Brown University

Senior Honors Thesis

A Multi-scale Ensemble Model of Chromatin Conformation

Author:
Benjamin Siranosian

Supervisor:
Dr. Nicola Neretti

A thesis submitted in fulfilment of the requirements for the degree of Bachelor of Science

in the

Center for Computational Molecular Biology

May 2015

Declaration of Authorship

I, Benjamin Siranosian, declare that this thesis titled, 'A Multi-scale Ensemble Model of Chromatin Conformation' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Acknowledgements

I would like to thank my advisor, Nicola Neretti, who has been an excellent mentor and teacher for the past two years. Sorin Istrail provided sage advice throughout this project, both about mathematics and the research process as a whole.

Members of the Neretti lab, including Steven Criscione, Feifei Ding and Alan Hwang have assisted at various points along the way.

Finally, I would like to thank my friends and housemates, especially Austin Draycott, Ben Chesler and Lizzy Kinnard. Also my parents and anyone else who has helped me on this thesis journey...

Contents

List of Figures

Chapter 1

Introduction

1.1 Background on chromatin structure

A single copy of the human genome contains over 3 billion nucleotide base pairs. Diploid somatic cells, which make up most of the human body, each contain two copies of this genetic information. Stretched end to end, the DNA in every diploid cell would reach almost 2m in length. However, all the genetic information must fit inside the cell nucleus with a diameter of 6μm - a difference of over 6 orders of magnitude. The packing and organization of cellular DNA has interested scientists since the discovery of metaphase chromosomes by Walther Flemming in 1882.

Genetic material is not packed into the nucleus in a random way. DNA packing is well organized and tightly regulated. DNA packing is also not static; it changes depending on the position in the cell cycle, with specific transcriptional paradigms and in response to external factors. DNA packing and organization is also crucial in any cellular process that involves transcription, underscoring it's importance in the basic function of a cell.

1.1.1 How is cellular DNA organized and how do we study it?

Cellular DNA is organized in a hierarchical manner, with each level building on the previous one. Additionally, each level of organization has a method specifically suited to studying it.

At the most basic level, DNA is a string of nucleotide bases. Although not typically considered with 3D organization, the linear arrangement of nucleotides determines protein sequences, transcription factor binding specificity and other processes. DNA sequencing is used to study this linear arrangement. The recent advances in high-throughput

sequencing technologies have lead to a decrease in cost an increase in amount of data produced, increasing their use in all parts of biology. DNA sequencing is also essential for the study of higher-order chromatin conformation.

In the next level of organization, linear DNA is wrapped around nucleosomes. Nucleosomes are proteins made up of eight subunits, called histones. 147bp of DNA are wrapped around each nucleosomes, and each nucleosome has an average of 50bp of linker DNA before the next, leading to units occupying roughly 200bp of DNA (Luger et al. 1997). The tails of histone proteins can be chemically modified, with varying consequences for transcription and chromatin conformation. Histone modifications and other DNA-associated proteins can be studied with Chromatin ImunoPrecipitation followed by sequencing (ChIP-seq). In ChIP-seq, proteins are cross-linked to DNA, the DNA is fragmented, proteins are pulled down with antibodies and DNA fragments are sequenced. This allows the location of histone modification and protein associations to be tracked across the genome (Johnson et al. 2007).

Nucleosomes are compacted into higher-order structures; although the exact organization at this level remains under debate. Previous work in structural biology suggested a 30nm fiber composed of a repeated structure of nucleosomes. Repeated experiments with more modern imaging techniques have been unable to conclusively identify this fiber, however. (Tremethick 2007)

Higher-order chromatin structure depends on several factors. Position in the cell cycle is the major determinant. In interphase, chromatin is dispersed and occupies much of the volume of the nucleus. Gene transcription and regulation is a hallmark of interphase, and the dispersed yet regulated chromatin structure reflects this. When the cell enters metaphase, chromatin becomes increasingly compact, eventually forming distinct the mitotic chromosomes that are visible with a light microscope. The exact structure of the metaphase chromosome is still unknown, despite extensive studies with microscopy and sequencing-based approaches (Naumova et al. 2013).

Interphase chromatin serves as both a repository for genetic information and a regulatory system for transcription. A major feature of interphase chromatin organization are chromosome territories (CTs). At a basic level, CTs can serve to segregate a particular chromosome or chromosome section to a specified part of the nucleus. Evidence for CTs initially came from light microscopy studies of roundworm and hamster cells (Cremer and Cremer 2010). With the development of more advanced imaging methods, such as Florescence In-Situ Hybridization (FISH) and 3D-FISH, individual chromosome territories could be visualized. This led to the conclusion that individual chromosomes occupy distinct territories in the nucleus. Furthermore, specific sections of individual chromosomes associate non-randomly within the nucleus (Cremer et al. 2008).

1.2 Chromosome Conformation Capture techniques

Imaging-based approaches have provided much of the evidence for higher-level chromatin organization. Newly developed methods based on anchoring cellular chromatin close in 3D nuclear space have greatly extended our knowledge of chromatin biology. Chromosome Conformation Capture, or 3C (Dekker et al. 2002), was the first method in this category. In 3C, formaldehyde is used to cross-link DNA segments to associated proteins and cross-link proteins with each other. This chemically links fragments of DNA that are close in 3D space. A restriction enzyme is then used to cut the DNA. Pieces that were cross-linked will remain in contact. DNA fragments are then ligated under conditions that favor forming circles with linked fragments. The cross-links are reversed by increasing the temperature, and the result is quantified with PCR or Quantitative PCR. In 3C, the sequences of all loci need to be known so that PCR primers can be developed. Because of the combinatorial nature of this, 3C is a "one-by-one" technology and can only investigate small numbers of pairwise interactions.

Circularized Chromosome Conformation Capture (4C) was developed soon after and only required a single site to be of a known sequence, increasing the resolution to "one-by-many" (Zhao et al. 2006). Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al. 2006) further improved the method and used microarray sequencing to create "many-by-many" resolution. 5C is still used for investigating high-resolution chromatin folding in specified genomic regions.

The high-throughput approach, Hi-C, allows for genome wide investigation of chromosome conformation in cis and trans (i.e. within and between chromosomes). Hi-C is an "all-by-all" method and the current state of the art for investigating genome-wide chromatin structure with a sequencing based method (Lieberman-Aiden et al. 2009). Hi-C starts with a pool of cells in culture and typically ends with a 2D contact map representing pairwise interaction frequencies between genomic loci. The three major steps in the protocol are experimental preparation, high-throughput sequencing and data processing.

The experimental steps in Hi-C are similar to the other 'C' methods. Sequencing of the library is next, and most approaches that allow for paired-end sequencing at reasonable read lengths (50-100 bp typically) will be sufficient. Data processing follows. The output of DNA sequencing is millions of paired-end sequencing reads. Ideally, each read corresponds to a single Hi-C ligation event, and therefore a 3D chromatin interaction within the nucleus. To understand which parts of the genome were interacting in the nucleus, it is necessary to map each sequencing read to the genome of the organism being studied. After mapping, reads are assigned to the restriction fragment they originated

from, because restriction fragments are the finest resolution possible in a Hi-C experiment. Interpreting Hi-C data at the restriction fragment level is currently not possible, however, because of the immense amount of sequencing reads that would be necessary. To decrease resolution to a level where interpretation is possible, the genome is divided into fixed-length bins. Restriction fragments (which differ in length and frequency across the genome) are assigned to bins based on their 5' position.

1.2.1 Resolution of Hi-C contact maps

Resolution is a persistent issue when working with Hi-C data. Increasing resolution (smaller bins) allows for greater analysis of local chromatin folding, but can produce contact maps that are noisy and sparse. Decreasing resolution is necessary to interpret interactions between genomic loci on different arms or chromosomes, as these interactions are much less likely to occur in 3D in the nucleus. The maximum resolution at which a Hi-C dataset can be analyzed typically depends on the quality of the library preparation and the depth of sequencing. Additionally, choosing a restriction enzyme that cuts more frequently along the genome allows for a higher resolution, provided it is coupled with a proportional increase in sequencing depth.

Hi-C is an all-by-all method: n loci can form n^2 possible interactions. Increasing resolution by a factor of 2 requires a $2^2 = 4$-fold increase in sequencing depth. Initial studies with Hi-C processed data with a binning resolution of 1 million bases, or 1Mb. Later studies could interpret chromatin interactions at increasingly higher resolutions, form 200kb to 40kb (Dixon et al 2012) all the way to 5kb or 1kb with the latest combined dataset (Rao et al 2014).

1.2.2 Biases of Hi-C contact maps

Although Chromosome Conformation Capture approaches hold promise, the results from such investigations have to be interpreted carefully. Chromosome conformation capture methods have inherent biases in the methodology. Biases associated with GC content, mappability and chromatin accessibility have been examined and methods of correction have been proposed (Yaffe and Tanay 2011, Hu et al. 2012). Iterative correction (Imakaev et al. 2012) tackles the problem by assuming each interval has equal "visibility" across the genome. The contact map is then corrected to meet this assumption as best as possible.

