatin, typically found in parts of the genome that stay silent through the cell cycle [30], and facultative heterochromatin which can alternate between a compact state and a loose state that allows for gene transcription [31]. It should be noted that the heterochromatin/euchromatin classification was initially defined at a coarse scale and does not represent unique chromatin states at a finer scale, nevertheless this classification is still in use as a common way of qualitatively summarizing several epigenetic states based on the packaging density of the chromatin structures they form.

Beyond the scale of a few nucleosomes (∼1 kbp), chromatin has been observed to generate three-dimensional loops, where distant parts of the genome are found closer than they would be otherwise expected from the polymer entropy alone (Fig. 1.2E). The formation of these loops (that will be later reviewed in more detail) have been shown to regulate gene expression [32], as well as the spread of euchromatic/heterochromatic regions [33] and may form the basis of the largest-scale structures [34].

Apart from specific loops, chromatin has also been shown to be organized into structural domains which are thought to work as independent functional units of the genome. A particular type of domain that has recently captured much attention in the literature is the topologically associated domain (TAD) [35–37], which is observed at the 40 kbp–3 Mbp scale. The genomic sites within these domains tend to be found in close proximity while they avoid interacting with neighbouring sites from other domains (Fig. 1.2F). These domains are typically associated with a dominant epigenetic signature and strong chromatin loops tend to be situated at their borders. TADs are highly conserved across different cell types and are considered to be the basic unit of chromosome folding [38,39].

At the megabasepair scale, long-range interactions generate compartments, which are comprised of multiple TADs (Fig. 1.2F). These compartments, initially classified into two types A and B [40], and later subdivided into six types [32], tend to interact with other compartments of the same kind. In contrast with TADs, compartments are not conserved across cell types. The functional distinction between TADs and compartments is not yet completely understood, however it is possible that compartments simply arise from the self-attraction between groups of TADs with similar epigenetic states [41].

At the largest scale, chromosomes form chromosomal territories inside the nucleus [42]. They consist of regions of the nucleus that are occupied by a single chromosome, rather
than an intermix of many (Fig. 1.2G). The interior of these territories display a network of channels through which molecules can travel. Chromosomal territories are largely conserved across cell types. Although the underlying mechanism that determines this structure is not completely understood, specific interactions with the nuclear envelope (lamina) may play a role [43].

1.3.1 Higher-order chromatin structure

Given the difficulty of directly observing higher-order chromatin conformation in detail, understanding its organization often requires the use of polymer models. Two distinct approaches exist. The first is the consensus view, which tries to find the best 3D structure in agreement with experimental observations [44–46]. The second is the ensemble view, where chromatin structure is explained as a population of possible conformations [47–50]. The consensus view is the most traditional approach, perhaps initially inspired by the observation of regular bands of heterochromatin and euchromatin in the microscope. This perspective has the advantage of providing a simple description of chromatin conformation and it can be easily visualized as the 3D shape of a single structure, allowing biologists to easily integrate the chromatin model with other existing data. In contrast, the ensemble view is a more realistic description since the polymer dynamics of DNA are stochastic in nature and each cell within a population takes a unique structure [51]. However, ensemble methods are substantially more complex to develop because of the large landscape of possible polymer structures that need to be sampled. The strength of ensemble methods is that they can provide both an average structure akin to the consensus structure, as well as describe the predicted likelihood of any deviation from it.

We previously discussed how an ensemble of random walks (the random chain), with $R(N) \sim N^{1/2}$, fails as a complete description of the packing of DNA inside a nucleus. Nevertheless, this model describes the ideal behaviour that any polymer converges towards whenever interactions between monomers and topological constraints are negligible. Adding steric repulsion between monomers, one obtains the swollen-coil model [52] (also known as self-avoiding walk, SAW), with an even larger root-mean-square distance $R(N) \sim N^{0.6}$ than the random chain. In contrast, adding attractive interactions between monomers (or repulsion from solvent) generates an ensemble known as the equilibrium globule with $R(N) \sim N^{1/2}$ [53]. This type of polymer is space-filling as the volume it occupies increases as the length of the polymer $V \sim R^3 \sim (N^{1/2})^3 \sim N$, in contrast to the random chain $V \sim N^{3/2}$. Interestingly and also in contrast with the random chain model, subsegments of the equilibrium globule of size $L \lesssim N^{1/2}$ do not scale as the whole chain, but instead they scale as a random chain $R(L) \sim L^{1/2}$. This effect can be understood by the fact that at the smaller scales, volume exclusion and polymer compression compensate each other. Although its space-filling properties combined with the less-compact packaging at smaller scales are desirable features for DNA packaging, simulations have shown that these structures are highly knotted, therefore
requiring extremely slow reptation for its equilibration [54]. See polymer visualizations and scalings in Fig. 1.3A and B.

These three ensemble models described so far are equilibrium models and they can be related to each other: increasing monomer attraction in the swollen coil we obtain the random chain, and further increasing attraction we obtain the equilibrium globule [55]. In contrast, methods that include topological constraints typically generate long-lived non-equilibrium polymers. The modelling of the chromosome as a homogeneous 30-nm fiber with excluded volume confined into a 5-µm diameter nucleus has successfully reproduced the observed $R(s)$ for yeast, human and *Drosophila* [57,58] by matching simulation time to the chromosome decondensation timings (after cell division) of such organisms. Other models involving melted polymer rings [59,60] and randomly spaced long-distance harmonic interactions between pairs of monomers [61] have also been successful at explaining experimental scalings, although their biological interpretation is not clear yet. Lastly, a non-equilibrium model known as fractal globule, or crumpled globule, originally proposed in 1988 [62] has gathered much attention. This is a long-lived state found when quenching (or strongly confining) a swollen coil before it reaches the equilibrium globule state. This leads the polymer to collapse, which in turn has the effect of generating topological constraints at all scales, resulting in a hierarchical organization of the polymer into a series of DNA crumpled of increasing size [63]. This model is particularly attractive because, in contrast to the equilibrium globule, the polymer has uniform space-filling density $R(N) \sim N^\frac{1}{3}$ regardless of scale (after 10-20 Kuhn lengths). In addition, the model is largely unknotted and composed of spatially segregated domains [56], (Fig. 1.3C-G). Most importantly, this polymer model is the only one so far consistent with the experimentally observed scaling $R(N)$ in microscopy experiments (in the case of inactive chromatin), and also the probability of contact $P(N)$ observed in DNA conformation capture data (more details in the section “Measuring chromatin” of this chapter).

### 1.4 Function follows form

After reviewing the basic hierarchy of chromatin structure, we now review the architectural elements known to contribute to its establishment, maintenance and regulation. In particular we will focus on how sequence can lead to structure, and how structure can lead to biological function.

As we previously mentioned, some proteins can bind specifically to other molecules. Some of them can target and bind to precise regions of the DNA through their attraction to specific sequence-motifs [64]. Interestingly, other proteins (known as cofactors) interact and bind with the DNA-bound proteins generating a network of interactions between different parts of the DNA. As a consequence, chromatin structure can be partially determined by
Figure 1.3: (A) View of the random walk (RW), self-avoiding walk (SAW), equilibrium globule (EG). Credit: Fudenberg and Mirny (2012) [55]. (B) Radius of gyration of RW, SAW and EG. Credit: Fudenberg and Mirny (2012) [55]. (C) View of fractal globule. Loop was opened (smaller globule) by removing attractive interaction of the monomers of the region. Credit: Mirny 2011, [56]. (D) View of equilibrium globule. Loop failed to open due to entanglement after having removed attractive interaction of monomers of the region. Credit: Mirny 2011, [56]. (E) End-to-end distance of equilibrium globule and fractal globule. Credit: Mirny 2011, [56]. (F,G) Left to right: Sub-chains of 100, 300, 1000, and 3000 monomers for the fractal globule and the equilibrium globule. The fractal globule forms globules at all scales, whereas subregions of the equilibrium globule diffuse inside the globule. Credit: Mirny 2011, [56]. — All images are reproduced with permission from the owners.
the DNA sequence itself. All of these chromatin-associated proteins have concentrations that may vary over time and position, allowing for complex dynamic chromatin regulation.

1.4.1 Promoter-enhancer loops

In the cell, there is a large biomolecular machine known as RNA polymerase that binds to DNA upstream of genes and transcribes the DNA to produce messages of RNA that may get translated into proteins. The binding of the RNA polymerase to the upstream region of a gene (known as a promoter) is a highly regulated process. In order to efficiently transcribe genes, RNA polymerase is able to recognize where to start and where to finish the transcription process. RNA polymerase repeatedly attaches to the DNA whenever it collides with it. However, only when it passes through regions called promoters it can adhere more tightly, open the double strand and continue sliding along the gene.

In eukaryotic cells, RNA polymerase is rarely able to bind and start transcription by itself because the DNA tends to be packaged into dense chromatin structures. Consequently, promoter regions need the presence of several protein complexes called transcription factors and chromatin remodelers to help RNA polymerase find the promoter, attach to the DNA and remain stably bound there while it separates the strand and slides along the genome [66].

The availability of transcription factors can be influenced by cascades of events in response to both internal and external signals sensed by the cell, thus allowing the cell to adapt to its environment [67].

Promoter sequences are 100–1000 bp long and typically contain a binding site for RNA polymerase as well as binding sites for transcription factors. In addition, promoters are rich in CG sequences, which can be methylated by enzymes resulting in the stable silencing of genes [68]. Nevertheless, promoters are sometimes not able to recruit all the necessary transcription factors by themselves and need the action of enhancers (distal promoters) that are located elsewhere in the genome up to a megabase away, which recruit additional transcription factors known as activators (and repressors) that regulate the binding of the RNA polymerase. Enhancers and promoters create DNA loops [69], which in turn stabilize the RNA polymerase attached to the DNA with the help of a protein complex known as Mediator [70], (Fig. 1.4A). Depending on the availability of specific transcription factors, a given enhancer can interact with a different number of promoters, and multiple enhancers can act in a coordinated fashion with a single promoter [71].

It is important to point out that enhancer-promoter looping does not imply transcriptional activity. In fact, some of these loops have been detected regardless of the gene being transcribed [69,72,73]. It has been suggested that in these loops, RNA polymerase is stabilized in position waiting to be released from a paused state [74].
Figure 1.4: (A) Schematic of the recruitment of transcription factors by the promoter region and activators by the enhancer region. The DNA folds into a loop enabling RNA transcription. (B) The promoter-enhancer blocking effect of insulators can be tested by inserting insulator sequences at various positions with respect to an enhancer. The mean level of expression of the cases where the insulator sequence is inserted in between the enhancer and the gene is compared to the level of expression when the insulator is inserted close to the enhancer but not in between the enhancer and the gene. Credit: J. E. Phillips-Cremins and V. G. Corces (2013) [65]. (C) The barrier activity of insulators can be tested by inserting a transgene flanked by insulators at random locations. The control is done by randomly inserting the transgene without the insulator sequences. The experimental result suggests that insulators stop the spreading of heterochromatin with time, maintaining the transgene’s expression level. Credit: J. E. Phillips-Cremins and V. G. Corces (2013) [65]. — All images are reproduced with permission from the owners.
1.4.2 Insulator loops

Chromatin loops are not restricted to an enhancer-promoter context, in fact they are ubiquitous in the genome and have been identified as key elements for the formation of chromatin structure beyond the kilobase scale. A particular group of chromatin factors called insulators are known to be the primary drivers of a large number of structures associated to specific functional effects.

Insulators are DNA-protein complexes initially defined by their ability to block enhancer-promoter interactions whenever they are placed between them (Fig. 1.4B). Furthermore, they are also defined by their effect as barriers against the spreading of heterochromatin [33] (Fig. 1.4C).

In mammals, a protein called CTCF is considered to be the primary insulator, whereas in *Drosophila* there are five of them: suppressor of hairy wing (Su[Hw]), *Drosophila* CTCF (dCTCF), Boundary-element associated factor of 32 kDa (BEAF-32), GAGA binding factor and Zeste-white 5 (Zw5) [33]. Each *Drosophila* insulator is defined by a protein complex common to all of them and an insulator-specific DNA-binding protein that recognizes a specific DNA sequence motif. It may seem paradoxical that a simpler organism has a larger variety of insulators than the more complex mammals. However, evidence indicates that multiple insulators may be necessary for regulating compact genomes with less spacing between genes [65,75]. Nevertheless, it is possible that mammalians achieve a similar level of regulatory complexity through recruiting different binding partners and post-translational modifications [76].

The mammalian CTCF insulator can form multimers in vivo [77], thereby allowing CTCF to bind other CTCF molecules, and thus create long-range loops [78]. A large body of evidence indicates that the knockdown of CTCF significantly reduces the intensity of contact of several long-range loops (see review in [65]). Between 50% and 80% of the genomic sites occupied by CTCF overlap with the presence of Cohesin [79–82], a protein known to hold sister chromatids together during cell division. This has led to a prevalent model where CTCF recruits Cohesin which in turn forms a ring-like structure that stabilizes long-range loops [83]. Experiments that deplete the concentration of Cohesin have shown a disruption of long-range loops at several locations [84–87]. Nevertheless, it should be emphasized that some CTCF loops are Cohesin dependent whereas some are not.

In *Drosophila*, the spatial distribution of insulators varies with genomic context [88–91] which may be an indicator that different insulators have distinct regulatory roles. For example, BEAF-32 is typically found in promoter regions whereas Su(Hw) tends to reside in intergenic regions and is associated with the transposable element known as gypsy. Interestingly, some locations known as “aligned insulator elements” contain various insulator elements bound in close proximity (100-300 bp) [92]. In contrast with the mammalian insulator system, *Drosophila* insulators have the capability of forming heterogeneous dimers.
For example, BEAF-32 has been observed to bind to Zw5 both in vivo and in vitro [93]. In addition, *Drosophila* insulators have also been observed to form protein complexes by heterodimerizing with cofactors such as Mod[mdg4] and the centrosomal protein CP190 [94,95]. These complexes, that may involve various insulator types, are known as “insulator bodies” and can interact with each other. In contrast with mammalians, Cohesin does not play an important function in the establishment or maintenance of loops in *Drosophila* [96], instead it appears that CP190 and Mod[mdg4] carry the role of mediating inter-insulator interactions.

The focus on insulators has recently been increased by the discovery that they tend to demarcate the boundaries of TADs. In mammals more than 75% of TAD boundaries contain CTCF [37], whereas in *Drosophila* the boundaries are enriched for BEAF-32, dCTCF, CP190 but not for Su[Hw] [35]. In addition, *Drosophila* TADs are enriched for aligned insulator elements composed of two or more insulator types plus CP190 [97]. Recently, a polymer effect known as loop extrusion [98,99] has been proposed to contribute to the formation of (so far mammalian) TADs [100]. In this model, Cohesin acts as a loop-extruding motor that forces the DNA to go through its ring-like structure. Loop extrusion stalls once that Cohesin encounters a boundary element, such as a CTCF sequence in a particular relative orientation.

A closer inspection of the traditionally defined roles of insulators as barrier elements for both promoter-enhancer interactions and heterochromatin spreading has recently casted several doubts about these properties. An example of the former is that more than 75% of human promoter-enhancer interactions pass over at least one CTCF-occupied site [101]. An example of the latter is a *Drosophila* experiment where the knock-down of insulator proteins showed only a slight spreading of the heterochromatic histone modification H3K27me3 into the surrounding regions in 75% of the cases and no effect at all in 25% of the cases [102]. Even their action as TAD barriers is not entirely predictable. For instance, in the mouse genome only 15% of the CTCF-occupied sites are at TAD boundaries while the rest are within TADs and may be involved in intra-TAD interactions [37]. Despite these findings, it remains clear that insulators are the primary regulators of intra- and inter-chromosomal chromatin loops [65]. Their regulatory functions may be a result of the complex interplay between the structures they create, the local concentration of site-specific binding partners, post-translational modifications and other epigenetic factors.

### 1.4.3 Other mechanisms

Chromatin conformation is affected by a variety of additional mechanisms that we will not review in detail. Some examples are the Polycomb-group proteins which create long-range clusters of repressed chromatin [103]; the complex nature of the protein Mediator which not only affects enhancer-promoter loops but also interacts with Cohesin [104]; non-coding RNAs that interact with insulators, cofactors and transcription factors [105–107]
and transcription factories consisting in protein-rich chromatin clusters that can contain up to 30 RNA polymerase molecules in a single spot [108].

1.5 Measuring chromatin

Recent advancements have allowed for the precise measurement of hundreds of different chromatin factors, as well as recurring chromatin structures in living cells. In this section we present the principal experimental techniques that constitute most of the higher-order structural evidence from which newer ways of classifying chromatin organization have recently emerged.

1.5.1 Chromatin Immuno-Precipitation (ChIP)

Chromatin Immuno-Precipitation (ChIP) is a method for detecting in a cell population the average occupation of DNA-bound factors along the genome through the use of antibodies that bind to them specifically [109]. This technique is usually followed by the identification of the genomic positions where the factors are bound through the use of microarrays (ChIP-chip) [110] or more recently the direct sequencing of the selected DNA fragments (ChIP-seq) [111].

This method, reviewed in [112] and summarized in Fig. 1.5, consists in first crosslinking DNA-binding proteins to the DNA using formaldehyde. Then the cells are lysed and the DNA is sonicated into small fragments 200-600 bp long. Next, antibodies that specifically target the protein of interest recognize and bind to the protein-DNA complex. These antibody-protein-DNA complexes are then isolated and the formaldehyde crosslinking is reversed, allowing for the assay of the DNA fragments by either microarrays or direct sequencing. Once the detected fragments are aligned to their genomic positions, statistical models are used to estimate the significance of protein enrichment with respect to experimental controls.

Multiple variations of this method exist. For instance, ChIP experiments for the detection of nucleosome positioning or specific histone modifications typically use the micrococcal nuclease enzyme (MNase) instead of sonication, since this enzyme digests the DNA fragments that are not wrapped in nucleosomes, thus allowing for a more precise sequence identification [113]. Another related technique known as DNA adenine methyltransferase identification (DamID) [114] is based on using DNA adenine methyltransferase to methylate adenines near the genomic sites occupied by the protein of interest. Fragments containing methylated adenines can then be detected by DNA microarrays. In addition, recent advances have enabled the measurement of chromatin markers at a single-cell level by tagging the DNA of individual cells through DNA barcoding in a microfluidics set-up [115]. Although this method generates sparser data, it is sufficient for characterizing cell-to-cell variability within a population.
Figure 1.5: Schematic of ChIP-seq protocol. (A) A strand of DNA with several bound proteins. (B) DNA and proteins are fixed in position by adding formaldehyde. (C) DNA is fragmented. (D) Bead-attached antibodies bind to the protein of interest, and the DNA-protein-antibody complex is precipitated. (E) DNA fragments are purified, sequenced and mapped to genomic locations. (F) View of the resulting ChIP-seq data. *Credit: Peter J. Park (2009) [112]. — All images are reproduced with permission from the owners.*
Comprehensive mapping of hundreds of DNA-binding proteins, histone modifications and other regulatory elements detected with this technique under various conditions can be found in The Human Encyclopedia of DNA Elements (ENCODE) [116] and in modENCODE [117] for other model organisms. The distribution of these factors forms the sequence that encodes information about how the DNA will fold inside a cell.

1.5.2 Fluorescence in situ hybridization (FISH)

Fluorescence in situ hybridization (FISH) is a method developed in the early 1980s to visualize the spatial distribution of genomic sites with fluorescent microscopy via fluorescently tagged probes that specifically bind to designated sequences of the genome [118] (FIG 1.6A). This method has been successful at unveiling the spatial organization of chromosomes inside the cell nucleus [51,119,120] and characterizing long-range regulatory contacts [69,121,122].

By combining FISH probes with super-resolution microscopy [124], a recent study [123] measured the spatial scaling of Drosophila genomic domains of sizes ranging from 10 kbp to 500 kbp associated to three distinct epigenetic states: active (enriched for histone modifications H3K4me2 or H3K79me3), repressed (enriched for histone modification H3K27me3 or Polycomb-group proteins) and inactive (depleted of histone modifications as well as Polycomb-group proteins and transcriptional activators). The results showed that active domains occupied larger volumes than inactive domains which were in turn always larger than repressed domains (FIG 1.6B). In addition, the study confirmed power-law scaling of volume with genomic distance $V(L) \sim L^a$ with different exponent values for each epigenetic state. Inactive domains had constant-density packing ($a \approx 1, \frac{N}{V} \sim constant$), active domains had tighter packing for smaller domains ($a \approx 1.3, \frac{N}{V} \sim N^{-0.3}$) and repressed domains showed denser packing for larger domains ($a \approx 0.8, \frac{N}{V} \sim N^{0.2}$) (FIG 1.6C). Interestingly, by looking at subsections of the larger-size domains, the study found that both active and inactive domains had a self-similar scaling, but repressed domains did not. Subsections of the repressed domains grew rapidly in size and then saturated for lengths only one-fifth of the parent domain length, thus implying that smaller repressed domains tend to occupy as much physical volume as larger repressed domains.

1.5.3 Chromosome conformation capture

Chromosome conformation capture (3C)-based methods are a family of experimental techniques (3C, 4C, 5C, HiC and ChIA-PET) for detecting pairs of DNA loci that are in close proximity (10-100 nm) inside the nucleus [125]. These methods are typically performed in a cell population, thus reporting population-average contact frequencies between distant sites of the DNA, currently at a resolution of 1–10 kbp.

All 3C-based methods are based on fixing the DNA in its three-dimensional conformation inside the nucleus by applying formaldehyde and then fragmenting the DNA by either restriction enzymes or sonication. Next, DNA fragments in proximity are ligated to form a
Figure 1.6: Schematic of FISH. (A) A fluorescently labeled DNA probe hybridizes specifically to a target region. The target region can then be visualized through fluorescent microscopy. Credit: http://www.abnova.com/; December 2017. (B) Super-resolution microscopy of three ∼ 100 kbp fragments identified as active, inactive and repressed by their chromatin factors. Credit: Boettiger et al. (2016) [123]. (C) Volume and radius of gyration of different-sized fragments coloured by their associated epigenetic state. Ph-KD denotes the knockdown of the Polyhomeotic Polycomb-group protein associated to the repressed epigenetic state. Credit: Boettiger et al. (2016) [123]. — All images are reproduced with permission from the owners.
circular hybrid DNA that contains two genomically distant but spatially close sequences, which can be purified and detected, generating a list of contacting DNA pairs. The main differences between 3C-based technologies, reviewed in detail in [126–128], are the type of sites each technology targets. In short, 3C measures contacts between two defined regions; 4C measures genome-wide contacts with a site of interest; 5C measures all-against-all contacts within a defined region; Hi-C measures all-against-all contacts genome-wide; ChIA-PET measures all-against-all contacts in locations where a given protein of interest is present.

In this thesis, we will obtain our chromatin conformation data from Hi-C, a high-throughput 3-C technique capable of simultaneously reporting DNA contact intensities genome-wide. As shown in Fig. 1.7A-G and in [40,129], the method consists of the following steps: Crosslinking the DNA with formaldehyde, digesting the DNA with a restriction enzyme, marking the end of the fragments with biotin, ligating the crosslinked fragments, pulling down the biotin-marked strands and lastly, sequencing and mapping fragments to the genome. After correcting for several experimental biases detailed in [130], one obtains a map of DNA contacts typically presented as a square matrix where rows and columns represent non-overlapping genomic bins of a fixed genomic size (Fig. 1.7H). The entries of the matrix are the number of detected unique fragments pairs belonging to the two genomic locations.

This matrix, often referred as the “raw contact map” still contains several biases that must be corrected for. Namely, the CG content of the bin, the mappability of the genomic region, and the length of restriction fragments within bins [131]. An option for normalizing the data is to explicitly model those biases and correct for them. An alternative normalization option is to simultaneously correct for all biasing factors by enforcing the assumption that, if no biases where present, each genomic site should have the same number of total counts. This implies finding the weights that must be applied to each row and column in order to obtain a normalized contact map in which all rows and columns have equal sums [32,132].

The resulting normalized Hi-C matrix reflects the probability that two sites of the genome are found in contact with each other in a cell population, and is thus a proxy for the average chromosome conformation in the experiment. This technique has revealed several properties of chromosomes, such as the scaling of the probability of contact as a function of genomic distance \( N \) ( \( P(N) \sim N^{-1.08} \) between 500 kbp and 7 Mbp [40], consistent with the fractal globule [56]) (Fig. 1.7I), the classification of the genome into TADs [35–37] and compartments [32,40], the identification of functional looping interactions [34,45,101] and the inference of 3D polymers consistent with Hi-C contacts [36,133–135].
Figure 1.7: Chromosome conformation capture (A) A DNA loop. (B) By adding formaldehyde the DNA is fixed in position. (C) The DNA is cleaved by the action of restriction enzymes. (D) Biotin-marked nucleotides are added to fill the end of the enzyme-digested fragments. (E) Genomically distant looping regions are ligated with DNA ligase. (F) The DNA is fragmented and purified. Bead-attached antibodies target the biotin markers and are immunoprecipitated. (G) DNA fragments are isolated and sequenced, generating a library of contacting regions. (H) Hi-C data can be presented as a contact map. Pictured: region of the mouse chromosome 18, ICE normalized. Credit: J. Dekker et al (2013) [34] (I) Hi-C data uncovers the distance-dependent probability of contact of genomic regions. The data in the 500 kbp–7 Mbp region is consistent with the fractal globule. Credit: J. Dekker et al (2013) [34] — All images are reproduced with permission from the owners.
1.6 Modelling chromatin data

The vast amount of data generated by the recently developed high-throughput techniques for measuring chromatin has allowed for simultaneously analyzing several chromatin factors and/or structures in the search for functionally relevant patterns. Consequently, a large body of work has focused on reducing the complexity of the data by generating one-dimensional genomic annotations that classify each locus into one of a few possible chromatin states, based on experimental observations. These methods are often referred as SAGA (semi-automatic genome annotation) since a human typically performs the functional interpretation of the labels after the annotation process [136].

Some chromatin classification problems are purely based on which chromatin factors are present at each location and aim to reduce the up-to-hundreds of bound elements into just a few bound states. This task can simply be done with an unsupervised clustering of factors [137], however a more popular approach is to apply (when necessary) dimensionality reduction techniques such as Principal Component Analysis (PCA) that take advantage of the highly correlated nature of chromatin factors, followed by the classification into a sequence of bound states by a Bayesian network such as a Hidden Markov Model (HMM) [138–143].

