EM ALGORITHM FOR RECONSTRUCTING 3D STRUCTURES OF HUMAN CHROMOSOMES FROM CHROMOSOMAL CONTACT DATA

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Yuxiang Zhang

Professor Jianlin Cheng, Thesis Supervisor

MAY 2016

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

EM ALGORITHM FOR RECONSTRUCTING 3D STRUCTURES OF HUMAN CHROMOSOMES FROM CHROMOSOMAL CONTACT DATA

presented by Yuxiang Zhang,

a candidate for the degree of Master of Science and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Jianlin Cheng

Professor Chi-Ren Shyu

Professor Zhihai He

ACKNOWLEDGMENTS

I would like to show my deepest gratitude to my advisor, Professor Jianlin Cheng, for his patient advice, guidance and inspiration over the years and giving me so much freedom and opportunity to explore different aspects of computational Optimization Methods. I am greatly indebted to Professor Cheng for all that I learnt from him both in terms of technical knowledge and research style, and this work would have been impossible without his support.

I would like to thank my friends who give great advices for my thesis research for their kindness of willing of helping, patient and inspiration. I would also like to thank all my teachers who have helped me to develop the fundamental and essential academic competence. Without their instruction, I could not have achieved today?s progress.

I would like to thank to my family for inspiring me to fulfill my dream. Their friendship and constructive suggestions constantly encouraged me when I felt frustrated with research works.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

ABSTRACT

Recent research suggested that chromosomes have preferred spatial conformations to facilitate necessary long-range interactions and regulations within a nucleus. So that, getting the 3D shape of chromosomes of a genome is very important for understanding how the genome folds and how the genome interact, which can know more about the secrete of life.

The introduction of the chromosome conformation capture (3C) based techniques has risen the development of construct the 3D structure of chromosome model. Several works have been done to build the 3D model, among which can be divided into two groups one is consensus methods in early work, the other is ensemble method.

In this paper I proposed an ensemble method for reconstructing the 3D structure of chromosome structure. First step is to process Hi-C data, and then do normalization. After that I applied the Bayesian inference model to get an objective function. Finally I used EM based algorithm along with using gradient descent method which is applied in expectation step. I applied the objective function and the optimization method to all 23 Hi-C chromosomal data at a resolution of 1MB.

Chapter 1

Introduction

In this chapter, I give an overview of chromosomes structure modeling problem that I will discuss in this thesis. I also describe some of the issues faced in computational approaches for chromosomes structure modeling that I will try to address, and give an outline for rest of this thesis.

1.1 Introduction To Three-dimensional Chromosomes Structure Modeling

The 3D organization of genomes was found to play an important role in gene-gene interaction [1] [2] [3], gene regulation and genome methylation, which can be used to define chromatin signatures. It is recognized that the three-dimensional organization of chromatin affects gene regulation and genome function. For example, it was shown that elements that lie far apart in the one-dimensional genomic sequence or on different chromosomes could functionally interact through physical contacts. So

understanding the 3D structures of chromosomes can provide important hints toward decoding the mechanisms of gene regulation and chromatin packing, as well as DNA replication, repair and modification [4].

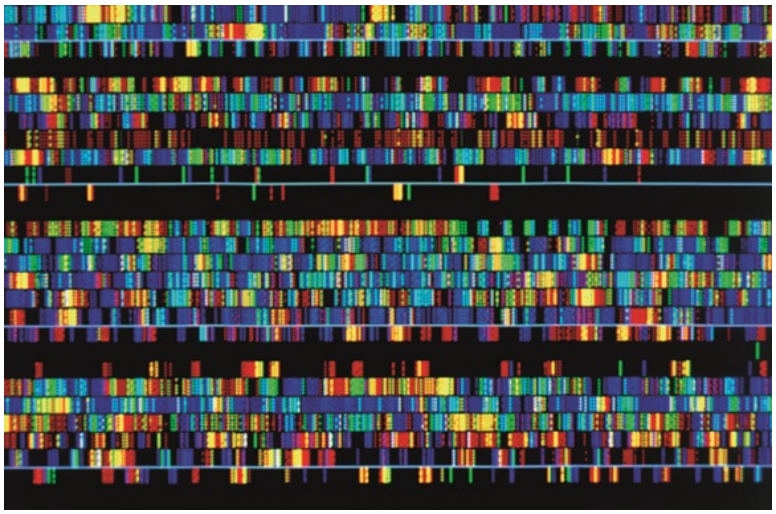


Figure 1.1: An example of 1D human genome sequencing.

However, owing to lack of experimental data, early work on chromatin structure modeling mainly focused on building up a theoretical model to describe the physical property of chromatins based on known knowledge on polymer physics [5] [6]. Little is known about the 3D organization of a genome and its largest discrete components, chromosomes [7] [8] [9].

Recently, chromosome conformation capture (3C) based techniques have emerged as powerful tools for capturing physical interactions between pairs of chromosomal regions on the same or two different chromosomes [10] [11]. Particularly, an advanced 3C technique, Hi-C, has been developed to determine both intra- and inter-chromosomal contacts at a genome scale rather uniformly and unbiasedly, which provides crucial information necessary for studying and reconstructing the 3D shape of a chromosome

or genome for the first time [12]. Hi-C technology made it possible to link chromatin structure to gene regulation [13], DNA replication timing [14], and somatic copy number alterations [15]. Furthermore, genome-wide conformation capture studies reveal conserved structural features that are now accepted as organizing principles of chromatin folding [16]. Hi-C data have also proved to be useful in many other applications, ranging from genome assembly to finding the coordinates of centromeres and ribosomal DNA (rDNA).

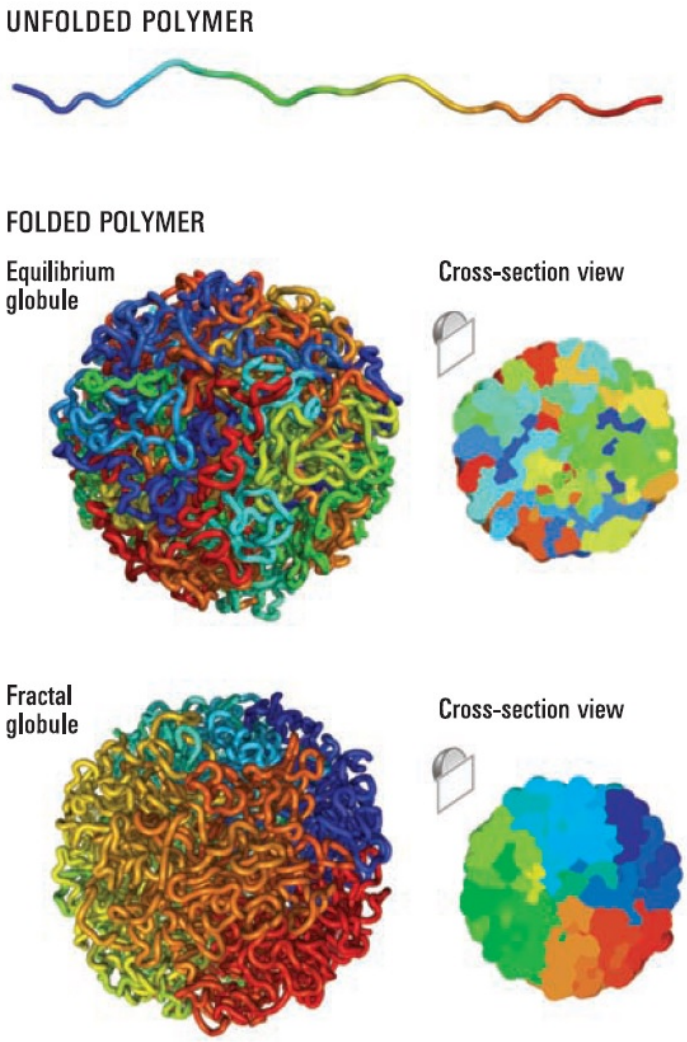


Figure 1.2: An example of 3D structure of human genome
(Top) An unfolded polymer chain, 4000 monomers (4.8 Mb) long.
(Middle) An equilibrium globule. The structure is highly entangled; loci that are nearby along the contour (similar color) need not be nearby in 3D.
(Bottom) A fractal globule. Nearby loci along the contour tend to be nearby in 3D, leading to monochromatic blocks both on the surface and in cross section. The structure lacks knots.

1.2 Existing Approaches And Challenges In Three-dimensional Genome Modeling

With the availability of genome-wide contact maps, the reconstruction of the three-dimensional chromatin structure that underlies the observed contacts became a fundamental problem. These observed contact maps made it possible to generate detailed three-dimensional models using the contact counts on the relative locations of loci with respect to each other [17]. Fittingly, these models are referred to as restraint-based models. Other terms used for these models include probabilistic, statistical, or inverse models, in contrast to polymer-based direct models. These restraint-based models can be further divided into two groups [18].

Consensus Methods

One of the most commonly used methods to infer consensus three-dimensional models from conformation capture data is multi-dimensional scaling (MDS) [19]. MDS is a classical statistical method that, given all pairwise distances between a set of objects, aims to find an N-dimensional embedding such that the pairwise distances are preserved as well as possible [20]. In this situation, objects are beads that represent chunks of DNA, and pairwise distances are computed by applying a transfer function on contact counts. Several studies use MDS with additional constraints to find a consensus structure. A recent method applies a semidefinite programming (SDP) approach to three-dimensional genome reconstruction [21]. The SDP approach guarantees perfect three-dimensional reconstruction if the input pairwise distances are noise-free. However, all MDS-based methods depend on a transfer function that con-

verts contact counts to pairwise spatial distances, and the methods are very sensitive to the selection of this transfer function [22].

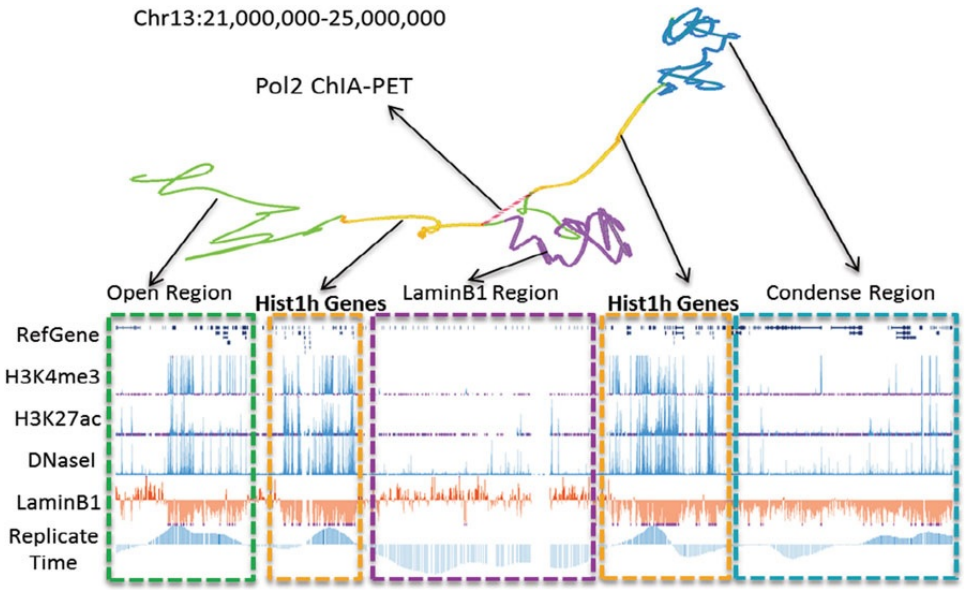


Figure 1.3: A high-resolution chromosomal 3D structure for the region chr13:21Mb-25Mb.