A Hi-C experiment is also done on a large population of cells. Different cells may be in different parts of the cell cycle or expressing different genes, leading to heterogeneous

underlying chromatin conformations. A Hi-C contact map only captures the population average conformation. Nagano et al (2013) recently attempted to tackle this issue through single-cell Hi-C, although the resolution of the resulting experiment limited the possible interpretation.

1.2.3 Characteristics of Hi-C contact maps

Hi-C contact maps from mammals all share characteristic features that are reflective of the shared chromatin structure through evolutionary history. A key global feature of Hi-C contact maps is a decrease in contact probability with increasing genomic distance between two loci. This leads to a higher number of contacts along the diagonal in any Hi-C map. A decrease in contact probability with increasing distance makes intuitive sense: packing of any polymer in a confined 3D space leads to a decrease in contact frequency with linear distance, on average. However, the rate of decrease with increasing genomic distance can give us some clues about the 3D structure of the chromatin polymer.

Plotting the probability of contact between two loci as a function of the genomic distance between them shows a negative linear relationship on a log-log plot. Although different datasets have produced varying numbers, the slope of this line typically falls in the range of -0.8 to -1.08. This finding has given evidence to the theory of a fractal globule conformation of chromatin structure. In the fractal globule model, chromatin is packed in a hierarchical manner - smaller crumples are consecutively packed into larger crumples which in turn are packaged into the large domains that give the final structure. This is in contrast to the much more mixed equilibrium globule model, where a subchain in the globule behaves like a random walk until it hits a boundary and another random walk is started. Further evidence for the fractal globule model is given by the fact that a polymer in an equilibrium globule configuration often contains knots, which would be detrimental to a cell because of the increased time and energy needed to un-knot the polymer before DNA replication (Mirny 2011).

Hi-C contact maps show consistent structure beyond a global probability scaling. An immediately apparent feature is the “checkerboard” structure of the map in sections off the diagonal. This feature immediately suggests chromatin is partitioned into two self-interacting classes that are isolated from each other. Early Hi-C studies confirmed this (Lieberman-Aiden 2009) and designated the two classes as A and B compartments based on the sign of the eigenvector after a spectral decomposition of the contact map.

A and B compartments correlate well with other features of genetic and chromatin structure. A compartments contain euchromatin, early replicating, gene-rich and actively transcribed regions. B compartments contain heterochromatin, late replicating,

gene-poor regions. A recent study suggested that the two-compartments could be further divided into subcompartments based on other genomic features (Rao 2014).

Another consistent feature of Hi-C contact maps is smaller, self-associating regions near the diagonal. These features show up as darkly colored squares near the diagonal and are smaller than A and B compartments (average size of <2Mb). These regions are termed Topologically Associating Domains (TADs) and can be thought of as the building blocks of genome architecture. A TAD typically contains genes and nearby regulatory elements necessary to maintain transcription.

A recent study found evidence for structure within TADs (Rao 2014). High contact frequencies between TAD boundaries point to a folding structure with a defined anchor point before local folding. These "loop domains" were found throughout high-resolution Hi-C contact maps of human cells. Loop domains were found to be anchored by convergent CTCF binding sites on either side of the domain, further supporting the role for this protein as a master regulator of genome organization.

Chromatin structure is well conserved across evolutionary time, and this finding is supported by Hi-C experiments in different mammalian organisms. Global features, such as probability scaling and the presence of compartments and TADs are found in both human and mice Hi-C contact maps. The boundaries of TADs are consistent in genomic blocks that are conserved between mice and human. These results show that genome organization, at least at the scale of TADs, is an evolutionarily old feature that has been persistent over time.

1.3 Reconstructing a 3D structure from a contact map

The features of chromatin structure described above were all interpreted from 2D contact maps which describe the pairwise contact frequency between loci across the genome. Although informative, a 2D contact map is inherently a poor representation of a polymer that occurs in three dimensions. Inferring a 3D structure from a 2D contact map is not an easy task, though. There are several issues that need to be considered with the methods of data processing and reconstruction of the structure. Several groups have proposed algorithms to accomplish this task. I will first review the challenges associated with 3D reconstruction, then discuss different approaches found in the literature with the advantages and shortcomings of each.

Before even thinking about 3D reconstruction, the Hi-C contact maps need to be processed in a way that eliminates as much bias and noise as possible. The data correction algorithms discussed above can be used for this. Data filtering, such as the removal of

genomic intervals that have no reads covering them or intervals that are noisy and have a majority of zero values is also an option and probably necessary for any reconstruction algorithm to perform well.

Contact map resolution is also an important point to consider. A high resolution contact map can allow for reconstruction of fine-level structure along the diagonal - the small, local folds in a polymer. However, the contact data becomes increasingly sparse and noisy farther away from the diagonal. A high-resolution yet sparse dataset cannot be used to reconstruct the 3D organization of large-scale features, such as partitioning into different genomic compartments. To reduce the resolution of a contact map, one can increase the size of genomic intervals represented by a bin. It is also possible to coarse grain a higher resolution dataset. This can be done with a naive approach, such as combining adjacent bin, which is basically equivalent to increasing the bin size. Coarse graining can also be done by combining adjacent bins with a high degree of correlation in a clustering-based approach (Kalhor et al. 2012).

A Hi-C experiment is conducted on a population of cells in vitro. Although attempts can be made to ensure the culture is pure and cells are synchronized, heterogeneity in either the population or position in the cell cycle can lead to the experiment capturing a population of different chromatin conformations. Even in a pure and synchronized culture, differences in signaling environments or gene expression can lead to differences in chromatin conformation. This means that Hi-C is not capturing a single "true" 3D structure. The resulting contact map is the population average across all cells in the culture.

Given that the contact map is an average of an experiment capturing different chromatin conformations, it is an impossible task to reconstruct a single 3D structure that can capture all the variability in the original dataset. Some approaches try to identify a single 3D structure by minimizing the distance between the structure and the experimental contact map; these are called consensus approaches. On the other hand, ensemble approaches attempt to build a population of 3D structures. In a good ensemble mode, the properties of structures on average match the experimental data. Ensemble approaches may be a better way to represent the heterogeneous cellular population that is used in a Hi-C experiment.

Chapter 2

Literature Review

2.1 The current state of 3D reconstruction literature

Inferring a 3D structure that is representative of chromatin interaction data is not a new concept – the first 3D reconstruction algorithm was published in 2009 (Fraser et al. 2009) for use on 5C data. In this section, I will review the current state of the art methods for reconstructing 3D structures from chromatin interaction data. This will only be a subset of the different models that have been developed and applied – for a more complete review, please see Segal et al. (2014) and Dekker et al. (2013). The different types of methods can be classified hierarchically (Figure 2.1). Definitions of the terms and evaluations of each method follow.

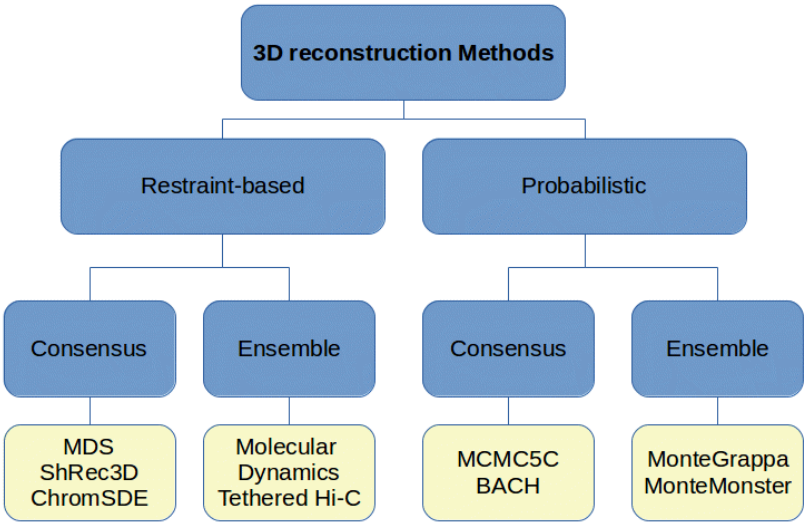


FIGURE 2.1: Methods to reconstruct 3D structures from Hi-C contact maps fall into a number of different categories. See below for a list of definitions and citations.

2.1.1 Definitions

Restraint-based models: Also known as optimization models, these algorithms attempt to place loci in 3D space in a way that is consistent with experimentally observed data. To do so, they typically represent genomic regions as particles with spatial restraints between them, optionally including restraints on things like interactions with the nuclear lamina or spindle pole body. After the model is designed, a scoring function is used to reduce the violations of the imposed constraints, leading to the final 3D structure. Restraint-based approaches can produce both consensus and ensemble solutions.