Other chromatin classification methods are purely based on categorizing structural data such as Hi-C. An eigenvalue decomposition of the correlations between rows (or columns) of the Hi-C matrices results in the A/B compartment segmentation of chromatin into two types of sites which tend to interact with other sites of the same compartment type while avoiding contacts with the opposite type [40]. Nevertheless, a further inspection of this segmentation scheme identified smooth transitions between compartments indicating the possibility that more than two compartment types exist [132]. Other studies have applied variations of the same type of Hi-C based classification. For instance, a K-means clustering of inter-chromosomal contacts has found a third compartment class [131]; a comprehensive clustering analysis (multi-variate HMM, hierarchical clustering and K-means) of inter-chromosomal contacts found five compartments that could be related to the initially discovered A/B compartments along with an additional compartment specific to chromosome 19 (the six compartments are known as A1, A2, B1, B2, B3 and 19*) [32]. Alternatively, at a more local genomic scale, one can identify borders between interacting domains by comparing the number of contacts that each genomic site has with their neighbours up- and downstream in the genome, thus building a contact directionality index. A strong transition in this directionality index indicates a TAD border, which can be identified by applying an HMM to the sequence of directionality index [37]. More recently, an increasing number of methods for the identification of TADs using more sophisticated techniques has become available [32,144–146].
So far we have described a number of one-dimensional genomic classifications based on either purely binding factors or purely structural data. The vast majority of such analyses link the two types of data by either looking at enrichment of contacts between bound states, or enrichment of bound states in compartments, domains or loops. However, a recent study [136] has been able to integrate chromatin structure into the genomic classification of bound chromatin factors by classifying pairs of sites with high contact frequency into the same bound state, based on the observation that large chromatin domains tend to co-localize with other domains of similar type [140,147].

Despite the success of these methods, looping interactions often skip large genomic regions, implying that the linear arrangement of chromatin states is an incomplete description of the functional and structural processes in the nucleus [34]. Our principal focus in this thesis is to bring together chromatin sequence and structure with the intention of modelling how one affects the other. In particular, we are interested in explicitly modelling both elements at the same time. Most progress on this problem has been achieved by constructing heterogeneous beads-on-a-string polymers whose types of beads correspond to different chromatin states and then, distances (or contacts) between beads are compared to the existing Hi-C data [50,61,148–160]. In contrast, our goal is to provide a direct link between the probability of contact observed in Hi-C data and the distribution of chromatin factors along the genome. In this thesis we will explore this connection from a variety of perspectives such as direct physical modelling, information-theory-based analyses, maximum-entropy modelling and state-of-the-art dense neural networks.

The work described in the following chapters presents a novel way of modelling chromatin conformation data that is based on probabilistic dependencies on chromatin sequence, rather than the sampling of polymer conformations from simulations. The models that we developed allow us to generate contact maps directly from information about the chromatin sequence with little computational cost and often solve the inverse problem, namely inferring chromatin sequence from chromatin contacts. Our work is thus able to highlight connections between particular chromatin sequence patterns and their corresponding conformational effect. Importantly, we can identify key determinants of chromatin conformation, as well as provide the field with new tools to study the conformational effects of mutating the sequence of chromatin states in specific genomic locations. Our models thus contribute to the understanding of how distant genomic regions can communicate with each other and, ultimately, how the underlying chromatin sequence orchestrates the folding of DNA inside the nucleus.

1.7 Thesis outline

In Chapter 2, we present our first approach to the problem of chromatin folding, published in 2015 [161]. In it, we apply Principal Component Analysis (PCA) to Hi-C data. PCA had
been used before to classify the genome into A/B compartments [40] and we extend its use as a data-normalization tool to filter out systematic biases and high-frequency noise in Hi-C contact maps. In order to validate the biological relevance of the PCA-based normalization procedure, we construct a model that connects the underlying chromatin sequence to the probabilities of contact as measured in Hi-C contact maps. Using just the bound sequence information at contacting sites, we were able to extract the interaction energies of the network of chromatin factors in Drosophila melanogaster. This pairwise interaction model served a basis to create a number of improved models that take into account non-local effects, which we present in the subsequent chapters.

In Chapter 3, we further inspect the connections between the chromatin sequence and chromatin contacts in a model-free manner. In particular we highlight a number of effects that the model from Chapter 2 model could not capture, namely non-local effects due to interactions with the neighbourhood of bound factors. We then use concepts from information theory to find which regions of the chromatin sequence influence chromatin contacts the most. Last, we show some of the specific biological contexts in which chromatin sequence modulates chromatin contacts. These results serve to inform further development of models for chromatin conformation by indicating the most influential features of the chromatin sequence.

In Chapter 4, we formulate the problem of chromatin folding in a Bayesian fashion, thus breaking down the probability of contact between chromatin sites into probability distributions that we can model from the data. In particular, we use the principle of maximum entropy to build a model for chromatin conformation that takes into account non-local information about the chromatin sequence. This model generates significantly improved predictions in comparison to the model of Chapter 2. Interestingly, the new model allows us to quantify the conformational consequences of mutating chromatin sequence, highlighting different structural effects depending on the genomic function of the mutated region. In addition, using the Bayesian approach we can invert the probability distributions to solve the inverse problem, namely predicting chromatin sequence from chromatin contacts.

In Chapter 5, we present the use of dense neural networks to solve the chromatin folding problem. The method is able to predict chromatin conformation from sequence (and sequence from chromatin conformation) with remarkable precision (better than the maximum entropy model of Chapter 4), as it can assimilate a great amount of information about sequence and structure. In particular, we design our neural networks with interpretability in mind, generating a biologically relevant chromatin state classification of sequence that is optimized to explain chromatin structure. In addition, we inspect how our neural networks map the two types of data to one another, highlighting the sequential and structural patterns of chromatin that are most relevant to the problem of chromatin folding.

In Chapter 6, we conclude with a brief summary of findings, as well as possible future directions for applying the ideas developed in this thesis.
Chapter 2

A pairwise model for chromatin contact free energies 1

At the smallest length scale, DNA is wrapped around histones to form nucleosomes that aid the packaging of the DNA. Histones can be chemically modified and depending on the type of modification can mark the chromatin as either being in a silent heterochromatic state or an active euchromatic state. These histone modifications are passed down from one cell to the next, thus contributing to the cell’s epigenetic regulatory machine [162,163]. On longer length scales, chromosomes fold into topological domains in the space of the nucleus [32,35–37,40]. Such organization may contribute to separate heterochromatin and euchromatin via the formation of chromatin loops (ranging from 1 to 500 kbp in length) within each type of domain [163]. In addition, the likelihood of long-range contacts between distant loci involves specific DNA-binding factors participating in chromatin organization [91,164]. Some of these proteins condense the DNA making heterochromatin regions [165] while others are associated with euchromatin [91,141,163,166–171]. It therefore remains unclear to what extent their impact on looping is a result of being in a particular epigenetic background, as euchromatin and heterochromatin may influence long range interactions as well.

The recent development of the Hi-C method has provided a valuable tool to study the 3D organization of chromosomes on a genome-wide level. This method measures the frequency of contact between any two segments along the genome. Using this data, a variety of methods have aimed to predict the underlying 3D structure of the chromosomes [44,45,47,48,172]. On large scales, DNA confinement plays a role in structuring the chromatin [56], and modeling has shown that such effects can lead to inheritable territories [57]. Nevertheless, many studies have also shown that the overall organization found in the Hi-C data correlates with

1The work presented in this chapter was done by Saeed Saberi, Pau Farré, Olivier Cuvier and Eldon Emberly and published in BMC Bioinformatics in 2015 [161]. PF performed the model fit, generated the figures and contributed to drafting the manuscript. The text of that paper has been edited to conform to the style of this thesis.
the underlying domain structure of the chromatin and the corresponding bound proteins in those domains [35, 40, 47] as identified by chromatin immunoprecipitation (ChIP-chip or ChIP-seq). It has thus been possible to intersect these two data sets to infer the influence of chromatin-associated proteins in long-range interactions. As such, it has been shown that in mammals, the insulator protein CTCF facilitates looping between distant sites provided the additional presence of Cohesin and/or Mediator complexes [173]. CTCF has been found to be enriched at boundaries between heterochromatic and euchromatic domains and sometimes aids the regulation of enhancer-promoter interactions [91, 173]. In *Drosophila*, a number of additional insulator proteins that bind insulator sequences have been identified: BEAF32, dCTCF, GAF, Zw5 and Su(Hw). They have been found to interact with each other thereby stabilizing long-range interactions among distant insulator sites [174]. Such looping involves further insulator protein cofactors such as CP190, Chromator or Mod[mdg4] [35, 173, 174]. These insulator proteins form a network of interactions that may contribute to structure and isolate active domains from inactive chromatin within the *Drosophila* genome [173].

The assembly of chromatin into silent domains has a similar network of interactions confined within such domains [35] or that involve long-range interactions between distant silent domains [175]. Important contributors to these interactions within heterochromatin are the Polycomb-group (PcG) proteins that play key roles in the spreading of the silent state upon binding of PcG and co-factors to specific DNA sequences called Polycomb Response Elements (PREs). Hence, analogous to how insulators aid the structuring of euchromatin domains, PcG proteins have their own associated set of interactions that aid the formation of heterochromatin.

Here, we aim to quantify the effective energetics of interaction between different chromatin regulators from the measured Hi-C contact frequencies. For this purpose, it is crucial to disentangle the effect of different contributing factors within the Hi-C data itself. The largest contributing factor to the observed frequency of contacts in Hi-C data is the distance-dependent likelihood of contact between loci due solely to the polymer nature of the DNA. This distance-dependent likelihood acts as a background that helps to hide the specific contacts that exist between chromatin regulatory factors. The distance-dependent scaling of this background contact frequency has been shown to be consistent with a confined polymer model [56, 57, 62, 176, 177]. Other contributing factors to Hi-C data are systematic biases introduced as a result of the nature of the experimental protocol. For example, it is known that the Hi-C procedure generates biases due to the sequence and length of contacting DNA segments [35]. Thus, depending on various DNA features, some loci may be observed more frequently than expected. Several normalization methods have been proposed to correct for these biases [35, 132, 178] (for a review see [179]). However, which normalization method provides the most significant information between the specific contact frequencies and the underlying bound factors has not been surveyed in detail before.
In this chapter, we introduce a method for transforming the measured Hi-C contact frequencies into free energies. The method is based on an equilibrium statistical mechanics approach, where we assume that the frequency of contact between two genomic locations is related to the free energy of forming that particular contact state. Due to the additive nature of free energy, Principal Component Analysis (PCA) provides a convenient tool for then decomposing these free energies into a set of independent modes of interaction. PCA identifies a length-dependent background looping energy, systematic biases, and then a series of modes of increasing spatial frequency with which to express the data. We can reconstruct the transformed Hi-C data using the PCs, leaving out those that are due to systematic biases as well as those which are high-frequency noise. Our approach is a data-driven method for normalizing Hi-C data.

We assess our normalization scheme, as well as two other methods by fitting the free-energy data to an interaction model involving the locations of known DNA-bound chromatin factors. The energy of interaction between two loci is modelled as a linear superposition of pairwise interactions between all the bound chromatin factors at those two locations. Given that the energy of interaction and the bound occupancies for the various factors are measured, the model can be fit to predict the couplings between factors. Our fitted couplings show a complex interplay of interactions between the chromatin factors, capturing many known biological relationships. We use the quality of fit of the model as a criterion to determine how many PCs should be filtered. Interestingly we find that other normalization schemes that correct for various biases are less well-fit by this pairwise model than our PCA-based normalization scheme.

2.1 Methods

2.1.1 Chromatin contact maps

We have used the Hi-C contact maps for *Drosophila melanogaster* reported in Sexton et al. [35]. This data consists of a list of genomic locations for pairs of sequences that were found to be in contact, and the number of times each sequence pair was sequenced. Using this data we have constructed a contact matrix \( n_{ij} \) at 10 kbp resolution for the *Drosophila Melanogaster* genome. Following the approach of Sexton et al., for each sequence pair we only count the contribution to a particular \( n_{ij} \) element once, rather than the number of times it was sequenced. This is argued to remove some of the sequence-dependent bias in the Hi-C protocol.

The contact matrix created from the raw Hi-C data is not corrected for any potential systematic biases (aside from counting each sequence pair only once). In addition, we also used two separate normalization procedures that correct for biases in the data. The first method, ICE, normalizes the contact matrix so that each bin has the same number of interactions as any other (see [132] for details). The second was introduced in Sexton et
al. [35] and uses a probabilistic model to correct for various systematic biases. This method
does not normalize all the bins to have the same number of interactions genome-wide.

2.1.2 DNA-bound factors

We have downloaded publicly available genome-wide binding profiles and enriched binding
regions from modencode.org for the following chromatin factors (insulators: BEAF, CP190,
dCTCF, GAF, ZW5; epigenetic marks: H3K27Me3, H3K4me3; dosage compensation complex:
MOF; Polycomb-group proteins: Pc, Pho, PCL2; Cohesin: SMC3; other factors: Nurf,
Chromator, PolII) [117].

We used the enriched regions of each chromatin factor $\mu$ to generate a sequence of bound
factors $S^\mu_i$, where $i$ denotes genomic position (at a 10 kbp resolution). A given enriched
region has a beginning and end genomic coordinate as well as a log-odds score which can
be thought of as a binding energy. For a given bin $i$ in the genome, the total binding energy
$E^\mu_i$ for factor $\mu$ is found by adding up the log-odds scores for all of its enriched regions that
overlap with the bin. Statistical physics gives a prescription for converting these binding
energies into occupancies via $S^\mu_i = 1/(1 + \exp[-(E^\mu_i - \bar{\epsilon}^\mu)/\sigma^\mu])$ where we take $\bar{\epsilon}^\mu$ to be the
average binding energy over the bound bins, and $\sigma^\mu$ the standard deviation which is related
to an effective temperature.

2.1.3 Principal Component Analysis

Principal Component Analysis (PCA) is a method that allows one to find the dominant
sources of correlation in high-dimensional datasets. This technique decomposes a set of
measurements of possibly correlated variables into one expressed in a basis of independent
variation. The orthonormal directions of the new basis are called principal components
(PCS) and they represent the principal patterns of variation of the experimental measure-
ments.

More specifically, suppose we have a set of multidimensional measurements $\vec{x}_i$, each of
length $m$, stored as the rows of a matrix $X$. By calculating the covariance matrix $C$
of the data columns, we find how the $m$ dimensions of our measurements covary. Then, by
diagonalizing $C$, we obtain a set of $m$ eigenvectors $V$ of length $m$, the PCs, that correspond
to directions of independent variation in the data. The $m$ eigenvalues obtained indicate how
much of the variance in the data is captured by each PC.

One can project the data $X$ onto the new basis by first subtracting the mean of the
columns $\bar{X} = X - \bar{x}$, and then calculating the dot product of the mean-subtracted data and
the new basis, $P = \bar{X}V$, thus obtaining a new data matrix with the same dimensionality as
the initial $X$, but with uncorrelated columns. Moreover, the eigenvalues equal the variance
of each column.

Often, one encounters that the PCs of high-dimensional data can be interpreted as vari-
ables of interest indirectly measured in the data as correlations between the $m$ original
variables. In addition, the lowly ranked PCs (i.e. those with small eigenvalues) often correspond to just noise in the signal. Consequently, one can leave some of these low-variance PCs out, effectively low-pass filtering the data. This can be done by reconstructing our set of measurements by adding only the $j$ PCs of interest,

$$\vec{x}_i' = \bar{x} + \sum_j p_{ij} \vec{v}_j,$$

where $p_{ij}$ are the values of the projection matrix $P$, and $\vec{v}_j$ are the PCs of interest.

### 2.1.4 Free energy matrix

Assuming the Hi-C measurements represent an equilibrium distribution, we can associate the contact frequency between bins $i$ and $j$ with a free energy, $F_{ij}$, via $n_{ij} \propto \exp(-F_{ij}/k_B T)$. Thus we can transform the above contact matrix into a matrix of free energies defined by,

$$\frac{F_{ij}}{k_B T} = -\log(n_{ij}) + F_0.$$  

(2.2)

We set $k_B T = 1$ for the sake of simplicity in the rest of this study, and set $F_0 = 0$, as it just defines a reference energy. This free energy contains both an energetic contribution, arising from specific interactions between DNA-bound factors, and an entropic contribution that is due to the assortment of conformations that the polymer of DNA can adopt. (We have added a pseudo-count of 1 to all $n_{ij}$ to fill in locations $i, j$ where the contact matrix was zero. Other methods to fill in missing values, such as interpolating between $F_{ij}$, yielded similar results).

### 2.1.5 Free energy decomposition: PCA-based normalization

The free energy, $F_{ij}$ between bin $i$ and $j$, can be decomposed into two terms,

$$F_{ij} = \bar{F}_d + \delta F_{ij},$$

(2.3)

where $\bar{F}_d$ is the average free energy at a fixed genomic distance, $d = j - i$, and is found by averaging over all such distances genome-wide, and $\delta F_{ij}$ is the free energy difference from this average that depends on the two interacting bins. The genome-wide average free energy, is computed via $\bar{F}_d = (1/N) \sum_i F_{i,i+d}$, where $N$ is the number of $F_{i,i+d}$ at a given separation $d$. We impose a fixed range on the genomic separation, namely $d = -d_c \ldots d_c$ with a separation cut-off $d_c$.

The average free energy, $\bar{F}_d$ represents the dominant distance-dependent free energy and results from entropic cost for making a loop in the DNA with genomic distance, $d$. (Additional distance dependences due to chromatin structure may still remain in $\delta F_{ij}$). Polymer physics suggests that $\bar{F}_d \sim a \log |d|$ [52], which grows logarithmically with distance.
This is akin to the probability of contact as a function of separation for a random polymer going as $p_d \sim |d|^{-\alpha}$, with $p_d \propto e^{-\bar{F}}$.

The free energy fluctuations away from the average, $\delta F_{ij}$, will contain additive contributions from specific interactions due to chromatin factors, biases due to the protocol and potentially additional distance-dependent energies arising from differences in the polymer nature of chromatin at different loci. PCA provides a method for decomposing data fluctuations into a linear combination of independent modes. In order to apply PCA, we need a set of observations. Here, the observations correspond to the set of fixed-length free energy profiles, one for each bin in the genome. For each bin, $i$, the corresponding free energy profile $F_i$ is the list of interaction energies $F_{i,i+d}$ where $d = -d_c \ldots d_c$ and has a fixed length of $2d_c + 1$.

Each PC represents a particular spatial pattern of interaction energy and its corresponding eigenvalue, the amount of variance in the free energy fluctuations that it accounts for. We find that the spatial frequency of a given PC increases with decreasing variance. The free energy between bin $i$ and $j$, separated a distance $d$, can be decomposed using PCA as

$$F_{ij} = \bar{F}_d + \frac{1}{2} \sum_\alpha [p_\alpha^i v_{\alpha,j-i} + p_\alpha^j v_{\alpha,i-j}],$$

where, $v_d^\alpha$ is the $\alpha^{th}$ eigenvector whose elements depend in genomic separation $d$. The coefficient, $p_d^\alpha$, is the projection of the $i^{th}$ free energy profile onto the $\alpha^{th}$ eigenvector, $p_d^\alpha = \sum_d v_d^\alpha (F_{i,i+d} - \bar{F}_d)$. In the analysis that follows, we have used a genomic separation cut-off of $k_c = 60$ bins, which at a resolution of 10 kbp corresponds to free energy profiles and eigenvectors that range from $[-600, \ldots, 600]$ kbp. Matrix elements corresponding to bins $i$ and $j$ that have $|i - j| > d_c$ are excluded from analysis, as Hi-C counts become too sparse.

We can use PCA to filter out principal components (PCs) that are identifiable with systematic biases or noise, leading to a smoothed set of interaction energies, $\delta F'_{ij}$. The specific interaction energy can be reconstructed via

$$\delta F'_{ij} = \frac{1}{2} \sum_\beta [p_\beta^i v_{\beta,j-i} + p_\beta^j v_{\beta,i-j}],$$

where the sum is over only the eigenvectors $\beta$ that are not identified with systematic biases and whose eigenvalues lie above the noise cut-off.

### 2.1.6 Calculating chromatin coupling energies

We model the specific energy of interaction, $\delta F'_{ij}$, between bins $i$ and $j$ as a sum of pairwise interactions between the bound chromatin factors at those two locations. This can be written
as
\[
\delta F'_{ij} = \sum_{\nu \geq \mu} J_{\mu\nu} [S_i^\mu S_j^\nu + S_i^\nu S_j^\mu],
\]  
(2.6)

where \(0 < S_i^\mu < 1\) is the occupancy of chromatin factor \(\mu\) at bin \(i\) (and can be determined from binding data), and \(J_{\mu\nu}\) is the symmetric coupling energy between chromatin factors \(\mu\) and \(\nu\).

Equation 2.6 presents a linear system that can be fit through least-squares to obtain the \(J_{\mu\nu}\).

### 2.2 Results

Using the analysis techniques described above, we determine the long-range coupling energies, \(J_{\mu\nu}\), between a set of chromatin factors by combining the frequencies of interaction as measured by Hi-C and the factors’ genome-wide binding locations. *Drosophila* makes an excellent model organism on which to test this analysis, as the measured Hi-C dataset [35] is of sufficient resolution (down to 10 kbp resolution) and there exist a number of measured binding sites for chromatin factors [117,171]. Here, we consider insulator-associated proteins (BEAF, dCTCF, GAF, Zw5 and CP190), as well as Polycomb-group proteins (Pc, PCL, Pho) that have, respectively, been shown to be responsible for setting up euchromatic [173,174] and heterochromatic [35,175] domains via looping interactions. We also include factors such as Cohesin, Chromator and PolII that are known to be associated with insulators. We will show that our PCA methodology can be used to filter out biases as well as high frequency noise in the Hi-C data. Using our interaction model, we assess our PCA normalization procedure against other normalizations methods based on how well it can fit the corrected Hi-C data. The result is a biologically meaningful set of predictions for the effective energetic couplings between chromatin factors.

**2.2.1 Distance-dependent free energy**

One of the challenges in analyzing Hi-C data is the existence of systematic biases due to the measurement protocol, and several normalization procedures have been put forward to correct for them. We wish to determine whether these normalization procedures have any effect on the predicted coupling energies between chromatin factors. Using the original published Hi-C dataset for *Drosophila* [35] we have constructed several different contact matrices at a resolution of 10 kbp (see Methods). The contact matrix gives the number of times that a given 10 kbp bin is in contact with another, non-overlapping 10 kbp bin in the genome. We have made a contact matrix based on the original observations, termed *raw* in what follows. In Sexton et al. [35], they presented a hierarchical probabilistic model to correct for various biases in the raw data (see Methods). We have applied this method to the
distance-dependent background free energy, \( F_{ij} \) that are at a fixed genomic separation \( d = j - i \) (see Methods). In Fig. 2.1, we plot \( F_{ij} \) for the three different free energy matrices used in the analysis. It can be seen that the free energy associated with this looping increases with the linear separation. We have fit each of the three average free energies to the prediction for that of a random polymer, namely that \( F_{ij} \sim \alpha \log |d| \), where \( \alpha \) is the scaling exponent. For an ideal random polymer in

\[ F_{ij} \sim \alpha \log |d| \]

This leads to three Hi-C contact matrices and they will be labeled as: raw, raw + ICE, and hierarchical. Each of these will be transformed into free energies and filtered by PCA to see if it can improve the fit to the interaction model we presented in Methods and detail below.

For each contact matrix, we apply our free energy transformation (see Eq. 2.2 in Methods), leading to three different free energy matrices that represent the energetics of interaction between genomic locations. Regardless of whether the contact matrix was normalized or not, the dominant contribution to the free energy is due to the distance-dependent entropic cost of looping the DNA polymer between two genomic locations. We determine this distance-dependent background free energy, \( F_{ij} \), by averaging together all free energy matrix elements \( F_{ij} \) by averaging together all free energy matrix elements

\[ F_{ij} \sim \alpha \log |d| \]

where \( \alpha \) is the scaling exponent. For an ideal random polymer in

\[ F_{ij} \sim \alpha \log |d| \]
3D, the scaling exponent would be predicted to be $\alpha = 3/2$. From the *Drosophila* Hi-C data, we find that the four matrices have average free energies that have roughly the same scaling ($\alpha = 1.1 \pm 0.1$). This result is in agreement with that found for other Hi-C datasets [40] and is consistent with the scaling of the fractal globule polymer model [56]. We next show how the free energy fluctuations around this distance average can be further decomposed into an independent set of interaction modes using PCA.

### 2.2.2 PCA of free energy profiles

From each free energy matrix $F_{ij}$, we create a list of free energy profiles $F_i$, one for each genomic bin. A given free energy profile shows how that particular genomic segment interacts with the surrounding region, $F_i = \{F_{i,i+d}\}$ for $d = -d_c \ldots d_c$. Besides the background free energy $\bar{F}_d$, each profile will have free energy fluctuations, $\delta F_i$, that are potentially due to interactions between bound factors, or systematic biases. We use PCA to identify the independent contributions to the fluctuations in the free energy. The top principal components (PCs) constitute common patterns of interaction that are present at many locations in the genome. The aim is to then identify which PCs represent systematic effects as well as those that are just associated with noise in the Hi-C data. These can then be filtered out to create a corrected set of free energy profiles.

We performed PCA on each of the three matrices. In Fig. 2.2 we show the top four PCs for the raw + ICE free energy matrix. Each PC shows the variation in the free energy as a function of the genomic separation from the bin located at $d = 0$. Positive free energies correspond to repulsive interactions whereas negative ones are attractive, and thus represent stabilizing interactions. It should also be noted that for each PC there is also the inverse interaction profile that is obtained by multiplying the PC by $-1$. These PCs can also be interpreted as a set of spatial modes with which to represent the data, akin to a Fourier decomposition. The characteristic spatial frequency of a PC increases as the corresponding eigenvalue (variance) associated with it decreases. Many of the PCs corresponding to small eigenvalues represent high-frequency noise. In what follows, we show that this noise can be filtered out by reconstructing the specific interaction energies (Eq. 2.5) without including them in the sum.

