Protein Design and Chromatin Structure: Novel Computational Approaches

BY

YUN XU
B.S., Shanghai Fisheries University (Shanghai Ocean University), 1995
M.S., Fudan University, 2004

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Jie Liang, Chair and Advisor
Yang Dai
Ao Ma
G. Ali Mansoori
Amy Kenter, Microbiology and Immunology

To my parents, Yougen Xu and Qunying Tian;

my brother, Zhixiang Xu;

and my sister, Lei Xu

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Prof. Jie Liang, for his continuous inspiration, encouragement, and support of my Ph.D study and research. This thesis would not have been possible without his guidance from the initial to the final phases. I am thankful to have him as my role model and mentor. His immense knowledge enabled me to develop an understanding of research and to work on diverse and exciting projects. His enthusiasm and motivation allowed me to build critical thinking skills and also help me to regain my passions on those difficult days. I could not have imagined having a better advisor.

In addition, I would like to thank the rest of my thesis committee: Prof. Yang Dai, Prof. Ao Ma, Prof. G. Ali Mansoori, and Prof. Amy Kenter, for their encouragement, insightful comments, and interest in this project. I am also grateful to Prof. Kenter for her generous support and guidance during my research.

I would like to express my gratitude for Prof. Weiyuan Qiu and Prof. Fuyao Ren. Prof. Qiu and Prof. Ren were the first to show me how beautiful a mathematical world is when I was in Fudan University. I would have never pursued a Ph.D in bioinformatics without their inspiration and encouragement.

I would like to thank lab alumni: Dr. Jinfeng Zhang, Dr. Xiang Li, Dr. Yan Yuan Tseng, Dr. Ronald Jackups, Jr, Dr. Sema Kachalo, Dr. Ming Lin, and Dr. Jian Zhang. It has been a great pleasure working with you all. Also I would like to thank my fellow colleagues: Gamze Gürsoy, Jieling Zhao, Yingzi Li, Marco Maggioni, Meishan Lin, Michael Montesano, Volga Pasuleti, Ying Wang, and Ke Tang for the stimulating discussions and all the fun we had. In particular, I am grateful to Dr. Larisa Adamian, Dr. Joe Dundas, Dr. David Jimenez Morales, Dr. Hammad Naveed, Dr. Youfang Cao, Dr. Joel Fontanarosa, Dr. Gang Feng, Dr. Lei Huang, Dr. Zheng Ouyang and Dr. Hsiaomei Lu for the constant support and friendship. My sincere

thanks also goes to my friends and all of those who supported me in any respect during the completion of this degree.

Last but not the least, I would like to thank my parents for their unconditional love and constant support throughout my entire life. Also I am grateful to have a wonderful brother and sister who have supported me in many ways.

Chapter 2 text, figures and tables are submitted to the journal PLOS ONE. It is under review process. The copyright permission is not required. Please see the appendices.

CONTRIBUTION OF AUTHORS

Chapter 1 is a literature review that places my dissertation question in the context of the larger field and highlights the significance of my research question. Chapter 2 represents a unpublished manuscript for which I was the primary author and major driver of the research. Changyu Hu assisted me in the designing experiments. I designed the experiment, wrote code, analyzed result, validated test. I generated all figures, tables and played a major role in the writing of the manuscript along with Dr. Dai and my research mentor, Dr. Liang guidance. The manuscript is submitted to PLOS ONE, it is under review process. Chapter 3 represents a unpublished manuscript for which I was the primary author and major driver of the research. Inspired discussions were made between Gamze, Dr. Kenter and Dr. Liang. I designed the experiment, wrote code, analyzed result, validated test. I generated all figures and played a major role in the writing of the manuscript along with Dr. Kenter and Dr. Liang guidance. I anticipate that this line of research will be continued in the laboratory after I leave and that this work will ultimately be published as part of a co-authored manuscript.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support vector machine
RSVM	Reduced support vector machine
RMSD	Root Mean Square Deviation
ChIP	Chromatin immunoprecipitation
ChIPSeq	ChIP with massively parallel sequencing
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
FISH	Fluorescence *in situ* hybridization
3C	Chromosome conformation capture
4C	Chromosome conformation capture-on-chip or circular Chromosome conformation capture
5C	Chromosome conformation capture carbon copy
Hi-C	Genome-wide Chromosome Conformation Capture
CTCF	11-zinc finger protein or CCCTC-binding factor
PCR	Polymerase chain reaction
TSS	Transcription start site
ENCODE	Encyclopedia of DNA Elements

LIST OF ABBREVIATIONS (Continued)

SIS	Sequential Importance Sampling
FDR	False discovery rate

SUMMARY

I studied three dimensional structural problems of biological systems as they related to two aspects of genetic information: protein design and folding chromatins.

I studied the problem of constructing the fitness landscape of inverse protein folding. Fitness landscapes have broad implications in molecular evolution, cellular epigenetic state, and protein design. Computational inverse protein folding, or protein design, aims to generate amino acid sequences that fold into an *a priori* determined structural fold for engineering novel or enhanced biochemical structures. For this task, a function describing the fitness landscape of sequences is critical to identify correct ones that fold into the desired structure. In this study, I showed that nonlinear fitness functions for protein design can be significantly improved relative to those published in the scientific literature. Using a rectangular kernel with a basis set of proteins and decoys chosen *a priori*, I obtained a simplified nonlinear kernel function via a finite Newton method. The full landscape for a large number of protein folds was captured using only 480 native proteins and 3,200 non-protein decoys. A blind test of a simplified version of sequence design was carried out to simultaneously discriminate 428 native sequences not homologous to any training proteins from 11 million challenging protein-like decoys. This simplified fitness function correctly classified 408 native sequences (20 misclassifications, 95% correct rate), which outperformed several other statistical linear scoring function and optimized linear functions. The performance was also comparable with results obtained from a far more complex nonlinear fitness function with $>$ 5,000 terms. The results further suggested that for the task of global sequence design of 428 selected proteins, the search space of protein shape and sequence can be effectively parametrized with just about 3,680 carefully chosen basis set of proteins and decoys. In

addition, I showed that the overall landscape was not overly sensitive to the specific choice of this set. These results may be generalized to construct other fitness landscapes.