Probabilistic models: Also known as generative models, these algorithms use assumptions about some probability distributions to specify the model for 3D structure. They then use various sampling methods, such as MCMC, to obtain solutions which are evaluated probabilistically. Probabilistic models can also produce both consensus and ensemble solutions.

Consensus solutions: A consensus solution is a single solution from a restraint-based or probabilistic model that can explain the experimental data. Although consensus solutions are easier to interpret, they fall short because of the problems discussed above.

Ensemble solutions: An ensemble solution contains a large number of 3D structures that, when taken together, can explain the experimental data. An ensemble solution is often interpreted in terms of it's statistical properties, for example,the average radius of gyration across structures in the ensemble. Although an ensemble solution can be difficult to visualize and directly interpret, they better represent the heterogeneous nature of chromatin conformation.

Molecular dynamics (MD) models: A MD approach models DNA conformation in terms of the physical energy of the system. Restraints can be imposed on the spatial distance between loci, the total volume of the structure and localization of centromeres and telomeres, to name just a few. Optimizations like simulated annealing are then used to minimize the restraints and optimize the structure. One advantage of MD models is that they can readily generate an ensemble of solutions by conducting many independent optimizations.

2.1.2 Models in the flowchart

Multi-dimensional scaling (MDS) is a general class of methods aimed at solving the following problem: given a matrix of distances between points and a dimension

n, place each point in n-dimensional space such that the distance matrix is optimally preserved. This method is readily applied to chromatin 3D reconstruction as a distance matrix can be calculated from the interaction frequencies. There have been several MDS implementations in the literature. The most usable software is the Pastis package by Varoquaux et al. (2014). The authors also included metric MDS and Poisson models as alternative reconstruction options.

Shortest path reconstruction in 3D (ShRec3D) is an adaptation of a typical MDS solution. Lesne et. al (2014) used the concept of shortest path in graph theory to construct the distance matrix from chromosomal contact frequencies. This method improved the distance values assigned to interactions with very low contact frequencies in sparse sections of the contact map, which would approach infinite distance with decreasing contact frequency. MDS is then used to construct a 3D structure from the shortest path distance matrix. This method claimed to be more consistent than typical MDS and less computationally intensive than methods like ChromSDE and BACH, but it's applicability to other datasets has proven to be problematic as the code is not easily customizable.

ChromSDE (Zhang et al. 2013) reformulates the problem of converting a distance matrix to a set of 3D points as a semi-definite programming technique. This guarantees a correct 3D structure in a noise-free case and can be computed in polynomial time. ChromSDE also attempts to optimize the contact frequency to distance function by finding the best scaling parameter α for the distance-to-contact-frequency function. It assumes the difference between the predicted contact frequencies and experimental contact frequencies is unimodal and depends on α. A golden section search is then carried out in the area of $0.1 \leq \alpha \leq 3,0$ to find the best parameter. This step is a unique feature of ChromSDE and is one of the biggest advantages of the algorithm.

ChromSDE also attempts to quantify if more than a single structure is necessary to explain the experimental contact frequencies (basically, if the assumption of a single consensus structure is violated). The authors propose a consensus index $0 < c < 1$ that quantifies how well a single 3D structure fits the input contact frequency matrix. Briefly, the consensus index considers if the calculated distance matrix satisfies the triangle inequality and how good a 3-dimensional representation is compared to a n-dimensional representation.

Tethered Hi-C (Kalhor et al. 2012) used a molecular dynamics (MD) approach to 3D modeling of the human genome. This was not the first MD approach to 3D modeling by far, although most other examples have been on simpler organisms like yeast (Tjong et al. 2012 and others). In Kalhor et al. 2012, the authors define the genome by the positions of spheres for each genomic loci. The scoring function is composed of nuclear

volume restraints, excluded volume restraints and contact restraints. This function is optimized using a simulated annealing with MD and conjugate gradient optimizations. When all restraints are satisfied, the scoring function is equal to zero. Starting the simulation from 10,000 random initial configurations and carrying out the optimization on each independently leads to an optimized population of 10,000 genome structures, which can then be analyzed for statistical properties.

Bayesian Inference of Spatial Organization of Chromosomes (BACH) is a MCMC-based probabilistic model developed by Hu et al. (2013). The authors use a more complex probabilistic interpretation of Hi-C data that is designed to remove systematic biases included in Hi-C data. This is based on the previous work by the authors (Hu et al. 2012). The details of this probabilistic model are defined in Chapter 4.

With this probabilistic interpretation, BACH uses a three-stage process to draw samples from the posterior distribution of structures. First, a Poisson regression approach is used to assign initial values for the β nuisance parameters in the probabilistic model. Then an initial structure conformation is generated based on the initial nuisance parameters with sequential importance sampling. Finally, the structure and nuisance parameters are refined with a Gibbs sampling, hybrid Monte Carlo and adaptive rejection sampling technique. Details of this model can be found in the supplementary text of Hu et al. (2013).

MCMC5C is one of the first ensemble MCMC-based probabilistic models developed in Rousseau et al. (2011). MCMC5C was designed to model the 3D conformation of specific genomic regions assayed with 5C (hence the name), but it is also generalizable to Hi-C data. MCMC5C represents each genomic loci by a point in 3D space with a random initial conformation. At each step in the Markov Chain, a single point is perturbed by moving it within a sphere of a certain radius. The probability of the structure is evaluated and the move is accepted according to the Metropolis-Hastings algorithm. After assessing mixing of the Markov Chain, the independent samples after a certain number of iterations form the members of the structure ensemble. Although this model is simple, it has several good ideas that I will draw from in developing my own model, namely the definition of the probability of a given structure.

2.2 The MonteGrappa algorithm

MonteGrappa is a MCMC-based ensemble model detailed in Giorgetti et al. (2014). Although designed for 5C, it is also applicable to Hi-C data at high resolution for small

genomic regions. Both the probabilistic model of chromatin contacts and the representation of genomic regions in space is different from the other models discussed so far. As this algorithm is a key step in my multi-scale 3D reconstruction process, I will discuss it in detail here.

To represent genomic loci in 3D space, MonteGrappa uses a "beads attached by a fixed linker" model. A certain genomic region is mapped on to each bead. Beads are connected by a linker of fixed length that sets the scale of the simulation. In the Giorgetti et al. (2014), each bead represents consecutive 5C restriction fragments summing to 3kb in total sequence length. The authors propose that the algorithm can be applied to high-resolution Hi-C data by mapping each bin in the contact map to a bead. However, MonteGrappa has not yet been applied to Hi-C data. Extending the algorithm to work on the new high-resolution Hi-C dataset is one of the contributions of this work.

In contrast to other models that attempt to convert euclidean distance between points to a measure of contact probability, MonteGrappa takes the view that two loci are "in contact" for a given conformation if the euclidean distance between two points, d_{ij}, is less than R. This definition of "contact" works well for this model because an ensemble of structures are generated. The contact probability for a pair of loci is defined as the proportion of structures in the ensemble where the loci are closer than R. The ensemble of structures functions much like a population of chromatin structures in a cell culture. If two loci are close enough in 3D space to facilitate protein-protein interactions, they are close enough to cross-linked by paraformaldehyde and contribute to the Hi-C contact map.

A spherical-well potential (Figure 2.2) is used to represent the interactions between beads in the model. If two beads are farther than the interaction distance R ($d_{ij} > R$), they do not interact. If the beads are closer than the hard-core repulsion radius r_{HC} ($d_{ij} < r_{HC}$), they interact with infinite energy, effectively disallowing interactions less than r_{HC} and ensuring beads do not overlap. If $r_{HC} < d_{ij} < R$, the beads interact with energy β_{ij}, which can be negative or positive depending if attraction or repulsion is modeled. The initial choice for the β values is determined by equation 2.1, but they are updated to be more accurate with each iteration of the algorithm.

2.2.1 Algorithmic steps

The MonteGrappa algorithm follows several iterative steps to generate an accurate ensemble of 3D chain conformations.

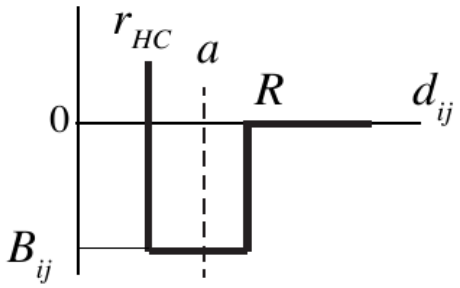


FIGURE 2.2: The interaction energy of two beads i and j is calculated according to a spherical-well potential. The energy can take on values of infinity, β_{ij}, or 0 depending on the distance d_{ij} between the beads. Figure reproduced from Giorgetti et al. (2014).