### 2.2.3 Interaction profiles and chromatin structure

Prior PCA analysis on Hi-C data highlighted the existence of A/B chromatin compartments, namely topological domains that tend to interact with other domains of the same compartment type while avoiding the other [40]. The insulator and Polycomb-group factors of interest in this work are thought to interact to generate such compartmentalization. These domains, which are strongly associated with euchromatin or heterochromatin, can exist on a range of scales. Our free energy PCs also show such a structure, with the first PC marking a transition between such compartments. With increasing number, and decreasing
Figure 2.2: Shown are the first four principal components (unitary vectors) and their sign-inverted versions calculated genome-wide from the $F_{ij}$ matrix created from the raw + ICE contact matrix (top plots). Below each principal component are heat maps of the genome-wide average binding profiles for the selected chromatin factors (see Text). The top heat map corresponds to the positive free energy interaction profile (blue curve), and the bottom heat map for that of the inverse profile (red curve). Red regions in the heat maps represent locations of higher occupancy and blue regions represent lower occupancy. The range of the heat maps goes from 0.0 (blue) to 1.0 (red).
variance in free energy, the PCs show such interactions at smaller and smaller spatial scales. As such, PCs form a basis with which to decompose a given energy interaction profile into various spatial scales.

In order to help clarify the interpretation of the energy interaction profiles represented by each PC, we look at the distribution of bound chromatin factors associated with them. In Fig. 2.2, we show the PCs calculated from the raw + ICE free energy matrix and the corresponding average binding profiles of our selected chromatin factors as heat maps. Each average binding profile was computed using only those genomic locations where the specific energy interaction profile had a significant projection onto the given PC. In particular, the energy profile can have either a negative or positive projection (see Fig. S1, B), and so we create two sets of bins: those bins, $i$, that have projections $p^\alpha_i > 2\sigma_\alpha$ and those with projections $p^\alpha_i < -2\sigma_\alpha$ (corresponding to the inverse PC profile), where $\sigma_\alpha^2$ is the eigenvalue (variance) of the $\alpha^{th}$ PC. We then extract binding profiles ($S^\mu_i$) from the genome-wide binding data of each chromatin factor that are centered on each set of locations. These then get averaged together to give the average binding profile that shows the underlying chromatin structure associated with the given principal component computed from the Hi-C data.

For example, for the raw + ICE free energy matrix, PC1 represents a domain boundary between euchromatin (marked by H3K4me3) and heterochromatin (marked by H3K27me3). Those genomic locations that have a positive projection onto PC1 (middle heat map) have euchromatin factors bound on the left and heterochromatin factors bound on the right. Looking at the associated free energy, euchromatic DNA shows a larger cost in free energy (positive values) associated with looping likely due to it having greater entropy, since it is more open and hence more disordered. As such, PC1 may represent the mutual exclusion of interactions between euchromatin and heterochromatin domains that are physically insulated from one another [35,40]. Fig. 2.2 shows that for the top PCs derived from the raw + ICE matrix, strong correspondences exist between the type of the interaction and the underlying bound factors (i.e. locations that are bound by insulator factors have attractive (negative) interactions with other domains bound by insulators).

2.2.4 Coupling energies between chromatin factors

Using PCA, we can choose to filter out various patterns of variation from the data (using Eq. 2.5), yielding corrected specific interaction energies $\delta F'_{ij}$ between locations $i$ and $j$ (see Methods). For matrices not treated with ICE, the first PC simply represents a constant-value offset that is present in each energy profile (not shown). This corresponds to the biases identified by ICE. In reconstructing the specific interaction energies, $dF'_{ij}$, we leave out this PC for the non-ICE matrices. Reconstructing the interaction energies using a subset of the remaining PCs, smooths the data and filters out noise. We now assess how much filtering to perform based on how well the pairwise interaction model fits the data.
Figure 2.3: Shown in (A, B, C) are the energies of interaction $\delta F_{ij} = F_{ij} - \bar{F}_{j-i}$ of a portion of chromosome 2L for three different $F_{ij}$ matrices: A) raw, B) raw + PC-filtered and hierarchical. (The first 35 PCs were used in reconstructing the raw + PC-filtered free energies). All have been aligned so that the zeroth column corresponds to $i = j$. Blue regions correspond to effective attractive interactions (negative) and red regions to effective repulsive interactions (positive). Figures (D, E) show the locations of pairwise self-contacts, $S_i^\mu S_j^\mu$ for the insulator factor BEAF and the Polycomb-group protein Pc (blue corresponds to $S_i^\mu S_j^\mu = 0$ and red to $S_i^\mu S_j^\mu = 1$). Comparing the interaction energies (A, B, C) with the locations of pairwise contacts (D, E) highlights how these contacts could be generating the observed interactions.
In Fig. 2.3(A-C) we show the specific interaction energies for a portion of chromosome 2L. As can be seen, PCA filtering dramatically smoothens the data, highlighting domains of attractive (blue) and repulsive interactions (red). As a comparison we show the energies computed from hierarchical normalized data for the same region. The two normalized energy matrices agree in many domains, but do possess differences, such as the size of the interacting domain situated around 9 Mbp. Many of these interactions are due to specific contacts between chromatin factors at the given loci. We highlight this connection by showing the pairwise self-contacts \((S_i^\mu \text{ with } S_i^\nu)\) in the same region for the insulator BEAF and the Polycomb factor Pc (Fig. 2.3D,E). For example, some of the attractive energies (blue region near 8Mb in the \(\delta F_{i,j}\) heat maps) are likely due to interactions between insulators (BEAF-BEAF domain in Fig. 2.3D), whereas other attractive interactions (region between 5Mb and 6Mb) could be due to interactions between the Polycomb group of factors (Pc-Pc domain in Fig. 2.3E). We now assess how well the interaction energies are fit to a model that takes the distribution of contacts between bound factors as input.

For each set of interactions energies, either filtered by PCA or some other normalization method, we fit Eq. 2.6 to determine a fitted set of coupling energies \(J_{\mu\nu}\). (We have fit all the chromosomes at once, as well as chromosome by chromosome, allowing us to determine how much the fitted interactions vary by chromosome). We use \(\chi^2\) and the Pearson correlation coefficient to determine how much PC filtering, if any, should be applied to the interaction energies. In Fig. 2.4, we show that for the interaction energies derived from the raw and raw + ICE matrices, PCA filtering can improve the quality of the fit. Fig. 2.4A shows that using the first 35 PCs (out of the total 120) leads to the best genome-wide fit of the data by the model (for the non-ICE matrix, we also left out the constant offset PC). Interestingly, using ICE reduced the overall quality of the fit compared to the raw matrix, though PC filtering was able to improve the fitting for both. In addition, we found that applying any form of PC filtering to the interaction energies derived from the hierarchical normalized matrix always made the fit worse. As a summary, in Fig. 2.4B,C we show the chromosome by chromosome \(\chi^2\) and Pearson correlation coefficient for the various fits of the model to both PC-filtered and unfiltered data. PC filtering of the energies computed from the raw matrix give the best overall fit. The distance-dependent scalings applied in the hierarchical normalization method lower the correlation between the interaction energies and the underlying bound chromatin factors, lessening the quality of the fit.

In Fig. 2.5, we show the fitted coupling energies from the fits to the raw, raw + PC filtered and hierarchical data. As mentioned, PCA filtering improved the fit, yet the resulting \(J_{\mu\nu}\) show an overall agreement between the different data sets. Here we show the average \(J_{\mu\nu}\) over all the chromosomes (left heat maps) and their associated standard deviations (right heat maps). The parameter error estimates show that many of the couplings are consistently predicted from one chromosome to the next. An inspection of the fitted couplings that are consistent across chromosomes show that many of the insulators and factors that are
Figure 2.4: There is an optimal number of PCs to use in reconstructing the energies of interaction \( \delta F_{ij} \). A) Pearson correlation coefficient and reduced \( \chi^2 \) of the genome-wide fits of the given data (see legend) to the model using Eq. 2.6 as a function of the number of PCs used in the filtering. For the energies derived from raw matrix, PC1 was excluded as it is simply a constant offset. Figures B, C) show the best fit results for the various datasets by chromosome. PCA filtering for the raw matrix leads to the best overall results.
Figure 2.5: The fitted coupling energies, $J_{\mu\nu}$, between chromosome associated factors (units of $k_B T$). The left heat maps show the average $J_{\mu\nu}$ over chromosomes, and the right heat map the associated standard deviations. The following free energy matrices were used: A) raw, B) raw + PC-filtered (optimal number of PCs used was 35) and C) hierarchical.
linked to euchromatic domains have attractive (negative) interactions, speaking to their ability to stabilize loops in such domains [173, 174]. Many of these have effective repulsive (loop hindering) interactions with Polycomb-group proteins (PCL, Pc), though some have attractive interactions with Pho. Other things that are shared between these sets of $J_{\mu\nu}$ are the associations between BEAF, Chromator and Cohesin and the transcriptional machinery factors, PolII and Nurf. Interestingly, the predicted interactions between CTCF and such factors are more complex, highlighted by the effective positive interactions. We should also point out that a given $J_{\mu\nu}$ represents a pair’s effect on looping and should not be interpreted as a prediction of whether they interact or not. Factors may very well interact (i.e. have attractive protein-protein interactions) but yet have a destabilizing effect on loop formation. We note that within both the insulator and Polycomb-group, some pairs of factors are predicted to effectively raise the energy of loop formation.

It should be recalled that these are interactions determined at a resolution of 10 kbp, so factors that might juxtapose side-by-side at boundaries that form on finer length scales would get grouped together. Experiments that probe at finer resolutions would be valuable in sorting out potential conglomerated interactions. Nevertheless, our findings highlight how using PCA can help improve the quality of fit of Hi-C data to a model for chromatin factor interactions and that a consistent set of couplings can be predicted, which can be explored experimentally.

2.3 Conclusion

In this chapter, we have described a method for normalizing Hi-C data using PCA. This technique decomposes the free energy into various contributions, including a distance-dependent entropic free energy, potential systematic biases, and specific energies of interaction potentially arising from DNA-bound factors. We assessed the performance of the PCA based normalization method, along with two others, by fitting the corrected data to a pairwise interaction model that took as input the locations of bound chromatin factors. This allowed us to determine the coupling energies between DNA-bound factors from the energies of interaction as determined from Hi-C data. As a test case, we calculated the couplings between insulators, Polycomb-group, other chromatin factors and some of the transcriptional machinery. These factors are responsible for setting up the domains/compartments in the DNA, yielding the two-compartment model that can broadly classify chromatin structure. Recent work has shown that a simple A/B interacting copolymer model can capture many of the observed patterns found in Hi-C data [149]. Polymer simulations including a simple insulator interaction have also shown how compartments can be formed [153]. Our work, is a first step toward trying to break apart the interactions within such compartments into their constitutive parts. The couplings found here could help further such simulations by including a richer set of interactions. Of course, this requires a reliable set of predictions for
interactions, and we have shown that correcting the Hi-C data using PC can improve the quality of the fit.

The methods presented here are readily applicable to the Hi-C and bound factor data obtained in other organisms and should provide a common framework in aiding the correction and ultimate functional analysis of such data. Our work may thus provide the community with a valuable tool not only to predict the strength of Hi-C interactions due to chromatin-associated factors, but also to better evaluate the specific variations encountered depending on cellular contexts and/or conditions.
Chapter 3

Mutual information between chromatin sequence and structure: quantifying the role of non-local effects

Statistical analyses of chromatin data are key to uncovering details about the biological processes that shape chromatin conformation. Nevertheless, how do we determine which particular statistic is the most relevant to the process? Or in other words, what statistical feature of the chromatin sequence is most informative about chromatin structure?

In the previous chapter, we presented a method based on Principal Component Analysis that was able to filter out various biases and noise present in Hi-C data, to facilitate the identification of which interactions between insulator proteins, Polycomb-group proteins and associated cofactors aid the formation of contacts between distant genomic sites. It is well known that many of these DNA-binding factors interact in the formation of DNA loops [41,65], and thus we developed a model where the energetics of contact between a pair of sites depended only on the pairwise interactions between the various factors bound to them.

Building on that model, here we further investigate the connections between DNA-binding factors and chromatin conformation, with the goal of developing a statistical model that can predict chromatin contact maps from information about the distribution of bound chromatin factors. This involves overcoming a main challenge posed by our previous model where the energy of contact between two sites only depended on what was bound to those sites and ignored the effect of interactions with other neighbours. Here, we explore the validity of this assumption and how we can best connect the distribution of DNA-bound factors on the genome to chromatin contacts.

In this chapter, we explore these connections from a model-free perspective by measuring how informative chromatin factors are about contacts. Our analysis takes advantage of
information-theoretic quantities that were initially developed to find the fundamental limits of signal transmission through noisy channels. These concepts have been useful in studying a great variety of biological systems [180–185]. Here, we will use them to quantify how much information is shared between the sequence of bound chromatin factors and chromatin contacts and specifically, which locations around a pair of genomic sites are informative of their likelihood of contact.

3.1 Methods

In this section, we present the methods used in this chapter for analyzing the chromatin folding problem using information-theoretic techniques. For this, we need to define the chromatin sequence and its conformation as input and output signals as well as how to discretize them into just a few possible states. We also review some useful information-theoretic quantities that can be calculated on those variables.

3.1.1 Chromatin sequence states

Hundreds of chromatin factors such as DNA-binding proteins and chemical modifications of histones and DNA nucleotides have been identified so far. The occupations of these factors along the genome are highly correlated, suggesting the existence of just a few bound states. Those states are generally associated to specific functional effects on chromatin conformation and/or gene expression, such as “silent chromatin”, “facultative and constitutive chromatin”, “active chromatin”, etc. We thus use the bound chromatin state of each genomic location as a random variable for our signalling system input.

In particular, we take advantage of a readily available chromatin classification of the *Drosophila melanogaster* genome into five states. This classification is known as the “chromatin colors annotation” and was generated by Filion et al. [141]. In it, each genomic site belongs to one of five possible chromatin states $\mu = \{\text{blue, black, green, red, yellow}\}$. This classification is based on the DamID profiles of 53 different chromatin factors in the embryonic *Drosophila melanogaster* cell line Kc167 (8-12 hours). The authors first compressed the 53 profiles into just three dimensions through Principal Component Analysis [186,187]. Then they assumed that this dimensionally reduced chromatin sequence is made of five hidden states that emit the three-dimensional signal with a certain probability. In addition, they assumed that the state of a given genomic site depended on the state of its nearest neighbour sites. They thus applied a Hidden Markov Model (HMM) [188] to derive a sequence of chromatin states that maximizes the likelihood that each site belongs to a particular state.

By looking at the characteristics of the associated factors, as well as features such as gene activity and gene type, the five states can be interpreted in the following way:
• Green chromatin (G): Heterochromatin involving the proteins SU(VAR)3-9, HP1, LHR and HP6. Enriched for the constitutively repressive histone modification H3K9me2. Small genomic coverage (<5%).

• Blue chromatin (B): Heterochromatin involving PcG proteins. Enriched for the facultatively repressive histone modification H3K27me3.

• Black chromatin (K): Prevalent type of strongly repressive chromatin (covering 48% of the genome).

• Red chromatin (R): Transcriptionally active chromatin. Enriched for insulator proteins and the histone marks H3K4me2 and H3K79me3. Strong enrichment for the origin recognition complex. Linked to specific genetic functions.

• Yellow chromatin (Y): Transcriptionally active chromatin. Enriched for the histone marks H3K4me2, H3K79me3 and H3K36me3. Linked to broad genetic functionality.

The sequence of chromatin states that we use in this chapter was generated by first discretizing the genome into 10 kbp long bins and then reporting as chromatin states the chromatin colors that occupied the largest genomic portion of each bin.

3.1.2 Chromatin contact affinities

For chromatin contacts, we used the publicly available Hi-C experiments done by Schuettengruber et al [189] (GSE61471), performed on 3000–4000 Drosophila melanogaster embryos, 16–18 hours after egg laying. This data is presented as a contact map, a matrix whose elements $n_{ij}$ are the number of times a particular pair of genomic bins (of resolution 10 kbp each) were found to be in contact.

In order to convert chromatin contact counts into a random variable that can take multiple states, we discretized the Hi-C counts $n_{ij}$ at each given distance $d = |j - i|$ into three contact affinity states $a_{ij} = \{\text{high}, \text{medium}, \text{low}\}$. This classification should satisfy the conditions that the average counts $\langle n_{ij} \rangle$ observed in high contact-affinity states are greater than those for medium affinity, and the average counts for medium affinity states are greater than low affinity. Instead of choosing an arbitrary threshold to separate chromatin counts at a given distance of contact into high, low and medium affinities, we used an HMM [188] to find the most likely classification in an unsupervised manner (Fig. 3.1A). In this HMM, we assume that each contact affinity state emits contact counts from a normal distribution, $\mathcal{N}(\langle n_{ij} \rangle, \sigma_{ij} | a_{ij})$. In addition, given the spatial correlation that exists in chromatin contact maps, the probability that a pair of sites at a given distance $(i, i + d)$ belongs to a particular affinity state likely also depends on the state of their nearest neighbours at that distance, namely, the pairs $(i + 1, i + 1 + d)$ and $(i - 1, i - 1 + d)$. After fitting the HMM to the Hi-C
Figure 3.1: (A) Pairs of sites in the genome were converted to a three-state classification of contact affinities based on their Hi-C contacts at each genomic distance using a Hidden Markov Model. On the left is a Hi-C matrix highlighting all the pairs of sites separated by a distance of 100 kbp. On the right are the counts (gray) and the means of the three affinity states (high, medium, low) that they were classified into (red). (B) Mean contact counts for pairs belonging to each affinity state as a function of distance. The distributions follow approximately a power-law. (B) Probability that pairs of sites belong to each contact affinity class, from the most likely state classification. Throughout all distances of contact, low affinity is the most probable, followed by medium affinity, and high affinity as the rarest class.
data, we calculate the posterior probability of each pair of sites being in each of the three affinity states and categorize the pair as being in the most probable affinity state.

In Fig. 3.1B we see that the mean number of contacts for each contact affinity state approximately follow a power law with distance as one would expect to find for the probability of a polymer looping. Moreover, we observe that the log-distance between the contact means of high, medium and low affinity states is approximately constant for all distances, suggesting a constant free-energy offset between these states. Fig. 3.1C shows that for all distances, low contact affinity is the most likely state to occur in the contact map, followed by medium contact affinity, and lastly, high contact affinity.

### 3.1.3 Information-theoretic quantities

Now we present a series of information theoretic quantities that will allow us to assess the amount of information transmitted between chromatin sequence and chromatin structure in experimental data.

**Shannon entropy**

If we have a random variable $X$ that can take discrete values $x$ with a given probability $P(x)$, Shannon entropy [190], or simply entropy, is a measurement of the inherent uncertainty of the random variable,

$$H(X) = - \sum_x P(x) \log P(x), \quad (3.1)$$

where the logarithm can be chosen to be base 2, so that the quantity is measured in bits.

To gain some intuition, suppose a binary $N$-long word made of 0s and 1s that occur with the same frequency, i.e. at each position within the word $P(0) = P(1) = 1/2$. The probability of each word is $(1/2)^N$ and there are $2^N$ of them, so their entropy is

$$H(X) = - \sum_x P(x) \log P(x) = -2^N \left(\frac{1}{2}\right)^N \log \left(\frac{1}{2}\right)^N = N. \quad (3.2)$$

The number of bits of the binary word (with equal probabilities) is equal to its length.

The concept of uncertainty is tightly related to information; the more uncertain a variable is, the more information one needs to describe it. Following with the binary word example, one needs to ask $N$ yes/no (1/0) non-redundant questions in order to know which $N$-long word it is. The answer to each question adds one bit of information to your knowledge about the word and at the same time reduces the uncertainty of the word by one bit.

Since $0 < P(x) < 1$ and $-\log P(x) > 0$, entropy is always positive. In addition, by using Jensen’s inequality (if $f$ is convex, $\langle f(X) \rangle \geq f(\langle X \rangle)$ [191]) it can easily be shown that entropy is maximized when the distribution is uniform [192], which fits the intuition.
that if all states of the variable are equally probable, the uncertainty about which state the
variable is in is the largest.

Entropy can be calculated for joint probability distributions $P(x, y)$ in a similar manner,

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y). \quad (3.3)$$

Using Jensen’s equality once again [192], it can be shown that the joint entropy is subad-
ditive, i.e. $H(X, Y) \leq H(X) + H(Y)$, where the equal sign holds for the case where $X$ and
$Y$ are completely independent variables.

**Conditional entropy**

The conditional entropy is the entropy of one variable $X$ when another variable $Y$ is fixed,

$$H(X|Y) = -\sum_{x,y} P(x, y) \log P(x|y). \quad (3.4)$$

This quantity is therefore the expectation of $-\log P(x|y)$ over all values of $X$ and $Y$, and
indicates the uncertainty of $X$ whenever $Y$ is known.

**Mutual Information**

For two random variables $X$ and $Y$, how much information about one variable is carried
by the other? Or in other words, how much is the uncertainty about one variable reduced
when the other is known? This quantity is the mutual information,

$$I(X; Y) = -\sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}. \quad (3.5)$$

Mutual information is symmetrical, $I(X; Y) = I(Y; X)$, and non-negative (shown through
Jensen’s inequality) [192]. By looking at the expression 3.5 we see that mutual information
is a measurement of how different the joint distribution $P(x, y)$ is from the case where $X$
and $Y$ are completely independent variables ($P(x, y) = P(x)P(y)$). In the independent case,
$I(X; Y) = 0$, i.e. one would not gain any information about $X$ by knowing $Y$.

One can further show from Eqs. 3.3, 3.4 and 3.5 that the mutual information between
$X$ and $Y$ is equal to the entropy of $X$ minus the conditional entropy of $X$ with respect to
$Y$,

$$I(X; Y) = H(X) - H(X|Y), \quad (3.6)$$

i.e. $I(X; Y)$ represents how much the uncertainty about $X$ is reduced when $Y$ is known, or
in other words how informative about $X$ is $Y$. From this equation we see again that if $X$
and $Y$ are independent from each other, $H(X) = H(X|Y)$ and thus $I(X; Y) = 0$. 

47
The additive relations between the information-quantities presented so far can be easily visualized in a Venn diagram (Fig. 3.2).

Figure 3.2: A Venn diagram illustrates the additive properties of the following information-theoretic quantities. $H(X)$ is the entropy of $X$, $H(Y)$ is the entropy of $Y$, $H(X,Y)$ is the joint entropy of $X$ and $Y$, $H(X|Y)$ is the conditional entropy of $X$ with respect to $Y$, $H(Y|X)$ is the conditional entropy of $Y$ with respect to $X$ and $I(X;Y)$ is the mutual information between $X$ and $Y$.

**Uncertainty coefficient**

We have seen that mutual information $I(X;Y)$ is the entropy shared between two variables $X$ and $Y$. Nevertheless, how do we quantify if the mutual information between two variables is significant? One strategy is to divide mutual information by the entropy of one of the two variables involved, a quantity known as the uncertainty coefficient,

$$U(X|Y) = \frac{I(X;Y)}{H(X)} = \frac{H(X) - H(X|Y)}{H(X)}, \quad (3.7)$$

which indicates how much of the total information in $X$ is captured by $Y$, a quantity that ranges from zero ($X$ and $Y$ are independent) to one (there is a one-to-one function between $X$ and $Y$).
3.1.4 Maximum-entropy method

Suppose we have a system composed of a number $N$ of discrete random variables $X = (x_1, x_2, \ldots, x_N)$. As the size of the system increases, the number of possible states in $X$ grows exponentially with $N$. This poses a challenge for estimating $P(X)$ directly from data, as the number of possible states in the system rapidly becomes too large.

Nevertheless, despite the impossibility of directly estimating $P(X)$, one may still be able to reliably measure the average value of certain functions of the system $f_i(X)$, $\langle f_i(X) \rangle_{\text{exp}}$, where $\langle \cdot \rangle_{\text{exp}}$ denotes the average over experimental data. Consequently, we can search for the least-structured probability distribution $\tilde{P}(X)$ that is constrained to match the given experimental averages [193,194],

$$
\sum_X f_i(X) \tilde{P}(X) = \langle f_i(X) \rangle_{\text{exp}}.
$$

(3.8)

As we have already seen, the entropy $H(X)$ is a measurement of uncertainty about a system. We thus look for the most entropic distribution $\tilde{P}(X)$ that satisfies Eq. 3.8 to find an expression for $\tilde{P}(X)$ that is consistent with what can be “known” about the system.

Maximum-entropy distribution

One can find the maximum-entropy distribution $\tilde{P}(X)$ by the method of Lagrange multipliers [193,194]. Specifically, we write a generalized entropy function $\Lambda$, which includes $M$ Lagrange multipliers $\lambda_i$ to enforce the experimental measurements (Eq. 3.8) as well as an additional Lagrange multiplier $\lambda_0$ to enforce the normalization of $\tilde{P}(X)$ to unity,

$$
\Lambda = -\sum_X \tilde{P}(X) \log \tilde{P}(X) - \sum_{i=1}^M \lambda_i [\sum_X f_i(X) \tilde{P}(X) - \langle f_i(X) \rangle_{\text{exp}}] - \lambda_0 [\sum_X \tilde{P}(X) - 1].
$$

(3.9)

We then proceed to maximize $\Lambda$ with respect to $\tilde{P}(X)$,

$$
0 = \frac{\partial \Lambda}{\partial \tilde{P}(X)} = -\log \tilde{P}(X) - 1 - \sum_{i=1}^M \lambda_i f_i(X) - \lambda_0,
$$

(3.10)

which gives

$$
\tilde{P}(X) = \frac{1}{Z} \exp \left( -\sum_{i=1}^M \lambda_i f_i(X) \right),
$$

(3.11)

where $Z = \lambda_0 + 1$. It can further be shown by either maximizing $\Lambda$ with respect to $\lambda_0$ or by normalizing $\tilde{P}(X)$ to unity that $Z = \sum_X \exp[-\sum_{i=1}^M \lambda_i f_i(X)]$.

Note that if we consider the measurements $x_k$ to be binary random variables $\sigma_k$ that can take the values 1 and $-1$, with $X = \vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$, and we chose as the experi-
mental measurements all one-spin averages $\langle \sigma_k \rangle$ and two-spin averages $\langle \sigma_k \sigma_l \rangle$, then Eq. 3.11 becomes the Ising model [193,194],

$$\tilde{P}(\sigma) = \frac{1}{Z} \exp \left( \sum_k h_k \sigma_k + \sum_{l>k} J_{kl} \sigma_k \sigma_l \right),$$  \hspace{1cm} (3.12)

where $(h_k, J_{kl})$ are the Lagrange multipliers and $Z$ a normalization constant.