The portion of this study related to chromatin folding was based on the data obtained from the Chromosome Conformation Capture (3C)-based technologies, which are used to detect pairs of loci located on the same chromosome or on different chromosomes that are in close spatial proximity. There are biases which may affect the 3C-based experimental procedure, including the non-alternative primer design and the distance between restriction sites. To overcome such biases, I developed a general novel constrained self-avoiding chromatin (C-SAC) model to remove non-specific physical interactions. I further developed a sequential importance sampling algorithm to rebuild 3D chromatin structures on 5C experiments, which are derived from 3C technology. I applied this approach to the ENCODE region ENm008 α-globin gene domain on human chromosome 16 for the lymphoblastoid cell (GM12878) and the chronic myelogenous leukemia cell (K562). I successfully removed non-specific physical interactions from the 5C reads for both two cells by our random ensemble generated by C-SAC model. My results showed that α-globin gene domain is a compact globule in the GM12878 cell, and it is formed by two separate domains in the K562 cell. My studies show that we can not only recover most of 5C indicated proximity interactions, but can also discover new proximity interactions which are not captured in the 5C experiments. Specifically C-SAC model can recover 77% of the known interactions after comparing with the results from independent ChIA-PET measurements. Based on the ensemble of the reconstructed 3D conformations, a novel mechanism was proposed, which may explain why the α-globin gene is inactive in the GM12878 cell but is active in the K562 cell.

SUMMARY (Continued)

In chapter 2, I will discuss the problem of protein design. In chapter 3, the studies of chromatin structures will be discussed.

CHAPTER 1

INTRODUCTION

1.1 DNA is a carrier of hereditary information of the cell

DNA (deoxyribonucleic acid) carries hereditary information of the cell. In the reproductive process of living cells, the parent cells' DNA is replicated and passed to the next generation. The DNA molecule stores genetic information in every living cell that determines the characteristics of a species as a whole as well as the individuals within each species.

The famous Hershey-Chase experiment conducted by Alfred Hershey and Martha Chase in 1952 provided the evidence for the role of DNA as carrier of genetic information (Hershey and Chase, 1952). Bacteriophages, the viruses that attack bacteria, consist of a protein coat surrounding a DNA core. During the process of infection, the radioactive protein coat of the bacteriophage was shown to remain outside of the bacteria cell, while the bacteriophage injected their radioactive genetic material, DNA, into the bacteria cells. Once inside the bacteria cells, the bacteriophage gene directs the synthesis of new bacteriophage and assembles more offspring of bacteriophages. This piece of evidence, together with Avery-MacLeod-McCarty experiment (Avery et al., 1944), firmly concluded that DNA is the agent of heredity.

DNA consists of two long polynucleotide chains, or strands. Each of DNA chains is composed of four type of nucleotide subunits: adenine (A), cytosine (C), guanine (G) and thymine (T). These four subunits in a DNA chain are covalently linked together by sugars and phosphates. Another complementary chain follows base-pairing rules: A is always paired with T, C is always paired with G. These two polynucleotide

chains running anti-parallel are held together by hydrogen-bonding between the bases on the different chains and interwined to form a double helix structure (Watson and Crick, 1953). This double helix structure of DNA allows it to carry information and to be faithfully duplicated and provides a mechanism of heredity.

1.2 DNA and Protein

Gene are short segments of DNA which contain the information to manufacture specific proteins. There are two important steps to process genetic information. The first step is transcription. The DNA double helix unwound, and RNA (ribonucleic acid) is synthesized from the beginning of the gene to the end of the gene by using one of DNA strands as the template. RNA is chemically similar to DNA except it contains a different sugar in its nucleotides and contains the closely related nitrogenous base uracil (U) instead of thymine (T). In contrast to double helix structure DNA, RNA is generally single stranded. In the transcription stage, the RNA based on the DNA template is called mRNA (messenger RNA). The second step is translation. The mRNA is transported out of the cell nucleus, and is moved into the cytoplasm in eukaryotes. The ribosome, a large and complex molecular machine, with the aid of tRNA (transfer RNA), recognizes the information encoded in the mRNA codons, and carries the proper amino acids to synthesize the coded protein. Each mRNA codon is constituted by a group of three base pairs and corresponds to a particular amino acid. The mRNA sequence is used as a template in units of three base pairs to assemble the chain of amino acids to form a protein. Different mRNA codons may be used to encode the same amino acid. There are three stop codons to guide ribosome to stop translation.

1.3 Protein Structure

There are four distinct levels of protein structure in the protein folding process.

The amino acid sequence constitutes the primary level of the protein structure. A protein molecule is made from a long chain of amino acid sequence. The backbone of the polypeptide chain is built by the formation of peptide bonds between amino acids. To form peptide bonds, different amino acids are linked together between their carboxylic acid groups and amino groups. The dipeptide molecule has a free amino end (N-terminal end) and a free carboxylic acid end (C-terminal end) such that it can be reiteratively linked to form peptide bonds and finally form polypeptide chains. The length of the gene that codes the protein determines the length of the polypeptide chain ranging from less than 50 amino acids to more than 10,000 amino acids.

The secondary structure refers to regular local stable sub-structures which includes the α-helix and the β-strand conformations commonly found in folded proteins. In the α-helix conformation, hydrogen bond interactions between the carbonyl oxygens and the nitrogens are in adjacent peptide bonds. The large amount of intra-helix hydrogen bonds between peptide bonds is a major contribution to stabilize the α-helix structures. The side-chain groups are on the outside of the α-helix and perpendicular to the α-helix axis. The β-strands are only stable when hydrogen bonds link two parallel or anti-parallel peptide bond segments.

The tertiary structure refers to three dimensional structure of a single protein molecule. The folding is driven by non-specific hydrophobic interactions. Dipole interactions with water molecules of the solvent lead to spatial arrangements in which hydrophobic amino acids are on the inside of the structure and residues with polar groups are on the outer surface. Hydrophobic interactions play a central role in determining the shape of a protein.