Ensemble Methods

Several probabilistic methods have been proposed to produce the ensemble of three-dimensional models [23] [24] [25] [26]. They give a set of structures representative of the observed contact data. These methods aim to find an ensemble that, in aggregate, optimally describes the bulk data. These methods can be further divided into two depending on whether they aim to find multiple solutions, each of which fits the bulk Hi-C data, or to find a true ensemble that, in aggregate, optimally describes the bulk data, which is similar to the consensus approach, but instead of inferring one locally optimal model, the optimization is run with multiple initializations resulting in multiple different models. Some studies use Markov Chain Monte Carlo (MCMC)

sampling to approximate the posterior probability of each model given the data from a large number of models that are independent of random initialization [23]. The other one is more complicated because it requires coordinated inference of a large number of models. One example is using MCMC with a mixture model component to determine whether a mixture of structures better explain the conformation of a locus than a single consensus structure [18].

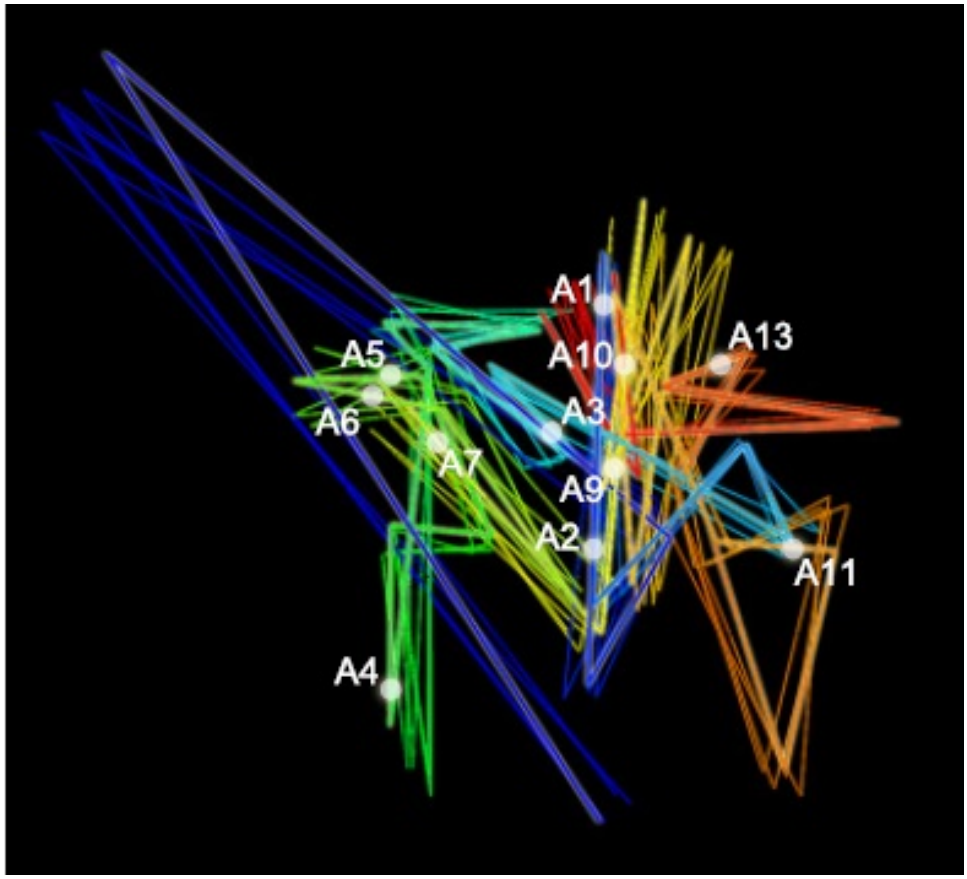


Figure 1.4: This is an example of ensemble method. The 'gold standard' structure is used as a reference structure to which structures from four different parallel MCMC5C runs on simulated data generated from the gold standard structure are aligned.

1.3 Thesis Outline

The remainder of this thesis is structured as follows. In chapter 2, I start introducing the Hi-C data, including how the data looks like, how to processing row Hi-C data and also the heat map for visualization. After that I introduced a normalization method to further process the Hi-C data and get ready for further use.

In chapter 3, I first apply the Bayesian inference model to get an objective function. And then I introduce EM algorithm to optimize the objective function, in which I use the gradient method for optimizing in the expectation step.

Chapter 4 is the result that I retrieve from the optimization method and the comparison of different methods. Appendix is the introduction of the development environment and source code.

Chapter 2

Data Processing

This chapter mainly talks about how the data processed, formed into contacted matrix, normalized for further use. The data that I use is raw Hi-C data from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18199. And I try to form the Hi-C data into two different resolutions which are 1MB and 200KB. After that a method called Sequential Component Normalization has been applied to the contacted matrix to normalize the data for reduce the biases in Hi-C experiments.

2.1 Introduction To Chromosome Conformation Capturing Techniques

Chromosome Capture Conformation (3C) was first introduced by Dekker et al [2]. The 3C technique aims in detecting physical contact between pairs of genomic loci and is now widely used to detect intrachromosomal and interchromosomal interactions between genes and regulatory elements [27]. The development of the 3C-based

techniques has changed our vision of the nuclear organization. With the development of high throughput analyses, and in particular second-generation sequencing, the 3C has been adapted to study in parallel physical interactions between many loci, and thus increase the scale at which interactions between genomic loci can be detected [28].

More recently, this technique was further extended to obtain detailed insights into the general three-dimensional arrangements of complete genomes [29] [2]. While the use of Hi-C techniques is expected to increase in the coming years, it also creates some new statistical and bioinformatics challenges. In this way, publicly available bioinformatics tools, as well as clear analysis strategy are still lacking.

Hi-C allows unbiased identification of chromatin interactions across an entire genome. We briefly summarize the process: cells are crosslinked with formaldehyde; DNA is digested with a restriction enzyme that leaves a 5overhang; the 5overhang is filled, including a biotinylated residue; and the resulting blunt-end fragments are ligated under dilute conditions that favor ligation events between the cross-linked DNA fragments. The resulting DNA sample contains ligation products consisting of fragments that were originally in close spatial proximity in the nucleus, marked with biotin at the junction [30]. A Hi-C library is created by shearing the DNA and selecting the biotin-containing fragments with streptavidin beads. The library is then analyzed by using massively parallel DNA sequencing, producing a catalog of interacting fragments. The overview of how Hi-C library is created is presented below followed by the overview of Hi-C analysis pipelines.

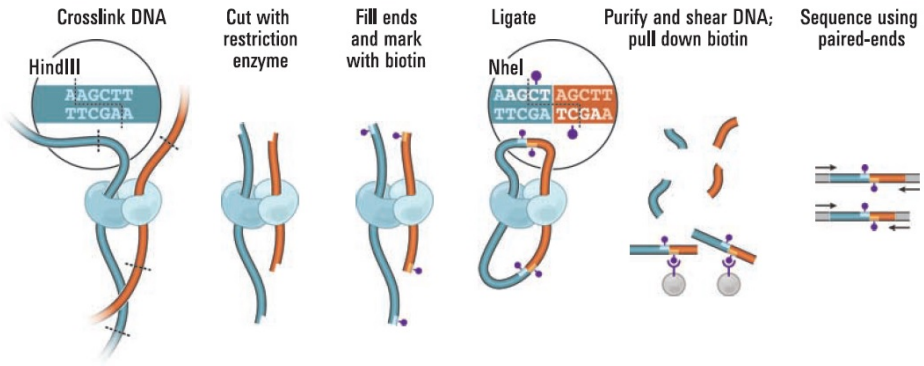


Figure 2.1: Cells are cross-linked with formaldehyde, resulting in covalent links between spatially adjacent chromatin segments (DNA fragments shown in dark blue, red; proteins, which can mediate such interactions, are shown in light blue and cyan). Chromatin is digested with a restriction enzyme (here, HindIII; restriction site marked by dashed line; see inset), and the resulting sticky ends are filled in with nucleotides, one of which is biotinylated (purple dot). Ligation is performed under extremely dilute conditions to create chimeric molecules; the HindIII site is lost and a NheI site is created (inset). DNA is purified and sheared. Biotinylated junctions are isolated with streptavidin beads and identified by paired-end sequencing.

Briefly, the traditional Hi-C assay consists of six steps: (1) crosslinking cells with formaldehyde, (2) digesting the DNA with a restriction enzyme that leaves sticky ends, (3) filling in the sticky ends and marking them with biotin, (4) ligating the crosslinked fragments, (5) shearing the resulting DNA and pulling down the fragments with biotin, and (6) sequencing the pulled down fragments using paired-end reads. This procedure produces a genome-wide sequencing library that provides a proxy for measuring the three-dimensional distances among all possible locus pairs in the genome.

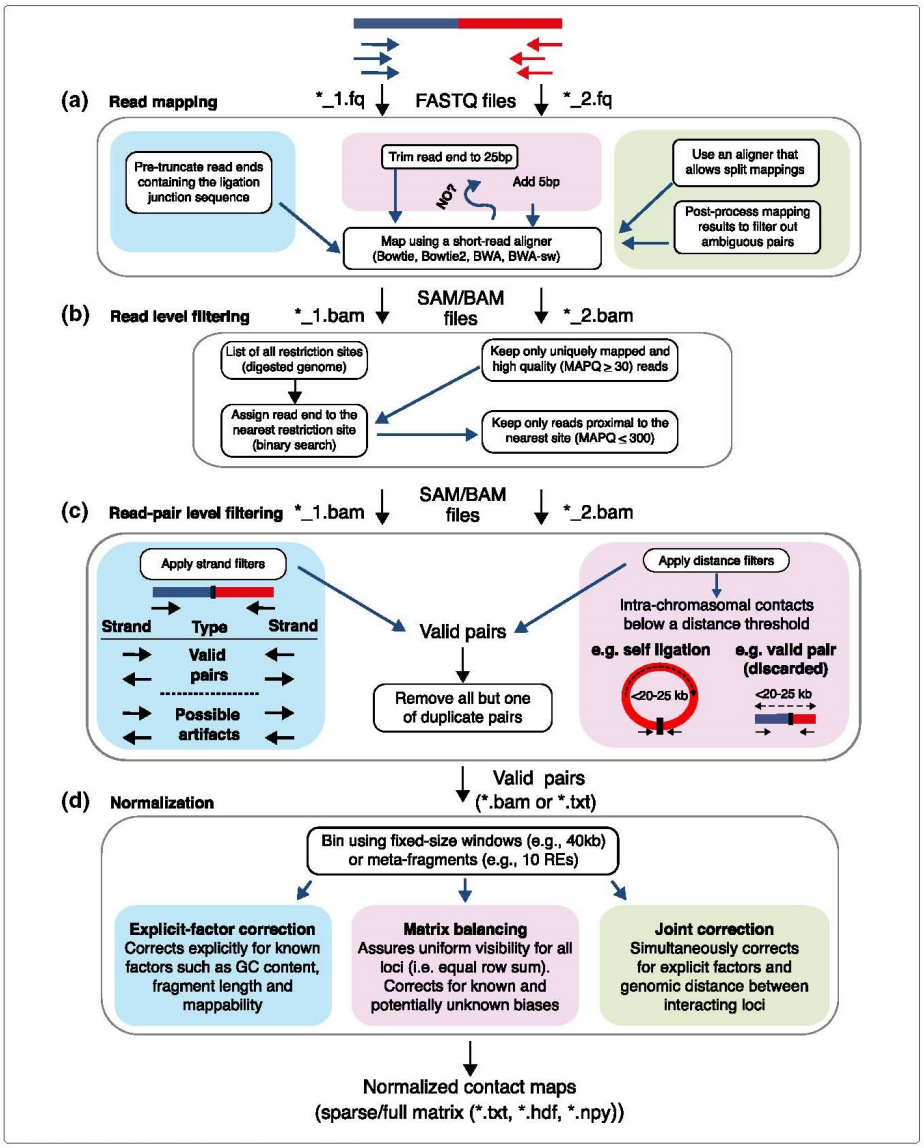


Figure 2.2: Overview of Hi-C analysis pipelines. These pipelines start from raw reads and produce raw and normalized contact maps for further interpretation. The colored boxes represent alternative ways to accomplish a given step in the pipeline. RE, restriction enzyme. At each step, commonly used file formats (.fq, .bam, and .txt) are indicated. **a**, The blue, pink and green boxes correspond to pre-truncation, iterative mapping and allowing split alignments, respectively. **b**, Several filters are applied to individual reads. **c**, The blue and pink boxes correspond to strand filters and distance filters, respectively. **d**, Three alternative methods for normalization.