**Fitting the maximum-entropy distribution**

There exists a unique set of Lagrange multipliers $\lambda_i$ that satisfy all constraints in Eq. 3.8 [195]. This is a well-known inverse problem in computer science often referred as “Boltzmann machine learning” [196,197].

In order to find the values of $\lambda_i$, one can maximize the generalized entropy function $\Lambda$ with respect to them, obtaining a set of equations,

$$0 = \frac{\partial \Lambda}{\partial \lambda_i} = \langle f_i(X) \rangle_{\exp} - \sum_X f_i(X) \tilde{P}(X).$$  \hspace{1cm} (3.13)

While these equations are generally hard to solve, they suggest a gradient descent algorithm with the following update rules:

$$\lambda_i \leftarrow \lambda_i + \eta \left( \langle f_i(X) \rangle_{\exp} - \sum_X f_i(X) \tilde{P}(X) \right),$$  \hspace{1cm} (3.14)

where $\eta$ is a small learning rate.

Each iteration of this gradient descent algorithm consists of two parts. First, one needs to calculate $\sum_X f_i(X) \tilde{P}(X)$. This involves summing over all states of $X$, which can be done by exhaustive enumeration of all states for small-size systems, however this becomes computationally costly as the size of the system increases. Nevertheless, a number of Monte Carlo methods for approximating the sum exist [198,199]. Second, the parameters are updated using Eq. 3.14 until convergence.

For the particular case described above where our system consists of binary random variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ and the constraining statistics are the one-spin averages $\langle \sigma_k \rangle$ and two-spin averages $\langle \sigma_k \sigma_l \rangle$, the fitting of the maximum entropy $\tilde{P}(\sigma)$ works as follows,

1. Start with random guess for $h_k$ and $J_{kl}$

2. Use Eq. 3.12 to calculate $\tilde{P}(\sigma)$ for every possible $\sigma$ and obtain $\langle \sigma_k \rangle_{\tilde{P}}$ and $\langle \sigma_k \sigma_l \rangle_{\tilde{P}}$.

3. Check for convergence, $|\langle \sigma_k \rangle_{\tilde{P}} - \langle \sigma_k \rangle_{\exp}| < \epsilon$ and $|\langle \sigma_k \sigma_l \rangle_{\tilde{P}} - \langle \sigma_k \sigma_l \rangle_{\exp}| < \epsilon$, where $\epsilon$ is the error tolerance (we use $\epsilon = 0.0001$). If it hasn’t converged yet, continue.

4. Update $h_k$ as $\Delta h_k = \eta (\langle \sigma_k \rangle_{\exp} - \langle \sigma_k \rangle_{\tilde{P}})$ and $J_{kl}$ as $\Delta J_{kl} = \eta (\langle \sigma_k \sigma_l \rangle_{\exp} - \langle \sigma_k \sigma_l \rangle_{\tilde{P}})$, where $\alpha$ is a small learning rate (we use $\alpha = 0.025$). Go back to step 2.
3.2 Results

3.2.1 The importance of neighbours in forming chromatin contacts

In the model presented in the previous chapter, the free energy of contact between two sites $i$ and $j$, $F_{ij}$, was separated into the sum of two independent terms: a free energy coming from the polymer entropy $S \propto \alpha \log|j - i|$, that treated the chromatin as a homogeneous polymer, and an energetic part $E_{ij}$ that consisted of the sum of pairwise interactions between the factors bound to $i$ and $j$. We thus consider this model to be local as it does not take into account the energetics of other sites that may contribute to $i$ and $j$ forming a contact.

Although the model can predict free energies of contact, the separation of the free energy of contact into an entropic contribution from the homogeneous polymer and the local binding free energy effect was compromised by the fact that chromatin factors in the genome have characteristic domain sizes, as can be seen from their spatial autocorrelation (FIG. 3.3). This implies that each combination of binding factors at $i$ and $j$ will tend to have specific combinations of binding factors in between them. Those factors have been found to affect the entropy of the polymer [35] and therefore one may expect to find deviations from the genome average $S \propto \alpha \log|j - i|$ linked to the factors present between $i$ and $j$.  

Figure 3.3: Spatial autocorrelation of DNA binding factors along the genome from ChIP-seq data measured by Roy et al [90].
In addition, we can directly measure from the data how the interactions with the neighbouring chromatin generate deviations from what the local model would predict. As an example, we consider the probability of contact $P(c_{ij})$ between two sites in the genome situated a certain distance apart $d = |j - i|$ that have a third site $k$ in the neighbourhood of $j$ that is energetically attracted to $i$, with a free energy of binding $E_{ik} < 0$. In this scenario, $k$ and $i$ will be in contact more often than one would expect for non-interacting sites, therefore altering the effective distance between $i$ and $j$ (see scheme in Fig. 3.4A).

In Fig. 3.4B we show this effect in real data by measuring how the presence of the self-associating BEAF protein in *Drosophila* alters the probabilities of contact of surrounding genomic sites that are not bound by BEAF. Specifically, we selected pairs of genomic sites that were each bound by BEAF and we labeled them as $i$ and $k$. Then we extracted from Hi-C data the probability of contact $P(c_{ij})$ between sites $i$, bound by BEAF, and other sites $j$ in the proximity of $k$ that were not bound by BEAF. This probability of contact was then divided by the genomic-background probability of contact $P(c_d)$ between sites situated at the same distance $d = |j - i|$, irrespective of chromatin factors. In Fig. 3.4B we show $P(c_{ij})/P(c_d)$ versus $|j - i|/|k - i|$, which is the distance of contact normalized by the distance between the BEAF-bound sites. This normalization allows us to display on the same graph the data measured at different distances between the BEAF-bound sites $i$ and $k$. For the data points where $|j - i|/|k - i| < 1$, $j$ resides between the BEAF-bound sites; for the data points where $|j - i|/|k - i| > 1$, $j$ resides outside the BEAF-bound sites. We observe that the probability that a site bound by BEAF, $i$, contacts a site not bound by BEAF, $j$, gets higher with respect to the background the more proximal that the latter site is to a BEAF-bound site, $k$. Note that if neighbour effects were nonexistent, as considered in a purely local model, $P(c_{ij})/P(c_d)$ would equal 1 except when $j = k$.

This non-local effect can be modelled by considering a polymer chain with only two attractively interacting beads situated at $i$ and $k$. In it, each segment of the polymer of size $N$ is an ideal chain [13], whose probability of generating a contact loop is

$$\phi_N = \int_0^{2\sigma} \left( \frac{3}{2\pi N} \right)^{3/2} 4\pi r^2 e^{-\frac{3r^2}{4N}} dr$$  \hspace{1cm} (3.15)

for a radius of contact $\sigma$. The probability of not generating a contact is therefore $\overline{\phi_N} = (1 - \phi_N)$.

Then, the probability that $i$ and $j$ are found in contact is calculated by adding the individual probabilities of all the events in which $i$ and $j$ are in contact (e.g. $i$ and $j$ will be in contact if $i$ and $k$ are in contact at the same time that $k$ and $j$ are also in contact). In addition, each configuration in which $i$ and $k$ contact each other, includes an extra Boltzmann factor $e^{-E_{ik}}$. The resulting probability is therefore

52
\[ P(c_{ij}) = \frac{e^{-E_{ik} \phi_{|k-i|}} \phi_{|k-j|} \phi_{|j-i|} + \phi_{|j-i|} \left( \phi_{|k-i|} + e^{-E_{ik} \phi_{|k-i|}} \right) + e^{-E_{ik} \phi_{|k-j|}}}{Z} , \]

where \( Z \) is a normalization constant that can be found by adding the Boltzmann weights of all the possible contact/not-contact configurations between sites \( i, j \) and \( k \) (8 in total).

Fitting this expression to the BEAF data (the points in in Fig. 3.4B), we obtain an energy of self-interaction \( E_{ik} = -0.72 \pm 0.01 \) for the BEAF insulator (see blue line in FIG. 3.4B).

This calculation thus shows that the energy of contact between two chromatin sites affects the rest of the polymer configuration. Consequently, a complete description of the probability that two genomic sites interact with each other must take into account the role of the neighbouring sites. However, moving from local energetics to non-local energetics dramatically increases the complexity of the problem.

In this section we modelled such non-local effects by only considering the interactions of three sites, treating the rest as a homogeneous polymer chain. A more complete description of the problem would need to consider the energetic effects of a larger number of neighbours \( n \). However, the number of configurations that would need to be taken into account grows as \( 2^{n(n-1)/2} \), so even for a modest number neighbours the problem becomes rapidly intractable.

Nevertheless, our intuition is that not all genomic neighbours may have the same level of influence on the contact probability between two chromatin sites. For instance, one may imagine that sites far away from the sites of contact may have little influence, allowing for the independent regulation of distant genomic regions. On the contrary, one may naively think that the energetics of the sites that are in the immediate vicinity of the contacting sites have the largest influence on probability of contact. In the following sections we explore the relationship between contact probabilities and the chromatin sequence at neighbouring sites in a model-free fashion.

3.2.2 Non-local chromatin information

Chromatin specificity of contacts

Starting from a local perspective, regardless of any particular dependence between chromatin sequence and chromatin conformation, a connection between the contact affinity \( a_{ij} \) between two sites \( i \) and \( j \) and the chromatin sequence states \( \mu_i \) and \( \mu_j \) at those sites would involve a certain amount of mutual information between the two quantities. We thus calculate Eq. 3.5 considering \( X = (\mu_i, \mu_j) \) and \( Y = a_{ij} \) as the random variables involved,

\[ I(a_{ij}; \mu_i \mu_j) = \sum_{a_{ij}} \sum_{\mu_i \mu_j} P(a_{ij}, \mu_i \mu_j) \log \frac{P(a_{ij}, \mu_i \mu_j)}{P(a_{ij}) P(\mu_i \mu_j)} . \]
Figure 3.4: Non-local effects in chromatin loops. (A) Schematic of how the probability of contact $P(c_{ij})$ between two sites $i$ and $j$ can be altered if site $i$ interacts favourably with another neighbouring site $k$. Whenever the energetically attractive $i$ and $k$ are in contact with each other ($c_{ik}$), the effective distance between $i$ and $j$ is altered in comparison to the cases where $i$ and $k$ do not contact each other ($\overline{c_{ik}}$). (B) Experimental evidence of this non-local effect can be seen by analyzing genomic sites bound by the self-attractive BEAF protein. In the plot, sites $i$ and $k$ are bound by BEAF whereas site $j$ is not. We show the probability $P(c_{ij})$ that site $i$ contacts site $j$ (extracted from Hi-C data) normalized by the genomic background probability $P(c_{d})$ that two sites situated a distance $d = |j - i|$ apart contact each other irrespectively of chromatin factors. The horizontal axis represents the distance between $i$ and $j$ divided by the distance between $k$ and $i$, allowing to show in a single plot data from BEAF-bound sites situated at different genomic distances. The line plot is an analytic fit to the data using an ideal chain as a model for the polymer with an energy of interaction $E_{ij}$ between $i$ and $k$ as a fit parameter ($E_{ik} = -0.72 \pm 0.01$).
A significant amount of mutual information would indicate that contact affinities are not statistically independent from chromatin sequence states, so we use the uncertainty coefficient \( U(X|Y) \), Eq. 3.7, presented in the Methods section to provide a reference magnitude to the mutual information. In particular, we calculate

\[
U(a_{ij}|\mu_i\mu_j) = \frac{I(a_{ij};\mu_i\mu_j)}{H(a_{ij})},
\]

(3.18)

which we will refer as the “local chromatin specificity of contacts”, since it indicates how much of the information in contact affinities between site pairs is explained by the chromatin sequence states of those sites. In addition, we also calculate

\[
U(\mu_i\mu_j|a_{ij}) = \frac{I(a_{ij};\mu_i\mu_j)}{H(\mu_i\mu_j)},
\]

(3.19)

which we will refer as the “local contact specificity of chromatin”, since it shows how much of the entropy of the chromatin states of a pair of sites can be explained by contact affinities between those sites.

These quantities were calculated at various distances of contact \( d \) in the genome by estimating the probability \( P(a_{ij}, \mu_i\mu_j) \) from the chromatin affinity sequence and chromatin color state sequence for all pairs \( i \) and \( j \) for which \( |j-i|=d \). Then \( P(a_{ij}, \mu_i\mu_j) \) was used to calculate \( U(a_{ij}|\mu_i\mu_j) \) and \( U(\mu_i\mu_j|a_{ij}) \) as specified in the Methods section of this chapter.

In Fig. 3.5 we see that the local chromatin specificity of contacts \( U(a_{ij}|\mu_i\mu_j) \) is 16% for the immediate neighbouring site pairs (\( d = 10 \) kbp). This quantity suddenly drops to 3% when we increase the distance between sites and then steadily increases with distance of contact until achieving a value of 6% at the longest distance calculated (800 kbp). We therefore find a rise of the influence of local chromatin as the distance between sites increase. The rest of the entropy of \( a_{ij} \) that is not accounted by the local chromatin, \( H(a_{ij}|\mu_i\mu_j) \), could be related to other effects such as the influence of the chromatin in other neighbouring sites and inherent noise in chromatin contacts that is independent of chromatin context.

Fig. 3.5 also shows the local contact specificity of chromatin states \( U(\mu_i\mu_j|a_{ij}) \). We see a similar trend as in the chromatin specificity of contacts but with smaller magnitude; 8% for nearest neighbours, and then an increasing trend that goes from 2% at short distances to 3% at the longest distance sampled. Comparing the two trends we thus find that chromatin states are more informative about chromatin contacts than the other way around. This finding is expected since in our parametrization, the entropy of the joint distribution of chromatin pairs, \( H(\mu_i\mu_j) \) is approximately twice the entropy of contact intensities \( H(a_{ij}) \).

The rest of the entropy in chromatin pairs that cannot be explained by the contact affinity between those sites, \( H(\mu_i\mu_j|a_{ij}) \), may represent information related to generating contacts between other locations, non-structural effects like gene regulation, as well as an inherent noise in chromatin sequence states that is independent of the rest of effects.
Figure 3.5: The local chromatin specificity of contacts as a function of distance (blue) shows what proportion of information about contact affinity $a_{ij}$ can be explained by the local chromatin state in $i$, $\mu_i$, and the local chromatin state in $j$, $\mu_j$, in the genome. The local contact specificity of chromatin sequence as a function of distance (red) shows how much of the entropy of the pairs of chromatin sequence sites in the genome can be explained by the contact affinity between them.
Keeping in mind that our primary goal is to assess the influence on chromatin contacts due to chromatin states that may not be situated in the contacting sites themselves, we now extend the chromatin specificity calculations to non-local sites. At each distance of contact between \(i\) and \(j\), we now calculate the uncertainty coefficient \(U(a_{ij}|\mu_{\Delta i}\mu_{\Delta j})\) between the contact affinity, \(a_{ij}\) and the chromatin sequence states \(\mu_{\Delta i}\), \(\mu_{\Delta j}\) located at a distance \(\Delta i\) and \(\Delta i\) from the sites \(i\) and \(j\), respectively.

\[
U(a_{ij}|\mu_{\Delta i}\mu_{\Delta j}) = \frac{I(a_{ij};\mu_{\Delta i}\mu_{\Delta j})}{H(a_{ij})}.
\]

(3.20)

We fix \(i < j\), so that when \(0 < \Delta i < d\), \(\mu_{\Delta i}\) is for a site that resides between \(i\) and \(j\) (inside the loop). In addition, since \(\Delta j\) is defined with respect to \(j\), when \(-d < \Delta j < 0\), \(\mu_{\Delta j}\) is for a site that resides inside the loop. These relative locations \(\Delta i\) and \(\Delta j\) form the vertical and horizontal axes of a heat map of \(U(a_{ij}|\mu_{\Delta i}\mu_{\Delta j})\) as shown in Fig. 3.6A.

In Fig. 3.6B-I we show the non-local chromatin specificity \(U(a_{ij}|\mu_{\Delta i}\mu_{\Delta j})\) calculated at various distances of contact. Overall we observe two distinct chromatin regions for which the chromatin specificity of contacts is high: the region inside the loop, and the sites of contact. Interestingly, we see that after a distance of contact \(d \approx 300\) kbp the inside of the loop ceases to be an important determinant of contact affinity, while the chromatin sequence at the sites of contact remain as the largest contributors.

In Fig 3.6J, we plot the chromatin specificity of three distinct regions, as a function of separation \(d\) between the contacting sites \(i\) and \(j\). As displayed on the scheme in the right, the three regions are the contacting sites (\(\Delta i = 0, \Delta j = 0\)) (the same sites as in the previously calculated local specificity), the pairs of sites situated at a quarter of the looping distance (\(\Delta i = d/4, \Delta j = -d/4\)) and the pairs of sites situated in the middle of the loop (\(\Delta i = d/2, \Delta j = -d/2\)). We see that for distances of contact smaller than 250 kbp the information from pairs of sites inside the loop contribute more to contact intensity than the chromatin state of the looping sites themselves. This indicates that small loops are formed by the neighbourhood of sites within the loop. As we increase the distance of contact, the chromatin specificity of the sites inside the loop strongly decays, while the chromatin specificity for the looping sites continues to increase, as we have previously noted.

**Detailed composition of chromatin contacts and states**

In the previous section, we identified that the chromatin regions that carry the largest informative weight in determining chromatin contacts are the regions between the contacting sites and the contacting sites themselves. We quantified this effect through the measurement of the chromatin specificity of contacts, \(U(a_{ij}|\mu_{\Delta i}\mu_{\Delta j})\). This quantity is a positive scalar bound between zero and one that gives us a sense of magnitude, but it leaves out much of the detail of how the different states of contact affinity \(a_{ij}\) are associated to particular combinations of chromatin factors \((\mu_{\Delta i}\mu_{\Delta j})\).
Figure 3.6: (A) Schematic of the arrangements of relative chromatin locations in the chromatin specificity heat maps. Vertical axis corresponds to the relative position $\Delta i$ of one chromatin location with respect to the contacting site $i$; horizontal axis corresponds to the relative position of the other location $\Delta j$ with respect to the contacting site $j$. (B-I) Chromatin specificity $U(a_{ij}|\mu_{\Delta i},\mu_{\Delta j})$ of the contacting sites for chromatin regions in their neighbourhood at different separations $d = |j - i|$ in units of 10 kbp. Blue indicates low values and red indicates high. (J) Chromatin specificity as a function of distance of contact for three neighbouring chromatin regions: The contacting sites (blue), pairs of sites at a quarter of the contacting distance $d$ (green) and pairs of sites at half of the contacting distance $d$ (red).
One can gain insight into which combinations of chromatin states are informative for forming contacts by looking at the log-ratio between the measured joint distributions of contact affinities and chromatin states $P(a_{ij}, \mu_{\Delta i} \mu_{\Delta j})$, and the joint distributions for the independent case $P(a_{ij})P(\mu_{\Delta i} \mu_{\Delta j})$. This quantity indicates how a particular combination of affinities and chromatin states are enriched with respect to the null hypothesis that the pair of chromatin states has no role in determining contact affinity. By further weighting these log-ratios by their occurrence in the data, $P(a_{ij}, \mu_{\Delta i} \mu_{\Delta j})$, one can see how much these particular enrichments are present in the data. We thus calculate the following vector quantity that will be referred as “detailed composition”:

$$I_{\Delta i, \Delta j} = \left\{ P(a_{ij}, \mu_{\Delta i} \mu_{\Delta j}) \log_2 \frac{P(a_{ij}, \mu_{\Delta i} \mu_{\Delta j})}{P(a_{ij})P(\mu_{\Delta i} \mu_{\Delta j})} \right\} \quad \forall (a_{ij}, \mu_{\Delta i} \mu_{\Delta j}). \quad (3.21)$$

which has as many entries as combinations of $a_{ij}$ and $\mu_{\Delta i}, \mu_{\Delta j}$. Note that the sum over all the vector elements is equal to the mutual information $I(a_{ij}; \mu_i \mu_j)$, and thus the notation choice of $I_{\Delta i, \Delta j}$. This vector however is not an entropy; its elements, which can be either positive (denoting enrichment) or negative (depletion) should be seen as simply weighted log-ratios of probabilities. At a fixed distance of contact $d = |j - i|$, one can measure the detailed composition vector for each combination of relative positions $\Delta i$ and $\Delta j$ by measuring the joint probability distributions $P(a_{ij}, \mu_{\Delta i} \mu_{\Delta j})$ similarly to what it was done in the previous section. We thus obtain a series of $I_{\Delta i, \Delta j}$ at each distance of contact.

Since we now we have vectors rather than scalar quantities, we cannot visualize them in a single heat map for each distance of contact as we previously did with chromatin specificity. Thus we use Principal Component Analysis (PCA), which was introduced in the previous chapter, to reduce the dimensionality of the data by finding common patterns of variation in these vectors. In particular, $I_{\Delta i, \Delta j}$ has a length of 45 (15 unique pairs of chromatin color states and 3 contact affinity states), and since we measure 50 $\Delta i$ locations and 50 $\Delta j$ locations, we have 2500 detailed composition vectors $I_{\Delta i, \Delta j}$ at each distance of contact. PCA is performed by diagonalizing the covariance matrix of the entries of $I_{\Delta i, \Delta j}$, thus obtaining a new basis of independent variance (eigenvectors, or PCs) to express the data and the corresponding eigenvalues, which represents how much variation of the data each PC captures when the original data is projected into the new basis (more details in Chapter 2). We observed that in this data, the first 3 PCs at each distance of contact capture $\sim 90\%$ of the variation. We can thus ignore the remaining PCs and effectively reduce the dimensionality of our detailed composition data from 45 to just 3 while keeping most of the variance in the signal.

By inspecting the values of a PC, which have the same dimensionality as the original vectors $I_{\Delta i, \Delta j}$, we find how much each entry of the original vector contributes to the pattern captured by the PC, allowing us to interpret what type of variation a PC represents, as we describe later. Fig. 3.7 shows as color-coded dot plots the weights of each PC; color indicates
Figure 3.7: Principal Component Analysis (PCA) of detailed composition $\tilde{I}_{\Delta i, \Delta j}$ for four distances of contact (in units of 10 kbp). The first three PCs of each distance of contact are displayed. Scatter plots on the left of the panels show the elements of the PC (unit vector) grouped by the pairs of chromatin states they represent. The color of the dots indicates their corresponding contact intensity (see legend at the bottom). Arrows in the vertical axis connect the PCs to their projections displayed as a heat map on the right of each panel (blue indicates negative values and red positive values). Positive PC entries (red arrow) are features enriched in locations with positive projections (red regions in heat maps). Negative PC entries (blue arrow) are features enriched in the locations with negative projections (blue locations in heat maps).
a particular contact affinity state $a_{ij}$ and the horizontal axis shows all of the possible combinations of pairs of chromatin sequence states $\mu_{\Delta i}, \mu_{\Delta j}$. By projecting $\vec{I}_{\Delta i, \Delta j}$ onto the PC vectors, we find how much the detailed composition of each pair of locations $\Delta i, \Delta j$ is contributed to by the pattern that a PC describes. That allows us to show the spatial distribution of every PC in a heat map (Fig. 3.7), similarly to our previous visualization of the scalar quantity $U(a_{ij}|\mu_{\Delta i}, \mu_{\Delta j})$. On the projection heat maps, regions with positive values (red) are regions enriched for the feature described by the PC whereas regions with negative values (blue) are patterns depleted for the feature described by the PC (or, alternatively, negative values indicate regions enriched for a sign-inverted version of the PC).

In Fig. 3.7, we find that the first three PCs calculated at different distances of contact show similar patterns of chromatin factors and contact affinities (dot plots). In addition, the PC projections (heat maps) at different distances show consistent spatial patterns: PC1 describes a variation between what is in between contacting sites and what is outside of them, PC2 describes a variation occurring mostly at the midpoint of the loop and PC3 describes a variation that distinguishes the sites of contact from the rest. This thus allows us to describe each PC in a general manner, regardless of distance of contact.

PC1 shows a consistent pattern where high contact intensities $a_{ij}$ (red dots) involve homogeneous pairs of heterochromatin with silent chromatin in the interior of the loop (B-B, B-K, K-K) whereas low contact intensities $a_{ij}$ (blue dots) involve having inactive chromatin located at the exterior of the loop. In addition, active chromatin shows the complementary pattern, where both homogeneous euchromatin pairs (R-R, R-Y, Y-Y) and heterogeneous euchromatin/heterochromatin pairs (B-R, B-Y, K-R, K-Y) tend to be situated outside the loop when contact intensity is high, and inside the loop when contact intensity is low. Interestingly, this trend is stronger for heterogeneous pairs that involve Yellow chromatin than for those with Red chromatin. This observed similarity between heterogeneous and homogeneous euchromatin pairs indicates that the presence of only one euchromatic site in the neighbourhood seems capable of altering contact intensity in the loop.

PC2 shows a pattern of variation that does not separate high $a_{ij}$ from low $a_{ij}$ (both have the same sign on the PC). It instead tends to distinguish homogeneous pairs of factors (G-G, B-B, K-K, R-R, Y-Y) from the rest, regardless of contact intensity. We also see that the location where this PC has the strongest projection is the midpoint of the loop, which is the location where the two relative positions $\Delta i$ and $\Delta j$ are the closest in the genome. As we previously showed in this chapter, chromatin factors have a certain degree of autocorrelation, which implies that nearby genomic locations are enriched for homogeneous pairs of factors, and depleted for heterogeneous pairs. We therefore interpret PC2 as the carrier of this background feature, which is non-informative in terms of how chromatin factors affect contact intensities, and will be ignored henceforth.

PC3 is related to a pattern of variation that differentiates what is bound around the contacting sites compared to all other sites. We find that high contact intensity is related
to the presence of euchromatic pairs around the contacting sites (Y-Y, R-Y, R-R), as well as silent chromatin (K-K) and a depletion of them in the other regions. In contrast, a heterogeneous combination of heterochromatin/silent chromatin (B-K) is strongly depleted in the contacting region when contacts are high. This is consistent with the fact that blue and black chromatin represent two different mechanisms of chromatin inactivation that do not interact with each other. In addition, for the longer distances of contact we see that features enriched in the loop tend to also be enriched at certain relative positions inside of it (FIG. 3.7(C,D)). This observation could indicate that the formation of larger loops can be aided by smaller loops nested inside of them.

This analysis is thus able to separate the influence of the two previously found informative chromatin regions. In summary, the presence of silent chromatin in between the contacting sites is related to high contact affinity, and the presence of either euchromatic factors, or pairs of silent chromatin in the contacting sites is also related to high contact affinity.