Multiple-subunit proteins possess a quaternary structure, which is a combination of two or more chains to form a complete unit. For example, hemoglobin contains four polypeptide subunits: two α-globin polypeptide chains and two β-globin polypeptide chains.

The order of amino acid sequence determines the protein three-dimensional structure. A protein generally folds into a single stable conformation in which the free energy is minimized, and this structure is critical fro a given protein's biological activity.

1.4 DNA and Chromatin

DNA as a hereditary agent, combined together with associated proteins, is highly folded in the nucleus. Chromosomes (chromatins) are the structures which can pack very long double stranded helix DNA molecules into. In a haploid human cell, there are 23 chromosomes, including about 3.2 billion base pairs long and containing about 20,000-25,000 distinct protein-coding genes. The fully extended length about 2 m of human DNA is confined into the only 5-20 μm in diameter cell nucleus. It is similar to folding 40 km of extremely fine thread into a tennis ball. DNA, bounded and folded by specialized proteins, can generate a series of coils and loops that provide increasingly higher levels of organization and prevent DNA from becoming a unmanageable tangle. Meanwhile, chromosome also allows DNA to remain accessible to all the enzymes and proteins to replicate it, repair it, and direct it to express genes.

During the interphase stage of cell, DNA is less tightly packed than DNA in the mitosis stage. So the DNA in interphase stage is highly packed but less condensed. While the chromatin is a lower order of DNA organization, chromosomes are higher order of DNA organization. During interphase, the chromatins are extended in the nucleus and cannot be easily distinguished with a light microscope. When the cell reaches mitotic phase, the DNA coils up and is tightly compacted, therefore it can be easily visualized.

Interphase chromosome contains both condensed (heterochromatin) and more extended forms of chromatin (euchromatin). Some of the DNA folds into heterochromatin and does not contain any genes. Heterochromatin is usually localized to the periphery of the nucleus. Euchromatin is rich of genes, compact, and often under active transcription (International Human Genome Sequencing Consortium, 2004).

1.5 Chromatin Structure

Histones, the most abundant proteins in chromatin, are small, positively charged proteins of five major types: H1, H2A, H2B, H3, and H4. All histones have a high percentage of arginine (Arg) and lysine (Lys). These positively charged amino acids give the histones a net positive charge which attracts the negative charges on the phosphates of DNA such that DNA can tightly interact with histones.

DNA wraps around the histones to form nucleosomes. The nucleosome is a basic unit in eukaryotes consisting of a segment of DNA wrapped 1.65 times around an octamer of histone proteins core (two of each H2A, H2B, H3, and H4). The fifth type of histone, H1, binds to the linker region of DNA where the DNA joins and leaves the octamer and help to lock the DNA into place. Then nucleosomes fold to form a dense, tightly packed structure and make up a fiber with a 30 nm diameter.

The 30 nm fiber is folded into a series of loops, and these loops are further packed and folded to produce a 250-nm-wide fiber. This compact fiber undergoes at least one more level of packing to finally form the mitotic chromosome.

1.6 Central dogma of molecular biology

The famous central dogma of molecular biology was stated by Francis Crick (Crick, 1958; Crick, 1970). The general transfer is one of which can occur in all cells:

- DNA $\rightarrow$ DNA

- DNA $\rightarrow$ RNA
- RNA $\rightarrow$ protein

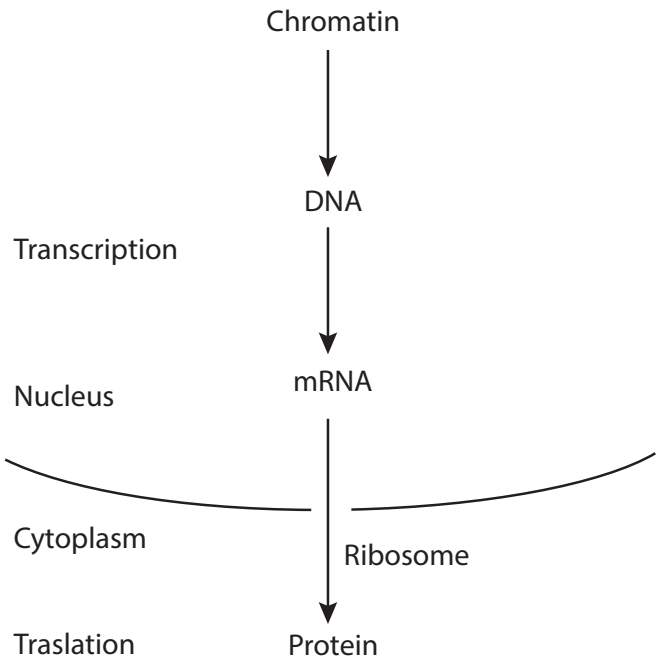

Figure 1: The general principle of genetic information flow in eukaryotes. The DNA self-copy is not shown. In the eukaryotes, chromatin carries DNA, DNA can be cloned by itself, DAN transcript to mRNA, and mRNA will eventually produce protein.

The central dogma of molecular biology is the general principle of genetic information flows which applies to both prokaryotes and eukaryotes. In eukaryotes, chromatin carries DNA. DNA is copied and

then is transcripted to mRNA in the nucleus, and mRNA will be transported to cytoplasm and eventually translated into protein with the aid of ribosomes and tRNA (Figure 1).

Gene regulation can occur at any of the stages leading from DNA to protein in eukaryotes. Chromatin unwinding DNA is regulated by chromatin remodeling. DNA transcription to mRNA is regulated by transcription, splicing and processing. mRNA translation to protein is regulated by transportation and translation. The final product protein may also be regulated by protein modifications. Under these complex and delicate regulation mechanisms, genes can be expressed at various level at specific time, in different cell types, and in response to complex environment.