1. Initial choice of β

The initial interaction energies for beads in the model are defined according to:

$$\beta_{nm} = \frac{\beta_0}{L} \ln(\lambda^{-3/2}(p_{Hi-C}(nm))^{-1}) - 1) \tag{2.1}$$

where L is the number of possible interactions between the two segments, λ is the length of the chain that separates the two segments and β_0 is the initial energy scale in k_BT units, where k_B is Boltzmann's constant (Giorgetti et al. 2014). Although the β values are updated in each iteration of the algorithm, good initial choices speed convergence.

2. Monte Caro sampling

A Markov Chain Monte Carlo sampling of the conformation space is used to generate an ensemble of structures. Initially, the beads are placed in a linear order. Each Monte Carlo move changes the organization of the beads. Several types of moves are allowed (Figure 2.3). The probability of each move type can be defined.

- A flip is the rotation of a backbone atom chosen at random around the axis defined from the preceding and the following one. It is efficient because it is local (i.e., it changes only the positions of few atoms of the chain).
- A pivot move changes a bond angle at random. This is not effective when sampling among compact conformations because it is a non-local move which is likely to produce clashes between atoms.
- A multiple pivot move is an extension of pivot moves, which changes at random a set of consecutive backbone bonds.

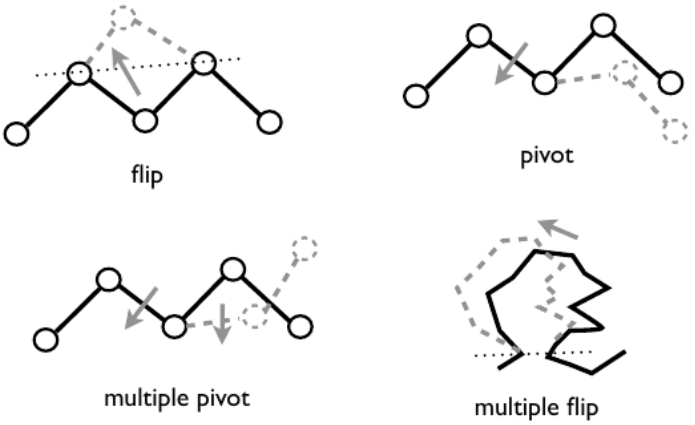


FIGURE 2.3: Moves (changes to the polymer chain) allowed in the MonteGrappa algorithm. These are described in detail below. Figure reproduced from the MonteGrappa software manual.

- A multiple flip consists in choosing two non-consecutive atoms of the backbone and moving the backbone atoms in between around the axis defined by the two.

Each potential move is applied to a random set of beads in the conformation. The energy of the polymer is calculated according to the spherical-well potential described above. A move is accepted according to the Metropolis criteria (Metropolis et al. 1953). The Metropolis algorithm dictates a potential move should always be accepted if the energy of the system decreases (the polymer better matches the experimental contact map). To prevent the sampling from getting stuck in local maxima, potential moves that increase the energy of the system (and therefore move the polymer further from the experimental contact map) are accepted with probability equal to the ratio of the energy before and after the potential move. Moves that make the system slightly higher in energy will usually be accepted, while moves drastically increasing the energy will only rarely be accepted. In each iteration of the sampling step, 5,000 conformations are generated by conducting 5×10^8 steps and recording the conformation of the polymer every 1×10^5 steps.

3) Minimization of the χ^2 function

After an ensemble of 5,000 conformations is generated, the algorithm measures how close the ensemble represents the experimental contact map driving the simulation. This is done with a χ^2 function to test how the simulated distribution of contact probabilities differs from the experimental data:

$$\chi^2 = \frac{1}{n} \sum \frac{(p_{Hi-C}(i,j) - p_{model}(i,j))^2}{\sigma(i,j)}$$

Where $p_{Hi-C}(i,j)$ is obtained from the experimental Hi-C counts and $p_{model}(i,j)$ is obtained from the generated ensemble of conformations. $\sigma(i,j)$ can be obtained from the variance in contact maps from duplicate Hi-C experiments.

The ensemble of conformations is also used to record how the B matrix could be improved to decrease the χ^2 function. A random update scheme is carried out where values of B randomly changed and the χ^2 function is recalculated. If χ^2 decreases, the change is kept and rejected otherwise. 1,000 updates are attempted on the current ensemble of conformations, after which the ensemble cannot be thought of as an accurate representation of the system and a new set of structures are necessary. This step makes the algorithm converge faster than having to re-sample the conformation space after every update to the β matrix (Norgaard et al. 2008).

After optimizing the β matrix, a new Monte Carlo sampling of the conformation space is conducted. This iterative algorithm is repeated until χ^2 converges to a minimum and the algorithm terminates. The ensemble of structures can then be analyzed for statistical properties of the conformations.

2.2.2 Comments on the MonteGrappa algorithm

I think the MonteGrappa algorithm is well-designed and suited for generating ensembles of 3D structures for Hi-C data. The algorithm is only applicable to very high-resolution Hi-C and 5C datasets because of the fixed-linker polymer model it uses. This model assumes adjacent loci are constrained to a fixed distance. This assumption is invalid for lower resolution Hi-C datasets - it's impossible to constrain the distance between loci that are tens or hundreds of kilobases apart. For the high-resolution data that I have been analyzing for this manuscript, the MonteGrappa algorithm performs extremely well (see 3). It can generate a ensemble of structures that accurately reconstruct the experimental contact map given.

Chapter 3

Applying MonteGrappa to a single loop domain

3.1 An ensemble analysis of chromatin conformation from a single loop domain

As a first step in the data analysis process of my thesis research, I chose to examine the performance of the MonteGrappa algorithm in generating an ensemble of structures from high-resolution Hi-C data. This is a novel analysis because the MonteGrappa algorithm was proposed for 5C data and has not yet been used to analyze Hi-C contact maps. Additionally, no group has published on 3D reconstruction using the latest data presented in Rao et al. (2014).

Through analyzing this data, I hoped to investigate a few questions. First, how does the MonteGrappa algorithm perform on Hi-C data? The authors provide an implementation of the algorithm tailored to 5C, but I needed to make some changes to get it to work on the data I was using. I was also interested in investigating how large of a structure the implementation could handle in a reasonable amount of time and how the results would differ when using the Hi-C data at different resolutions. Second, Giorgetti et al. (2014) present results that 3D structures from an ensemble calculated from 5C data show clustering with biologically relevant results. I was interested to test if clustering was also visible from Hi-C data, and whether clustering might differ in different parts of the chromosome or regions with different epigenetic marks. Finally, I am interested in methods to compare an ensemble of 3D structures and wanted to expand on the results presented in Giorgetti et al. (2014). The authors only use a simple clustering method (pairwise RMSD calculation followed by hierarchical clustering) and do not provide any

ways to visualize an ensemble of structures or compare regions of the conformations that are locally similar to each other. Although I have not yet accomplished all of these goals, they are still interesting topics for future research projects and should prove worthwhile for other members of our lab to investigate.

3.1.1 Description of the data

Rao et al. (2014) published the highest resolution Hi-C dataset to date. The authors combined experimental data from hundreds of individual Hi-C experiments in human GM12878 B-lymphoblastoid cells to produce a dataset with 4.9 billion pairwise contacts and a binning resolution up to 1kb. Previously published Hi-C datasets had a maximum binning resolution of around 20kb (Dixon et al. 2012), so this was a large improvement. The authors investigated a number of interesting features of the contact maps, including the presence of 10,000 chromatin loops – pairs of regions that are closer in 3D space with each other than the loci between them.

Chromatin loops are identified by "peaks" in the Hi-C contact maps, or small regions off the diagonal that have a significantly enriched contact frequency when compared to the surrounding background (Figure 3.1). The presence of some chromatin loops was confirmed through 3D-FISH, indicating that contact map peaks predict biologically relevant features. Although the resolution of the contact maps can be pushed as high at 1kb, I have been evaluating it in my analysis from 5-25kb. Reducing the resolution decreases the noise in the data and allows 3D reconstruction programs to run faster with less points to update.

3.1.2 Applying MonteGrappa to Hi-C data

To test the performance of the MonteGrappa algorithm and implementation, I first looked at a single chromatin loop domain on human chromosome 4 using the GM12878 combined dataset. This domain was chosen for its clear loop definition and proximity to loci our lab has previously investigated on chromosome 4. The region chr4:12,820,000-13,580,000 was chosen because of a peak identified in the upper corner of this region (Figure 3.2). I examined the data at 5, 10, 25 and 50kb resolution, with 152, 77, 31 and 15 bins for the domain, respectively. The number of bins was equal to the number of beads in 3D space optimized in the model. This domain will be referred to as the "single" domain because it has a single contact peak.