**Chromatin specificity of multiple neighbours**

In the previous sections, we inspected the relation between chromatin contact affinity $a_{ij}$ and the chromatin state of pairs of sites in the genomic neighbourhood $\mu_{\Delta_i} \mu_{\Delta_j}$. In particular, the chromatin specificity $U(a_{ij} | \mu_{\Delta_i} \mu_{\Delta_j})$ tells us how much the state of the chromatin at the pair of sites $\Delta_i$ and $\Delta_j$ can explain contact affinity between $i$ and $j$. Interestingly, the quantity $1 - U(a_{ij} | \mu_{\Delta_i} \mu_{\Delta_j}) = H(a_{ij} | \mu_{\Delta_i} \mu_{\Delta_j}) / H(a_{ij})$ also contains useful information, as it represents what fraction of entropy of contact is not explained by the chromatin state at that pair of locations. The rest of the entropy in contacts could thus be explained by other chromatin locations as well as an inherent noise in chromatin contacts $\xi(a_{ij})$ that is independent of chromatin sequence.

In this section, we extend our chromatin specificity calculation to take into account a neighbourhood $N$ of chromatin sites instead of just a pair of sites as done above. We thus define the chromatin vector $\vec{\mu}$ of size $N$ containing the chromatin state of $N$ neighbouring locations. This requires characterizing the joint probability distribution $P(a_{ij}, \vec{\mu})$, which would contain $3 \times 5^N$ entries if we followed our previous parametrization of three affinity states $a_{ij}$ and five states for $\mu_k$. In this case, the number of possible states would surpass the number of experimental samples (size of the genome) for a neighbourhood of just $N = 5$.

In order to reduce the complexity, we thus consider using a binary categorization for contact affinity $a_{ij} = \{\text{low}, \text{high}\}$ by running an HMM with two hidden states, and a binary classification of the chromatin sequence by grouping the original five states into two states based on their functionality $\mu_k = \{H = K \cup B, E = G \cup R \cup Y\}$, where $H$ indicates heterochromatin and $E$ indicates euchromatin. Taking advantage of the fact that both the contact affinities and the chromatin sequence states are now binary, we group them using the same notation, calling all binary variables “spins” $\sigma_k = \{-1, 1\}$ in analogy to...
the physics of ferromagnets. The size of the joint distribution $P(a_{ij}, \vec{\mu}) = P(\vec{\sigma})$ under this binary parametrization is thus reduced from $3 \times 5^N$ to $2^{N+1}$ entries.

Although we considerably reduced the size of the system by introducing binary variables, the number of experimental samples continues to be too small to characterize the occurrence of all possible $2^{N+1}$ states, even for modest $N$. We thus resort to approximating $P(\vec{\sigma})$ with a maximum-entropy distribution $\tilde{P}(\vec{\sigma})$ that is constrained by the experimental one-spin average $\langle \sigma_k \rangle$ and two-spin average $\langle \sigma_k \sigma_l \rangle$ which we can reliably measure from the data. As shown in the Methods section, the maximum-entropy distribution of $\tilde{P}(\vec{\sigma})$ consistent with such spin-statistics is the Ising model,

$$\tilde{P}(a_{ij}, \vec{\mu}) = \tilde{P}(\vec{\sigma}) = \frac{1}{Z} \exp \left( \sum_k h_k \sigma_k + \sum_{l>k} J_{kl} \sigma_k \sigma_l \right),$$  \hspace{1cm} (3.22)$$

where $(h_k, J_{kl}, Z)$ are parameters that can be fit to reproduce the experimental averages [200].

In order to check whether the binary-state maximum-entropy distribution is an acceptable approximation for the distribution of contact affinities and chromatin states, we repeated the calculations previously done for just two neighbours $(\Delta i, \Delta j)$ shown in Fig. 3.8A, except now using just a binary classification of the data where we estimated $P(\vec{\sigma})$ directly from the data (see Fig. 3.8B) as well as by fitting the maximum-entropy distribution $\tilde{P}(\vec{\sigma})$ given by Eq. 3.22 (see Fig. 3.9). Reassuringly, we find that the order of magnitude of the chromatin specificity $U(a_{ij}|\mu_\Delta i \mu_\Delta j)$ calculated with the three different methods is very similar. We also observe the same qualitative trend where the chromatin specificity for the chromatin factors situated in the contacting sites increases with distance, whereas the chromatin specificity for what is inside the loop decays with distance. Some differences are that for the two distributions calculated using just a binary classification, we measure a smaller chromatin specificity for the contacting sites at the shorter distances, which displaces the crossover of the two trends (inner loop and contacting sites) from $d = 250$ kbp (FIG. 3.8A) to $d = 350$ kbp (FIG. 3.8A,B). We also see that the Ising model in Fig. 3.9 displays more chromatin specificity for what is bound outside the loop than the previous calculation based on direct sampling (FIG. 3.8B, $d = 41 - 71$). Nevertheless, we find similar qualitative behaviour in all three approaches, which allow us to take advantage of characterizing our data using a maximum-entropy distribution that is constrained to match a few measures rather than a comprehensive sampling of all possible states. This gives us the chance to explore how larger neighbourhoods (beyond just two sites) inform the connection between sequence and contacts.

We now define the chromatin neighbourhood vector, $\vec{\mu^w}$, to be the sequence of chromatin states that includes the contacting sites $i$ and $j$ together with positions that flank them on
Figure 3.8: Chromatin specificity of contacts for (A) three contact affinities and five chromatin states, (B) two contact affinities and two chromatin states. Top heat maps show an example of the joint probability distribution of contact affinities and chromatin states. Central panels show chromatin specificities of neighbouring sites for different distances of contact (units of 10 kbp). Blue denotes low values and red denotes high. Bottom line plots display the chromatin specificity vs distance for three selected relative neighbours (red: loop midpoint, green: a quarter of the loop size, blue: contacting sites).
maximum-entropy distribution

\[ P(a_{ij}, \mu_{\Delta_i} \mu_{\Delta_j}) \]


Figure 3.9: Chromatin specificity of contacts using the maximum-entropy distribution \( P(a_{ij}, \mu_{\Delta_i} \mu_{\Delta_j}) \) constrained to reproduce the one-spin and two-spin statistics of a binary classification of contact affinities and a binary classification of chromatin states. Top heat map show an example of the joint probability distribution of contact affinities and chromatin states. Central panels show chromatin specificities of neighbouring sites for different distances of contact (units of 10 kbp). Blue denotes low values and red denotes high. Bottom line plots display the chromatin specificity vs distance for three selected relative neighbours (red: loop midpoint, green: a quarter of the loop size, blue: contacting sites).
either side out to a distance $w$ (Fig. 3.10A). For a given $w$, the spin vector, $\vec{\sigma}$ now becomes
\[
\vec{\sigma} = \{a_{ij}, \vec{\mu}_w\} = \{a_{ij}, \mu_{i-w}, \cdots, \mu_{i-1}, \mu_i, \mu_{i+1}, \cdots, \mu_{i+w}, \mu_{j-w}, \cdots, \mu_{j-1}, \mu_j, \mu_{j+1}, \cdots, \mu_{j+w}\},
\]
with size $3 + 4 \times w$ ($a_{ij}$, $i$, $j$, and flanking neighbours). Note that for $w = 0$, $\vec{\sigma}$ becomes that of the two-site local model.

We now calculate the maximum-entropy distribution $\tilde{P}(\vec{\sigma})$ using these larger neighbourhoods by fitting the Ising model (Eq. 3.22) to the one- and two-spin statistics and then use it to calculate $U(a_{ij}|\vec{\mu}_w)$. The time and resources needed to fit an Ising model are proportional to the number of possible states in $\vec{\sigma} (2^{3+4\times w})$. This thus limits our capability of extending the length of $\vec{\sigma}$ indefinitely. In particular, our computational capacity allows us to fit an Ising model up to a neighbourhood of $w = 5$ flanking sites by enumerating all the possible states $\vec{\sigma}$ in the system. Beyond $w = 5$ one would need to resort to Monte Carlo methods [199], as the number of states grows as $16^w$.

In Fig. 3.10B we show $U(a_{ij}|\vec{\mu}_w)$ as a function of $w$ and $d$. As expected, we see that the larger the neighbourhood size $w$ is, the larger the chromatin specificity as more information is introduced into the system. In particular, we see that short-range contacts ($\sim 80 – 200$ kbp) and long-range contacts ($\sim 500 – 780$ kbp) are the most chromatin-sequence specific given our choice of neighbourhood vectors $\vec{\mu}_w$. Interestingly, $U(a_{ij}|\vec{\mu}_w)$ for short-distance contacts continues to grow with increasing $w$; for middle-distance contacts, $U(a_{ij}|\vec{\mu}_w)$ grows at a slow steady rate with $w$; for long-distance contacts, $U(a_{ij}|\vec{\mu}_w)$ seems to have saturated with increasing $w$.

The small value and slow increase of $U(a_{ij}|\vec{\mu}_w)$ for intermediate distances ($\sim 200 – 500$ kbp) can be explained by looking back at Fig. 3.6E and noticing that intermediate distances still show chromatin specificity at the loop midpoint, which is not captured yet by $w = 5$. It therefore remains unclear at this moment how much it would increase at the large $w$ limit when all informative neighbours are included in the calculation.

For short-range contacts, we see that $U(a_{ij}|\vec{\mu}_w)$ continues to grow with $w$, whereas $U(a_{ij}|\vec{\mu}_w)$ for long-range contacts saturates at a value of 12%. Thus, we conclude that short-range contacts are most dependent on the surrounding sequence when non-local chromatin is taken into account. Interestingly, for longer-range contacts which are governed by the chromatin sequence immediately at the sites of contact, when we increase from $w = 0$ (a strictly local neighbourhood) to $w = 5$ (a larger neighbour window centered around contacting sites), chromatin specificity doubles. Therefore we see that although long-distance contacts are governed by the sequence at the contacting sites, they also depend on the chromatin in a small genomic window around the sites of contact.

By increasing the size of the neighbourhood window $w$ until $U(a_{ij}|\vec{\mu}_w)$ saturates at all distances, one could account for most of the entropy in chromatin contacts that can be explained by the surrounding chromatin states. In that case, one would be measuring the
Figure 3.10: (A) Schematic of the arrangement of chromatin neighbourhoods for $w$ flanking neighbours. (A) Chromatin specificity $U(a_{ij}|\vec{\mu}^w)$ for the maximum-entropy distribution of binary contact intensities $a_{ij}$ and binary-chromatin neighbourhoods $\vec{\mu}^w$. Central heat map simultaneously shows $U(a_{ij}|\vec{\mu}^w)$ at each distance of contact $d = |j - i|$ and each $w$ used. Blue indicates low values and red high. Top and right panels display the same data, but show a line for each $w$ or for each $d$, respectively.
inherent noise of chromatin contacts $\xi(a_{ij}) \approx 1 - U(a_{ij} | \vec{w})$ when other effects such as inter-chromosomal, distant intra-chromosomal, and nuclear envelope interactions are ignored. We leave for future work the measurement of this quantity by expanding neighbouring windows $w$ with Monte Carlo sampling methods [199].

### 3.3 Conclusion

This chapter summarized a number of statistical analyses done to quantify the importance of nearby chromatin states in influencing the likelihood of two sites being in contact. We used a series of information-theoretic calculations to pinpoint which regions of the neighbouring chromatin are the most informative about DNA contacts. The results presented in this chapter serve as a basis for the modelling presented in the next chapters of this thesis, where we construct statistical models for chromatin conformation that can predict contact maps from sequence neighbourhoods.

Our results show that chromatin contacts are influenced by the underlying chromatin sequence differently depending on their distance of contact. At the shorter distances $d < 250$ kbp the contact affinity is mainly driven by the factors situated inside the loop, as opposed to the factors situated at the contacting sites or outside the loop. As we increase the distance of contact, factors inside the loop cease to be informative about contact affinity while the factors situated at the contacting sites strongly dominate the levels of contact affinity.

By analyzing the individual contributions of pairs of chromatin factors to their mutual information with DNA contacts, we found that the elements inside the loop that are associated with high contact affinity are silent/heterochromatin factors while active/euchromatin factors tend to be situated in the contacting sites as well as outside the loop. In addition, we also found that euchromatin at the contacting sites increases contact affinity, as well as the presence of silent chromatin whenever these factors are absent in the rest of regions.

Apart from highlighting mechanisms of genome conformation, our findings may have functional consequences for gene expression, since the communication between enhancers and promoters at a distance smaller than $\sim 300$ kbp may be strongly regulated by the epigenetic state of the chromatin in between. One may imagine exploring further the functional effects of chromatin conformation by extending the type of spatial analysis presented in this chapter to include gene-expression and genome annotation data as additional sources of entropy that could influence or be influenced by chromatin conformation.
Chapter 4

A maximum-entropy model for predicting chromatin contacts

In higher-complexity organisms, the packaged molecule of DNA inside the nucleus of a cell consists not only of the DNA polymer but also a large number of DNA-associated protein and RNA complexes that together form what is known as chromatin. These complexes bind to the DNA by either sequence affinity or by interactions with other bound factors [33], and their presence is known to be correlated with the 3D conformation of the DNA polymer [65]. They impact the structure of the DNA on many length scales: from the packing of nucleosomes at the smallest scale (∼200 basepair) to the stabilization of loops and genomic domains on much larger scales (∼10–100 kilobasepair). As a result, they regulate a host of important cell functions from DNA replication to gene expression [202,203]. Developing models that can predict how chromatin folds given only the locations of bound factors is of key importance to better understand how they regulate such processes by shaping DNA structure.

The average structure of DNA over a population of cells can be measured genome-wide using high-throughput DNA-sequencing methods such as Hi-C [40]. Briefly, contacting sites in the genome are cross linked; then the DNA is fragmented and the contacting pairs are sequenced. These sequenced pairs are then used to construct a contact map at a given spatial resolution that gives the number of times any pair of sites along the genome were found to be in contact. Analysis of the contact maps have shown that chromatin can be classified into structural types such as A/B compartments [40,204] or topologically associated domains (TADs) [37] just from the spatial distribution of contacts.

These structural features are known to be strongly correlated with different types of DNA-bound factors. The locations of these bound factors along the genome constitute a form of sequence that helps drive the folding of the DNA, similar to the specific sequence

---

1The work presented in this chapter was published in *PLOS Computational Biology* in 2018 [201]. The text of that paper has been edited to conform to the style of this thesis.
of amino acids that drives a protein to fold. High-throughput methods can provide the binding locations of such factors on a genome-wide scale [90]. Interestingly, despite tens to hundreds of different chromatin-associated factors, clustering of their binding locations shows that there are only few unique bound states [139–142,171] (similar to the grouping of amino acids into just hydrophobic and polar types). With the richness of this structural and sequence data for chromatin, predictive models that aim to solve the chromatin folding problem — namely predicting the structure of DNA inside a cell given only the locations of the DNA-bound chromatin factors — are now being developed.

Recent modeling efforts have used polymer-based models whose parameters can be tuned to reproduce experimental observations, such as the contact map from Hi-C. One set of approaches tries to find the best 3D polymer structure that is consistent with the constraints imposed by the Hi-C contact map [56,58–60,205]. Other methods include bound factors by adding sequence-specific interactions to a given polymer model for the DNA [50,61,148–160]. These approaches have been successful in showing how interactions between factors together with topological constraints may be responsible for the observed chromatin structures. Challenges involve continuing to improve the physics of the interacting polymer model using data-driven methods and the time-consuming process of carrying out the polymer simulations.

Complementing the polymer simulation methods are purely statistical approaches that aim to predict a contact map instead of a full 3D structure. Sexton et al. [35] developed a statistical model based on site-specific scaling factors that could predict a matrix of expected counts. Other work, presented in Chapter 2 and in [161], included sequence information by fitting a pairwise interaction model for the DNA-bound factors that could then predict the probability of contact between pairs of sites given only the sequence at those locations. However, as we highlighted in Chapter 3, the pairwise interaction model did not include an important effect that is present in polymer simulations, namely the role of neighbours.

A growing body of experimental evidence supports the importance of the local sequence neighbourhood of bound factors in mediating contacts between pairs of genomic sites [206]. In particular the probability of contact between two sites $i$ and $j$ at a certain genomic distance apart is altered if some site $k$ in the neighbourhood of $j$ is attracted to $i$. Sites $k$ and $i$ will spend a fraction of the time in contact, thus altering the effective polymer distance between $i$ and $j$ (see Fig. 3.4 in the previous chapter). Here, our aim is to take such effects into account and predict the probability of two sites being in contact (i.e. the contact map) given a local sequence neighbourhood. The results presented in the following sections demonstrate how these probabilities can be modelled by a maximum-entropy distribution. In addition, by formulating the problem in a Bayesian fashion, we are not only able predict probabilities of contact from an experimentally measured sequence of bound factors, but also predict the probability of a site having a particular bound factor using only the measured local Hi-C contact map.
4.1 Results

4.1.1 Maximum-entropy model

Fig. 4.1 shows a schematic of our model and resulting calculation. First, the genome is discretized into non-overlapping sites of fixed size. A particular chromatin state \( \sigma_k \) is assigned to each site \( k \) based on the bound chromatin factors there. As mentioned in the introduction, only a few bound states exist and for simplicity we classify each site into only one of two possible sequence states corresponding to active/euchromatic and inactive/heterochromatic DNA, respectively labeled as spin-up \( (\sigma_k = 1) \) and spin-down \( (\sigma_k = -1) \) analogous to the physics of ferromagnets (see Fig. 4.1A and the Methods section for details).

\[ d = |j - i| \]

Figure 4.1: Schematic of model for predicting chromatin contacts from sequence. (A) The DNA was discretized into sites of finite size and the bound-factor sequence at each site was categorized into one of two bound states (spin-up or spin-down). The probability of contact, \( P(c|\vec{\sigma},d) \), between two sites separated by a genomic distance, \( d = |j - i| \), depends on the sequence neighbourhood \( \vec{\sigma} \) situated around them (shown as gray line). (B) Hi-C data measures the number of times pairs of sites were found in contact (shown as heat map above). For each pair of sites \( i \) and \( j \) at a given genomic distance \( d = |j - i| \), one can extract a local sequence neighbourhood \( \vec{\sigma} \) around the sites. (C) Schematic of data used for fitting the maximum-entropy model. The collection of all sequence neighbourhoods at a given genomic distance \( d \) was extracted by scrolling along the genome. The statistical weights associated to each neighbourhood were given by either the number of times they were observed in contact from Hi-C (left column), or occurrence in the genome (right column). These weights were implemented in the calculation of the statistical averages used as constraints on the maximum-entropy distributions, \( P(\vec{\sigma}|c,d) \) and \( P(\vec{\sigma}|d) \), respectively.

With respect to structure, Fig. 4.1(A,B) shows two sites \( i \) and \( j \) separated by a genomic distance, \( d = |j - i| \), that have a certain chance of forming a contact that we assume is
determined by the sequence neighbourhood, $\vec{\sigma}$, situated around them. Hi-C data gives an estimate of this chance by reporting the number of times $n_{ij}$ that this pair formed contacts, and hence the likelihood that their sequence neighbourhood will form a contact (Fig. 4.1B). We denote this probability of contact as $P(c|\vec{\sigma},d)$ which depends on the genomic distance $d$ between sites and sequence neighbourhood $\vec{\sigma}$ (here $c$ indicates that the two sites are in contact; $c$ is therefore one of the values of the random variable $z$: $z = c$ if contact and $z = \bar{c}$ otherwise). Inspired by our findings shown in Chapter 3, where we identified the regions of the local chromatin sequence which influence probabilities of contact the most, we built the neighbourhood $\vec{\sigma}$ as a vector containing $N$ sites from the union of sequence windows situated around the contacting sites $i$ and $j$ and thus depends on $d$ (see Fig. 4.1(B,C), Methods and Fig. 4.6 for further details).

Using Bayes’ theorem we can rewrite $P(c|\vec{\sigma},d)$ as

$$P(c|\vec{\sigma},d) = \frac{P(\vec{\sigma}|c,d)P(c|d)}{P(\vec{\sigma}|d)},$$

(4.1)

where $P(\vec{\sigma}|c,d)$ and $P(\vec{\sigma}|d)$ are the probability of observing the given sequence neighbourhood $\vec{\sigma}$ given that it has formed a contact and regardless of contact, respectively. $P(c|d)$ is the sequence-independent probability of contact at a separation $d$ and can be estimated directly from the Hi-C experiments. In principle, Hi-C data could also be used to directly estimate $P(\vec{\sigma}|c,d)$ since the contact map gives the number of times that a given neighbourhood was found to form a contact (see Fig. 4.1C). Similarly, $P(\vec{\sigma}|d)$ could be estimated from the frequency of occurrence of that sequence neighbourhood in the genome. However, because the number of different sequence neighbourhoods goes as $2^N$, even for modest $N$ there is not enough data to make accurate estimates.

To overcome this we utilize the principle of maximum entropy, previously introduced in Chapter 3, to calculate probability distributions, $P(\vec{\sigma}|c,d)$ and $P(\vec{\sigma}|d)$, that reproduce a few reliably estimated statistics of $\vec{\sigma}$ extracted from experimental Hi-C data [193,194]. The statistics we use to constrain both of the above distributions are the average spin $\langle \sigma_k \rangle$ at a given position $k$ in the neighbourhood (which represents the average chromatin state at that neighbourhood position) and the average spin-spin correlation $\langle \sigma_k \sigma_l \rangle$ between sites $k$ and $l$ in the neighbourhood (which represents the average correlation between the chromatin states at the two neighbourhood positions). The averages denoted by $\langle \cdot \rangle$ are calculated over their respective ensembles of sequence neighbourhoods $\{\vec{\sigma}\}_d$ that surround pairs of sites at a distance of contact $d = |j - i|$. The ensembles for both distributions can be extracted by scrolling along the sequence of binary states in the genome. When characterizing $P(\vec{\sigma}|c,d)$, each neighbourhood has a weight equal to the number of contacts $n_{ij}$ observed between sites $i$ and $j$; when characterizing $P(\vec{\sigma}|d)$, each neighbourhood has a unit weight for every time it appears in the genome (as shown in Fig. 4.1 and in Methods section).
With the above statistics as constraints, the maximum-entropy distribution can be found using the method of Lagrange multipliers (see Chapter 3) and has the form of a Boltzman distribution for the Ising model at $k_B T = 1$,

$$P(\vec{\sigma} | \cdot) = \frac{e^{\sum_k h_k \sigma_k + \sum_{l>k} J_{kl} \sigma_l \sigma_k}}{Z(\cdot)},$$

(4.2)

where $h_k$ and $J_{kl}$ are Lagrange multipliers that constitute the fitting parameters of the model. The partition function $Z(\cdot)$ is a normalization constant obtained by summing the numerator of Eq. 4.2 over all possible $\vec{\sigma}$ neighbourhoods, $Z(\cdot) = \sum_\sigma \exp (\sum_k h_k \sigma_k + \sum_{l>k} J_{kl} \sigma_l \sigma_k)$. In the above, we use “|” to summarize in a single notation two different conditions at each distance of contact, namely $P(\vec{\sigma}|c, d)$ and $P(\vec{\sigma}|d)$ (“|” substitutes “c, d” or “d”). For neighbourhoods of size up to $N \sim 23$, exact enumeration can be used to evaluate Eq. 4.2 in minimal time. Beyond that, one would need to estimate this distribution by Monte Carlo sampling [199]. We have thus restricted our neighbourhoods to a maximum size $N = 20$ so that we can conveniently calculate Eq. 4.2 by enumerating the $2^N$ possible neighbourhoods at a given $d$ (see Methods section).

At a given genomic distance $d$, the distribution in Eq. 4.2 can be fit to reproduce the experimental statistics $\langle \sigma_k \rangle$ and $\langle \sigma_k \sigma_l \rangle$ calculated from the sequence ensemble at that distance (see Methods). We now detail how we applied this method to estimate these distributions and ultimately the contact probability, Eq. 4.1, from real experimental sequence and structural data.

### 4.1.2 Model parameters

As a test of the method, we used the measured structure and sequence data for *Drosophila Melanogaster*. For structure, we used a Hi-C contact map from *Drosophila* embryos, generated using sites that were 10 kbp in size [189]. For sequence, we used our own binary classification (spin-up or spin-down) of the measured *Drosophila* chromatin-associated factors at each 10 kbp site, which was based on the “chromatin colors” classification [141] (see Methods section). To avoid overfitting the model, we divided the data from *Drosophila* chromosomes 2 and 3 into a training set used for fitting and a test set for prediction (see Methods).

For genomic distances ranging from $d=10$ kbp to 800 kbp in 10 kbp steps, we used the above maximum-entropy method to estimate the two conditional probability distributions, $P(\vec{\sigma}|c, d)$ and $P(\vec{\sigma}|d)$. Fig. 4.2(A-D) shows that the fitted distributions successfully predicted the experimental statistics $\langle \sigma_k \rangle$ and $\langle \sigma_k \sigma_l \rangle$ on the test data. They also captured three-point correlations $\langle \sigma_k \sigma_l \sigma_m \rangle$ (see Fig. 4.2(E,F) ) despite not having incorporated them into the fit. In Fig. 4.2(G,H) we also see that the predicted sequence neighbourhood probabilities $P(\vec{\sigma}|\cdot)$ from the model agreed with their frequencies as seen in the data. Thus these second-order maximum-entropy distributions seemed to be adequate approximations to the
true distributions (first-order maximum-entropy distributions were also tested and failed to reproduce experimental statistics, see Methods and Fig. 4.7).

**Figure 4.2**: Spin statistics of model versus experiment at three different contact distances, $d$. Top row is for models conditioned on contact, $P(\vec{\sigma}|c,d)$, whereas the bottom row is for models that are regardless of contact, $P(\vec{\sigma}|d)$. (A, B) Site average statistics, $\langle \sigma_k \rangle$. (C, D) Pairwise correlation statistics, $\langle \sigma_l \sigma_k \rangle$. (E, F) Three body correlation statistics, $\langle \sigma_k \sigma_l \sigma_m \rangle$. (G, H) The probability of $\vec{\sigma}$ from the model versus the observed frequency from contact maps (G) and from the genome (H). For reference the insert compares experimentally measured spin-vector probabilities from training versus testing data.