1.7 Living creature lives in three dimensional space

As Stephen Hawking said, we can only live in the three dimensional space (Hawking, 1996). Two dimensional space is too simple to have capacity to accommodate our body. For example, if a two dimensional creature eats something like an apple into its body, its digestive system will have to form a passage throughout its body. Then the two dimensional creature will be divided into two separate parts and it will be difficult to digest any food. This also impossible to live in more than three dimensional space. The gravitational force between two bodies would decrease much more rapidly with distance in higher than three dimensional space, and the earth would either be thrown away from or toward the sun. We would either freeze or be burned up. Studying three dimensional structure is therefore very important.

In my thesis work, I studied three dimensional structures at the two ends of general genetic information flow. The first portion of this work involved studying factors influencing the ultimate product of the gene expression: protein. The other component researched in this work was genetic information carrier via chromatin modifications and regulation.

1.8 Protein design

The three-dimensional structure of protein allows for diverse functions. For example, an essential element of cellular metabolism is that hemoglobin and myoglobin proteins transport oxygen to cells via structural rearrangement. Structural proteins, such as collagen and keratin, comprise a large proportion of the skin, connective tissue, and hair of organisms. Immunoglobuline proteins, produced by plasma cells in response to an immunogen, is essential in the immune system of vertebrates. Histones, which bind to DNA in eukaryotic organisms, combine with transcription factors to regulate gene expression. Many researchers focus on predicting protein structure from amino acid sequence (Becker et al., 2006; Qian et al., 2007; Jagielska et al., 2008; Zhu et al., 2008). The inverse problem is also equally important. It can be stated as follows: given a protein structural template, can we identify the protein sequences that would fold into this structural template. This problem represents a major ongoing challenge in computational biology research and bioinformatics.

This problem was first formulated 30 years ago (Drexler, 1981; PABO, 1983). Also known as the inverse protein folding problem, it addresses the fundamental problem of designing proteins to facilitate engineering of enhanced or novel biochemical functions. Protein design has been the focus of intense theoretical, computational, and experimental studies (Desmet et al., 1992; Yue and Dill, 1992; Shakhnovich and Gutin, 1993; Li et al., 1996; Deutsch and Kurosky, 1996; Dahiyat and Mayo, 1997; Kleinberg, 1999; Hill et al., 2000; Siegel et al., 2010)

A key component for designing protein is a fitness function: it detects whether a solution has been found, and guides the search of viable sequences. An ideal fitness function can characterize the properties of fitness landscape of many proteins simultaneously. Such a fitness function would be useful for designing

novel proteins and novel functions, and for studying the global evolution of protein structure and protein functions.

Many protein design studies employ a linear fitness function in the form of weighted linear sum of pairwise contacts, sometimes with additional solvation terms derived from exposed surface area (Shakhnovich and Gutin, 1993; Deutsch and Kurosky, 1996; Yue and Dill, 1992). Such functions can be obtained from statistical analysis of a database of protein structures (Miyazawa and Jernigan, 1985), by perceptron learning/linear programming (Vendruscolo et al., 2000; Tobi et al., 2000; Wagner et al., 2004), or by gradient descent (Bastolla et al., 2000; Bastolla et al., 2001). Another approach is to use a force field modeled after those used in molecular mechanics simulation (Dahiyat and Mayo, 1997). Often these fitness functions have their roots in protein folding studies. However, they do not provide global characterization of fitness landscapes for protein design. They also often have poor performance in blind tests when challenged with the task of simultaneously designing many different proteins (Hu et al., 2004), or are unsuitable for high-throughput testing.

A promising alternative approach is to use a nonlinear function to capture the complex design fitness landscape. In a recent study, a nonlinear Gaussian kernel function was constructed by maximizing soft margins between native proteins and decoy non-proteins (Hu et al., 2004). This fitness function significantly outperforms linear functions in a blind test of identifying 201 native proteins from 3 million challenging protein-like decoys.

However, it is parametrized by about 350 native proteins and 4700 non-protein decoys and its form is rather complex. This makes the evaluation of the fitness of a candidate sequence demanding. Although obtaining a good answer at high computational cost is acceptable for some tasks, it is difficult to incorporate

this type of a complex function in a search algorithm. It is also difficult to characterize landscape properties of protein sequence design using a complex function.

In Chapter 2 , I will discuss the challenge of protein design problem by the reduced support vector machine approach.

1.9 Chromatin structure

DNA is not naked *in vivo* but associated with basic proteins, the histones, to form a nucleoprotein complex named "chromatin". Chromatin inevitably holds a significant part of the regulatory information encoded in the nucleus, sometimes collectively referred to as "epigenetic" information. After more than 30 years of intense experimental and modeling efforts, chromatin structure remains one of the major unsolved problems in molecular biology (Widom, 1989; van Holde and Zlatanova, 1996; van Holde and Zlatanova, 2007).

The highly folded three-dimensional chromatin structure, a carrier of genetic material, also plays a essential gene regulation role while it is unwound. Chromosomal rearrangements underlie a variety of malignant cancers and congenital diseases. For example, aggressive mature B cell lymphomas often carry translocations involving the immunoglobulin heavy chain (IgH) locus which coupled to C-MYC (Hoffman et al., 2012). Most of the common mutation groups (over 75%) among the known cancer genes in somatic cells involve chromosomal translocations that creates a chimeric gene or apposes a gene to the regulatory elements of another gene (Futreal et al., 2004). Therefore it is important to predict the chromatin structure to shed some light on how gene regulation happened during in the chromatin remodeling.

Since the development of the 3C technology (Dekker et al., 2002), chromatin interaction data is considered to be employed to model the 3D chromatin structure. And its high-throughput modifications that

include several related methods: 4C (Duan et al., 2010; Zhao et al., 2006; Simonis et al., 2006; Schoenfelder et al., 2010), 5C (Dostie et al., 2006), Hi-C (Lieberman-Aiden et al., 2009), and ChIA-PET (Fullwood et al., 2009). These 3C-based methods use formaldehyde cross-linking to capture interacting loci. It is followed by DNA fragmentation and after which ligation on cut DNA is performed to obtain unique ligation products from interacting loci. Combining these methods with pair-end sequencing for PCR amplification and next generation sequencing techniques have enabled of determination of long-range interactions on a locus or genome-wide scale (Sachs et al., 1995; Bohn and Heermann, 2010; Rippe, 2001).