2.2 Visualization Of Hi-C Data

Visualization of Hi-C data is important for understanding the genomic structure. Heat map is one of the methods for visualization, which is an ingenious display that simultaneously reveals row and column hierarchical cluster structure in a data matrix. It consists of a rectangular tiling with each tile shaded on a color scale to represent the value of the corresponding element of the data matrix. The heat map is a synthesis of several different graphic displays developed by statisticians over more than a century. The heat map is well-known for visualization in the natural sciences and one of the most widely used graphs in the biological sciences. In the case of gene expression data, the color assigned to a point in the heat map grid indicates how much of a particular DNA or protein is expressed in a given sample. The gene expression level is generally indicated by red for high expression and either green or blue for low expression. I will use the heat map to visualize my data and normalized data in result.

Several web browsers are used for visualizing thousands of data tracks for human, mouse and other organisms. However, these browsers are mainly designed for visualization of one-dimensional signals and are not easily extensible to visualizing two-dimensional Hi-C or any conformation capture data. Furthermore Hi-C data can be used for three-dimensional modeling, which requires tools not only for two-dimensional but also for three-dimensional visualization [18].

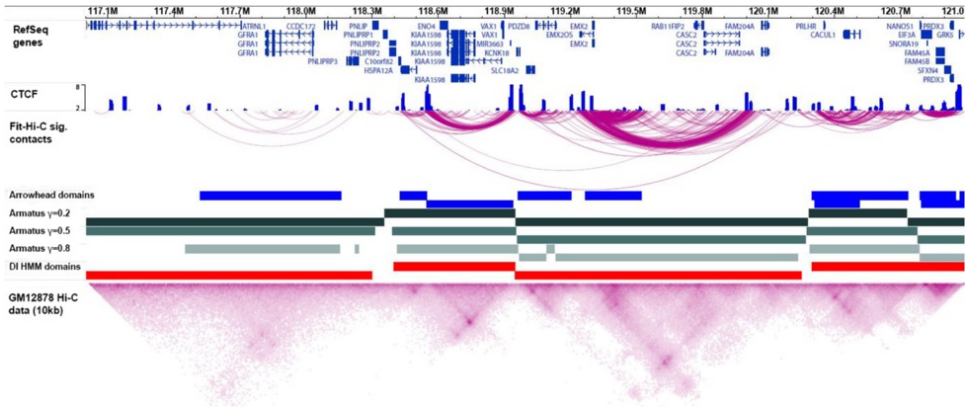


Figure 2.3: An Epigenome Browser snapshot of a 4 Mb region of human chromosome 10.

2.3 Contact Maps

Before discussing the method, it is necessary to describe how the data are represented in matrix form. A contact map is a matrix with rows and columns representing non-overlapping bins across the genome. Each entry in the matrix contains a count of read pairs that connect the corresponding bin pair in a Hi-C experiment. The resolution here that I use to build the contact map is 1MB and 200KB.

	Mb 1	Mb 2	Mb 3	Mb 4	Mb 5	Mb 6	Mb 7	Mb 8	Mb 9	Mb 10
Mb 1	2767	527	113	88	123	190	166	109	118	117
Mb 2	527	3826	440	239	261	183	63	43	23	54
Mb 3	113	440	3522	948	341	156	44	24	25	44
Mb 4	88	239	948	5156	876	139	35	21	19	30
Mb 5	123	261	341	876	5703	492	76	42	27	71
Mb 6	190	183	156	139	492	3854	372	173	132	192
Mb 7	166	63	44	35	76	372	2684	501	342	231
Mb 8	109	43	24	21	42	173	501	2311	530	259
Mb 9	118	23	25	19	27	132	342	530	2096	385
Mb 10	117	54	44	30	71	192	231	259	385	2766

Figure 2.4: Sample of the Hi-C count matrix (Lieberman-Aiden et al., 2009), recording the DNA-DNA contacts made between megabase intervals 1 to 10 in Chromosome 14.

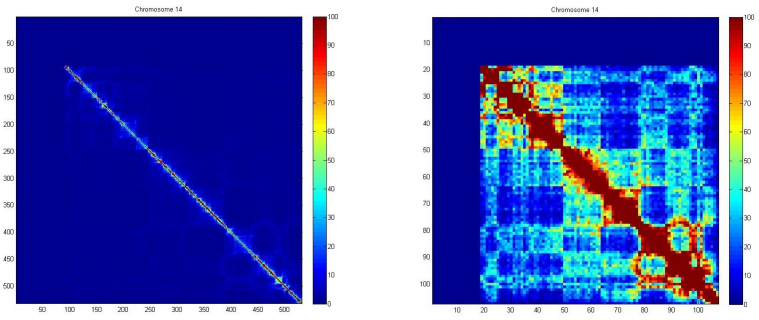


Figure 2.5: Heat maps of chromosome 14 in different resolutions before normalization. Left: Resolution of 200KB. Right: Resolution of 1MB.

2.4 Data Normalization

Not long after the first Hi-C datasets became available, several sequence-dependent features were shown to substantially bias Hi-C readouts. These include biases that are

associated with sequencing platforms and read alignment, and those that are specific to Hi-C, such as cutting frequencies of restriction enzymes, GC content and sequence uniqueness. Discovery of these biases led to several normalization or correction methods for Hi-C data.

The method that I use is called Sequential Component Normalization (SCN) methodology [31]. Firstly, normalization will give an equal weight to each fragment in the contact map. Therefore, restriction fragments with very low number of reads, which could not be properly detected, are likely to introduce noise in the normalized contact map and have to be removed. In order to identify these fragments, I computed the distribution of reads in the contact map. This distribution is roughly Gaussian, with a long tail corresponding to low interaction fragments. Based on this distribution, I cut the tail of the distribution.

Once low interacting fragments are removed, I wish to normalize all rows and columns of the contact map to one so that the matrix remains symmetric. This was done through the following simple procedure. Firstly, each column vector was normalized to one, using the Euclidian norm. Then each line vector of the resulting matrix was normalized to one. The whole process was repeated sequentially until the matrix become symmetric again with each row and each column normalized to one. Usually, two or three iterations are sufficient to insure convergence. Since it involves a sequential normalization of column and line vectors of the matrix, this method was named Sequential Component Normalization (SCN). This normalization can be viewed as a sequence of extensions and shrinking of interaction vectors so that they tend to reach the sphere of radius one in the interactions space. A similar and faster approach is to divide all the matrix elements C_{ij} by the product of the norms

of row i and column j : $C^*_{ij} = \frac{C_{ij}}{|C_{ik}||C_{kj}|}$. This method yields to a normalized contact map overall very similar to SCN. However since the sum of each component is not necessarily equal using this method, it may bias further analysis such as assessing the 3D colocalization of genomic elements. An alternative normalization method has been used so far by other groups, that use the sum of the components instead of the euclidian norm : $C^*_{ij} = \frac{C_{ij}}{\sum_k C_{ik} \sum_k C_{kj}}$. This method yields to a contact map with lower contrast than the SCN and therefore recommend SCN use in further works. The normalization using the sum will give more weight to fragments which makes fewer interactions.

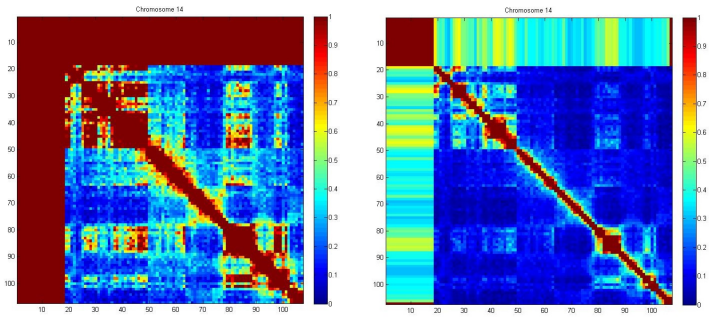


Figure 2.6: Normalization by using the SCN method. Left: A faster approach for SCN normalization. Right: A more recommended SCN method for normalization.

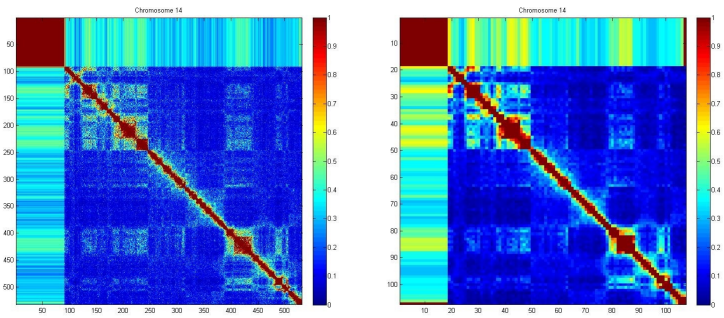


Figure 2.7: Heat maps of chromosome 14 in different resolutions after normalization. Left: Resolution of 200KB. Right: Resolution of 1MB.

The normalized maps overall are similar to those observed before. Since the probability of interaction between monomers along a polymer is decreasing with the linear distance between them, the diagonal which represents neighboring restriction fragments present the highest interactions score. The method that described consists in an easy and convenient way to normalize and represent genomic 3C data. The SCN normalization procedure proposed here will be helpful and adapted to any other organisms. Increasing the resolution of these contact maps will likely reveal more features, and can be addressed either through alternative protocols addressing the invisible zones of the genome (for instance by increasing the length of the sequenced reads or using various restriction enzymes), or through increasing the number of reads. Now I have the normalized contact maps for all chromosome, and I can use them to do further work for reconstruct the 3D chromosome model.

Chapter 3

Chromosomes Structure Modeling

3.1 Bayesian Inference Of A Chromatin Structure

Recently, with the development of 3C based techniques for getting the interaction frequencies between pairs of loci, lots of computational methods have been introduced to reconstruct the 3D structures of chromosomes [32]. As mentioned in the introduction chapter, there are two groups of methods that been proposed to model the structure. Because of uncertainty in the Hi-C experiment and the dynamic of the chromosome structure, I believe it is more reasonable and natural to use a probabilistic model to describe the chromosome structure [33] [34]. To solve the optimization problem by using a probabilistic model, here I introduce a method called Bayesian inference model to compute chromosome structures that describes how to convert the contact map into the distance between gene loci.

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics. Bayesian updating is particularly important in the dynamic analysis of a sequence of data. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability, often called 'Bayesian probability'.