To extend MonteGrappa to work on Hi-C data, some adjustments had to be made. First, I normalized the Hi-C data by the method suggested in Giorgetti et al. (2014): contact

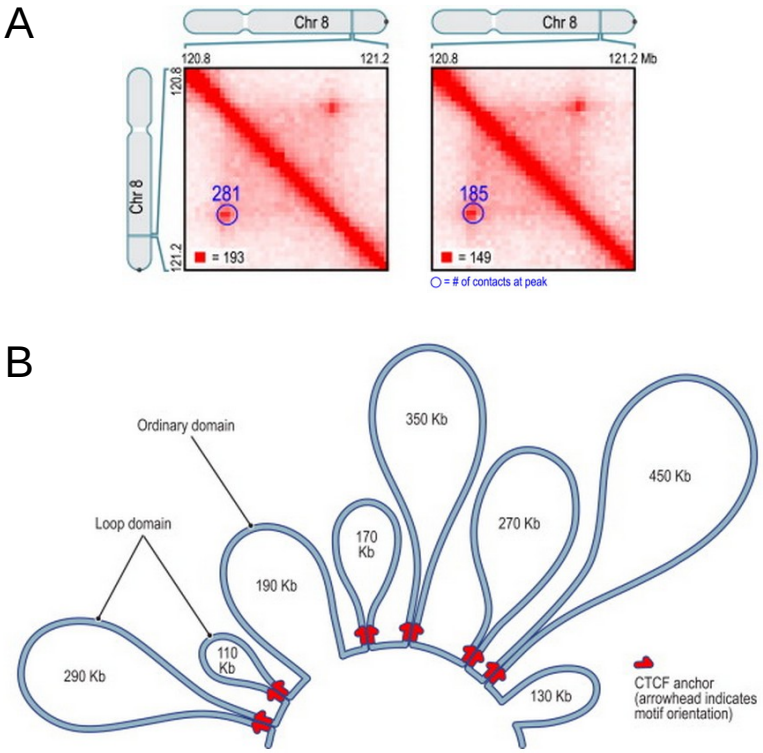


FIGURE 3.1: A) Loop domains are defined by a "peak" with contact frequency significantly enriched above the surrounding background. An automated algorithm was used to call peaks in this dataset. B) Schematic of higher-order DNA folding that contributes to the contact maps observed for normal domains and loop domains. Figure reproduced from Rao et al. (2014).

values were divided by the mean contact frequency between two adjacent loci. Adjacent loci are assumed to be "in contact" in all cells, and the mean value of this contact frequency can provide an estimate of the relative contact frequency of off-diagonal interactions. It is also necessary to consider some parameters in the simulation. Giorgetti et al. (2014) optimize the values for bin contact and hard-core repulsion, R and r_{HC} respectively and find values that minimize the χ^2 value between replicate simulations. I found the default values to work reasonably well on Hi-C data, so they were not modified. It would be interesting to see how changing the R and r_{HC} parameters changes the results of the simulation, especially when using Hi-C data at different resolutions. However, the χ^2 values in my simulations converged to a value very close to 0, indicating that optimization of these parameters might not be necessary.

I ran the MonteGrappa algorithm on the single domain at multiple resolutions. Only the 25kb and 50kb trials had an χ^2 value converge within a reasonable amount of time

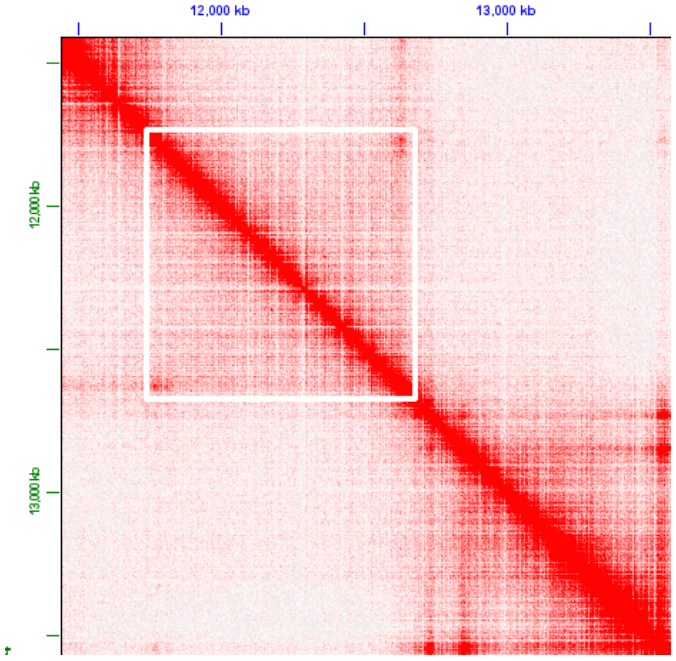
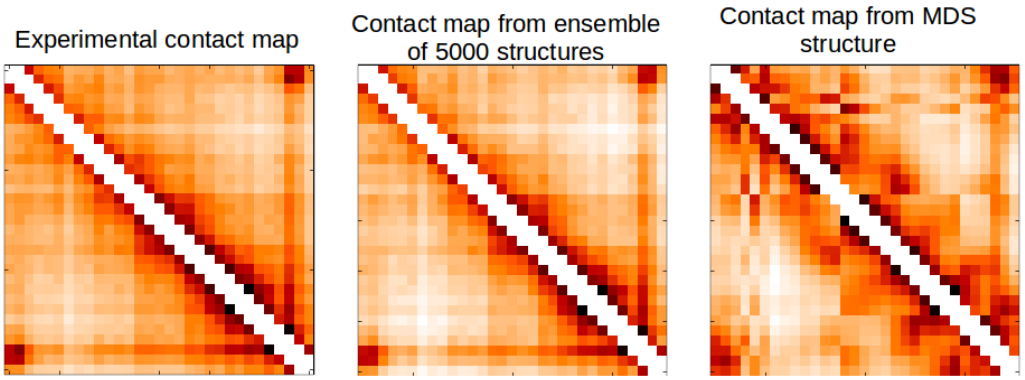


FIGURE 3.2: The "single" domain analyzed in this study. The genomic region of chr4:12,820,000-13,580,000 is highlighted in white. Notice the peak with increased contact frequency relative to the background in the upper corner.

(two days runtime on a personal computer, using a single core on a 4.5 GHz processor). The 5kb and 10kb resolution contained too many data points to produce reasonable structures within the time allowed. I investigated the heatmaps recomputed from the ensemble of structures, they did not represent the experimental input remotely. Hence, I will limit my analyses of the single domain to 25kb resolution data for the following points.

3.2 Performance of the MonteGrappa algorithm

I generated an ensemble of 500 3D structures representing the single domain using the MonteGrappa algorithm. To compare the ensemble of structures to the original Hi-C data, I calculated an average contact map from the ensemble. Each point c_{ij} in the average contact map is the proportion of the structures where the distance between points d_{ij} is less than R. Visually, the two contact maps are very similar (Figure 3.3). I also ran a MDS 3D reconstruction algorithm (Varoquaux et al. 2014) for comparison. A contact map was calculated from the MDS solution according to equation 4.2. The MDS solution does not match the experimental data well, either by visual inspection or mathematical comparison.



Sum of absolute differences between normalized contact maps	
Experimental - ensemble	1.153
Experimental - MDS	7.441

FIGURE 3.3: Contact maps from the experimental Hi-C data, an ensemble of 5000 structures generated with MonteGrappa and a single MDS structure are shown. Each is normalized so that the rows in the contact matrix sum to 1. The contact map from the ensemble of structures matches the experimental data very closely. Even small details, like increased contact frequency in the lower left corner of the "peak" are captured. The MDS reconstruction captures the overall structure of the domain and has a "peak" region, but the finer details are missed. To quantify the difference between the methods, I calculated the sum of the absolute difference between the normalized contact maps. The ensemble of structures performs much better under this measure.

The result in Figure 3.3 shows that an ensemble 3D reconstruction algorithm is necessary to capture all the variation in a high-resolution Hi-C dataset. A single 3D structure cannot reconstruct the fine details that are present in the experimental data, likely arising from several clusters of chromatin conformations present in the cell culture used in the experiment. This result prompts several interesting areas for future research:

- Clustering (Section 3.3) of 3D structures identifies several "classes" of structures with similar properties. A cluster of structures might define a particular feature of the contact map, such as a square of increased contact along the diagonal or off-diagonal looping interactions. I would generate a contact map only from the structures in a given cluster and compare it to the experimental contact map to investigate this point further.

- Recent research (Williamson et al. 2014) has identified discrepancies between data generated with chromosome conformation capture technologies and FISH imaging. An ensemble of 3D structures could be used to resolve these discrepancies. Some of the clusters of structures might be consistent with the FISH data presented, but other clusters might show structural properties that were not captured with

imaging. The contact map from Hi-C or 5C is an average across all chromatin conformations present in the cell culture, and an ensemble of 3D structures could help tease out the heterogeneity in the population.

3.3 Clustering and investigation of structure ensembles

The MonteGrappa algorithm produces a large ensemble of 3D structures, each of which are independent draws from the Markov Chain. The obvious next step is to compare the structures to each other and analyze properties of the ensemble. To compare 3D structures, a measure of distance is necessary. A proper distance function for 3D structures should satisfy the properties of distance functions: non-negativity, identity of indiscernibles, symmetry and the triangle inequality.