Inspection of the associated fit parameters $h_k$ and $J_{lk}$ as a function of genomic distance, $d$, showed that they naturally clustered into two groups (through both K-means and Principal Component Analysis), with a transition from one to the other occurring at $d \approx 390$ kbp (see Methods section and Fig. 4.8). Averaging the parameters within each group together we found that $\langle h_k \rangle$ of the two sites in contact ($i$ and $j$) were positive, especially for the group above transition. Positive values for $h_k$ favour the spin-up active/euchromatic state and thus, the above finding shows that contacting sites had a tendency to be in such state. Interestingly, the sites in between the contacting sites had an inactive/heterochromatic preference (i.e. negative values for $\langle h_k \rangle$) in the group below the transition, and no chromatin preference in the group above. Therefore at short genomic separations, a heterochromatic region between contacting sites favored contact. We also found that the interaction terms $\langle J_{lk} \rangle$ for the neighbours between contacting sites were ferromagnetic for the group below the transition (i.e. $\langle J_{lk} \rangle > 0$ which favors sites $l$ and $k$ to be in the same state) and were antiferromagnetic for the group above (i.e $J_{lk} < 0$ which favours $l$ and $k$ to be in opposite states). Thus at short genomic distances, the sequence between the contacting sites was...
favored to be homogeneous, whereas at longer distances heterogeneity or a more random sequence seemed to be the case (see Methods and Fig. 4.8).

4.1.3 Contact map prediction

Figure 4.3: Contact prediction in test set. Diagonal elements of contact maps (shown as heat maps) from the 5 Mbp test regions situated at the ends of chromosomes 2 and 3 of *Drosophila melanogaster*. The maps span genomic distances from 10 kbp to 800 kbp for sites that are 10 kbp in size. (The midline of each contact map corresponds to $i = j$). Predicted contact maps are shown as red heat maps, the experimentally measured maps from Schuettengruber et al [189] Hi-C data are shown as blue heat maps. Both the experimental and predicted maps were distance normalized. The column on the right shows the predicted versus experimental distance-normalized contact likelihood for each pair of sites in the test data as a scatter plot.

We used the fitted maximum-entropy distributions, $P(\tilde{\sigma}|c, d)$ and $P(\tilde{\sigma}|d)$ at each distance, $d$, along with Eq. (4.1) to make predictions for distance-normalized contact maps $P(c|\tilde{\sigma}, d)/P(c|d)$ on the test data from *Drosophila* chromosomes 2 and 3 (red heat maps in Fig. 4.3). Each $(i, j)$ pair had an associated $\tilde{\sigma}$ from which we could then evaluate Eq. 4.1 using the fitted distributions. (We normalized by $P(c|d)$ so as to remove the strong decay with distance of the probability of making a contact). These were then correlated to the normalized experimental Hi-C contact maps $n_{ij}/\langle n(d) \rangle$ of the test data (blue heat maps in Fig. 4.3), where $\langle n(d) \rangle$ is the average number of Hi-C contact counts at a given distance.
that also decays strongly with distance. The correlations with the *Drosophila* chromosomes were 0.39 (chromosome 2L), 0.53 (chromosome 2R), 0.53 (chromosome 3L), and 0.40 (chromosome 3R). These correlations are remarkable given that only a binary model for the sequence was used. The areas of discrepancy in Fig. 4.3 may be due to the fact that the binding factors were measured from different cell lines than the Hi-C ones, so it is possible that in those regions the actual underlying sequence that generated the observed contact counts may be different than what was used in making the prediction.

### 4.1.4 Structural changes due to sequence mutation

The fitted maximum-entropy model that can predict contact maps from chromatin sequence provides an opportunity to also predict structural changes that might arise due to mutations in the underlying sequence of bound factors. We thus set to identify which genomic locations are expected to disrupt the local structure the most in the event that their chromatin state is flipped. This analysis was performed in genomic regions where the locally predicted contact probabilities agreed the most with experimental measurements (correlation $C > 0.6$ between predicted and experimental distance-normalized probabilities of contact). The chromatin state, $\sigma_k$, of each of these well-predicted locations was individually inverted and the correlation, $C'$, between the newly predicted probabilities and the experimental ones was calculated (see Methods section for details). We interpret the change in correlation $\Delta C = C' - C$ to be a measure of structural-sensitivity to chromatin sequence mutations for that position.

In Fig. 4.4, we show histograms of $\Delta C$ for genomic locations categorized by genomic feature. Regardless of genomic feature, the vast majority of sites had a negative $\Delta C$ after their chromatin state was inverted, indicating that the local predicted structure tended to depart from the experimental structure when the sequence state was mutated. In Fig. 4.4A, we found that the inactive/heterochromatic (spin down) sites were significantly more structurally sensitive than the active/euchromatin (spin up) sites (average $\Delta C$ was -0.11 for spin-down whereas -0.04 for spin-up sites). In other words, a change from inactive to active DNA tended to produce a greater structural change than the opposite. In Fig. 4.4B, we classified genomic sites into three non-overlapping categories: sites where no genes were present, sites that contained gene promoters, and “gene body” sites corresponding to locations occupied by genes but with no promoters. We found that sites with no genes were the most structurally sensitive, followed by “gene body” sites and lastly, the least structurally sensitive were sites containing promoters (average $\Delta C$ was -0.15 for no genes, -0.10 for “gene body” and -0.06 for promoter sites). These findings highlight that non-coding regions of the genome have the greatest capacity to alter DNA structure, whereas mutations in gene-rich regions are less likely to cause significant alterations to structure.
Figure 4.4: Structural changes by chromatin mutation. Single-site chromatin mutations were performed by inverting the chromatin state of genomic locations where the correlation between the experimental and model-predicted probability of contact was initially high (correlation $C > 0.6$). The change in correlation with experimental probabilities after mutation $\Delta c$ was measured for different non-overlapping genomic classifications and shown here as histograms. (A) Spin-up (euchromatin) and spin-down (heterochromatin) sites. (B) Sites that did not contain genes, sites that had promoter regions and “gene body” sites that contained genes but no promoters.

4.1.5 Predicting sequence from structure

Here we show that it is possible to solve the inverse problem, namely, given a Hi-C map of contact counts, $\{n_{ij}\}$, and a model for the probability of a sequence neighbourhood $\vec{\sigma}$ to be in contact, determine the probability of each site $k$ to be in a particular sequence state ($\sigma_k = 1$ or $\sigma_k = -1$). We denote this probability as $P(\sigma_k|\{n_{ij}\})$ where $\{i, j\}$ is the set of all contacting pairs that contain genomic site $k$ in their sequence neighbourhood. By applying Bayes’ theorem we can write it as

$$P(\sigma_k|\{n_{ij}\}) = \frac{P(\{n_{ij}\}|\sigma_k)P(\sigma_k)}{P(\{n_{ij}\})},$$

(4.3)

where $P(\sigma_k)$ is a prior on the sequence state at site $k$, and $P(\{n_{ij}\})$ the probability of the data that simply acts as a normalization constant. As we show later in the Methods section, Eq. 4.15 can be rewritten as

$$P(\sigma_k|\{n_{ij}\}) = \frac{P(\sigma_k)^{1-M}}{P(\{n_{ij}\})} \prod_{i,j} \sum_{\vec{\sigma}} P(n_{ij}|\vec{\sigma}, d)P(\vec{\sigma}|d)\delta_{\sigma_k'},\sigma_k,$$

(4.4)

where $P(n_{ij}|\vec{\sigma}, d)$ is the probability of observing a particular number of counts between sites $i$ and $j$ given their sequence neighbourhood, $\vec{\sigma}$, $k'$ is the position of genomic site $k$ in that neighbourhood, and $M$ is the number of $\{i, j\}$ pairs. We take $P(n_{ij}|\vec{\sigma}, d)$ to be
a Gaussian distribution \( N(\lambda_{\bar{\sigma},d}, \zeta_{\bar{\sigma},d}^2) \) with mean equal to the average number of contact counts for a given sequence \( \bar{\sigma} \), \( \lambda_{\bar{\sigma},d} = \lambda(c|\bar{\sigma}, d) \), which is proportional to the probability of contact \( P(c|\bar{\sigma}, d) \) that can be evaluated from Eq. 4.1. The variance, \( \zeta_{\bar{\sigma},d}^2 \) can be estimated from the Hi-C data (see Methods for more details).

Using just the observed Hi-C counts \( n_{ij} \) for the test data and our prior fitted maximum-entropy models for \( P(c|\bar{\sigma}, d) \) and \( P(\bar{\sigma}|d) \), we calculated the probability of each site in the test set being spin-up. Fig. 4.5 shows this probability as a function of genomic location along the four test chromosome regions. Applying a threshold to these probabilities (\( P(\sigma_k = 1) = 0.5 \)), we predicted a sequence with a percent agreement with the original test sequence of 78% (chromosome 2L), 72% (chromosome 2R), 77% (chromosome 3L) and 77% (chromosome 3R). Thus structure alone can yield an important amount of information about the underlying sequence of bound chromatin factors. We feel that this could have significant impact in determining important regions for regulating structure. In particular, if the maximum-entropy model is reliably capturing the essential aspects that connect sequence to structure and if one believes that these aspects are conserved across different cell types, one could use a fitted model from one cell type to predict sequence in another for which only structural information is known.

Figure 4.5: Prediction of bound factors from contact map. Sequence of bound factors as a function of site along the four 5 Mbp test regions at the end of chromosomes 2 and 3 in Drosophila. In blue are shown the locations of spin-up sequence sites, \( \sigma_k = 1 \), as classified from the experimental measurements of bound factors. The predicted probability of each site being spin-up, \( \sigma_k = 1 \) using the inverse method is shown in red.
4.2 Conclusion

Our results demonstrate that a model for the probability of contact between two genomic sites given just their neighbourhood of bound factors can be estimated using a maximum-entropy method. The resulting model did well at predicting the contact map of a test set using only sequence information.

This model connects sequence to structure and thus offers the capability of testing the structural effect of mutating the chromatin sequence. Our analysis highlights that the alteration of DNA conformation by mutating chromatin sequence is particularly strong in sites containing no genes, while only minimal alterations to structure tend to occur when mutating active gene-rich regions.

Although we applied our fit model to data from Drosophila embryos, we envision that it could be useful in predicting contact maps from sequence data taken from other cell types from the same species (e.g. different developmental times or from mutated cell lines), or potentially other nearby species where the contact model might expect to hold.

We also showed how a fitted model connecting sequence to structure could be used to solve the inverse problem, namely that a sequence of bound factors could be predicted using just measured counts from a Hi-C contact map and a model for the probability of contact given sequence. One could imagine that making such sequence predictions from Hi-C measurements on individual cell lines may be more efficient than carrying out the potentially tens to hundreds of experiments needed to map the locations of all the bound factors.

The classification of the genome into chromatin states used to fit our model was based on the occupation patterns of specific binding factors rather than structure. Therefore, we consider the possibility that other classifications may have a greater ability to predict structure from sequence than the one presented. In the subsequent work presented in Chapter 5, we consider other possible ways of defining chromatin sequence states from data optimized for the special purpose of predicting structure.

Despite the great variety of DNA-binding chromatin factors identified so far, this method is able to reproduce DNA contact maps from binary classifications that group all of the observed factors into only two classes. It is therefore tempting to speculate that at the > 10 kbp scale, once DNA-bound factors are positioned on the genome, they tend to behave alike when it comes to generating DNA contacts. If correct, this would imply that the utility of such variety of chromatin factors is not entirely related to stabilizing complex DNA structures where a large number of different architectural elements is required. Instead, our findings favour the idea that the observed diversity of chromatin states is primarily related to alternative functions other than creating complex large-scale structures, such as the programming of chromatin states at different developmental times, or genetic regulation.
Overall our model has the potential for efficiently making predictions about both chromatin structure and sequence depending on which type of data is available.

4.3 Methods

4.3.1 DNA contact maps

DNA contact maps were obtained from the publicly available Hi-C experiments done by Schuettengruber et al [189] (GSE61471), performed on 3000-4000 *Drosophila melanogaster* embryos, 16-18 hours after egg laying. The contact map is an array whose elements are the number of times a particular pair of genomic sites were found to be in contact. The contact map we used defined sites to be of a fixed size (10 kbp) and were non-overlapping across the genome. The counts, $n_{ij}$, at a particular pair of sites $i$ and $j$ in the contact map were determined by counting up all sequenced pairs from the Hi-C measurement that fell into those sites. We normalized the contact map using the ICE method [204] so that total number of counts along each row across the contact map was the same.

4.3.2 Binary chromatin sequence classification

For every 10 kbp site, $k$, in the genome, we classified the genomic distribution of its DNA-bound factors into two possible chromatin sequence states corresponding to bound states that are associated with euchromatin (active DNA regions), $\sigma_k = 1$, and heterochromatin (inactive DNA regions), $\sigma_k = -1$. In order to make these sequence assignments, we used the “chromatin colors” classification of the *Drosophila melanogaster* genome by Filion et al [141] (GSE22069) which categorizes the distribution of bound factors into five different bound states (black, blue, yellow, green and red) based on the DamID binding profiles of 53 different chromatin factors in the embryonic *Drosophila melanogaster* cell line Kc167 (8-12 hours). We then grouped the five “chromatin colors” into just a spin-up and spin-down state based on the biological functions of their associated DNA-bound factors: spin-down = black and blue, corresponding to inactive/heterochromatin and spin-up = yellow, green and red that correspond to active/euchromatin. The chromatin color data gives the coordinates of the regions of the various colours. In order to assign a particular sequence state to each of our 10 kbp sites, we took the majoritarian color group (either active, or inactive) to be the unique sequence state of that site.

4.3.3 Training and test data sets

For the purpose of validating our model, we divided the *Drosophila melanogaster* autosomal chromosomes, 2 (comes as 2L and 2R) and 3 (comes as 3L and 3R) into a training set and test set. The training dataset consisted of genomic locations 2L: 1Mbp–17.02Mbp, 2R: 1Mbp–15.15Mbp, 3L: 1Mbp–18.55Mbp, 3R: 1Mbp–21.91Mbp. The testing dataset consisted

4.3.4 Extracting sequence neighbourhoods

Figure 4.6: Schematic of how the neighbourhoods $\vec{\sigma}$ of the contacting sites $(i, j)$ were built from the genomic distribution of chromatin states. First, the genome was binned into 10 kbp sites that could either be in a spin-up or spin-down state. (A,B) If $d = |j - i| < 12$, we defined the neighbourhoods as $\vec{\sigma} = (i - 4, \ldots, i, \ldots, j, \ldots, j + 4)$. (C) For $d \geq 12$ we defined the neighbourhoods as the union of the ten neighbours surrounding both $i$ and $j$ $\vec{\sigma} = (i - 4, \ldots, i, \ldots, i + 5, j - 5, j, \ldots, j + 4)$ thus keeping them at a maximum size of $N = 20$.

We constructed sets of sequence neighbourhoods $\{\vec{\sigma}\}_d$ for genomic distances between pairs $d = |j - i|$ ranging from $d = 1$ to $d = 80$ in units of the site size, 10 kbp. For a pair of sites $i$ and $j$ in the genome separated by $d$, we took the particular sequence neighbourhood, $\vec{\sigma}$, to be the union of sequence windows centered around each site. Specifically, we took 10 sites centered around site $i$ and 10 sites centered around site $j$ to create a neighbourhood
size of $N = 20$ (see S3 Fig.) where the given sequence neighbourhood consists of sites $\vec{\sigma} = (\sigma_{i-4}, \ldots, \sigma_i, \ldots, \sigma_{i+5}, \sigma_{j-5}, \ldots, \sigma_j, \ldots, \sigma_{j+4})$. For distances, $d < 11$, it was not possible to take sequence windows of size 10 sites centered on each site, so for these cases, we defined the neighbourhoods as $\vec{\sigma} = (\sigma_{i-4}, \ldots, \sigma_i, \ldots, \sigma_{j}, \ldots, \sigma_{j+4})$ which had a size $N = d + 9$ (Fig. 4.6). For a fixed genomic distance between sites $d$, we then scrolled through all pairs of sites, $\{i, j\}$ in both the training and test regions of the genome extracting their corresponding neighbourhoods $\vec{\sigma}$. Since the (left-right) directionality of the sequence neighbourhood should not influence the contact probability between $i$ and $j$, we also added the inverted sequence for each $\vec{\sigma}$ to the collection of neighbourhoods. Each genomic distance, $d$, thus had its own unique ensemble of sequence neighbourhoods spin-vectors for both the training and test data.

4.3.5 Calculating sequence-neighbourhood statistics

The maximum-entropy distributions, $P(\vec{\sigma}|c, d)$ and $P(\vec{\sigma}|d)$, were constrained to match several statistics of the extracted sequence neighbourhood ensembles. The statistics used are $\langle \sigma_k \rangle$ and $\langle \sigma_i \sigma_k \rangle$. For the distribution $P(\vec{\sigma}|c, d)$, these statistics were calculated as weighted averages of the extracted ensemble of sequence neighbourhoods, whereas for each neighbourhood we used as a weight the number of times, $n_{ij}$, that its contacting sites $i$ and $j$ were observed to be in contact. For $P(\vec{\sigma}|d)$ the above statistics were calculated as weighted averages where the weight of each sequence in the ensemble was equal to the number of times it appeared in the genome.

4.3.6 Fitting the maximum-entropy distributions

The parameters $h_k$ and $J_{lk}$ in

$$P(\vec{\sigma}|\cdot) = \frac{e^{\sum_k h_k \sigma_k + \sum_{i>j} J_{ki} \sigma_i \sigma_k}}{Z(\cdot)}$$

were fit to reproduce the experimental statistics $\langle \sigma_k \rangle_{\text{exp}}$ and $\langle \sigma_k \sigma_l \rangle_{\text{exp}}$ by following the method described by Tkačik et al. [207] that we summarize next. The term “$|$” simultaneously denotes the two conditions that we characterized at each distance of contact: The probability of observing $\vec{\sigma}$ given contact $P(\vec{\sigma}|c, d)$ and the probability of observing $\vec{\sigma}$ regardless of contact $P(\vec{\sigma}|d)$. We therefore fit a set of parameters $\{h_k, J_{lk}\}$ for each of the two conditions denoted by $|$ and for each distance of contact $d$.

Briefly, given one of the two conditions above, we used the corresponding collection of spin vectors and their associated weights (as described in the previous section) to calculate the experimental statistics $\langle \sigma_k \rangle_{\text{exp}}$ and $\langle \sigma_k \sigma_l \rangle_{\text{exp}}$. Then we implemented the following iterative scheme:

1. Start with random guess for $h_k$ and $J_{lk}$

82
2. Use Eq. 4.5 to calculate \( P(\vec{\sigma}|\cdot) \) for every possible \( \vec{\sigma} \) and obtain \( \langle \sigma_k \rangle_{\text{model}} \) and \( \langle \sigma_k \sigma_l \rangle_{\text{model}} \).

3. Check for convergence, \(|\langle \sigma_k \rangle_{\text{model}} - \langle \sigma_k \rangle_{\text{exp}}| < \epsilon\) and \(|\langle \sigma_k \sigma_l \rangle_{\text{model}} - \langle \sigma_k \sigma_l \rangle_{\text{exp}}| < \epsilon\), where \( \epsilon \) is the error tolerance (we used \( \epsilon = 0.0001 \)). If it hasn’t converged yet, continue.

4. Update \( h_k \) as \( \Delta h_k = \alpha (\langle \sigma_k \rangle_{\text{exp}} - \langle \sigma_k \rangle_{\text{model}}) \) and \( J_{lk} \) as \( \Delta J_{lk} = \alpha (\langle \sigma_k \sigma_l \rangle_{\text{exp}} - \langle \sigma_k \sigma_l \rangle_{\text{model}}) \), where \( \alpha \) is a small learning rate (we used \( \alpha = 0.025 \)). Go back to step 2.

This scheme is guaranteed to converge towards a unique solution regardless of the initial choice of \( h_k \) and \( J_{lk} \) [195].

### 4.3.7 Fitting first-order maximum-entropy distributions

Here, we present a first-order maximum-entropy model only constrained to reproduce one-spin statistics \( \langle \sigma_k \rangle \) of the experimental distributions of neighbourhoods \( \vec{\sigma} \). Similarly to the second-order model presented above, the first-order maximum-entropy distribution can be derived using the method of Lagrange multipliers [193,194,200] and has the following form:

\[
P(\vec{\sigma}|\cdot) = \frac{e^{\sum_k h_k \sigma_k}}{Z(\cdot)},
\]

where \( h_k \) are Lagrange multipliers that constitute the fitting parameters of the model. The partition function \( Z(\cdot) \) is obtained by summing the numerator over all possible \( \vec{\sigma} \) neighbourhoods. (In the above, we use “\( |\cdot| \)” to summarize that we were fitting two different conditions, namely \( P(\vec{\sigma}|c,d) \) and \( P(\vec{\sigma}|d) \).)

By noting that \( Z(\cdot) = \prod_k e^{h_k} + e^{-h_k} \), we can rewrite Eq. 4.6 as the product of terms that only depend on \( k \)

\[
P(\vec{\sigma}|\cdot) = \prod_k \frac{e^{h_k \sigma_k}}{e^{h_k} + e^{-h_k}}.
\]

The terms in the product are normalized to unity and thus find that \( P(\vec{\sigma}|\cdot) \) simply becomes the product of the independent probabilities for each \( \sigma_k \) in \( \vec{\sigma} \),

\[
P(\sigma_k|\cdot) = \frac{e^{h_k \sigma_k}}{e^{h_k} + e^{-h_k}}.
\]

The value of \( h_k \) that matches the \( \langle \sigma_k \rangle \) statistic can be found by solving

\[
\langle \sigma_k \rangle = \sum_{\sigma_k} \sigma_k P(\sigma_k|\cdot),
\]

83
which gives

\[ h_k = \frac{1}{2} \log \frac{1 + \langle \sigma_k \rangle}{1 - \langle \sigma_k \rangle}. \]  

(4.10)

Figure 4.7: Spin statistics of first-order maximum-entropy model versus at three different contact distances, d. Top row is for models conditioned on contact, \( P(\vec{\sigma}|c, d) \), whereas the bottom row is for models that are regardless of contact, \( P(\vec{\sigma}|d) \). (A, B) Site average statistics, \( \langle \sigma_k \rangle \). (C, D) Pairwise correlation statistics, \( \langle \sigma_l \sigma_k \rangle \). (E, F) Three body correlation statistics, \( \langle \sigma_k \sigma_l \sigma_m \rangle \). (G, H) The probability of \( \vec{\sigma} \) from the model versus the observed frequency from contact maps (G) and from the genome (H). For reference the insert compares experimentally measured spin-vector probabilities from training versus testing data.

Fig. 4.7(A,B) shows that the distributions calculated from Eq. 4.8 successfully predicted the experimental statistics \( \langle \sigma_k \rangle \) on the test data. However, this first-order maximum-entropy distributions failed to capture both the second- (Fig. 4.7(C,D)) and third- (Fig. 4.7(E,F)) order statistics that were not incorporated into the fit. In Fig. 4.7(G,H) we also see that the predicted sequence neighbourhood probabilities \( P(\vec{\sigma}|\cdot) \) from the model did not agree with their frequencies as seen in the data.

We thus conclude that the first-order maximum-entropy distribution in Eq. 4.8 is not a sufficiently good description of the experimental distributions of \( \vec{\sigma} \), and further-order moment distributions need to be considered.

4.3.8 Inspection of model parameters

For each distance of contact \( d = |j - i| \) we obtained two sets of parameters for \( P(\vec{\sigma}|\cdot) \), one describing the probability of a given neighbourhood \( \vec{\sigma} \) given contact between \( i \) and \( j \),
\[ P(\vec{\sigma}|c,d) = f(h_k^{c,d}, J_{kl}^{c,d}) \], and another set describing the probability of a given neighbourhood regardless of contact (which we term here as background), \[ P(\vec{\sigma}|d) = f(h_k^{bg,d}, J_{kl}^{bg,d}) \].

Fig. 4.8A shows that the Shannon entropy \[ [190], \]

\[ S[P(\vec{\sigma}|\cdot)] = -\sum_{\vec{\sigma}} P(\vec{\sigma}|\cdot) \log_2 P(\vec{\sigma}|\cdot), \] (4.11)

of the sequence distribution regardless of contact was greater than the entropy of the distribution of contacts at short distances \( (d < 330 \text{ kbp}) \) and the opposite happened for longer distances. In particular, the plot displays the average number of neighbourhood configurations encoded in the probability distributions, \( 2^S \).

Fig. 4.8B displays the Kullback-Leibler (K-L) divergence \[ [192] \] between neighbourhoods given contacts and background neighbourhoods at each distance \( d \),

\[ D[P(\vec{\sigma}|c,d)||P(\vec{\sigma}|d)] = \sum_{\vec{\sigma}} P(\vec{\sigma}|c,d) \log_2 \frac{P(\vec{\sigma}|c,d)}{P(\vec{\sigma}|d)}, \] (4.12)

that can be interpreted as a “distance” between the two probability distribution of neighbours, or the information gain when including information about contacts into the background distribution of neighbours, therefore going from \( P(\vec{\sigma}|d) \) to \( P(\vec{\sigma}|c,d) \). This quantity diminished with distance and saturated at its lowest value at a distance of \( \sim 300 \text{ kbp} \).

We then explored the similarity between the distributions of contacting neighbourhoods at different distances. Specifically, we subtracted the K-L divergence of the background from the K-L divergence of the contacting distributions (correcting for background sequence effects),

\[ \Delta D[P(\vec{\sigma}|c,d_1)||P(\vec{\sigma}|c,d_2)] = D[P(\vec{\sigma}|c,d_1)||P(\vec{\sigma}|c,d_2)] - D[P(\vec{\sigma}|d_1)||P(\vec{\sigma}|d_2)]. \] (4.13)

Fig. 4.8C shows two regimes of similarity between the distributions at different distances, one for \( d_1, d_2 < 390 \text{ kbp} \) and another for \( d_1, d_2 \geq 390 \), where \( \Delta D[P(\vec{\sigma}|c,d_1)||P(\vec{\sigma}|c,d_2)] \) takes the lowest values (blue denotes low \( \Delta D \) whereas red corresponds to high \( \Delta D \)).