The basic formula for reconstruct chromosome structure according to Bayesian probability can be described as follows:

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)} \tag{3.1}$$

Here S stands for a chromosome structure and D represents the data that derived from Hi-C experiment, which is also an observed structure. That's how I using the observed data forming the chromosome structure. Since there are no extra constraints put on the structure so the probability of the observed data P(D) can be considered as a constant with the respect to P(S). So I can replace P(S)/P(D) by using a constant ζ :

$$P(S|D) = \zeta \cdot P(D|S) \tag{3.2}$$

Now I am mainly discussing how to compute P(D|S) to get the probability of a chromosome structure. Note that:

$$P(D|S) = \prod_{i=1}^{n} P(D_i|S) \tag{3.3}$$

Here D_i stands for the i-th data in the contacted matrix that I have discussed before and n represents the total number of the data records. And here I assume that the spatial distance between two loci is inversely to their contact count, i.e. d=1/IF, for which I will discuss in the next section. Here d stands for the spatial distance between two loci and IF represents the interaction frequency between two loci. In addition, I assume that the Hi-C data is independent to each other. So I apply the normal distribution to model the probability, which is:

$$P(D_i|S) \sim \frac{1}{\sigma\sqrt{2\pi}} \cdot exp(-\frac{1}{2\sigma^2}(D_i^s - D_i)^2) \tag{3.4}$$

Here D_i^s stands for the distance from the chromosome structure and σ stands for the standard deviation of the normal distribution noise. In my research I assume that the Hi-C experiment data follows normal distribution. In addition, other distributions like Poisson distribution can also be applied to describe the Hi-C data. Now I can integrate the equation above into:

$$P(S|D) \sim (\frac{1}{\sigma\sqrt{2\pi}})^n \cdot exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i^s - D_i)^2) \tag{3.5}$$

Since the logarithm of the equation is much more easier to calculate than directly calculate the probabilistic function, I take the logarithm on both sides of the equation and note that $\sqrt{2\pi}$ is a constant here which does not affect our probability that can be ignored. And I have:

$$L(S|D) \sim -\frac{\sum_{i=1}^{n}(D_i^s - D_i)^2}{2\sigma^2} - n \cdot \lg \sigma \tag{3.6}$$

So I use equation 3.6 as an objective function to calculate the probability of a chromosome structure. And the method that I use is call EM algorithm which also require to use gradient descent method to maximize the objective function to get the highest probability of a chromosome structure.

Spatial distance

As I mentioned before that the conversion factor of convert the interaction frequency into distance matrix is a challenge. And the scale between the converted distance and the real distance is not matter here since the relevant structure is what I need. There is one method that has been used in the ChromSDE method can search for a correct conversion factor [21]. In fact the conversion factor is not a fixed constant for different data. And the predicted 3D structure is quite sensitive to the conversion factor. Given the same contacted matrix, different conversion factors cause different distances and finally goes to very different 3D structures. Therefore, estimating the correct conversion factor for a contacted matrix is important.

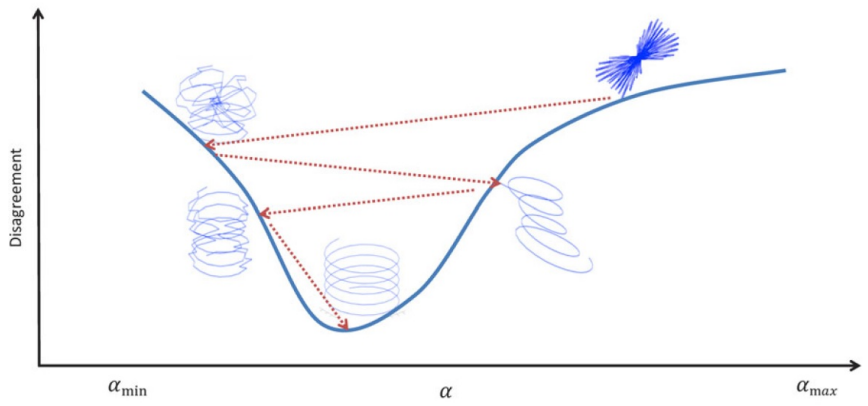


Figure 3.1: Illustration on how important to identify the correct conversion factor.

One of the methods use the fluorescent in situ hybridization (FISH) chromosomal distance data to estimate spatial distances. The FISH data contain spatial distances between some regions of the human chromosomes at various genomic distances. Then turn the problem into a constrained spatial optimization problem [24].

The method that I use can be described as follows, firstly, I randomly assign a value between 0 and 3, then compute the objective function which gives me a value which stands for the probability of a chromosome structure. After that I change the value of the conversion factor then compute the objective function again. Then repeat the process until I find the highest score for the probability. After many tries that I find the conversion factor always float around 1. So I set the conversion factor to 1 as a constant for easy computing.

3.2 EM Algorithm for Reconstructing 3D Structures

As long as I have the equation 3.6, now I can compute the probability of the objective function. However, there are unknown parameters that need to be addressed, such as the noise in the normal distribution. Because of these unknown parameters it is difficult to compute the probability of a chromosome structure. In this situation, I apply an EM algorithm to deal with the unknown parameters problem.

EM Algorithm

The EM algorithm has been used in many cases to find maximum likelihood parameters of a statistical model where the equations cannot be calculated directly. Typically these models involve unknown variables in addition to unknown parameters and unknown data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values and simultaneously solving the resulting equations. In statistical models with unknown variables, this usually is not possible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the unknown variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that the following is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It's not obvious that this will work at all, but in fact it can be proven that in this particular context it does, and that the derivative of the likelihood is zero at that point, which in turn means that the point is either a maximum or a saddle point. In general there may be multiple maxima, and there is no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e. nonsensical maxima. For example, one of the solutions that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

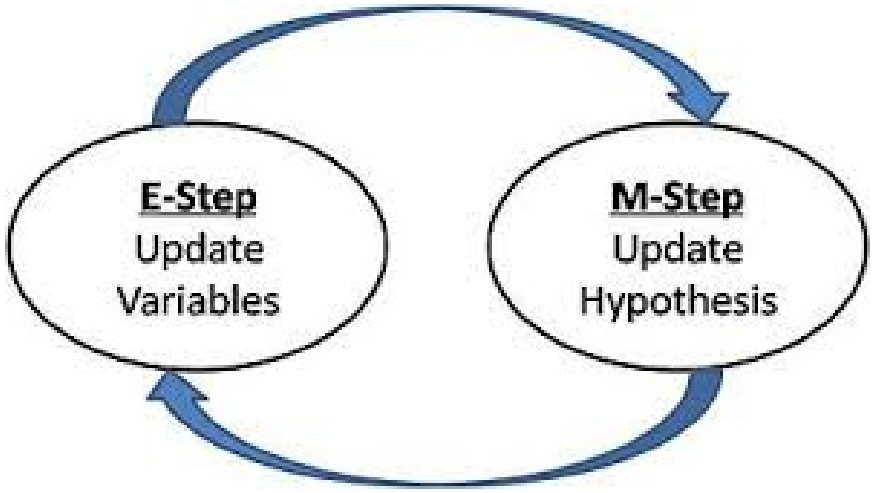


Figure 3.2: Illustration on how EM algorithm works.

In this situation, I apply the EM algorithm on optimizing the objective function which is equation 3.6. Note that the unknown value here is the σ , which is the

noise in the normal distribution. I follow the algorithm described below to find the maximum likelihood.

1.Initialize σ

2.**Repeat**

3.E-step: Maximizing objective function

4.M-step: Learning from the likelihood and assign σ to a new value.

5.**Until converge**

So far I have separated the problem into three sub-problems:

1.How to apply optimization method on the objective function to find the maximum likelihood.

2.How to set the new value of σ.

3.What is the converge condition.

The method that I use to find the maximum likelihood for the objective function is called gradient descent method, which I will discuss below.

Gradient Descent

Gradient descent is an optimization algorithm to find a local minimum of a function [35]. Taking the steps proportionally to the negative of the gradient of the function at the current point. If instead taking the steps proportionally to the positive of the gradient which is known as gradient ascent.

Given a function defined by a set of parameters, gradient descent starts with an initial set of parameter values and iteratively moves toward a set of parameter values that minimize the function. This iterative minimization is achieved using calculus, taking steps in the negative direction of the function gradient.

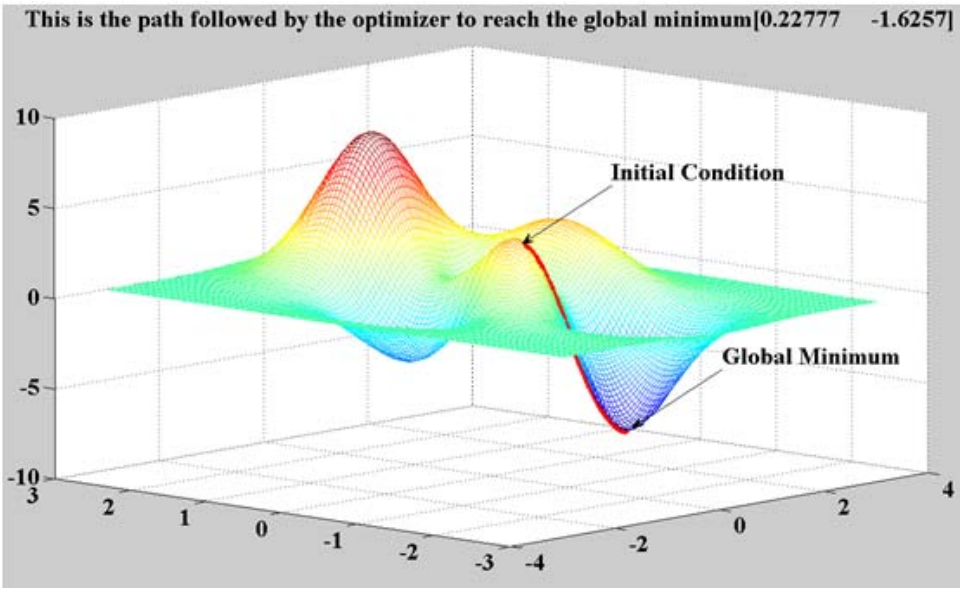


Figure 3.3: Illustration on gradient descent method.

Now I have the equation 3.6, so I can take the partial derivative of each coordinate which is :

$$\frac{\partial L(x_i, y_i, z_i)}{\partial x_i} = -\sum_{i,j=1;i\neq j}^{n} \frac{(x_i - x_j) \cdot (1 - \frac{D_{ij}}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}})}{\sigma^2} \tag{3.7}$$

$$\frac{\partial L(x_i, y_i, z_i)}{\partial y_i} = -\sum_{i,j=1;i\neq j}^{n} \frac{(y_i - y_j) \cdot (1 - \frac{D_{ij}}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}})}{\sigma^2} \tag{3.8}$$

$$\frac{\partial L(x_i, y_i, z_i)}{\partial z_i} = -\sum_{i,j=1;i\neq j}^{n} \frac{(z_i - z_j) \cdot (1 - \frac{D_{ij}}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}})}{\sigma^2} \tag{3.9}$$

After I have the partial derivative of all coordinates, I can easily apply gradient decent method on the E-step, maximizing the objective function.

$$f(x) = f(x) - \sigma \cdot \partial f(x) \tag{3.10}$$

Here σ is a gradient descent step which I take is 0.001. By doing iterations until converge which I will discuss later, I can get a better score here.