A commonly used method of comparing 3D structures is the Root Mean Squared Deviation, or RMSD. RMSD is a pairwise measure of distance. Theoretically, the algorithm attempts to find the optimal translation, rotation, and scaling of two structures such that the root mean squared pairwise distance between points in the structures is minimized:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{3.1}$$

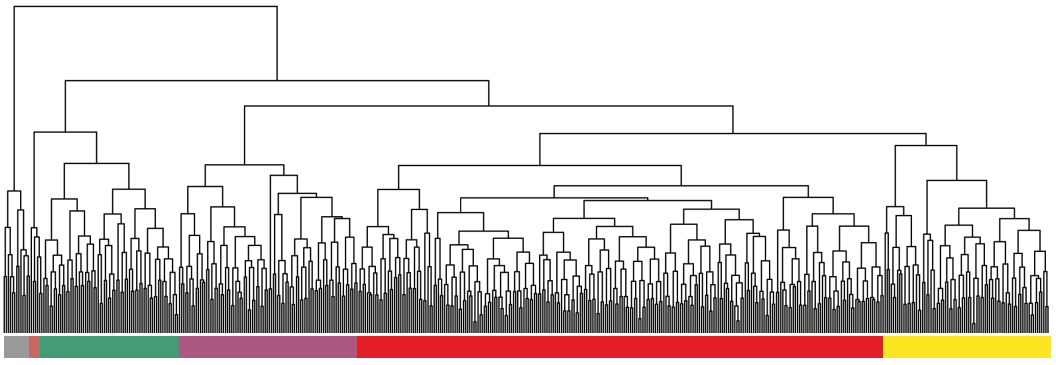
Where δ is the pairwise distance between points after the structures have been superimposed and N is the number of pairwise comparisons considered. RMSD is used extensively in protein structure analysis and has been applied to comparing Hi-C 3D reconstruction methods.

Other measures of distance between 3D structures operate on contact maps calculated from the structures. One possibility is the Contact Map Distance. With this method, a binary contact map is calculated from the structure information. Points closer than a threshold distance are assigned as in contact and points further are defined not to be in contact (Vendruscolo et al. 1997). Two contact maps can then be compared by taking the sum of entries where they are different. Another possibility is Maximum Contact Map Overlap (CMO) which attempts to find the maximum number of shared edges between contact maps (Andonov et al. 2011, Caprara et al. 2004). However, CMO is designed for situations when the structures also need to be locally aligned. Each member of the ensemble has the same number of points, so local alignment is not necessary.

After defining a measure of distance, the next step is to calculate a matrix of pairwise distances between structures. Hierarchical clustering with complete linkage is a logical

solution to this problem, as I am trying to find . The resulting dendrogram shows clear clusters, but it is difficult to define what the clustering cutoff or the members for each cluster should be.

One solution is to define a height in the dendrogram to cut clusters at. Another option is to chose a value k for the desired number of clusters and set the cutoff height to a value that results in k clusters. Both of these methods are relatively brute and require manual input, however. There have been ideas for clustering different protein structures presented in the literature, including an algorithm to minimize the number of clusters and the spread across each cluster presented in Kelley et al. (1996). In this first example, I set k to 6 and analyzed the resulting clusters (Figure 3.4). After examining the structures in each cluster, clear physical properties were apparent. Each cluster theoretically defines a "class" of structures that are necessary to capture the true nature of a chromatin loop domain.



Cluster	Characteristic	%
	elongated center, folded left end	0.03%
	elongated structure	0.01%
	folded at both ends	13%
	folded in center and at right end	17%
	very folded, typical loop domain	50%
	ends always close	16%

FIGURE 3.4: An ensemble of 500 structures was generated with the MonteGrappa algorithm to reconstruct the single domain. Clustering was done with agglomerative complete linkage on RMSD, $k = 6$ clusters were split and analyzed for structural properties. Structures in each cluster displayed distinct properties that were well conserved across the cluster. Properties in the small clusters (gray and brown) were the best defined.

After defining clusters of structures, it is desirable to visualize them in an informative way. Once again, ideas from protein structure investigation are useful here. Several methods have been developed to visualize ensemble NMR protein structures, including simple superposition and visualization of structures that vary in width according to the

variability of the structure at a given point (Kelly et al. 1996, Sutcliffe 1993). A naive approach is to simply superposition the structures such that the distance between them is minimized and visualize the ensemble (Figure 3.5). Another goal of ensemble analysis would be to determine which parts of the structure are most conserved and which parts are most variable. This question could be answered by computing RMSD in a sliding window across a number of structures. Windows with the lowest RMSD values would be the most conserved across the ensemble, and windows with the highest RMSD values would be the most variable.

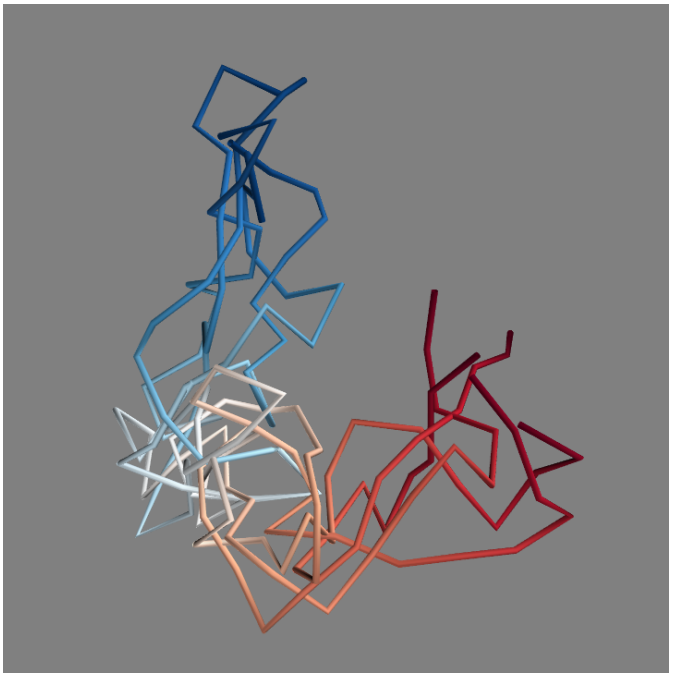


FIGURE 3.5: 5 structures that cluster close to each other are superpositioned and visualized. Each structure is colored from blue to red along the length of the polymer. The structures all have similar shapes overall and appear to be tightly folded in the center.

I have demonstrated that the MonteGrappa algorithm can be applied to high-resolution Hi-C data. The ensemble of structures generated shows interesting properties and clustering.Further investigation is necessary to extract true biological insight beyond this first pass analysis. Tests like the proportion of ensemble members where two enhancer/promoter loci are in contact could to be related to the percentage of cells in the Hi-C experiment experiencing that looping interaction.

Chapter 4

A Multi-Scale Ensemble Model of Chromatin Conformation

4.1 A high-level overview of the MonteMonster algorithm

In the previous section, I extended the MonteGrappa ensemble 3D reconstruction method to work with Hi-C data in local genomic regions. Ideally, a 3D reconstruction method would be applicable to entire chromosomes or the entire genome, generate an ensemble of structures that explain the experimental data and take advantage of the fact that Hi-C data is very high-resolution near the diagonal, but degrades in quality with longer-range interactions. In this section, I present a theoretical framework to perform an ensemble-based 3D reconstruction which includes multiple scales, from local high-resolution regions to long-range low-resolution interactions.

The 3D reconstruction method I have developed uses Markov Chain Monte Carlo sampling, and tackles a number of monster problems. I have therefore decided to name the algorithm **MonteMonster**.

MonteMonster is a two-step algorithm to solve the 3D reconstruction problem for a single chromosome. Step 1 involves breaking the chromosome into domains, which can be defined with a number of external algorithms, running the MonteGrappa algorithm on each domain and generating an ensemble of solutions. Step 2 involves using the ensemble of structures generated from each domain to capture the 3D structure of the entire chromosome. A member from the ensemble from each domain is placed in 3D space. The structures are then perturbed through translation, rotation, scaling and swapping members of the domain ensemble in a MCMC sampling scheme. The probability of the experimental data given the conformation of structures is tested after each proposed

move, and the move is accepted according to the Metropolis criterion (Metropolis et al. 1953).

4.2 Definitions of elements and parameters

A Hi-C experiment produces an experimental contact matrix for a given chromosome $\hat{C}$. $\hat{C}$ is a $n \times n$ square and symmetric matrix with entries of 0 along the diagonal. n is determined by the length of the chromosome in basepairs divided by the resolution (or bin size) of the contact matrix, r. Each element in $\hat{C}$, $\hat{c}_{ij}$ represents the pairwise interaction frequency of locus i and j.