In recent decade, new approaches to model 3D genomes and genomic domain structures have been developed. All these approaches have in common that, to the largest possible satisfy the experimental interaction data, they developed diverse experiments (3C, 4C, 5C, HiC) (Duan et al., 2010; Fraser et al., 2009; Jhunjhunwala et al., 2008) and computation to build 3D chromatin structure (Baù et al., 2010)

The chromatin is highly dynamic in these methods, therefore these static 3D models represent different substantial in the reflection of the overall trends for chromatin folding and particular path of chromatin fiber in a given cell (Baù et al., 2010). These models did not provide accurate structures and cannot be applied to larger genomic segments.

I will discuss this problem based on the partial 5C experimental data to deduce the three dimensional structure of the α-globin gene domain for two different cell types in chapter 3.

CHAPTER 2

PROTEIN DESIGN

2.1 Introduction

Protein design has been the focus of many experimental, theoretical, and computational studies (Desmet et al., 1992; Yue and Dill, 1992; Shakhnovich and Gutin, 1993; Li et al., 1996; Deutsch and Kurosky, 1996; Dahiyat and Mayo, 1997; Kleinberg, 1999; Hill et al., 2000; Siegel et al., 2010). Despite significant challenges, important progresses have been made, with profound implications in biotechnology and biomedicine (Bolon and Mayo, 2001; Röthlisberger et al., 2008; Jiang et al., 2008; Lazar et al., 2006; Joachimiak et al., 2006; Shifman et al., 2006).

Here we studied the problem of designing a protein sequence that is compatible with an *a priori* specified three-dimensional template protein fold. This problem was first formulated 30 years ago (Drexler, 1981; PABO, 1983). Also known as the inverse protein folding problem, it addresses the fundamental problem of designing proteins to facilitate engineering of enhanced or novel biochemical functions.

A key component for designing a protein sequence is a fitness function: it can detect if a solution has been found, and can also guide the search of viable sequences. An ideal fitness function can characterize the properties of fitness landscapes for many proteins simultaneously. Such a fitness function would be useful

---

[1]**THIS CHAPTER TEXT, FIGURES, AND TABLES ARE SUBMITTED TO THE JOURNAL PLOS ONE, IT IS UNDER REVIEW PROCESS. THE COPYRIGHT PERMISSION IS NOT REQUIRED. PLEASE SEE THE APPENDICES.**

for designing novel proteins and novel functions, as well as for studying the global evolution of protein structure and protein functions.

The development of a fitness function for protein design is closely related to the development of a scoring function for protein structure predictions, protein folding, and protein-protein/ligand docking (Huang and Zou, 2006; Huang and Zou, 2008; Huang et al., 2010; Wagner et al., 2004; Májek and Elber, 2009; Ravikant and Elber, 2010). There are many different approaches in constructing the fitness function. Several studies employ a linear fitness function in the form of weighted linear sum of pairwise contacts, with sometimes additional solvation terms derived from exposed surface area (Yue and Dill, 1992; Shakhnovich and Gutin, 1993; Deutsch and Kurosky, 1996). Such functions can be obtained from statistical analysis of a database of protein structures (Miyazawa and Jernigan, 1985), or from perceptron learning/linear programming (Wagner et al., 2004; Vendruscolo et al., 2000; Tobi et al., 2000), or by gradient descent (Bastolla et al., 2000; Bastolla et al., 2001). Another approach is to use a force field such as those used in molecular dynamics simulations (Dahiyat and Mayo, 1997; Jacak et al., 2012; Pokala and Handel, 2005; Liang and Grishin, 2004). However these functions often do not provide global characterization of the overall fitness landscape for protein design. They also often have poor performance in blind test when challenged with the task of simultaneously designing many different proteins (Hu et al., 2004), or they are so complex that they cannot be used in high-throughput testing. Inaccurate fitness functions would lead to low success rates in protein design (Li et al., 2013).

A promising alternative approach is to use nonlinear function to capture the complex design fitness landscape. In the study of (Hu et al., 2004), a nonlinear Gaussian kernel function was constructed by maximizing soft margins between native proteins and decoy non-proteins. This fitness function significantly

outperforms linear functions in a blind test of identifying 201 native proteins from 3 million challenging protein-like decoys (Hu et al., 2004). However, it is parametrized by about 350 native proteins and 4,700 non-protein decoys and its form is rather complex. It is computationally expensive to make the evaluation of the fitness of a candidate sequence. Although obtaining a good answer at high computational cost is acceptable for some tasks, it is difficult to incorporate this type of complex function in a search algorithm. It is also difficult to characterize global landscape properties of protein sequence design using a complex function.

In this study, we showed how to significantly improve nonlinear function for characterizing fitness landscape of protein design. Using a rectangular kernel with proteins and decoys chosen *a priori*, we obtained a nonlinear kernel function via a finite Newton method. The total number of native proteins and decoy conformations included in the function was reduced to about 3,680. In the blind test of sequence design to discriminate 428 native sequences from 11 million challenging protein-like decoy sequences, this fitness function misclassified only 20 native sequences (correct rate 95%), which far outperformed (Miyazawa and Jernigan, 1996) (87 misclassification, correct rate 57%) and linear optimal functions (Tobi et al., 2000; Bastolla et al., 2001) (44–58 misclassification, correct rate 78%–71%) both of which were tested on a smaller scale to discriminate 201 native sequence from 3 million challenging protein-like decoy sequences. It is also comparable to the results of 18 misclassifications (correct rate 91%) using far more complex nonlinear fitness function with $>$ 5,000 terms (Hu et al., 2004).

This paper is organized as follows. We first describe our model and theory for sequence design. We then discuss computational details. Results of a blind tests are then presented. We conclude with discussion and remarks.