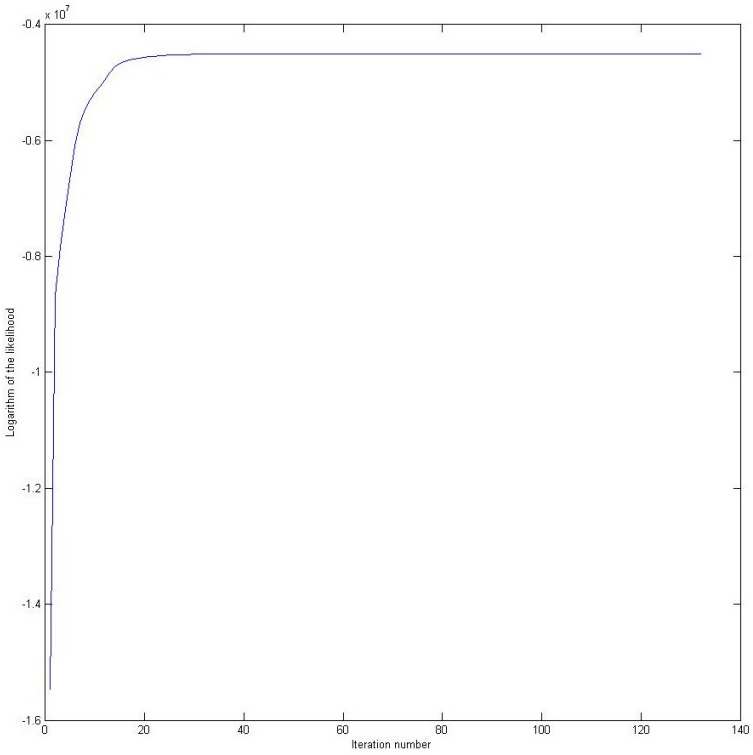
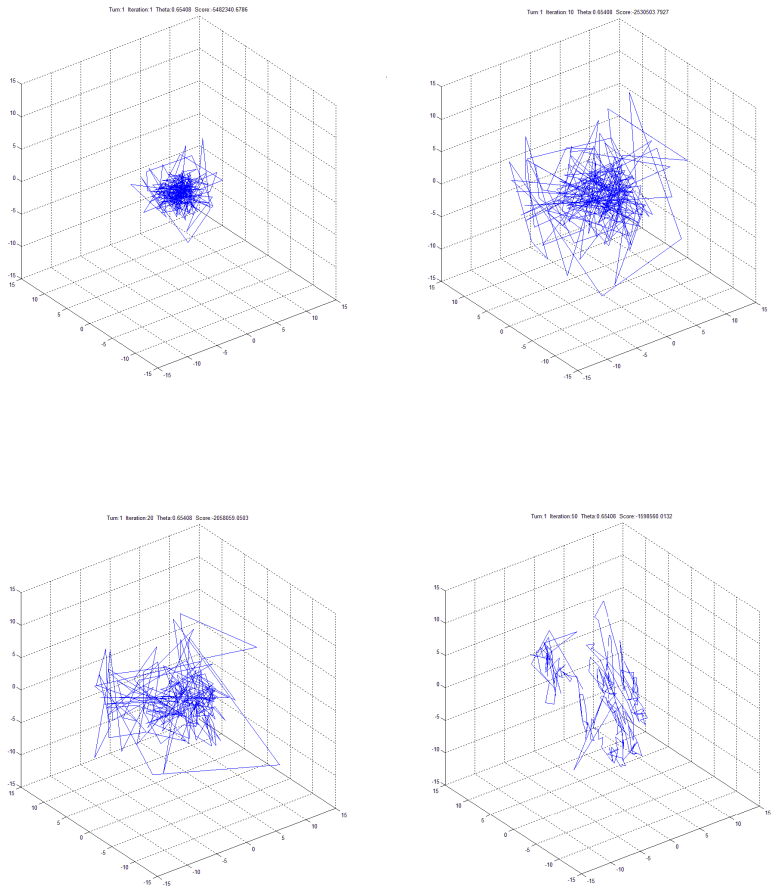


Figure 3.4: Illustration on how likelihood goes with the iterations by using gradient descent.

This figure shows how gradient descent works by the iteration goes. As we can see that the score changes as I move toward the maximum. A good way to ensure that

gradient descent is working correctly is to make sure that the error increases for each iteration. So that the step of gradient descent is important. I tried different steps for the gradient descent, the criterion here is converge happens not too quick nor too slow. After many tries I decide to use step=0.001, which is a reasonable value to run the gradient descent algorithm.



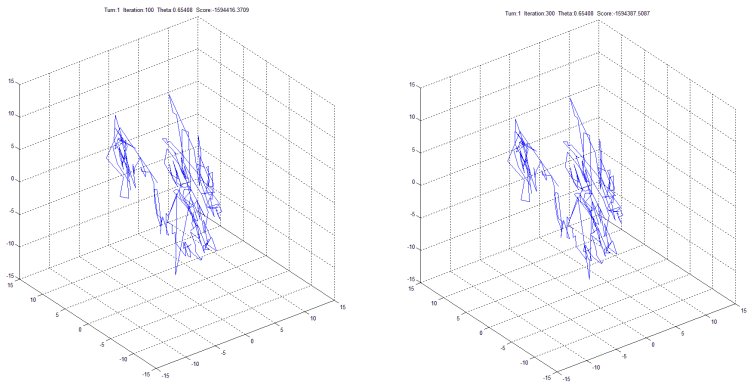


Figure 3.5: Different iterations: 0, 10, 20, 50, 100 and 300 in the gradient descent process.

These show how gradient descent works, from initial to 300 iteration. These figures show that gradient descent start quickly at the beginning and then slowly at the end. Gradient descent dynamically take the step and finally goes to a stable situation.

In conclusion, gradient descent is a very powerful method and most common way to find the local maximum or minimum, which has been successfully used on many optimization problem. I have a framework on the whole optimization problem and use gradient descent as a part of which turns out to be a nice algorithm to use.

Parameter Optimization

The second challenge there is how to set the new value of σ. Since the assumption that I made in the probability of single chromosome is normal distribution, which is:

$$P(D_i|S) \sim \frac{1}{\sigma\sqrt{2\pi}} \cdot exp(-\frac{1}{2\sigma^2}(D_i^s - D_i)^2) \tag{3.11}$$

Because σ is the variance in normal distribution, from which I can know that the unknown variable, in the objective function σ is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(D_i^s - D_i)^2}{n}} \tag{3.12}$$

Now I can set up the new value of σ. After doing E-step, I have a better chromosome structure rather than a randomly sample, by using which I can calculate the new σ. Comparing to the initial one, I can choose the one that gives the better score for the objective function.

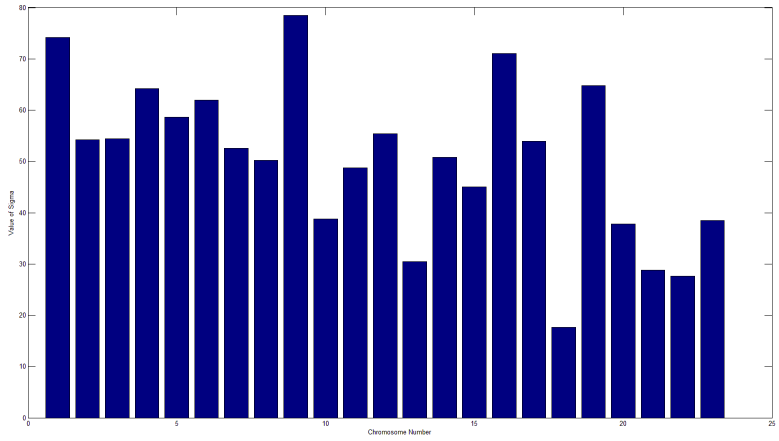


Figure 3.6: σ value among all chromosome.

Different chromosome has different value of σ. The smaller the σ value is the more convinced the chromosome structure is.

Converge Condition

There are two converge conditions need to be considered in this algorithm. First is the converge condition in gradient descent method, the second one is for the EM

algorithm. Basically they have the same converge condition, which both depend on the score of the objective function. If the current value of objective function is bigger than the previous one, then update the previous one to the current one, if else then end the loop. The likelihood converged to a stable situation within a small number of iteration. But in each iteration, it takes large number of optimization steps by running the gradient descent method. Different initial variables and different resolutions may cause different speed of converge. For 200KB it takes more than 1 hour to converge.

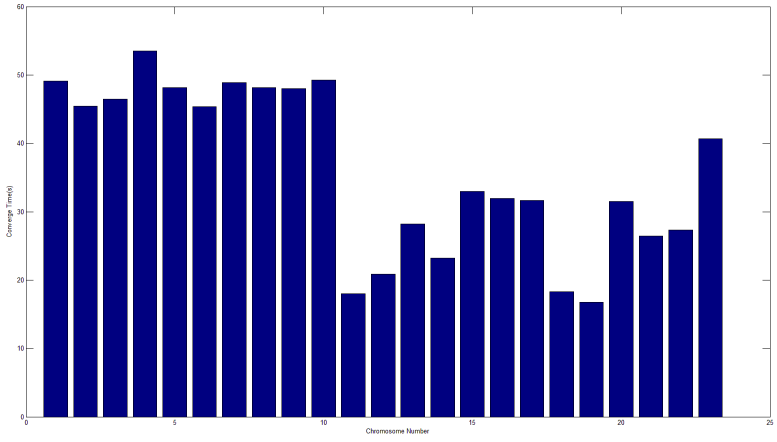


Figure 3.7: Converge time among all chromosome.

This figure shows that different chromosomes has different converge time. But big chromosomes like chromosome 1 to 10 converge slower than small chromosomes 16, 17, 19, 20, 21, and 22 which is reasonable.

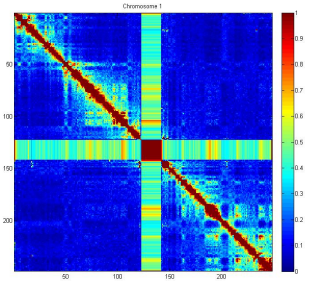
Chapter 4

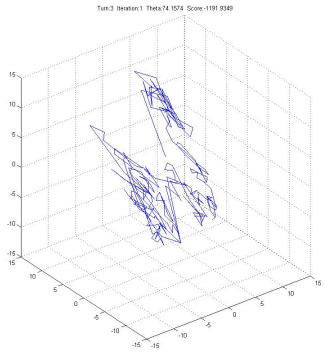
Results And discussion

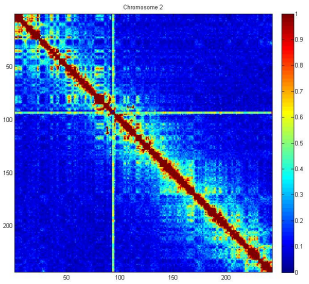
This chapter mainly shows the result of the 3D chromosome structure. And after that I will discuss some of the important methods that have been used in this area and compare these important methods to get a better understanding of Hi-C technology and some essential optimization methods.

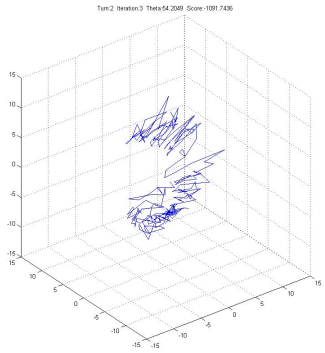
4.1 Results

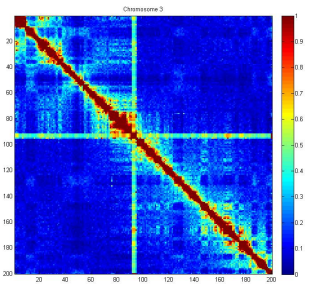
By using the EM based algorithm and gradient descent method for optimization, here I present the 3D structure for all chromosome in resolution of 1MB. The left figure is the heat map of the normalized contacted map; the right one is the chromosome structure.