We further compared models at different distances by analyzing the similarity of the energetic coefficients of enrichment defined as \( \Delta h_k^d = h_k^{c,d} - h_k^{bg,d} \) and \( \Delta J_{kl}^d = J_{kl}^{c,d} - J_{kl}^{bg,d} \) since these were the parameters involved in the distance-normalized contacts as it can be derived from Eq.4.1 and Eq. 4.2:

\[ \frac{P(c|\vec{\sigma},d)}{P(c|d)} = \frac{P(\vec{\sigma}|c,d)}{P(\vec{\sigma}|d)} = \frac{Z_{bg,d}}{Z_{c,d}} e^{-\sum_k \Delta h_k^d \sigma_k - \sum_{l>k} \Delta J_{kl}^d \sigma_l \sigma_k} \] (4.14)

For every distance of contact, the parameters \( \Delta h_k^d \) and \( \Delta J_{kl}^d \) were concatenated into a vector, and the set of vectors corresponding to all distances was then clustered into two groups with K-means \[ [208] \]. The coefficient vectors naturally separated at the distance of
Figure 4.8: (A) Average number of sequences ($2^S$) stored in the probability distributions of neighbours given contact between $i$ and $j$, $P(\tilde{\sigma}|c,d)$, and irrespectively of contact, $P(\tilde{\sigma}|d)$, as a function of distance of contact $d = |j - i|$. (B) Kullback-Liebler divergence between probability distribution of neighbours given contact and irrespectively of contact as a function of distance. (C) Kullback-Liebler divergence between the contacting distributions at different distances of contact minus Kullback-Liebler divergence between the background distributions: $\Delta D[P(\tilde{\sigma}|c,d_1)||P(\tilde{\sigma}|c,d_2)] = D[P(\tilde{\sigma}|c,d_1)||P(\tilde{\sigma}|c,d_2)] - D[P(\tilde{\sigma}|d_1)||P(\tilde{\sigma}|d_2)]$. Blue indicates low values whereas red indicates high values. (D) A vector concatenating the difference between energetic coefficients for neighbourhoods in contact and background, $\Delta h^d_k = h^c_k - h^{bg}_k$ and $\Delta J^d_{kl} = J^{c}_{kl} - J^{bg}_{kl}$, was built at every distance of contact. K-means clustering of the set of coefficient vectors naturally separated them into two clusters, one for $d < 390$ kbp and another for $d \geq 390$ kbp. (E) Average of energetic coefficient vectors in K-means cluster 1. Blue indicates negative values whereas red indicates positive values. (F) Average of energetic coefficient vectors in K-means cluster 2. Blue indicates negative values whereas red indicates positive values. (G) Principal Component Analysis was applied to the set of energetic coefficient vectors. PC1 separates vectors into two clusters, delimited at 390 kbp, the same distance that separates K-means clusters. (H) Coefficient vectors projected into PC1 highlight the differences between the coefficients in the two clusters. Blue indicates negative values whereas red indicates positive values.
contact of 390 kbp (Fig. 4.8D). In Fig. 4.8E and Fig. 4.8F we show the average energetic coefficients of enrichment of the two K-means clusters, which differ both in $\Delta h_k$ and $\Delta J_{kl}$.

In addition, we applied Principal Component Analysis (PCA) [186, 187] to the same set of coefficient vectors as above, and the first principal component (PC1) clearly separated the same clusters previously found by K-means delimited at a distance of contact of 390 kbp (Fig. 4.8G). PC1 shows positive scores for the shorter distances of contact ($d < 390$ kbp) and negative scores for the longer distances of contact ($d \geq 390$ kbp). Therefore, projecting the vectors of coefficients onto PC1, we found how the short-distance cluster differs from the long-distance cluster (Fig. 4.8H), which corresponded to an increase of ferromagnetic interactions between the sites situated inside the loop.

4.3.9 Structural changes due to sequence mutation

First, we identified the sites $k$ for which the local predicted contact map was in good agreement with the experimental contact map. The local contact map consists of only the subset of genomic pairs of sites $\{i, j\}$ whose contact probability $P(c|\sigma, d)$ is influenced by the state of site $k$ (i.e. $k$ is one of the spins in the neighbourhood $\sigma$ of $i$ and $j$). At each genomic site, we calculated the correlation $c$ between the local predicted distance-normalized contacts, $P(c|\sigma, d)/P(c|d)$, and the local experimental distance-normalized contacts counts, $n_{ij}/\langle n(d) \rangle$, and selected the best correlating locations of the genome ($C > 0.6$) for further analysis. Next, at each of these selected locations the chromatin state was flipped and the correlation $c'$ between predicted and experimental distance-normalized counts was measured. We then defined $\Delta C = C' - C$ as a measurement of structural disruption due to sequence mutation at a given site.

4.3.10 Predicting sequence from structure

Given fitted maximum entropy distributions, $P(\sigma|c, d)$ and $P(\sigma|d)$, over a range of genomic distances $d$, we now work out how to solve the inverse problem, namely finding the probability of a genomic site $k$ being in a particular binary state $\sigma_k$, given only structural data from a set of Hi-C counts $\{n_{ij}\}$. We denote this probability by $P(\sigma_k|\{n_{ij}\})$, where $\{n_{ij}\}$ is the set of counts between all $(i, j)$ pairs of sites considered to be neighbors of $k$ in our model.

As in the Results section, Bayes’ theorem gives

$$P(\sigma_k|\{n_{ij}\}) = \frac{P(\{n_{ij}\}|\sigma_k)P(\sigma_k)}{P(\{n_{ij}\})},$$

(4.15)

where $P(\{n_{ij}\}|\sigma_k) = \prod_{ij} P(n_{ij}|\sigma_k, d)$. $P(n_{ij}|\sigma_k, d)$ is the probability of observing exactly $n_{ij}$ contact counts between a pair of sites a distance $d = |j - i|$ apart, given that the genomic site $k$ in their sequence neighborhood is in a particular state, $\sigma_k$. (Note that $d$ is a redundant variable whenever $i$ and $j$ are specified, i.e. $P(n_{ij}|\sigma_k) = P(n_{ij}|\sigma_k, d)$.) We
nevertheless introduce it here for consistency with the rest of our notation. \( P(\sigma_k) \) is the prior on \( \sigma_k \) and is the probability for site \( k \) to be in one of the two states (here we take it to be a constant over the genome, with the same value as measured in the training set \( P(\sigma_k = 1) = 0.31 \)). \( P(\{n_{ij}\}) \) is simply a normalization constant and is found by summing the numerator over \( \sigma_k \). Rewriting Eq. (4.15), we have,

\[
P(\sigma_k|\{n_{ij}\}) = \frac{P(\sigma_k)}{P(\{n_{ij}\})} \prod_{ij} P(n_{ij}|\sigma_k, d). \tag{4.16}
\]

Using \( k' \) to label the position that the genomic site \( k \) takes in the particular neighborhood of \((i,j)\), \( \bar{\sigma} = \{\sigma_1, \cdots, \sigma_{k'}, \cdots, \sigma_N\} \), and considering \( P(\sigma_k|d) = P(\sigma_k) \) we then rewrite \( P(n_{ij}|\sigma_k, d) \) as

\[
P(n_{ij}|\sigma_k, d) = \frac{P(n_{ij}, \sigma_k|d)}{P(\sigma_k)} = \frac{P(n_{ij}, \sigma_k|d)}{P(\sigma_k)} = \frac{\sum_{\delta} P(n_{ij}, \bar{\sigma}, \sigma_k|d)}{P(\sigma_k)} = \frac{\sum_{\delta} P(n_{ij}, \bar{\sigma}|d) P(\bar{\sigma}|d) \delta_{\sigma_k, \sigma_{k'}}}{P(\sigma_k) \delta_{\sigma_k, \sigma_{k'}}} = \frac{\sum_{\delta} P(n_{ij}|\bar{\sigma}, d) P(\bar{\sigma}|d) \delta_{\sigma_k, \sigma_{k'}}}{P(\sigma_k)} \tag{4.17}
\]

where the Kronecker delta \( \delta_{\sigma_k, \sigma_{k'}} \) ensures that we sum all possible sequences of the neighborhood \( \bar{\sigma} \) that have genomic site \( k \) held fixed in a particular state \( \sigma_k \).

Next, by combining Eqs. 4.16 and 4.17, we obtain

\[
P(\sigma_k|\{n_{ij}\}) = \frac{P(\sigma_k)^{1-M}}{P(\{n_{ij}\})} \prod_{ij} \sum_{\delta} P(n_{ij}|\bar{\sigma}, d) P(\bar{\sigma}|d) \delta_{\sigma_{k'}, \sigma_k}, \tag{4.18}
\]

where \( M \) is the number of \((i,j)\) pairs that have sequence neighborhoods that contain genomic site \( k \). The distribution, \( P(\bar{\sigma}|d) \) is taken to be the fitted maximum entropy distribution at a distance \( d \). We assume that the probability of observing \( n_{ij} \) Hi-C counts given a sequence state \( \bar{\sigma} \), \( P(n_{ij}|\bar{\sigma}, d) \), is as a Gaussian distribution \( \mathcal{N}(\lambda_{\bar{\sigma}, d}, \zeta_{\bar{\sigma}, d}^2) \) with a mean number of counts, \( \lambda_{\bar{\sigma}, d} \), proportional to the fitted probability of contact for the given sequence neighborhood \( \bar{\sigma} \),

\[
\lambda_{\bar{\sigma}, d} = \lambda(c|\bar{\sigma}, d) = KP(c|\bar{\sigma}, d) = \frac{P(c|\bar{\sigma}, d) \lambda(c|d)}{P(\bar{\sigma}|d)}, \tag{4.19}
\]

where \( K \) is a constant that depends on experimental details such as the number of cells used and the efficiency of contact detection, and \( \lambda(c|d) = KP(c|d) = \langle n(d) \rangle \) is the experimental average of Hi-C counts at a distance \( d \). With this and the two fitted maximum entropy distributions, we can calculate the mean number of counts for a given sequence neighborhood
from Eq. 4.19. The variance $\zeta^2_{\vec{\sigma},d}$ is sampled from the train set as a function of $\lambda$. Specifically, we calculated the rates $\lambda_{ij}$ associated to all Hi-C counts $n_{ij}$ from the train set. Then, for various values of $\lambda$ we collected the Hi-C counts $n_{ij}$ that our model had assigned a rate $\lambda_{ij}$ between $0.9 \times \lambda$ and $1.1 \times \lambda$. Lastly, we calculated the variance of the rate-associated counts $\zeta^2(\lambda)$ and fitted a polynomial curve to it (we found $\zeta^2 \approx \lambda^2$).

Everything in Eq. 4.18 is now determined and so the probability of a particular sequence state (either $\sigma_k = 1$ or $\sigma_k = -1$) at every site $k$ can be calculated if given a Hi-C contact map, $\{n_{ij}\}$. 

Chapter 5

Neural networks for chromatin conformation

In eukaryotic cells, the DNA-protein complex known as chromatin folds into specific 3D structures that bring distant parts of the genome into spatial proximity. These conformations can modulate the expression of genetic information by altering the frequency of interaction between genomic regions and regulatory proteins. The recent advent of high-throughput genomic technology has allowed the genome-wide measurement of both chromatin structure via Hi-C contact maps [40,129] as well as the bound locations of a great number of chromatin-associated factors through ChIP-seq methods [111,112].

A large body of evidence supports the hypothesis that the spatial arrangement of chromatin factors strongly influences the probability of chromatin contacts between distant genomic regions [41]. In particular, megabase-sized genomic compartments with similar chromatin states tend to interact with each other [32,40]. At a finer sub-megabase scale, topologically associated domains [35–37] flanked by boundary chromatin elements [37] work as independent genomic units characterized by self-interaction and repulsion with other genomic regions [38,39]. Consequently, models that aim to resolve how a sequence of chromatin factors directs the folding of chromosomes are now being developed – a problem akin to the folding of proteins based on their one-dimensional amino acid sequence.

Most of the progress towards predicting chromatin conformation from chromatin sequence relies in constructing heterogeneous beads-on-a-string polymers whose bead types correspond to different chromatin states [50,61,148–160]. After generating an ensemble of conformations, the distances (or contacts) between beads are compared to the existing Hi-C data. These simulations have been successful in corroborating that interactions between factors together with topological constraints may be responsible for driving chromatin conformation. Nevertheless, calculating the probability of contact between two genomic sites relies on sampling a vast number of polymer configurations. Consequently, exploring the conformational effects of mutating the chromatin sequence in various ways is computationally challenging. An alternative Bayesian approach, presented in Chapter 4 and in [201],
has been successful in predicting the local contact maps from chromatin states without the need for simulations, a method that can rapidly calculate how contact probabilities change when chromatin states are altered.

Despite the recent success of computational techniques for predicting chromatin conformation, several important questions remain unanswered. First, the labeling of chromatin states used as inputs for the prediction of chromatin folding is typically based on previous chromatin classifications, namely the unsupervised clustering of chromatin factors, or the clustering of interaction profiles. It is therefore unlikely that these classifications constitute the best 1D description of the sequence that determines chromatin structure. One may expect to achieve better predictive power by generating a conformation-specific annotation of chromatin sequence based on the architectural effects that we are only now able to predict. Second, although current models can predict chromatin structures from interactions between genomic sites, little is known about how chromatin conformation is encoded in the chromatin sequence. For instance, how would a specific frequency of contact be modulated when altering the chromatin sequence around it? To what degree are chromatin contacts governed by the states of the contacting sites or the states of their genomic neighbours?

In recent years, the field of machine learning has seen a major breakthrough. Multi-layer neural networks that have been around for decades [209] can now take advantage of current processing power and vast amounts of training data to solve pattern recognition problems with an accuracy comparable, and sometimes superior, to that of humans. Some examples of their success are object detection [210], speech recognition [211], defeating humans at the game of Go [212], painting art [213] or driving cars [214]. They do so by learning the best features that describe an object, usually combining a hierarchy of sub-features. In the context of Hi-C data analysis, multi-layer neural networks have been used to generate statistical confidence estimates for chromatin contacts [215] and to enhance the resolution of contact maps [216]. In this chapter, we apply multi-layer neural networks for the first time to solve the problem of chromatin folding, namely the prediction of chromatin contacts from a sequence of chromatin states. We show that using multi-layer neural networks, one is not only able to predict chromatin conformation from sequence, but also predict sequence from chromatin conformation. In addition, the model generates a biologically relevant 1D chromatin sequence optimized to explain chromatin conformation. Furthermore, by exploring how the model relates chromatin sequence and conformation, we unveil key features behind their connection.

5.1 Dense neural networks for connecting conformation to sequence

An artificial neural network consists of a series of units known as neurons that receive a number of inputs, add them together with some weights and optionally pass the result
through a non-linear function, known as the activation function, to generate an output. The weights connecting neurons are trainable parameters that are fit to reduce a defined cost function of the output. When the neural network is feedforward, as in our case, the neurons are organized in a hierarchy of layers, where the output from the neurons in one layer act as the inputs to the next. feedforward neural networks are often referred to as dense neural networks (DNNs) when they feature a high number of neurons at each layer. In addition, neural networks are considered to be “deep” when they feature a high number of layers.

The universal approximation theorem states that under mild assumptions, a feedforward neural network with a single layer and a finite number of neurons can approximate any continuous function \[217\]. This gives DNNs the power of exploring the space of non-linear models that can best approximate the target function. This capability is one of the main reasons of their great success at modelling complex problems with minimal design input by humans. At the same time, neural networks hide their inner workings behind large combinations of matrix weights that are often hard to interpret. For this reason, DNNs tend to be seen as black boxes that can perform a great variety of tasks but offer little mechanistic explanation of how the inputs of the model are being used to generate the output target function. Nevertheless, in recent years a great amount of effort in the DNN community been directed to developing techniques to infer how information is processed in these models \[218–224\].

Convolutional neural networks (CNNs) are a special type of neural network typically used when the input data has some spatial structure relevant to the problem (e.g. images, or a sequence of characters). CNNs feature a special type of layer known as convolutional layer, or filter, where each filter consists of a neuron that scrolls through the spatial arrangement of inputs, and at each position it performs the same calculation on a local subset of its inputs. Usually, by inspecting the trained filter, one can infer what type of pattern the filter is detecting, thus highlighting the features that are relevant to the specific problem that the CNN has been trained to solve.

In this chapter, our principal goal is to predict chromatin contact maps from chromatin sequence data. Specifically, the chromatin sequence data is made of \(M\) different genomic-occupancy measurements of chromatin factors (with \(M = 50\) for our \emph{Drosophila melanogaster} dataset) at \(N\) non-overlapping bins of a fixed size (10 kbp) covering the genome (\(N = 9663\) for \emph{Drosophila}'s autosomal chromosomes). The resulting input sequence data has a total dimensionality of \(M \times N\) (for further details see Methods). With respect to chromatin conformation, using the same genomic binning, a chromatin contact map with dimensions \((N \times N)\) can be constructed where each element represents how many times a genomic bin has been observed in contact with another genomic bin (see Methods). The contact map is thus a symmetric matrix. Sequences and contact maps of smaller regions of the genome structure can be visualized by simply using a subset of bins. In particular, we are interested in predicting contact maps of subregions of size \(w\) (we choose \(w = 80\)).
Based on our prior findings from Chapter 3 and Chapter 4, we assume that the sequence of length $w$ that makes up that subregion as well as sequences flanking that region are the key determinants of the local contact map. Here we choose an input sequence of size $3w$, consisting of the same $w$ bins of the contact maps, plus the two immediately flanking regions of size $w$ each. We thus aim to predict $w \times (w+1)/2$ outputs, (i.e. the unique values of the $w \times w$ local contact map) from $M \times 3w$ inputs (Fig. 5.1).

Figure 5.1: Schematic of the forward model convolutional neural network (CNN). This neural network is trained to predict chromatin contacts maps (Hi-C, top) from various chromatin-sequence factors (Chip-seq profiles, bottom). The Hi-C data to be predicted as an output is the upper diagonal of the Hi-C matrix of a $w$-wide genomic window (with $w = 80$ bins of 10 kbp each). The input to the CNN is a $3w$-long sequence that includes the $w$-long region of the Hi-C matrix (inner sequence) as well as two $w$-long sequences on each side (flanking sequences). The CNN is made of two parts. First, a sigmoid-activated convolutional layer reduces the $M \times 3w$ input reducing its dimensionality to a one-dimensional $3w$-long vector (Fig. 5.1). This vector can be interpreted as a one-dimensional sequence of chromatin states (with values between 0 and 1) that will be used as input to the rest of the neural network to predict contact maps. More specifically, if we denote by $\vec{x}_i$ the $i$th position of the input chromatin sequence, with dimension $M$ equal to the number of chromatin factors, the value of the 1D chromatin annotation at that position, $y_i$, is obtained from

Our CNN for predicting chromatin conformation from chromatin sequence, which will be referred as “the forward model”, has interpretability in mind. First, a convolutional filter of width equal to one and a sigmoidal output function acts on the $(M \times 3w)$ input reducing its dimensionality to a one-dimensional $3w$-long vector (Fig. 5.1). This vector can be interpreted as a one-dimensional sequence of chromatin states (with values between 0 and 1) that will be used as input to the rest of the neural network to predict contact maps.
\[ y_i = \frac{e^{E_i}}{1 + e^{E_i}}, \quad (5.1) \]

with
\[ E_i = \sum_j W_j^0 \cdot x_{i,j} + \beta_0, \quad (5.2) \]

where \( W_j^0 \) and \( \beta_0 \) are the trainable weights of the convolutional filter. The index \( j \) corresponds to each of the \( M \) chromatin factors. The fitted filter thus denotes the weights applied to each of the \( M \) bound factors for classifying the chromatin into a single sequence that is the best predictor of structure. (Later in the text, this 1D chromatin sequence, \( y_i \), is referred as \( \sigma_i \) in analogy to our previous annotation from Chapter 4. Nevertheless, this classification is continuous whereas the previous was binary).

Next, the resulting 1D chromatin sequence of size \( 3w \) is fed to a DNN with multiple layers of increasing size, where the last layer has \( w \times (w + 1)/2 \) outputs corresponding to the values of a local contact map for \( w \) bins. The value obtained at neuron \( i \) of layer \( n \), \( y_n^i \), is calculated using the values of all neurons \( k \) in the previous layer \( y_{n-1}^k \),
\[ y_n^i = \text{ReLU} \left( \sum_k W_k^n \cdot y_{n-1}^k + \beta_i \right), \quad (5.3) \]

where \( W_k^n \) is a matrix of weights applied to each neuron of the previous layer, \( \beta_i \) is a constant, and \( \text{ReLU} \) is the rectified linear unit function, namely \( f(x) = \max(0, x) \), which helps to introduce non-linearities and sparse activation (50% of neurons are activated) while remaining easily computed and differentiated \[225\]. Both \( W_k^n \) and \( \beta_i \) are trainable parameters.

The cost function to be minimized during the fitting procedure was defined as the mean squared error between experimental and predicted distance-normalized contact maps, along with L2 regularization of the filter weights. Overfitting was controlled using the technique of dropout regularization, which consists of setting the output of randomly selected neurons to zero with a given probability \[226\]. Dropout regularization effectively forces the fitting procedure to utilize the entire network to solve the problem by randomly switching certain pathways to be off while training. Since all operations are linear (or easily differentiated exponentials) parameters can be efficiently updated using stochastic gradient descent. By using the Python package Keras on top of TensorFlow, the parameters of the model were trained until convergence in approximately 30 minutes on a personal laptop (more details are provided in Methods section).

In addition, we also built a dense neural network (DNN) that solves the inverse problem. Namely, it is trained to predict the previously found 1D chromatin annotation from contact maps alone. The architecture of this network resembles an inverted version of the forward-model, and we thus name it “the backward model”. This network outputs a \( w \)-long vector of
1D chromatin states from the $w \times (w - 1)/2$ contacts between pairs of sites in the sequence window. Note that the output of the backward model is $w$-long in contrast to the $3w$-long sequence used as input of the forward model. This is because the flanking regions are hard to predict using just the contact map from the $w$-long interior region, and trying to predict the $3w$-long sequence leads to convergence errors in the procedure. The network is comprised of multiple ReLu-activated dense layers, except for the last output layer which is sigmoid-activated.

5.2 Results

5.2.1 Model predictions

The forward model described above was fitted using a training dataset, along with a validation dataset used to check for convergence and to avoid overfit (Methods). With the fitted forward model, a test set of sequences was then given as input, resulting in a set of predicted local contact maps. We found a correlation of 0.61 between the individual pairs of predicted and experimental contact maps in the test set. Nevertheless, since our data consists of overlapping sliding sequence windows, a contacting pair is predicted multiple times as output from different sequences. This offers the opportunity of averaging the outputs from multiple windows and thus increasing the predictive power of the model (akin to bootstrap aggregating, also known as bagging [227]). Upon doing so, the correlation between the predicted and experimental maps increased to 0.68 (Fig. 5.2A). This correlation is substantially better than the 0.46 correlation found from the maximum entropy model in the previous chapter.

By training the forward model, the convolutional filter has been fit to optimize the predictive power of the high-dimensional input sequence to the contact map outputs. The learned weights in the convolutional filter, $W_j^0$ in Eq. 5.2, represent the strength of each chromatin factor in determining structure. Fig. 5.2B shows the sorted distribution of these weights for all chromatin factors. The resulting distribution contains a significant amount of biologically relevant information. We find that one set of factors (positive weights) are associated with inactive/heterochromatic factors (H1, H3K27me3, H3K9me1), whereas the other set of factors (negative weights) corresponds to active/euchromatic factors (H3K36me3, NURF, H3K4me1, RNAPolII). Interestingly, applying the filter to the input sequences and histogramming the resulting 1D chromatin annotation, we find that it has a bimodal nature (Fig. 5.2C) where the inactive mode is bell-shaped around a value of $\sim 0.8$, whereas the active mode is peaked at zero (Fig. 5.2). By looking at Eqs. (5.1) and (5.2), we see that the peak of the inactive state corresponds to a fine-tuned “energy” $E \approx 1.4$ whereas the active state that peaks at zero is composed of arbitrarily negative values of $E$, which may reflect a heterogeneous mixture of active chromatin factors. Thus the fitted filter has naturally grouped the multitude of different chromatin factors into two groupings based on their
Figure 5.2: Model predictions. (A) Distance-normalized Hi-C contacts predicted from the forward-model CNN in a 5 Mbp region of the test data set that was left out of the fitting procedure (chrom 3R 15–20 Mbp). A correlation of 0.68 between the original and predicted counts was obtained. (B) Weights of the sigmoid-activated convolutional filter applied to the chromatin factor sequences in order to generate the 1D chromatin sequence. (C) Histogram of the values of the 1D chromatin sequence obtained after the convolutional layer. (D) An independent DNN (backward model) was built to predict the 1D chromatin sequence from Hi-C data. From bottom to top, multiple chromatin factors were converted to this 1D chromatin sequence by running them through the convolutional filter of the forward model. Then, the backward model was used to predict those 1D chromatin sequence from Hi-C contacts. The predicted 1D sequence (red) and the original 1D sequence (below) showed a correlation of 0.73. (Top) A Gaussian-smoothed version of the original sequence showed a correlation of 0.93 with the predicted sequence from the backward model. The genomic region shown in (D) is the same 5 Mbp region of the test set shown in (A).
contribution to the 1D chromatin annotation, with the heterochromatic factors playing a more dominant singular role in shaping structure and the euchromatic factors having a more heterogeneous influence.

We then fit the backward model (structure to sequence) to predict the 1D chromatin sequence learned from the forward model using just the local Hi-C contact map as input. Thus the backward model predicts a likely chromatin sequence that could form that map. Applying the trained backward model to the test set, the individual sequences predicted from the local contact maps had a correlation of 0.66 with the original sequences and a correlation of 0.73 after averaging overlapping windows (Fig. 5.2D). The predicted profiles visually resemble a smoothed version of the original 1D chromatin sequence from the convolutional filter. This was corroborated by performing a Gaussian smoothing of the original 1D sequence, with the resulting smoothed sequence now having a correlation of 0.93 with the backward model prediction (Fig. 5.2D). Based on this finding, we hypothesize that the forward model may in fact be doing smoothing internally in the DNN by predicting chromatin contacts based on a local average of the chromatin sequence. We tested this hypothesis by feeding the Gaussian smoothened 1D chromatin sequence from the convolutional filter as input into the DNN layers of the forward model. The predicted contacts from the smoothed sequence showed a correlation of 0.98 with the previously predicted contacts derived from the non-smoothed sequence (shown in Fig. 5.3B), indicating that the forward-model network generates a similar contact map output from a smoother description of chromatin factors.