2.2 Model and Theory

We use a d-dimensional vector $\boldsymbol{c} \in \mathbb{R}^d$ to represent both the sequence and structure of a protein (Mintseris and Weng, 2003). One possible choice is the vector of the number count of non-bonded pairwise contacts of each of the $\binom{20+2-1}{2} = 210$ contact types (Miyazawa and Jernigan, 1985) between the 20 types of amino acid residues in a protein structure. Given a protein amino acid sequence $\boldsymbol{a}$ and its structure $\boldsymbol{s}$, the contact vector $\boldsymbol{c}$ is largely determined by the contact definition $f : (\boldsymbol{s}, \boldsymbol{a}) \mapsto \mathbb{R}^d$.

2.2.1 Inequality criterion.

In protein design, the native amino acid sequence $\boldsymbol{a}$ of a protein should have better fitness score on the native structure $\boldsymbol{s}$ of this protein than any other competing sequences taken from proteins of different fold. This leads to the requirement that the native sequence $\boldsymbol{a}_N$ mounted on its native structure $\boldsymbol{s}_N$ should have the best fitness score (lowest "energy") compared to a set of decoys $\mathscr{D} = \{D | \boldsymbol{c}_D = f(\boldsymbol{s}_N, \boldsymbol{a}_D) \text{ for all } \boldsymbol{a}_D\}$ derived from mounting unrelated alternative sequences $\boldsymbol{a}_D$ on the native protein structure $\boldsymbol{s}_N$:

$$H(\boldsymbol{c}_N) < H(\boldsymbol{c}_D) \quad \text{for all } D \in \mathscr{D}, \tag{2.1}$$

where $\boldsymbol{c}_D = f(\boldsymbol{s}_N, \boldsymbol{a}_D)$ is the contact vector of a decoy sequence $\boldsymbol{a}_D$ mounted on its native protein structure $\boldsymbol{s}_N$, and $\boldsymbol{c}_N = f(\boldsymbol{s}_N, \boldsymbol{a}_N)$ is the contact vector of a native sequence $\boldsymbol{a}_N$ from the set of native training proteins $\mathscr{N}$ mounted on the native structure $\boldsymbol{s}_N$. Here $\mathscr{D}$ is a set of sequence decoys mounted on native protein structures. $H(\boldsymbol{c}_N)$ and $H(\boldsymbol{c}_D)$ are the energy score for native sequence structure pair and for non-native sequence structure pair, respectively. Equivalently, the native sequence will have the highest probability to

fit into its native structure, and other sequences will have lower probability. This is the same principle described in (Shakhnovich and Gutin, 1993; Li et al., 1996; Deutsch and Kurosky, 1996).

A commonly used form for fitness function $H(\boldsymbol{c})$ is the weighted linear sum of pairwise contacts (Miyazawa and Jernigan, 1985; Tobi et al., 2000; Tanaka and Scheraga, 1976; Vendruscolo and Domany, 1998; Lu and Skolnick, 2001):

$$H(\boldsymbol{c}) = \boldsymbol{w} \cdot \boldsymbol{c}, \tag{2.2}$$

where "$\cdot$" represents inner product of two vectors. For this linear function, the basic requirement for protein fitness is then:

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) < 0. \tag{2.3}$$

We can further require that the difference in fitness must be greater than a constant $d > 0$:

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + d < 0. \tag{2.4}$$

2.2.2 Relation to support vector machines.

There may exist multiple $\boldsymbol{w}'$s if $\mathscr{P}$ is not empty. We can use the formulation of a support vector machine to find some weight vector $\boldsymbol{w}$. Let all vectors $\boldsymbol{c}_N \in \mathbb{R}^d$ form a native training set and all vectors $\boldsymbol{c}_D \in \mathbb{R}^d$ form a decoy training set. Each vector in the native training set is labeled as -1 and each vector in the decoy

training set is labeled as $+1$. Then solving the following support vector machine problem will provide an optimal solution to inequalities (Equation 2.3) :

$$\begin{aligned} \text{Minimize} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & \boldsymbol{w}\cdot\boldsymbol{c}_N + b \le -1 \\ & \boldsymbol{w}\cdot\boldsymbol{c}_D + b \ge 1. \end{aligned} \tag{2.5}$$

Note that a solution of the above problem satisfies the system of inequalities (Equation 2.3), since subtracting the second inequality from the first inequality in the constraint conditions of (Equation 2.5) will give us $\boldsymbol{w}\cdot(\boldsymbol{c}_N - \boldsymbol{c}_D) \le -2 < 0$.

2.2.3 Nonlinear fitness function.

A fundamental reason for this failure is that the functional form of linear sum of pairwise interaction is too simplistic. We can obtain a nonlinear fitness function for sequence design using an alternative functional form (Hu et al., 2004):

$$H(\boldsymbol{c}) = \sum_{D\in\mathscr{D}} \alpha_D K(\boldsymbol{c},\boldsymbol{c}_D) - \sum_{N\in\mathscr{N}} \alpha_N K(\boldsymbol{c},\boldsymbol{c}_N) + b, \tag{2.6}$$

where $\alpha_D \ge 0$ and $\alpha_N \ge 0$ are coefficients to be determined. This functional form is reminiscent of the linear fitness function $H(\boldsymbol{c}) = \boldsymbol{w}\cdot\boldsymbol{c}$, which can be written alternatively as an expansion around positive and negative contact vectors, as used in perceptron learning: $\boldsymbol{w} = -\sum_{N\in\mathscr{N}} \alpha_N \boldsymbol{c}_N + \sum_{D\in\mathscr{D}} \alpha_D \boldsymbol{c}_D$. A convenient kernel function K is:

$$K(\boldsymbol{c}_i,\boldsymbol{c}_j) = e^{-\gamma\|\boldsymbol{c}_i-\boldsymbol{c}_j\|^2} \text{ for any vectors } \boldsymbol{c}_i \text{ and } \boldsymbol{c}_j \in \mathscr{N}\cup\mathscr{D}, \tag{2.7}$$