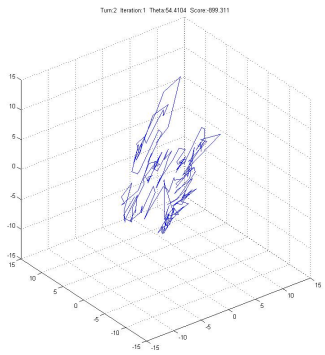

Chromosome 1

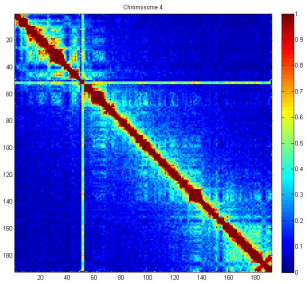

Turn:3 Iteration:1 Theta:74.1574 Score:-1191.9349

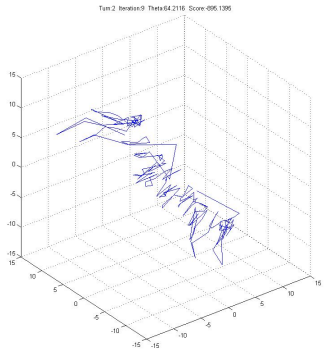

Chromosome 2

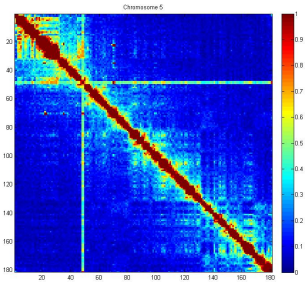

Turn:2 Iteration:3 Theta:64.2049 Score:-1091.7436

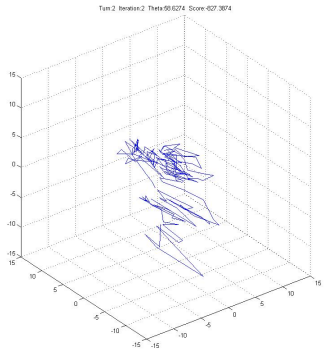

Chromosome 3

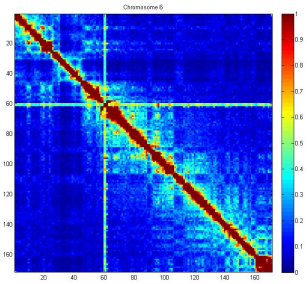

Turn:2 Iteration:1 Theta:54.4104 Score:-899.311

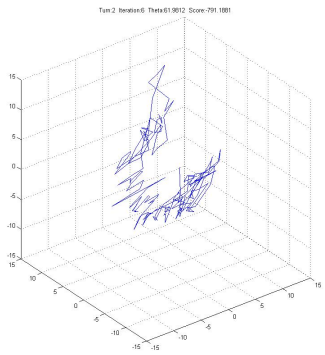

Chromosome 4

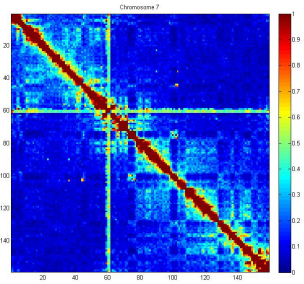

Turn:2 Iteration:9 Theta:64.2116 Score:-895.1395

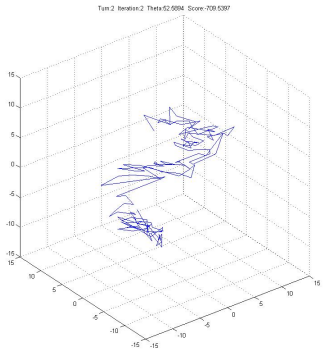

Chromosome 5

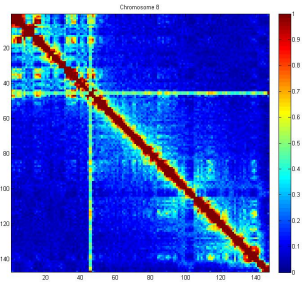

Turn:2 Iteration:2 Theta:68.6274 Score:-627.3874

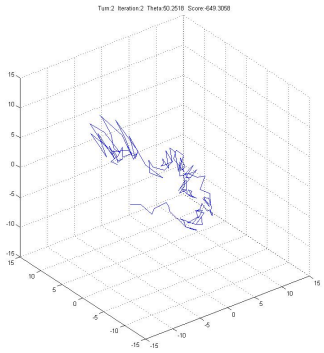

Chromosome 6

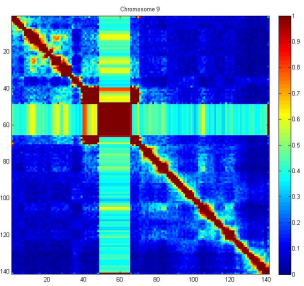

Turn:2 Iteration:6 Theta:61.9812 Score:-791.1881

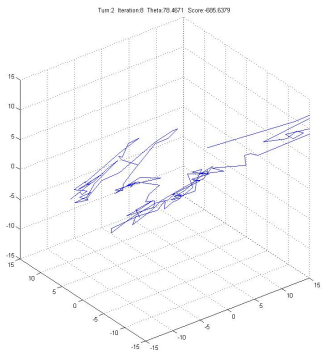

Chromosome 7

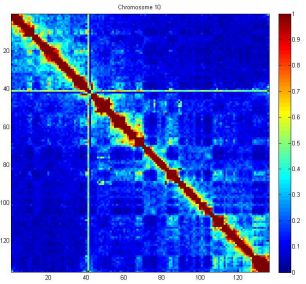

Turn:2 Iteration:2 Theta:62.6894 Score:-709.5397

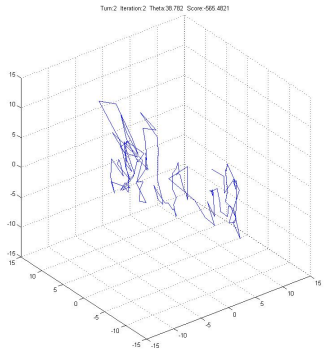

Chromosome 8

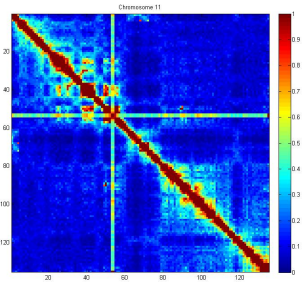

Turn:2 Iteration:2 Theta:60.2518 Score:-649.3068

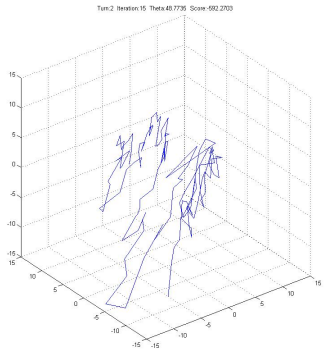

Chromosome 9

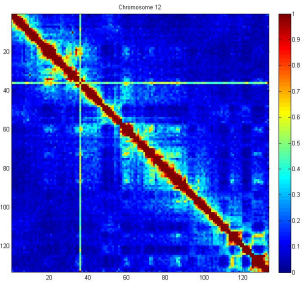

Turn:2 Iteration:8 Theta:70.4571 Score:-695.6379

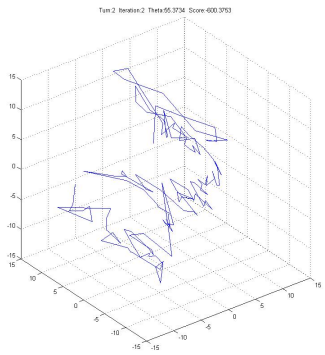

Chromosome 10

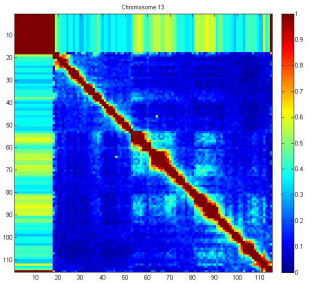

Turn:2 Iteration:2 Theta:36.782 Score:-565.4821

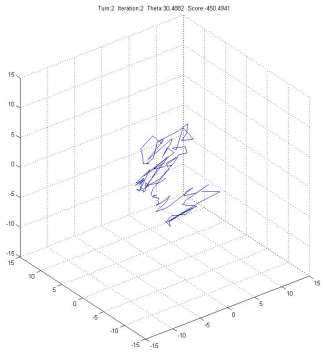

Chromosome 11

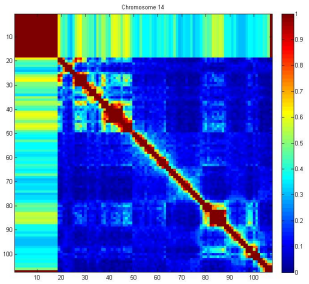

Turn:2 Iteration:15 Theta:48.7735 Score:-592.2703

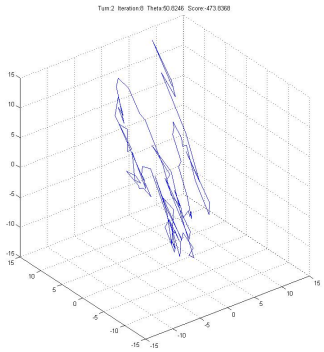

Chromosome 12

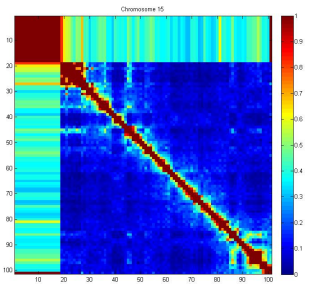

Turn:2 Iteration:2 Theta:55.3734 Score:-600.3753

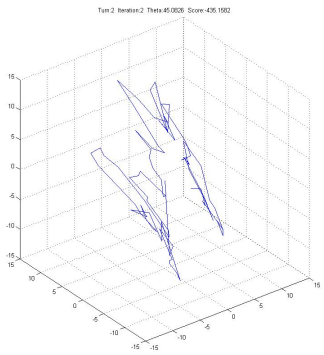

Chromosome 13

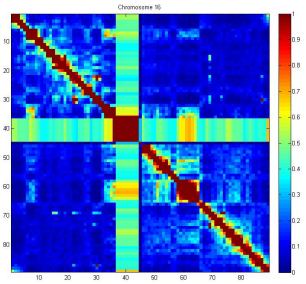

Turn:2 Iteration:2 Theta:30.4992 Score:-450.4941

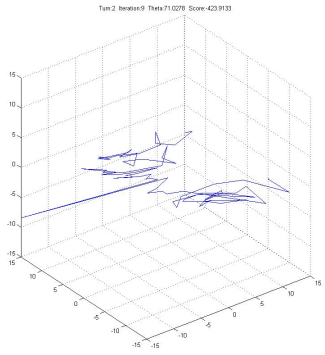

Chromosome 14

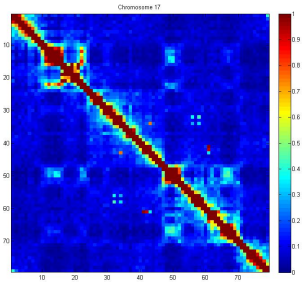

Turn:2 Iteration:8 Theta:60.6246 Score:-473.8368

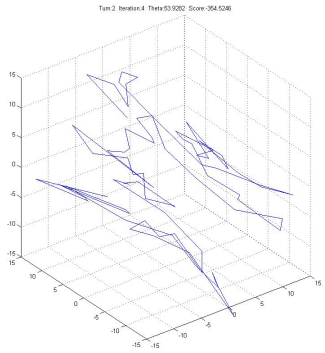

Chromosome 15

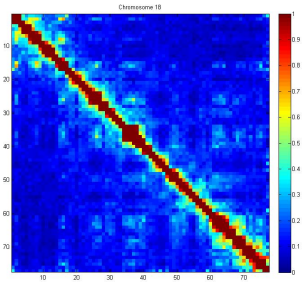

Turn:2 Iteration:2 Theta:45.0526 Score:-435.1582

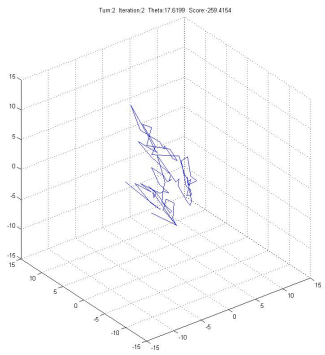

Chromosome 16

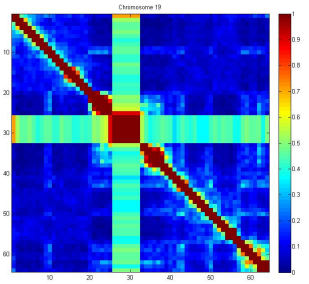

Turn:2 Iteration:9 Theta:71.0278 Score:-423.9133

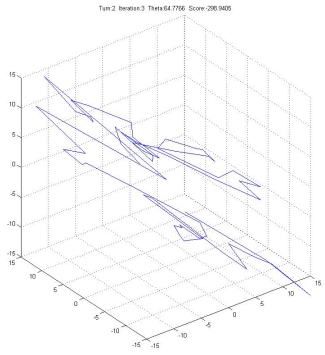

Chromosome 17

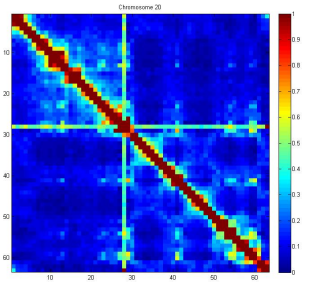

Turn:2 Iteration:4 Theta:53.9282 Score:-354.5246

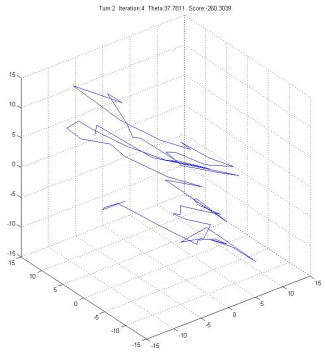

Chromosome 18

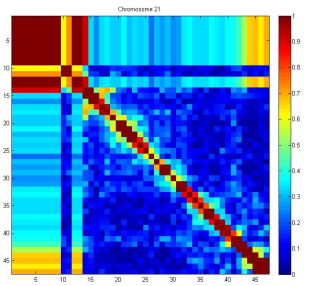

Turn:2 Iteration:2 Theta:17.6199 Score:-259.4154

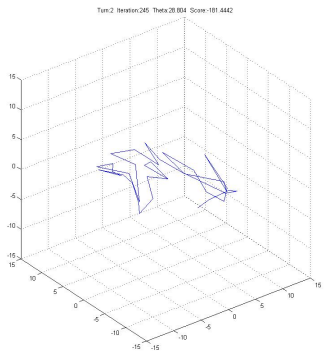

Chromosome 19


Turn:2 Iteration:3 Theta:64.7766 Score:-298.9405


Chromosome 20


Turn:2 Iteration:4 Theta:37.7011 Score:-260.3039


Chromosome 21


Turn:2 Iteration:245 Theta:28.804 Score:-181.4442

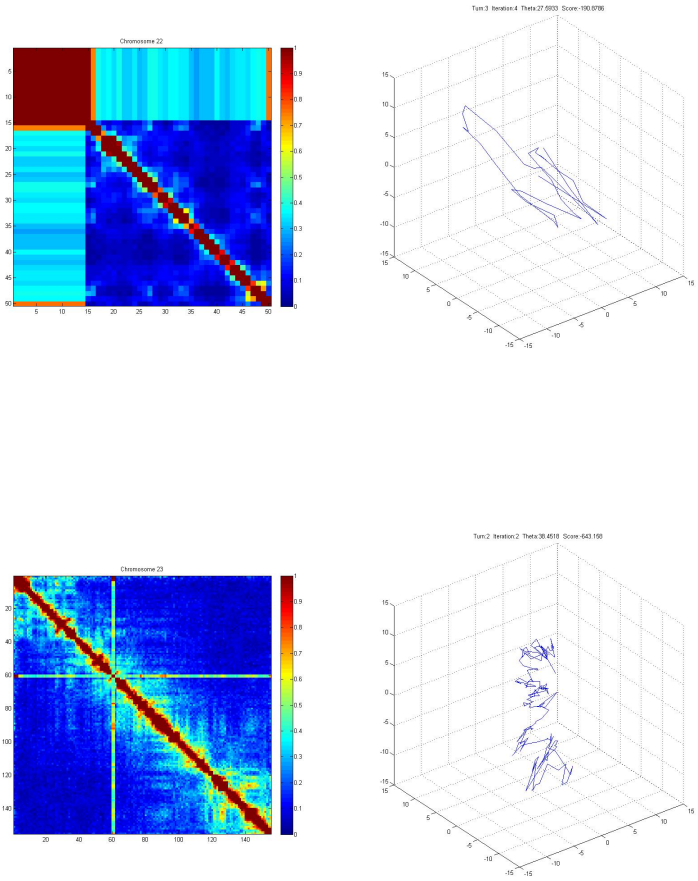


Figure 4.1: 3D structure for all chromosome.

I reconstructed chromosomal structure models in 1MB resolution, using EM based algorithm along with gradient descent method. For each chromosome, I generated an ensemble models and select a representative model for each ensemble. The criteria to select the representative is neither converged too slow nor too fast and the logarithm of the likelihood should be as high as possible. The higher the score the preciser the structure is.

4.2 Method Comparison And Discussion

As mentioned early, several researches have been done to reconstruct 3D structure for chromosome model. I would like to compare some of these important methods. There are many important methods have been developed, among which can be divided into two groups: consensus methods and ensemble methods. I will pick two methods for each group to compare.

4.2.1 ShRec3D

The ShRec3D method is based on multidimensional scaling (MDS). An important step in MDS-based methods of chromosome reconstruction is the derivation of a complete set of distances from a contact map. A weighted graph has been introduced whose nodes are the N loci detected in the experiment. The length of a link is determined as the inverse contact frequency between its end nodes. Then taking for the distance between any two nodes the length of the shortest path relating them on the graph, computed using the Floyd-Warshall algorithm. It offers a way to achieve the pre-processing step common to all 3C-based techniques of converting observed contact frequencies into a complete set of distances, independently of the downstream reconstruction method. This algorithm, which is called shortest-path reconstruction in 3D (ShRec3D), combines this shortest-path distance with MDS to achieve chromosome reconstruction.

The algorithm ShRec3D involves first the translation of a contact map into a distance matrix using a graph-theoretic method then the reconstruction of a 3D structure

using standard results from distance geometry and classical multidimensional scaling (MDS).