Since many of the available contact maps are often measured in different experimental conditions, cell types and/or developmental times than those of the bound chromatin factors, one may expect that the chromatin sequence used as input does not reflect the true sequence that determined the measured local contact map and vice versa. We therefore looked at the correlation between the forward 1D sequence (derived from ChIP-Seq) and the backward 1D sequence (derived from contacts) in the test set (Fig. 5.3B). We found that regions where these sequences differed the most were regions where the correlation between predicted Hi-C counts and original counts tended to also be poor (0.35 correlation between the two trends), which suggests a discrepancy between the state of the cells used for measuring chromatin sequence and chromatin contact maps. We thus used the sequences predicted with the backward model as the input to the dense layers of the forward-model, obtaining a new set of predicted contact maps (see schematic in Fig. 5.3A). We find that the correlation between the predicted and original contact maps improves from 0.68 to 0.71 using the sequence from the backward model as input to the DNN of the forward model, compared to the sequence from the convolutional filter applied to the ChIP-seq profiles (Fig. 5.3B). This finding thus highlights two aspects: First, 2D contact maps can be efficiently encoded into a 1D vector and decoded back. Second, a backward model that predicts chromatin sequence from structure can be used to identify locations in the genome where sequential
Figure 5.3: Evaluation of discrepancies between sequential data and structural data. (A) Schematic of the three sequence-structure predictions performed. On the left, the original ChIP-seq is fed to a CNN (forward model) and outputs both a 1D chromatin sequence and a predicted Hi-C. In the centre, the original Hi-C is fed to a DNN (backward model) to predict the 1D chromatin sequence found in the forward model. We name this prediction the backward sequence. On the right, the backward sequence is fed to the dense neural network of the forward model (without fitting it again) to generate a new Hi-C prediction (Hi-C prediction 2), based on the backward chromatin sequence derived from the original Hi-C. (B) In the testing set, we calculated the correlations between the data generated by the models in (A) along its \( w \)-wide genomic windows. Genomic regions where the backward sequence differs from the original sequence tend to coincide with regions where the predicted Hi-C and the original Hi-C differ (0.35 correlation). By predicting contact maps based on the backward-sequence instead of the initial 1D chromatin sequence we find an improvement between the predicted Hi-C and the original Hi-C (correlations with original test data increased from 0.68 to 0.71).
and structural datasets likely differed from one other. One could imagine this to be a powerful technique to determine variable sequence regions by using contact maps from cells of differing tissue, developmental time or disease states.

### 5.2.2 Spatial analysis of conformational effects

In this section, we focus on determining what characteristics of the chromatin sequence influence contact maps the most, and vice-versa, how contact maps are used to infer the chromatin sequence. The following analysis thus serves a double purpose: it answers specific questions about chromatin folding while making sure that the predictions from our neural networks come from a correct representation of the underlying biological mechanisms rather than the exploitation of data artifacts.

First, we explore what type of chromatin sequences tend to generate high and low probabilities of contact. Using the reduced 1D chromatin sequence output from the convolutional filter on the test data, we looked at the average chromatin sequences involved in high and low-probability contacts. More specifically, we separated chromatin sequences based on whether the sites $i$ and $j$ situated at a distance $d = |j - i|$ and centred in the sequence (see legend in Fig. 5.4) had either a high probability of contact $P(c_{ij})$ or a low $P(c_{ij})$ (here we call probability of contact the distance-normalized chromatin contacts, as we explain in Methods). These averages were performed for different distances $d = 0, \ldots, 80$, and can be visualized as heat maps (Fig. 5.4). We find that high probability of contact tends to involve inactive chromatin between sites $i$ and $j$, whereas low probability of contact tends to involve active chromatin in between sites. We further see that these particular trends are stronger at shorter distances of contact ($d < 300$ kbp) than at larger distances of contact (recall the same trend seen in Chapter 3 and Chapter 4). Interestingly, for highly contacting sites over long distance, there is a strong presence of active chromatin states at the contacting sites $i$ and $j$.

Second, we measured how sensitive is the probability of contact $P(c_{ij})$ to the chromatin state, $\sigma_k$, of every position $k$ in the surrounding $3w$-long sequence window. We did so by calculating the gradient of the probability of contact, $\partial P(c_{ij})/\partial \sigma_k$ at each site of the $3w$-long chromatin sequence. This quantity can be obtained by the method of back-propagation (see Methods) and highlights how the probability of contact would change upon increasing the value of $\sigma_k$, i.e. making the chromatin more inactive at that location $k$. Alternatively, gradient values can also be interpreted as how chromatin states must be altered in order to increase the probability of contact. A negative value of the gradient would imply that to increase the contact probability one would have to decrease the value of $\sigma_k$, making the state more active. This thus highlights the conformational effect that would be expected upon mutating the chromatin sequence at each particular location of the genome (Fig. 5.5A).

To examine the general effects of how varying the chromatin sequence affects the probability of contact, we averaged $\partial P(c_{ij})/\partial \sigma_k$ over the test data. Specifically, for each pair of
Figure 5.4: Averaged chromatin sequences based on whether their sites $i$ and $j$ separated at a given distance $d = |j - i|$ and centred in the sequence had either a high or a low probability of contact $P(c_{ij})$. The contact map on the right shows how the vertical axis represents separation between contacting pairs. Horizontal axis is separated into three regions corresponding to an inner sequence with same width (800 kbp) and alignment as the contact maps, and two (800 kbp) flanking sequences situated on each side of the contact map. On average, $P(c_{ij})$ involves inactive chromatin between the sites of contact, whereas low $P(c_{ij})$ tends to have active chromatin between the sites of contact. High $P(c_{ij})$ also tends to have active chromatin right at the contacting sites, especially in long-distance contacts.
Figure 5.5: (A) The gradient of the probabilities of contact with respect to the surrounding chromatin states was evaluated at every genomic region. As an example, we show a contact map $P(c_{ij})$ together with its respective chromatin states $\sigma_k$ (chromosome 2L 9.15–11.55 Mbp). The gradient was evaluated in four different pairs of sites $i$ and $j$ marked as red circles, and it shows how chromatin states should be altered in order to increase the probability of contact between those sites. The gradient profiles thus suggests the activation or inactivation of regions depending on the location of $i$ and $j$ in the contact map and the sequence of chromatin states. (B) Genome-wide average of probability gradients $\partial P(c_{ij})/\partial \sigma_k$ sorted by distance of contact between $i$ and $j$. An increase of inactive states between pairs of sites typically increases the probability of contact, especially at shorter distances of contact ($d < 300$ kbp). An increase of active states on the sites of contact itself typically increases the probability of contact between sites. The inset shows the average squared gradients, which are indicative of the average magnitude of the gradients, therefore highlighting the chromatin regions with the largest weight in determining probability of contact.
sites \(i\) and \(j\) in the genome separated by distances \(d = 0, \ldots, 80\), we calculated \(\partial P(c_{ij})/\partial \sigma_k\) and averaged them together. In Fig. 5.5B we show the \(\langle \partial P(c_{ij})/\partial \sigma_k \rangle_d\) at each distance \(d = |j - i|\) as a function of site \(\sigma_k\). We find that making the regions between contacting sites more heterochromatic tends to favour more contact \((\langle \partial P(c_{ij})/\partial \sigma_k \rangle_d > 0)\). This is strongest at shorter distances of contact \((d < 300\ \text{kbp})\) and becomes weaker as the distance increases. On the other hand, making the contacting regions more euchromatic also increases the contact probability \((\langle \partial P(c_{ij})/\partial \sigma_k \rangle_d < 0)\), regardless of distance of contact. The heat map also shows that active chromatin immediately outside \(i\) and \(j\) in the shorter range of contact can also bring up the probability of contact. The heat map in Fig. 5.5B was also calculated separating sites with high \(P(c_{ij})\) from low \(P(c_{ij})\) but no qualitative difference was observed between the two, so these features can be generalized to sites with both high and low probabilities of contact.

While the previously calculated \(\partial P(c_{ij})/\partial \sigma_k\) highlights changes in the probability of contact due to sequence mutation, it does not clearly answer the question of which locations in the chromatin sequence are the strongest determinants of contact probability. However, the gradient squared \((\partial P(c_{ij})/\partial \sigma_k)^2\) highlights where the magnitude of the gradient is the greatest, therefore indicating which sequence locations dominate contact probabilities. In the inset of Fig. 5.5B we observe that, overall, the chromatin sequences at the sites of contact are the strongest determinants of contact probability. Nevertheless, the contact probability of sites situated at a shorter distance \((d < 300\ \text{kbp})\) are also strongly determined by the state of the chromatin in between and outside the sites. These results are in agreement with our calculations of chromatin specificity of contacts in Chapter 3.

Third, we performed a similar gradient analysis on the backward model that predicts chromatin states from contact maps. For the backward model the gradient corresponds to \(\partial \sigma_k/\partial P(c_{ij})\), and indicates how a change in contact between two sites \(i\) and \(j\) would be reflective of a change in the chromatin state at site \(k\). We evaluated \(\partial \sigma_k/\partial P(c_{ij})\) at each genomic location, and it can be visualized as a map of gradients with the same size as the contact map. In FIG. 5.6A we show at the top the contact maps and chromatin sequence of a particular genomic location. At the bottom of FIG. 5.6A, we evaluate the gradient \(\partial \sigma_k/\partial P(c_{ij})\) for different \(\sigma_k\) positions in the same genomic location. The heat maps indicate how the chromatin state \(\sigma_k\) would change when increasing \(P(c_{ij})\). In FIG. 5.6B, we averaged genome-wide the gradients separating the locations whose central chromatin state was inactive chromatin from locations whose central chromatin state was active chromatin. This calculation highlights that when the chromatin state is inactive, an increase of contacts between sites situated at the left and the right of the inactive state would reflect that the state is more inactive. A similar pattern is observed for the active state — a decrease of contacts between the sites at the right and left of the active state makes it more inactive. Last, by looking at the average squared gradient we find that the contacts between sites on
Figure 5.6: Gradient analysis of probabilities of chromatin states. (A) At the top, a contact map and inner sequence of a genomic window used as an example to evaluate gradients (chromosome 2L 9.95–10.75 Mbp). At the bottom, the gradient of the chromatin states of three different locations \( \sigma_k \) of the chromatin sequence (marked as red) with respect to the probabilities of contact (in the same genomic window as above). The heat maps indicate how the probabilities of contact \( P(c_{ij}) \) would need to change in order to increase value of the chromatin state \( \sigma_k \) (make it more inactive), or equivalently, how \( \sigma_k \) would change when increasing \( P(c_{ij}) \). (B) Genome-average of gradients for subsets of sites where the central chromatin state is either inactive \( (\sigma_k \approx 1) \) or active \( (\sigma_k \approx 0) \), and genome-average gradients of the central chromatin state. On average, when \( \sigma_k \) is an inactive state \( (\sigma_k \approx 1) \), an increase of contacts between the sites surrounding \( \sigma_k \) makes \( \sigma_k \) more inactive (positive gradient). For the active state \( (\sigma_k \approx 0) \) an increase of contact probabilities between sites surrounding \( \sigma_k \) tends to make the sites more active (negative gradient). The mean-square gradient highlights that the contacts that are most informative about the chromatin sequence state at a given location are the contacts between sites on the left and right of the location.
the left and right of the site of interest are the main determinants of the chromatin sequence at that site.

5.3 Conclusion

In this chapter we have presented a method for predicting chromatin contact maps from sequences of chromatin factors. Notably, although a certain amount of human-guided design choices went into structuring the DNN, a large part of the model selection behind it was done automatically. This is thanks to both the universal approximation theorem, which allows dense neural networks to approximate any non-linear function of the inputs, and to stochastic gradient descent which allows the network to find the locally optimal parameters in an efficient manner. Here we have shown that DNNs can be used to solve computationally hard problems in biophysics such as determining local polymer configurations based on the properties of monomers. One may imagine generalizing these methods to the problem of RNA secondary structure prediction and maybe protein folding.

By fitting a convolutional filter, we were able to reduce the high-dimensional input of multiple chromatin factors to a single 1D chromatin state sequence that was most predictive of structure. Therefore, our method was able to generate a chromatin classification based on both the local occupancy of chromatin factors and the Hi-C structures that these sites create with no additional assumptions. Furthermore, by building an inverse model that predicts the chromatin state sequence from contact maps, we could show that the 2D contact maps can effectively be compressed into a 1D sequence and decompressed back. This supports the theory that local chromatin structure is strongly determined by a 1D sequence of chromatin states.

By analyzing how varying the inputs to the two neural networks (forward and backward models) changed the outputs, we highlighted that chromatin conformation is a non-local problem; the probabilities of contact between each pair of sites depends on the chromatin states in the larger neighbourhood. In general, the presence of inactive chromatin between two contacting sites or active chromatin on the sites themselves and outside their flanks increases the probability of contact. Nevertheless, the probability of contact between sites situated at a larger distance (> 300 kbp) is largely determined by the chromatin states of the contacting sites themselves. These results are consistent with both our findings from Chapter 3 where we analyzed the chromatin sequence specificity of contacts and our findings from inspecting the parameters of the maximum entropy distribution in Chapter 4.

The models presented in this chapter generated better predictions for chromatin conformation than the models from Chapter 2 and 4. This can be attributed to the great number of parameters that can be currently fit in neural networks with little computational cost. Remarkably, instead of using a small sequence of binary chromatin states as we previously did in the maximum-entropy model of Chapter 4, we could now use a much longer chromatin

104
sequence which in turn features a larger number of chromatin factors. In addition, our 1D chromatin annotation can take continuous values between 0 and 1, in contrast with the binary-valued chromatin “spins”, thus providing more information about chromatin states to the model. This also allowed us to calculate the gradient of the output with respect to the input and account for more sensitive mutational effects.

The work presented in this chapter is a proof-of-concept that can easily be extended to capture more biological features of interest. For instance, one could use a larger number of convolutional filters, which would provide a richer biological description of chromatin states by inferring multiple chromatin sequences (i.e. not just a simple “inactive” versus “active” state reduction as was done here). One could also introduce additional types of sequence as inputs, such as genomic annotations, gene expression measurements, genomic mappability and other relevant information to the problem. In addition, one may also be able to introduce non-sequential inputs, that may allow the inclusion of experimental details such as developmental times, cell types or temperatures, and thus allow the modelling of a heterogeneous mixture of cells.

5.4 Methods

5.4.1 ChIP-seq data

For the chromatin sequence profiles, we used the enriched genomic regions of 50 chromatin factors measured with ChIP-seq in 14-16 hour *Drosophila melanogaster* embryos, collected by Roy et al [90]. Specifically, we downloaded the following factors: BEAF, H3K23ac, H3K79Me1, HP1, POF, CP190, H3K27Ac, H3K79me2, HP1b, Pc, CTCF, H3K27me2, H3K79me3, HP1c, Psc, Chro, H3K27me3, H3K9acS10P, HP2, RNA Pol II, GAF, H3K36me1, H3K9me1, HP4, RPD3, H1, H3K36me2, H3K9me2, JHDM1, SU(HW)-HB, H2AV, H3K36me3, H3K9me3, LSD1, Su(var)3, H2B-ubiq, H3K4me1, H4, MDB-R2, ZW5, H3, H3K4me1, H4K16ac(M), MOF, dMi, H3K18Ac, H3K4me3, H4K20me1, NURF301, dRING. This data is publicly available in [http://www.modencode.org/](http://www.modencode.org/) as part of the modENCODE project.

Using a genome binning at a resolution of 10 kbp, we built chromatin sequence profiles by calculating what fraction of the bin was enriched for the chromatin factor. Therefore, the values of the chromatin sequence profiles range from 0 (factor is not present) to 1 (the bin is fully occupied by the factor). Our data was then arranged into \((M \times N_i)\)-dimensional arrays for each *Drosophila* chromosome, where \(M\) is the number of chromatin factors (50) and \(N_i\) is the length of each chromosome \(i\) binned at a 10 kbp resolution.

5.4.2 Hi-C data

DNA contact maps were obtained from the publicly available Hi-C experiments done by Schuettengruber et al [189] (GSE61471), performed on 3000-4000 *Drosophila melanogaster* embryos, 16-18 hours after egg laying. The contact map is an array whose elements \(n_{ij}\) are
the number of times a particular pair of genomic sites were found to be in contact. The contact map we used defined sites to be of a fixed size (10 kbp) and were non-overlapping across the genome. The counts, \( n_{ij} \), at a particular pair of sites \( i \) and \( j \) in the contact map were determined by counting up all sequenced pairs from the Hi-C measurement that fell into those sites (each unique sequence pair was only counted once to reduce experimental bias, as suggested in [35]).

We normalized the contact map using the ICE method [204] so that total number of counts along each row across the contact map was the same. Then, we measured the average number of contacts at each distance of contact and divided the Hi-C counts by it. This correction removed the strong decaying signal as a function of the distance between contacting sites due to the entropic polymer effect. Our final contact maps, which we label as \( P(c_{ij}) \), thus correspond to contact enrichments at a given distance and are proportional to the actual probabilities of contact when the polymer entropy is taken into account.

### 5.4.3 Datasets for training, validating and testing

The genomic data that constitutes our input and output to the models was obtained by scrolling along the genome while capturing overlapping genomic widows of size \( w \) (with \( w = 80 \), corresponding to 800 kbp) for the chromatin contact data, and windows of size \( 3w \) for the chromatin sequence data. More specifically, for each genomic window of width \( w \) we sampled from the Hi-C matrices described above the distance-normalized counts \( P(c_{ij}) \) between all pairs of genomic sites in that window \((w \times (w + 1))/2 \) in total). The sequence data associated to each genomic window consists of the \( M \) chromatin features in the same genomic window, as well as those for the two flanking genomic windows with the same width (thus generating an array of size \( M \times 3w \)). Our datasets included left-right inverted versions of the data, as the directionality of the genome should not influence the relationship between chromatin contacts and sequence. This thus allowed us to build a dataset of sequences and structures with a size approximately twice the length of the binned genome.

For training and validating the models, we have used the data from the *Drosophila* chromosomes 2L, 2R, 3L and the first half of chromosome 3R (from 1 to 12.95 Mbp). From these regions we obtained 13814 pairs of local sequences and structures. We randomly subsampled 80\% of this data as a training set (11052 pairs) and 20\% as a validation set left out of the parameter fitting procedure (2768 pairs). For testing the predictions of the model, we used the remaining second half of chromosome 3R (from 14.95 to 26.91 Mbp) as our testing set, which contained 2112 pairs of sequences and structures.

### 5.4.4 Neural network model details

The two neural networks (forward and backward model) used in this chapter where built with Keras [228], which is a Python-based high-level neural network API that runs on top of TensorFlow [229].
The forward model takes as an input \( M \) chromatin sequences (with \( M = 50 \)) of length \( 3w \) (with \( w = 80 \)) in length and returns contact maps with size \( w \times (w + 1)/2 = 3240 \). A first sigmoid-activated convolutional layer transforms the input into a single \( 3w \)-long sequence whose values range from 0 to 1. This sequence then feeds four ReLu-activated layers of exponentially increasing size, where the last layer is the output layer with same size as the output chromatin map data (Layer 1: 460 neurons, Layer 2: 881 neurons, Layer 3: 1690 neurons, Output Layer: 3240 neurons). The total number of trainable parameters is \( 7,486,472 \). The cost function was defined as the mean-square-error between the predicted and expected output plus a L2 regularization cost for the convolutional filter weights. A dropout of 0.1 was applied to the dense layers during training. This method sets neuron values to zero with a given rate at each iteration in order to avoid overfitting [226]. The model was then fit through stochastic gradient descent [230]. More specifically, the model was fit by dividing the training set into 30 batches and only evaluating the cost function on one batch at a time. At the end of each epoch (i.e. when the 30 batches were used) an independent validation set (left out of the training procedure) was used to evaluate the cost function independently to avoid overfit. The fitting procedure ended when the cost function calculated in the validation set converged. Results were then calculated on the test data set.

The backward model takes as input contact maps with size \( w \times (w + 1)/2 \) and returns one-dimensional chromatin sequences of length \( w \). This dense neural network is made of three layers of exponentially decreasing size, the first two ReLu-activated, and the last layer (output layer) is sigmoid-activated (Layer 1: 943 neurons, Layer 2: 274 neurons, Output Layer: 80 neurons). The total number of trainable parameters is \( 3,336,919 \). The cost function was defined as the mean-square-error between the predicted one-dimensional chromatin states and the original chromatin states. A dropout of 0.1 was applied to the dense layers during training. Similarly to the forward model, the backward model was fit through stochastic gradient descent by dividing the data into 30 batches and only evaluating the cost function on one batch at a time. At the end of each epoch, an independent validation set (left out of the training procedure) was used to evaluate the cost function independently. The fitting procedure ended when the cost function calculated in the validation set converged. Results were then calculated on the test data set.

5.4.5 Gradient analysis of DNNs

We calculated gradients by using a method known as sensitivity analysis [231, 232]. In particular, we followed the DeepTaylor tutorial in \( \text{www.heatmapping.org} \) to calculate for a given output neuron the gradient of the output function with respect to the input variables. First, we exported the values of every layer of our trained neural network to text files. Then, for each neuron in the last layer that we wanted to calculate its gradient, we rebuilt the trained neural network and only included the neuron of interest in the last layer. These neu-
ral networks were built using a minimal neural network implementation that can be found in the script “modules.py” from www.heatmapping.org/tutorial. The back-propagation of the gradient of the neuron in the last layer with respect to the input data was done using the methods in “utils.py” from www.heatmapping.org/tutorial as described in the website tutorial.
Chapter 6

Conclusion

In recent decades, thanks to the advent of high-throughput genomic assays able to measure chromatin features at an unprecedented resolution, it has become increasingly evident that the packaging of DNA inside the nucleus is a tightly regulated process. Genes do not work as single, independent units: they interact with each other, and their communication is mediated by a number of chromatin modifications along the genome, often situated at distant positions from the genes. In the work presented in this thesis, we have connected two types of data: chromatin conformation and chromatin sequence. In particular, we have built several models of increasing complexity that predict chromatin conformation data from chromatin sequence, and we solved the inverse problem, namely predicting chromatin sequence from chromatin conformation data. Furthermore, we have taken advantage of these models to highlight key properties of this biological system.

In Chapter 2, we presented a method for normalizing Hi-C data based on decomposing chromatin contact maps into their principal components and then subtracting modes of variation associated to systematic biases as well as high-frequency noise. We validated this normalization method against other normalization methods and found that our method can best improve the connection between chromatin contacts and bound factors. Interestingly, we evaluated this relationship by building a free-energy model for chromatin conformation, where the free energy of contact (directly linked to the probability of contact) depends on the entropic contribution of bending the polymer into a loop, as well as the energetic interaction between chromatin factors that help stabilize (or disrupt) the loop. Furthermore, our model generated a list of free-energy couplings between chromatin factors that was in agreement with previous findings from other studies.

Chapter 3 followed up on the free-energy model for chromatin conformation by exploring how this model could be improved to include a series of chromatin folding effects that were not yet taken into account. In particular, we showed that our previous model was incomplete because it modelled the probability of contact between pairs of chromatin sites based solely on the chromatin sequence state at those sites. By using a number of information theoretic quantities, we found that other regions of the chromatin sequence are important...
determinants of the probabilities of contact. Specifically, we showed that for short-distance contacts $d \lesssim 300$ kbp, the chromatin sequence inside the loop is a stronger determinant of contact probability than the chromatin at the looping sites themselves. In addition, we showed that regardless of distance of contact, a high contact probability between two sites is associated with heterochromatin factors in between the sites and euchromatin situated at the sites outside of them. Furthermore, the presence of either euchromatin or silent chromatin at the sites of contact was associated with high contact probability whenever those factors are depleted elsewhere.

In Chapter 4, we built a Bayesian model for calculating chromatin conformation as a function of chromatin sequence. In it, we used the principle of maximum entropy to take into account large neighbourhoods of chromatin factors, in contrast with our previous local model presented in Chapter 2 that was based on just two chromatin sequence locations. The chromatin conformation predictions that this model generated were significantly more accurate than the ones derived from the model in Chapter 2. By inspecting the model parameters we found, in agreement with the previous chapter, that distances of contact $d \lesssim 300$ kbp involve a different predictive model than longer distances of contact. In addition, we took advantage of this model to check for the conformational consequences of altering chromatin sequence. We found that a change from inactive to active chromatin is more structurally disrupting than the opposite change. Interestingly, we also found that the non-coding regions of the genome which do not contain genes are the ones that have the largest conformational effect upon being altered. In contrast, changing the chromatin state of gene-promoter regions had little influence on conformation. This finding is in agreement with the notion that non-coding regions have structural roles in the genome. Furthermore, the minimal change to conformation when the chromatin sequence is disrupted in promoter regions suggests that chromatin conformation would not be largely altered when genes are regulated. In this chapter, we also used Bayes’ theorem to invert the probabilities and calculate chromatin sequence from chromatin conformation. This result would allow researchers to infer the underlying chromatin sequence from a single Hi-C experiment, instead of the large number of experiments needed to map locations of bound factors.

Chapter 5 presented a novel way to approach the problem of chromatin folding using state-of-the-art dense neural networks. Dense neural networks have the capability of processing vast amounts of data, and we have used this property to build a model for chromatin conformation that depended on a rich description of chromatin sequence. This model was designed to generate its own biologically relevant classification of chromatin optimized to link the two types of data. This neural network, as well as an inverse-problem version which predicts sequence from structure, generated remarkably precise predictions compared with previous models. In addition, we used gradient analysis to find how a variation in sequence would alter chromatin conformation, as well as how a change in a Hi-C contact map would be reflective of a change in the sequence. This analysis allowed us to test for sequence-
alteration effects in a site-specific manner, as well as the identification of general properties of the chromatin folding mechanism. We found, in agreement with our previous studies, that inactive chromatin between two sites, as well as active chromatin in the sites themselves, increases their probability of contact. Furthermore, we found again that the probability of contact at shorter distances $d \lesssim 300$ kbp depends on larger neighbourhoods than at longer distances.

Overall, the models presented in this thesis have approached the problem of chromatin folding at an increasing level of complexity with consistent results. An extension of our methods would be to generalize them to other species to highlight key regulatory differences in genomes of different length, number of genes, and different chromatin factors. As richer datasets measured at different developmental stages start to become available [233], one may be able to use our models to study how chromatin structure is organized as a function of time. Additionally, using models such as the ones presented in Chapter 5, one could include additional sources of information such as gene expression to highlight details in the mechanisms that relate sequence to structure and structure to function, potentially discovering mechanisms in which gene regulation is able to feed back into the system, altering structure and sequence [74,234].
Bibliography


[38] J. Dekker, “Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture,” *Epigenetics & Chromatin*, vol. 7, no. 1, p. 25, 2014.