If $\hat{C}$ is an *unnormalized* contact matrix, $\hat{c}_{ij}$ is the number of reads from a Hi-C sequencing experiment that mapped to the interaction between i and j. If $\hat{C}$ is an *normalized* contact matrix, $\hat{c}_{ij}$ is the interaction probability between i and j.

I define an in-silico or "true" contact matrix, C, from the simulated 3D structures. C has the same properties of $\hat{C}$ (square, symmetric, diagonal 0).

A structure s is defined by a set of points in three-dimensional space. To convert a structure s to an in-silico contact matrix C, the function:

$$c_{ij} = f(D_s(i,j)) \tag{4.1}$$

is used, where $D_s(i,j)$ is the euclidean distance between points i and j in the structure s and $f(\cdot)$ is a function of the form:

$$f(D_s(i,j)) \propto \frac{1}{D_s(i,j)^\alpha} \tag{4.2}$$

With α defining the inverse relationship between euclidean distance and contact probability. $0.1 \leq \alpha \leq 3.0$, but most commonly $\alpha = 1.0$.

An experimental contact matrix is divided into a set of domains, $D = \{d_1, d_2, ..., d_m\}$. A set of domains must cover the entire chromosome and not overlap with each other. This means that domain d_1 must begin at locus 0, and domain d_m must end at locus n. If d_1 ends at locus i, domain d_2 must begin at locus $i + 1$.

In step 2 of the MonteMonster algorithm, I define a chromosome structure S as a set of structures $S = \{s_1, s_2, ..., s_m\}$ that represent the m domains in the chromosome. Associated with these structures are a set of:

Translations: $T = \{t_1, t_2, ..., t_m\}$

Rotations: $\rho = \{(\rho_x, \rho_y, \rho_z)_1, (\rho_x, \rho_y, \rho_z)_2, ..., (\rho_x, \rho_y, \rho_z)_m\}$

and scales $\Theta = \{\Theta_1, \Theta_2, ..., \Theta_m\}$

That modify the underlying coordinates in S to produce the final chromosome structure. These are all of the elements in the multi-scale model.

4.3 Probabilistic interpretation of Hi-C data

Hi-C is not a perfect experimental protocol. The observed experimental contact map $\hat{C}$ is distributed according to some probability function function. Several examples are in the literature, I will cover the normal distribution function discussed in Rousseau et al. (2011) and the Poisson function proposed in Hu et al. (2013).

In Rousseau et al. (2011), the authors propose $\hat{C}$ is distributed according to the "true" chromatin contact matrix C:

$$Pr(\hat{c_{ij}}|c_{ij}) = N(\hat{c_{ij}}; c_{ij}, c_{ij} + k) \tag{4.3}$$

where N is the normal distribution $N(x; \mu, \sigma^2)$ and k is a small constant, 10 in their algorithm, that prevents entries in C with low read numbers being assigned too low of a variance.

In Hu et al. (2013), the authors propose that Hi-C reads follow a Poisson distribution according to the distance between points and other parameters designed to correct systematic biases present in Hi-C data. This assumes c_{ij} follows a Poisson distribution with rate θ_{ij} where:

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log(d_{ij}) + \beta_{enz} \log(e_i e_j) + \beta_{gcc} \log(g_i g_j) + \beta_{map} \log(m_i m_j) \tag{4.4}$$

β_0 is a constant, β_1 measures the negative association between contact frequency and distance (equivalent, but not equal to, the α parameter discussed above) and $\beta_{enz}, \beta_{gcc}\beta_{map}$ are the coefficients to correct for the enzyme effect, GC content effect and mappability effect biases discussed in the manuscript. Once θ_{ij} is defined:

$$Pr(\hat{c_{ij}}|\theta_{ij}) = \frac{e^{-\theta_{ij}} \theta_{ij}^{\hat{c}_{ij}}}{c_{ij}!} \tag{4.5}$$

Other sources (Segal et al. 2014) note that the correction of biases from enzymes, GC content and mappability should be treated as pre-processing steps of the experimental

heatmap and need not be incorporated into the probabilistic model. In my implementation of the Poisson model, I will only use the β_1 term in the definition of θ_{ij}. I will test both of these probabilistic interpretations of Hi-C data in the MonteMonster algorithm.

4.4 A Bayesian approach to sampling 3D structures

The MonteMonster algorithm samples 3D structures from the sample space of all possible structures. I am interested in the posterior distribution of structures, given the experimental contact map data. We cannot compute this distribution exactly, however, and a MCMC sampling algorithm is used instead. Using Bayes rule, we can transform the posterior distribution:

$$Pr(S|\hat{C}) = \frac{Pr(\hat{C}|S)Pr(S)}{Pr(\hat{C})} \tag{4.6}$$

Applying an uninformative prior (we have no knowledge about the distribution of structures, therefore we assume they are all equally likely) and assuming that the probability of the experimental data is constant, we obtain:

$$Pr(S|\hat{C}) = Pr(\hat{C}|S) \tag{4.7}$$

This probability can be calculated exactly under one of the probabilistic interpretations of Hi-C data described above. For the Normal distribution:

$$Pr(S|\hat{C}) = \prod_{1\leq i\leq j\leq n} Pr(\hat{c}_{ij}|c_{ij},\sigma_{ij}) = \prod_{1\leq i\leq j\leq n} N(\hat{c_{ij}}; c_{ij}, c_{ij}+k) \tag{4.8}$$

And for the Poisson distribution:

$$Pr(S|\hat{C}) = \prod_{1\leq i\leq j\leq n} Pr(\hat{c}_{ij}|c_{ij},\sigma_{ij}) = \prod_{1\leq i\leq j\leq n} e^{-\theta_{ij}}\theta_{ij}^{\hat{c}_{ij}} \tag{4.9}$$

4.5 A two-step model

To generate an ensemble of chromatin conformations for an entire chromosome, I will use a two-step process.

4.5.1 Step 1

Step 1 is used to generate an ensemble of structures for individual domains - small sections of the chromosome. I apply the MonteGrappa algorithm as described above to the data from each domain individually. This results in a number of structures (the number is up to the user, I typically sample either 500 or 5,000) for each domain. The structures are all relatively the same scale because they are constrained by the fixed distance between beads in the model. The relative positioning of the structures can be quite variable.

4.5.2 Step 2

Step 2 is the novel part of this work. I leverage the information generated in the previous step to create a 3D structure representing several domains. I begin by initializing a Markov Chain. For each domain d_i, an initial structure from the ensemble, s_i, is chosen. These structures are translated so that the first point in the structure lies on the x axis, spaced uniformly apart. This is done by updating the translation parameter t_i

Proposed MCMC moves

A proposed move in the Markov Chain is defined as one of four options. Each move is applied to the structure from a single domain.

- Translate: Move the structure s_i linearly by a given amount. If v is a translation vector, $s_i' = s_i + v$. This updates the translation parameter t_i.
- Rotate: Rotate the structure s_i by a value around the x, y and z axis. If R is a rotation matrix, $s_i' = Rs_i$. This updates the rotation parameters $(\rho_x, \rho_y, \rho_z)_i$.
- Scale: Scale the structure s_i. If A is a scaling matrix, $s_i' = As_i$ This updates the scaling parameter Θ_i.
- Swap: There are a large number of potential structures in the ensemble for a domain. Each individual conformation (ensemble member) could be part of the greater solution. At a given point in the Markov Chain, only a single conformation is used. Therefore, there needs to be a way to change the structure of a domain for another member of the ensemble. Implementation of this move has some problems, see the section below.

A proposed move updates the values of the parameters for a structure. A matrix of distance values is calculated from the updated structure, which is then converted into the in-silico contact map C_{t+1} via equation 4.2. The probability of the structure given the experimental data, $Pr(S_{t+1}|\hat{C})$ is then calculated, and the move is accepted if it meets the criteria. The $t+1$ index indicates the progression of the Markov Chain.

Acceptance criteria

A proposed move is accepted according to the Metropolis algorithm (Metropolis et al. 1953). If the probability of the new structure at time $t+1$ is greater than the probability of the old structure at time t, the move is automatically accepted. However, if the probability of the new structure is less than the old, the move is accepted with probability a equal to the ratio of the two choices:

$$a = \frac{Pr(S_{t+1}|\hat{C})}{Pr(S_t|\hat{C})}$$

Otherwise, the proposed move is rejected and the structure returns to the previous state. The probabilistic acceptance of "bad" moves is the key to the Metropolis algorithm. If only moves that increased the probability were accepted, the algorithm would get stuck in local maxima and never explore the whole probability space. Occasionally accepting moves that make the probability lower allows the algorithm to escape local maxima.