The advantage of this method is accuracy at a large number of data sets. But the runtime for this method ranged from tens of seconds for small data sets (1,000 points) to 50 h for the largest one (26,538 points). The limiting step for ShRec3D computation time is the Floyd-Warshall algorithm computing shortest paths on the contact map, whose worst-case performance scales as $O(N^3)$ [10].

4.2.2 ChromSDE

ChromSDE applies semi-definite programming techniques to find the best structure fitting the observed data and uses golden section search to find the correct parameter for converting the contact frequency to spatial distance. This method aims to minimize the errors between the embedded distances and the expected distances, and reformulate equations as linear and quadratic semidefinite programming (SDP) problems by relaxing the solution space from R^3 to R^n. By solving the SDP problems, the solution can be obtained as a positive semidefinite kernel matrix K. By computing the eigenvalue decomposition of K, the coordinates can be recovered from K.

The advantage of this method is to find a more accurate conversion factor. It has been shown that the conversion factor changes with different resolutions. For a frequency matrix F, the goodness of a conversion factor can be determined by comparing the predicted frequency matrix and the input frequency matrix. So the goal is to compute the conversion factor that maximizes the goodness function. And this algorithm performed well to find a proper conversion factor.

To the best of their knowledge, ChromSDE is the only method that can guarantee recovering the correct structure in the noise-free case. They showed that ChromSDE is much more accurate and robust than existing methods.

4.2.3 BACH

BACH means Bayesian 3D constructor for Hi-C data. In the BACH algorithm, they assume that the local genomic region of interest exhibits a consensus 3D chromosomal structure in a cell population, and employ efficient Markov chain Monte Carlo (MCMC) computational tools to infer the underlying consensus 3D chromosomal structure. They also assume that the number of sequencing reads spanning two genomic loci follows a Poisson distribution, where the Poisson rate is negatively associated with the corresponding spatial distance between them and is also affected by a few other factors.

Compared to other published methods, BACH has the following advantages: It explicitly models and corrects known systematic biases associated with Hi-C data, such as restriction enzyme cutting frequencies, GC content and sequence uniqueness; It utilizes a Poisson model that better fits the count data generated from Hi-C experiments than the Gaussian model used in MCMC5C, and performs more robustly when applied to several experimental datasets; It employs advanced MCMC techniques, such as Sequential Monte Carlo and Hybrid Monte Carlo, that significantly improve the efficiency in exploring the vast space of possible models.

4.2.4 MCMC5C

The MCMC5C method is a probabilistic model linking 5C/Hi-C data to physical distances and describe a Markov chain Monte Carlo (MCMC) approach to generate a representative sample from the posterior distribution over structures from interaction frequency data. The MCMC5C method fits the data into a normal distribution model. The Markov chain Monte Carlo algorithm is a method used to sample from a complex distribution, resulting in an ensemble of solutions.

They believed that tools like MCMC5C are essential for the reliable analysis of data from the 3C-derived techniques such as 5C and Hi-C. By integrating complex, high-dimensional and noisy datasets into an easy to interpret ensemble of three-dimensional conformations, MCMC5C allows researchers to reliably interpret the result of their assay and contrast conformations under different conditions.

4.2.5 Conclusion

Many researches have been done to reconstruct the 3D structure for the chromosome model. Compared to these methods that discussed above, my method has these advantages as follow. I have the same Bayesian inference structure as the BACH, which means that I can corrects many systematic biases that created in the Hi-C experiment, such as restriction enzymes, GC content and sequence uniqueness. Admittedly Poisson model might better fits the count data generated from Hi-C experiments than the normal distribution, but in the EM algorithm performs more robustly when doing the maximizing process. Gradient descent is a good method for finding a local maxima, the advantage is the converge speed. In my research for 1MB resolution the average converge speed is approximately 60 seconds, while in

the ShRec3D method it takes 50 hours. One of the key advantages of my approach, compared to non-probabilistic or maximum likelihood approaches like ChromSDE or ShRec3D, is its ability to estimate the distribution of various structural properties, and thus to report both averages and confidence intervals for the selected properties.

The disadvantage of my method is not finding an accurate conversion factor compared to ChromSDE. And also the gradient descent method sometimes stuck in the local minima, which might not be a situation that I want. So the solution here is to run multiple times and discard local minima which can be detected easily, and integrate the result.

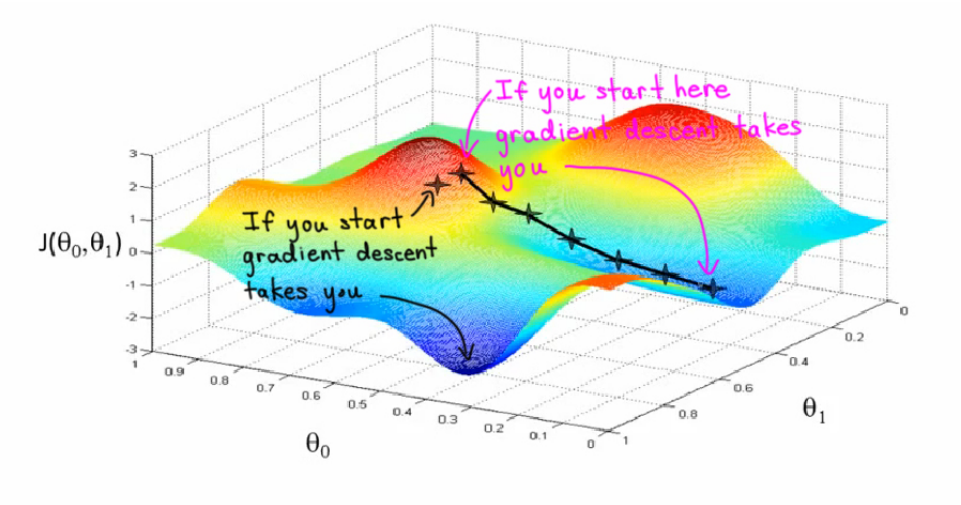


Figure 4.2: Gradient descent might stuck in local minima which relay on different initialization.

Chapter 5

Summary and concluding remarks

Understanding the structure of human chromosome is very important for knowing what we are. Recently, chromosome conformation capture based techniques have been developed rapidly, which lead to a development of 3D chromosome structure conformation. Additionally, an advanced 3C technique, Hi-C, has been developed to determine the interaction between chromosomes loci, which provides important information for understanding and reconstructing the 3D model of a chromosome.

The steps of processing the Hi-C data includes reading, classifying and normalization. I classified the data into 1MB resolution, which gives me a better view and quicker responding time rather than other resolution. For normalization, I used Sequential Component Normalization (SCN) methodology for normalization. This normalization can be viewed as a sequence of extensions and shrinking of interaction vectors so that they tend to reach the sphere of radius one in the interactions space.

After finishing data processing and normalization, I proposed a method that can actually reconstruct the 3D model of chromosome structure. The method that I

used is based on EM algorithm, which is an important method in both machine learning and data mining. I made an assumption that the Hi-C noise follows a normal distribution and apply Bayesian inference model on that, which gives me the objective function of a chromosome structure likelihood. After that I can apply the EM algorithm by using gradient descent method in expectation step and using the definition of the variance in maximization step. By doing this algorithm iteratively, I can finally retrieve the 3D structure of a chromosome model from chromosomal contact data.

Appendix A

Software and Source Code

A.1 Software and Environment

I use MATLAB, which is a really strong tool for matrix manipulation in this research. The version that I use is 7.11.0.584 (R2010b) on Windows 7 OS.

A.1.1 Input

The input of my software should be the row data of Hi-C, which indicates the chromosome contact in each row. And also the chromosome number should be given.

A.1.2 Output

The output is a figure for the given chromosome number. The left figure is the heat map for the chromosome after normalization; the right one is the 3D structure of the chromosome.

A.2 Source Code

A.2.1 main.m

```
clear;
clc;

name=input('Please input chromosome number: ');

contact_matrix;

normalization;

figure(1);
% set(gcf, 'position', [200 0 1000 2000]);
subplot(1,2,1),imagesc(contact);
axis image;
caxis([0 1]);
title(['Chromosome ',num2str(name)]);
```

```
colorbar;

%figure(2);
method_normal;
```

A.2.2 contact_matrix.m

```
filename = ['normal/', num2str(name)];
% resolution=200000;%resolutions of 200KB
resolution=1000000;%resolutions of 1MB


fp=fopen(filename,'r');
number=fscanf(fp,'%d %d %d %d',[4 inf]);%read data
number=number';
maxnumber=max(max(number));
regionnumber=ceil(maxnumber/resolution);
contact=zeros(regionnumber);
row=size(number,1);
for i=1:row
    n=ceil(number(i,2)/resolution);
    m=ceil(number(i,4)/resolution);
     contact(n,m)=contact(n,m)+1;
     contact(m,n)=contact(m,n)+1;
end
clear number;
```

```
fclose(fp);
```

A.2.3 normalization.m

```
n=size(contact,1);
contact=contact+1;
norms=sum(contact);
for i=1:n
    for j=1:n
        contact(i,j)=contact(i,j)*row/(norms(i)*norms(j));
    end
end
```

A.2.4 method_normal.m

```
format long;
hold off;
%contact=contact.^3;
Data=1./(contact);
n = size(Data, 1); % initial random samples
xyz=2*rand(n,3)-1;
theta=rand;
epsilon=0;
scale=3.38;
step=0.001; % step of gradient descent
```

```
count=0;
t=1;

goal_old=normal_goal(n,Data,xyz,theta);
goal_start=goal_old;
xyz_temp=xyz;
while t<1000
    % maximum coordinates
    count=0;
    while count<1000
        for i=1:n
            gradient_x=0;
            gradient_y=0;
            gradient_z=0;
            for j=1:n % calcuate gradient
                if(i~=j)
                    d=norm(xyz_temp(i,:)-xyz_temp(j,:));
                    gradient_x=gradient_x-(xyz_temp(i,1)-
                        xyz_temp(j,1))*(1-Data(i,j)/d)/theta
                        ^2;
                    gradient_y=gradient_y-(xyz_temp(i,2)-
                        xyz_temp(j,2))*(1-Data(i,j)/d)/theta
                        ^2;
```

```
                gradient_z=gradient_z-(xyz_temp(i,3)-
                    xyz_temp(j,3))*(1-Data(i,j)/d)/theta
                    ^2;
            end
        end
        gradient=[gradient_x,gradient_y,gradient_z];
        %gradient=gradient/norm(gradient);
        xyz_temp(i,:)=xyz_temp(i,:)+step*gradient; %
            gradient ascent find maximun
    end
    goal_new=normal_goal(n,Data,xyz,theta);
    if(goal_new-goal_old<epsilon)
        break;
    end
    goal_old=goal_new;
    xyz=xyz_temp;
    count=count+1;

    subplot(1,2,2),plot3(xyz(:,1),xyz(:,2),xyz(:,3));
    axis image;
    title(['Turn:',num2str(t),'  Iteration:',num2str(
        count),'  Theta:',num2str(theta),'  Score:',
        num2str(goal_old)]);
    xlim([-15 15]);
```

```
        ylim([-15  15]);
        zlim([-15  15]);
        grid on;
        pause(0.01);
    end

    % maximum parameters
    score_old=goal_old;
    theta_new=0;
    for i=1:n-1
        for j=i+1:n
            d=norm(xyz(i,:)-xyz(j,:));
            theta_new=theta_new+(d-Data(i,j))^2;
        end
    end
    theta_new=sqrt(theta_new/n);
    score_new=normal_goal(n,Data,xyz,theta_new);
    if(score_new-score_old<epsilon)
        break;
    end
    theta=theta_new;
    score_old=score_new;
    t=t+1;
end
```

Bibliography

[1] J. M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–76, 2012.

[2] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.

[3] T. B. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159):731–4, 2013.

[4] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–11, 2002.

[5] S. Ben-Elazar, Z. Yakhini, and I. Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Res*, 41(4):2191–201, 2013.

[6] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–7, 2010.

[7] J. Paulsen, G. K. Sandve, S. Gundersen, T. G. Lien, K. Trengereid, and E. Hovig. Hibrowse: multi-purpose statistical analysis of genome-wide chromatin 3d organization. *Bioinformatics*, 30(11):1620–2, 2014.

[8] C. Peng, L. Y. Fu, P. F. Dong, Z. L. Deng, J. X. Li, X. T. Wang, and H. Y. Zhang. The sequencing bias relaxed characteristics of hi-c derived data and implications for chromatin 3d modeling. *Nucleic Acids Res*, 41(19):e183, 2013.

[9] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–72, 2012.

[10] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nat Methods*, 11(11):1141–3, 2014.

[11] N. L. van Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander. Hi-c: a method to study the three-dimensional architecture of genomes. *J Vis Exp*, (39), 2010.

[12] L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–63, 2014.

[13] M. A. Ferraiuolo, M. Rousseau, C. Miyamoto, S. Shenker, X. Q. Wang, M. Nadler, M. Blanchette, and J. Dostie. The three-dimensional architecture of hox cluster silencing. *Nucleic Acids Res*, 38(21):7472–84, 2010.

[14] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 20(6):761–70, 2010.

[15] S. De and F. Michor. Dna replication timing and long-range dna interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol*, 29(12):1103–8, 2011.

[16] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80, 2012.

[17] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–13, 2012.

[18] F. Ay and W. S. Noble. Analysis methods for studying the 3d architecture of the genome. *Genome Biol*, 16:183, 2015.

[19] E. A. Coutsias, C. Seok, and K. A. Dill. Using quaternions to calculate rmsd. *J Comput Chem*, 25(15):1849–57, 2004.

[20] N. Varoquaux, I. Liachko, F. Ay, J. N. Burton, J. Shendure, M. J. Dunham, J. P. Vert, and W. S. Noble. Accurate identification of centromere locations in yeast genomes using hi-c. *Nucleic Acids Res*, 43(11):5331–9, 2015.

[21] Z. Zhang, G. Li, K. C. Toh, and W. K. Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *J Comput Biol*, 20(11):831–46, 2013.

[22] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80, 2014.

[23] M. Rousseau, J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC Bioinformatics*, 12:414, 2011.

[24] T. Trieu and J. Cheng. Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res*, 42(7):e52, 2014.

[25] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–65, 2011.

[26] Z. Wang, R. Cao, K. Taylor, A. Briley, C. Caldwell, and J. Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PLoS One*, 8(3):e58793, 2013.

[27] F. Ay, E. M. Bunnik, N. Varoquaux, S. M. Bol, J. Prudhomme, J. P. Vert, W. S. Noble, and K. G. Le Roch. Three-dimensional modeling of the p. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res*, 24(6):974–88, 2014.

[28] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C. A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–4, 2013.

[29] D. Bau, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 18(1):107–14, 2011.

[30] J. E. Lemieux, S. A. Kyes, T. D. Otto, A. I. Feller, R. T. Eastman, R. A. Pinches, M. Berriman, X. Z. Su, and C. I. Newbold. Genome-wide profiling of chromosome interactions in plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. *Mol Microbiol*, 90(3):519–37, 2013.

[31] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, 2012.

[32] N. Varoquaux, F. Ay, W. S. Noble, and J. P. Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–33, 2014.

[33] M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, 9(1):e1002893, 2013.

[34] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30(1):90–8, 2012.

[35] K. J. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, and P. Meinicke. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics*, 9:217, 2008.

VITA

Yuxiang Zhang was born in the city of Shanghai, China. He finished his undergraduate studies in 2013 from East China University of Science and Technology, China, majoring in Computer Science. In Fall 2013, Yuxiang joined the Department of Computer Science in University of Missouri - Columbia to pursue his M.S. studies under the advice of Professor Jianlin Cheng. His research is focused on applying machine learning and data mining techniques to analyze big biomedical data and address fundamental problems in biomedical sciences and he enjoys taking approaches that combine computational optimization and statistical methods with bioinformatics and systems biology.