Range of MCMC moves

It is desirable to tune the acceptance rate of the Metropolis algorithm to be in a certain range. Suggestions in the literature are varied, one possibility is a 23% acceptance rate for an N-dimensional distribution (Roberts et al. 1997). Acceptance rate is important for the convergence of the Markov Chain to the stationary distribution. If the moves made result in small changes, they are likely to be accepted very frequently. If the moves result in large changes, they will be accepted rarely. A acceptance rate that is too high will slow convergence because the moves are too small, while an acceptance rate that is too low will slow convergence because the structure rarely changes. The acceptance rate in this model can be tuned by varying the amount the structure is changed with each move. The parameters to translate, rotate or scale by should be limited to a certain range; this range can be determined experimentally by monitoring the acceptance rate. There also needs to be a parameter that defines the probability of making a given move which could also play a role in the acceptance rate.

Assessing convergence and drawing from the Markov Chain

To generate an ensemble of structures MonteMonster needs to make independent draws from the Markov Chain after it has converged to the stationary distribution. The first step is to assess when the Markov Chain has converged. There are statistics designed for this purpose I have found. A more intuitive approach is to start many Markov Chains from the same initial point and decide when they become indistinguishable from each other. Say this takes k moves. Starting a single Markov Chain will then take k moves to converge to the stationary distribution; these initial moves can be discarded as burn-in moves. After convergence, independent samples can be drawn from the Markov chain by recording the structure at $k, 2k, 3k$, etc number of moves. These structures will theoretically be independent draws from the Markov chain and will become the ensemble of structures. A more rigorous assessment of convergence of my model will be necessary in the future.

Implementation of the swap move

I have had difficulty coming up with a reasonable implementation for the swap move, although something that changes the members of the ensemble for a domain will be necessary in the model. There are two options I have been considering for swap:

- Implement swap as a regular move in the Markov Chain. If this move were selected for a domain, the current conformation would be swapped out for a different member of the ensemble. Some modification would be necessary, the 3D coordinates for one conformation could not be simply exchanged for another conformation. At the very least, I would translate the first point in the new conformation to match the first point in the old conformation. It could also be desirable to rotate and scale the new conformation to better match the old. These modifications could keep the global arrangement of structures similar while changing more local interactions through swapping the conformation.

  It would be best to keep the probability of choosing this move low to limit the extreme perturbations to the structure. However, swapping a conformation for an entirely different member of the ensemble could drastically change the structure that had been optimized for the last several moves. This could result in a very low probability of swap moves being accepted or a very chaotic structure resulting every time a swap move is performed.

- Change conformations when independent recordings from the Markov Chain are made. With this option, the conformations of every domain would be changed after

a structure is recorded. The structure could then be optimized through translation, rotation and scaling without swapping conformations until a structure is recorded. Afterwards, new conformations would be randomly selected for each domain and the optimization would be run again.

Swap is an important concept to keep in the model because it leverages the information from the previous step. Properties of conformations that appear more frequently in the ensemble for a single domain should appear more frequently in the whole structure.

4.6 Comments on the MonteMonster algorithm

The MonteMonster algorithm as described above tackles some of the problems I have identified with 3D reconstruction algorithms. First, it generates an ensemble of chromatin structures that, together, can represent the experimental Hi-C contact map. Second, it treats the data as two-scaled and uses a different step to tackle reconstruction at each scale. Constructing a structure for an individual domain is done with the polymer model of MonteGrappa - beads on a string attached by a fixed linker. The assumptions in this model (constraints on maximum polymer size, binary contact possibilities, for example) are only valid at these short-range, fine-scale interactions within a domain. Step 2 is better suited to interactions beyond individual domains. It does not assume anything about maximum structure size and uses a more general definition of contact probability (equation 4.1 and 4.2)

I am currently working on implementing and testing the MonteMonster algorithm as described. Of the issues I have discussed, implementation of the swap concept and the runtime of the algorithm have been the most problematic. The framework for the algorithm has been laid, though, and implementing and testing the ideas presented should make an interesting longer-term research project.

4.6.1 Issues that require further consideration

There area number of topics that I need to explore further in the development of this model. As I finish the implementation and testing of MonteMonster, the answers to some of these will become clear.

- The probabilistic interpretation of Hi-C data to use. I am curious how changing from the normal distribution to the Poisson would affect results of the algorithm.
- The best choice to use for the swap concept, clarified above.

- To be a true multi-scale algorithm, MonteMonster would ideally leverage the Hi-C data at more than one resolution. This is desirable because the quality of the signal in the data decreased with genomic distance between interacting loci. For calculations far from the diagonal I would like to use lower resolution data to avoid noise and decrease the computational load. However, I would need to work this idea into the mathematical formulation of the model.

- Runtime. MonteMonster is currently implemented in Python. The program runs too slow to be applied to many domains and once. The slowest part is the calculation of the pairwise distances between points and calculating the probability of the new structure after each move. Although I have identified several ways to increase the performance, I will still be limited by the python implementation.

- Single-cell HiC. The data presented in Nagano et al.(2013) is very valuable because it comes from singular chromatin conformations. I would like to validate the results of my ensemble model with this data.

Bibliography

[1] Rumen Andonov, Noël Malod-Dognin, and Nicola Yanev. Maximum Contact Map Overlap Revisited. *Journal of Computational Biology*, 18(1):27–41, January 2011.

[2] Alberto Caprara, Robert Carr, Sorin Istrail, Giuseppe Lancia, and Brian Walenz. 1001 Optimal PDB Structure Alignments: Integer Programming Methods for Finding the Maximum Contact Map Overlap. *Journal of Computational Biology*, 11(1):27–52, January 2004.

[3] Marion Cremer, Florian Grasser, Christian Lanctôt, Stefan Müller, Michaela Neusser, Roman Zinner, Irina Solovei, and Thomas Cremer. Multicolor 3d fluorescence in situ hybridization for imaging interphase chromosomes. *Methods in Molecular Biology (Clifton, N.J.)*, 463:205–239, 2008.

[4] Thomas Cremer and Marion Cremer. Chromosome Territories. *Cold Spring Harbor Perspectives in Biology*, 2(3), March 2010.

[5] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 14(6):390–403, 2013.

[6] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing Chromosome Conformation. *Science*, 295(5558):1306–1311, February 2002.

[7] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012.

[8] Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D. Green, and Job Dekker. Chromosome Conformation Capture Carbon Copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, October 2006.

[9] James Fraser, Mathieu Rousseau, Solomon Shenker, Maria A. Ferraiuolo, Yoshihide Hayashizaki, Mathieu Blanchette, and Josée Dostie. Chromatin conformation signatures of cellular differentiation. *Genome Biology*, 10(4):R37, April 2009.

[10] Luca Giorgetti, Rafael Galupa, ElphègeP. Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell*, 157(4):950–963, May 2014.

[11] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, 9(1):e1002893, 2013.

[12] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, December 2012.

[13] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, October 2012.

[14] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502, June 2007.

[15] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30(1):90–8, 2012.

[16] Lawrence A. Kelley, Stephen P. Gardner, and Michael J. Sutcliffe. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Engineering*, 9(11):1063–1065, November 1996.

[17] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature Methods*, advance online publication, September 2014.

[18] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S.

Lander, and Job Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, October 2009.

[19] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8Å resolution. *Nature*, 389(6648):251–260, September 1997.

[20] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[21] Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, January 2011.

[22] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, October 2013.

[23] Natalia Naumova, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R. Lajoie, Leonid A. Mirny, and Job Dekker. Organization of the Mitotic Chromosome. *Science*, 342(6161):948–953, November 2013.

[24] Anders B. Norgaard, Jesper Ferkinghoff-Borg, and Kresten Lindorff-Larsen. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophysical Journal*, 94(1):182–192, January 2008.

[25] SuhasS. P. Rao, MiriamH. Huntley, NevaC. Durand, ElenaK. Stamenova, IvanD. Bochkov, JamesT. Robinson, AdrianL. Sanborn, Ido Machol, ArinaD. Omer, EricS. Lander, and ErezLieberman Aiden. A 3d Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*.

[26] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, February 1997.

[27] Mathieu Rousseau, James Fraser, Maria A. Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1):414, October 2011.

[28] Mark R. Segal, Hao Xiong, Daniel Capurso, Mariel Vazquez, and Javier Arsuaga. Reproducibility of 3d chromatin configuration reconstructions. *Biostatistics*, 15(3):442–456, July 2014.

[29] M. J. Sutcliffe. Representing an ensemble of NMR-derived protein structures by a single structure. *Protein Sci*, 2(6):936–44, 1993.

[30] Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Research*, 22(7):1295–1305, July 2012.

[31] David J. Tremethick. Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. *Cell*, 128(4):651–654, February 2007.

[32] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, June 2014.

[33] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, October 1997.

[34] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059–1065, November 2011.

[35] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan . C. Toh, and Wing-Kin . K. Sung. 3d Chromosome Modeling with Semi-Definite Programming and Hi-C Data. *Journal of Computational Biology*, 20(11):831–846, 2013.

[36] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11):1341–1347, November 2006.