where γ is a constant. The fitness function $H(\boldsymbol{c})$ can be written compactly as:

$$H(\boldsymbol{c}) = \sum_{D\in\mathscr{D}} \alpha_D e^{-\gamma||\boldsymbol{c}-\boldsymbol{c}_D||^2} - \sum_{N\in\mathscr{N}} \alpha_N e^{-\gamma||\boldsymbol{c}-\boldsymbol{c}_N||^2} + b = K(\boldsymbol{c},A)D_s\boldsymbol{\alpha} + b, \tag{2.8}$$

where A is the matrix of training data: $A = (\boldsymbol{c}_1^T, \cdots, \boldsymbol{c}_{|\mathscr{D}|}^T, \boldsymbol{c}_{|\mathscr{D}|+1}^T, \cdots, \boldsymbol{c}_{|\mathscr{D}|+|\mathscr{N}|}^T)^T$, and the entry $K(\boldsymbol{c},\boldsymbol{c}_j)$ of $K(\boldsymbol{c},A)$ is $e^{-\gamma||\boldsymbol{c}-\boldsymbol{c}_j||^2}$. D_s is the diagonal matrix with $+1$ and -1 along its diagonal representing the membership class of each point $A_i = \boldsymbol{c}_i^T$. Here $\boldsymbol{\alpha}$ is the coefficient vector: $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{|\mathscr{D}|}, \alpha_{|\mathscr{D}|+1}, \cdots, \alpha_{|\mathscr{D}|+\mathscr{N}|})^T$.

Intuitively, the fitness landscape has smooth Gaussian hills of height α_D centered on location $\boldsymbol{c}_D$ of decoy contact vector $D \in \mathscr{D}$. The depth of smooth Gaussian cones α_N is located on the center of native contact vectors $\boldsymbol{c}_N$ ($N \in \mathscr{N}$). For contact vector $\boldsymbol{c}_N$ of native proteins, the corresponding value of fitness function is set to -1, and for contact vector $\boldsymbol{c}_D$ of decoys, the corresponding value of fitness function is set to $+1$.

2.2.4 Optimal nonlinear fitness function.

To obtain such a nonlinear function, our goal is to find a set of parameters $\{\alpha_D, \alpha_N\}$ such that, for native proteins, the values of the fitness function of $H(\boldsymbol{c})$ close to -1, for decoys, the values of the fitness function of $H(\boldsymbol{c})$ close to $+1$.First, we note that we have implicitly mapped each protein and decoy from $\mathbb{R}^d, d = 210$ to another high dimensional space where the scalar product of a pair of mapped points can be efficiently calculated by the kernel function $K(.,.)$. Second, we look for a hyperplane such that the hyperplane has equal distance between the closest native proteins and the closest decoys. Such a hyperplane has good performance in discrimination (Vapnik, 1999). It can be found using support vector machine

by obtaining the parameters $\{\alpha_D\}$ and $\{\alpha_N\}$ from solving the following primal form of the quadratic programming problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^m_+,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^m} \quad & \frac{C}{2}\boldsymbol{e}\cdot\boldsymbol{\xi}+\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\alpha} \\ \text{subject to} \quad & D_s(K(A,A)D_s\boldsymbol{\alpha}+b\boldsymbol{e})+\boldsymbol{\xi} \;\geq\; \boldsymbol{e} \\ & \boldsymbol{\xi} \;\geq\; \mathbf{0}, \end{aligned} \tag{2.9}$$

where m is the total number of training points: $m = |\mathscr{D}| + |\mathscr{N}|$, C is a regularizing constant that limits the influence of each misclassified conformation (Vapnik, 1999; Burges, 1998; Schölkopf, 2002; Vapnik and Chervonenkis, 1974), and the $m \times m$ diagonal matrix of signs D_s with $+1$ or -1 along its diagonal indicating the membership of each point A_i in the classes $+1$ or -1; and $\boldsymbol{e}$ is an m-vector with 1 at each entry. The variable ξ_i is a measurement of error for each input vector with respect to the solution: $\xi_i = 1 + y_i H(\boldsymbol{c}_i)$, where $y_i = -1$ if i is a native protein, and $y_i = +1$ if i is a decoy.

2.2.5 Rectangle kernel and reduced support vector machine (RSVM).

The use of nonlinear kernels on large datasets typically demands a prohibiting size of the computer memory in solving the potentially enormous unconstrained optimization problem and the use of large data set not only need costly storage and expensive time to evaluate new contact vector $\boldsymbol{c}$. To attack these difficult problems, the reduced support vector machines (RSVM) (Lee and Mangasarian, 2001) use a very small random subset of the training set to build a rectangular kernel matrix, instead of using the conventional $m \times m$ kernel matrix $K(A,A)$ in (Equation 2.9).

This model can achieve about 10% improvement on test accuracy over conventional support vector machine with random data sets of sizes between $1-5\%$ of the original data (Lee and Mangasarian, 2001).

The small subset can be regarded as a basis set in our study. Suppose that the number of contact vectors in our basis set is $\bar{m}$, with $\bar{m} \ll m$. We denote $\bar{A}$ as an $\bar{m} \times d$ matrix, and each contact vector from the basis set is represented by a row vector of $\bar{A}$. The resulting kernel matrix $K(A,\bar{A})$ from A and $\bar{A}$ has size $m \times \bar{m}$. Each entry of this rectangular kernel matrix is calculated by $K(\boldsymbol{c}_i, \bar{\boldsymbol{c}}_j)$, where $\boldsymbol{c}_i^T$ and $\bar{\boldsymbol{c}}_j^T$ are rows from A and $\bar{A}$ respectively. The RSVM is formulated as the following quadratic program:

$$\begin{array}{lll} \min\limits_{\bar{\boldsymbol{\alpha}} \in \mathbb{R}_+^{\bar{m}},\ b \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}^{\bar{m}}} & \dfrac{C}{2}\boldsymbol{\xi} \cdot \boldsymbol{\xi} + \dfrac{1}{2}(\bar{\boldsymbol{\alpha}} \cdot \bar{\boldsymbol{\alpha}} + b^2) & \\ \text{subject to} & D_s(K(A,\bar{A})\bar{D}_s\bar{\boldsymbol{\alpha}} + b\boldsymbol{e}) + \boldsymbol{\xi} \geq \boldsymbol{e} & \\ & \boldsymbol{\xi} \geq \boldsymbol{0}, & \end{array} \tag{2.10}$$

where $\bar{D}_s$ is the $\bar{m} \times \bar{m}$ diagonal matrix with $+1$ or -1 along its diagonal, indicating the membership of each point $\bar{A}_i$ in the classes $+1$ or -1; and $\boldsymbol{e}$ is an m-vector with 1 at each entry. As shown in (Lee and Mangasarian, 2001), the zero level set surface of the fitness function is given by

$$H(\boldsymbol{c}) = K(\boldsymbol{c},\bar{A})\bar{D}_s\bar{\boldsymbol{\alpha}} + b = \sum_{\boldsymbol{c}_D \in \bar{A}} \bar{\alpha}_D e^{-\gamma\|\boldsymbol{c}-\boldsymbol{c}_D\|^2} - \sum_{\boldsymbol{c}_N \in \bar{A}} \bar{\alpha}_N e^{-\gamma\|\boldsymbol{c}-\boldsymbol{c}_N\|^2} + b = 0, \tag{2.11}$$

where $(\bar{\boldsymbol{\alpha}}, b) \in \mathbb{R}^{\bar{m}+1}$ is the unique solution to (Equation 2.10). This surface discriminates native proteins against from decoys. Besides the rectangular kernel matrix, the use of 2-norm for the error $\boldsymbol{\xi}$ and an extra term b^2 in the objective function of (Equation 2.10) distinguish this from the conventional support vector machine formulation.

2.2.6 Smooth Newton method.

In order to solve equation (Equation 2.10) efficiently, an equivalent unconstrained nonlinear program based on the implicit Lagrangian formulation of (Equation 2.10) was proposed in (Fung and Mangasarian, 2003), which can be solved using a fast Newton method. We modified the implicit Lagrangian formulation and obtain the unconstrained nonlinear program for the unbalanced RSVM in equation (Equation 2.10). The Lagrangian dual of (Equation 2.10) is now (Mangasarian, 1994):

$$\min_{\bar{\boldsymbol{\alpha}}\in\mathbb{R}^{\bar{m}}_{+}} \frac{1}{2}\bar{\boldsymbol{\alpha}}\cdot(Q+\bar{D}_s(K(A,\bar{A})^T K(A,\bar{A})+\boldsymbol{e}\boldsymbol{e}^T\bar{D}_s)\bar{\boldsymbol{\alpha}}-\boldsymbol{e}\cdot\bar{\boldsymbol{\alpha}}, \tag{2.12}$$

where $Q = I/C \in \mathbb{R}^{\bar{m}\times\bar{m}}$, and $I \in \mathbb{R}^{\bar{m}\times\bar{m}}$ is unit matrix. Note that $\mathbb{R}^{\bar{m}}_{+}$ is the set of nonnegative $\bar{m}$-vectors. Following (Fung and Mangasarian, 2003), an equivalent unconstrained piecewise quadratic minimization problem of the above positively constrained optimization can be derived as follows:

$$\begin{aligned} & \min_{\bar{\boldsymbol{\alpha}}\in\mathbb{R}^{\bar{m}}} L(\bar{\boldsymbol{\alpha}}) \\ = \quad & \min_{\bar{\boldsymbol{\alpha}}\in\mathbb{R}^{\bar{m}}} \frac{1}{2}\bar{\boldsymbol{\alpha}}\cdot Q\bar{\boldsymbol{\alpha}}-\boldsymbol{e}\cdot\bar{\boldsymbol{\alpha}}+\frac{1}{2}\beta(\|(-\beta\bar{\boldsymbol{\alpha}}+Q\bar{\boldsymbol{\alpha}}-\boldsymbol{e})_{+}\|^2-\|Q\bar{\boldsymbol{\alpha}}-\boldsymbol{e}\|^2). \end{aligned} \tag{2.13}$$

Here, β is a sufficiently large but finite positive parameter to ensure that the matrix $\beta I - Q$ is positive definite, where $I \in \mathbb{R}^{\bar{m}\times\bar{m}}$ is a unit matrix, and the plus function $(\cdot)_{+}$ replaces negative components of a vector by zeros. This unconstrained piecewise quadratic problem can be solved by a Newton method in a finite number of steps (Fung and Mangasarian, 2003). Newton method requires the information of the

gradient vector $\bigtriangledown L(\bar{\boldsymbol{\alpha}}) \in \mathbb{R}^{\bar{m}}$ and the generalized Hessian $\partial^2 L(\bar{\boldsymbol{\alpha}}) \in \mathbb{R}^{\bar{m}\times\bar{m}}$ of $L(\bar{\boldsymbol{\alpha}})$ at each iteration. They can be calculated using the following formulae (Fung and Mangasarian, 2003):

$$\begin{aligned} \bigtriangledown L(\bar{\boldsymbol{\alpha}}) &= (Q\bar{\boldsymbol{\alpha}} - \boldsymbol{e}) + \tfrac{1}{\beta}(Q-\beta I)((Q-\beta I) - \boldsymbol{e})_+ - \tfrac{1}{\beta}Q(Q\bar{\boldsymbol{\alpha}} - \boldsymbol{e}) \\ &= \tfrac{(\beta I - Q)}{\beta}((Q\bar{\boldsymbol{\alpha}} - \boldsymbol{e}) - ((Q-\beta I)\bar{\boldsymbol{\alpha}} - \boldsymbol{e})_+), \end{aligned} \tag{2.14}$$

and

$$\partial^2 L(\bar{\boldsymbol{\alpha}}) = \frac{\beta I - Q}{\beta}(Q + diag((Q-\beta I)\bar{\boldsymbol{\alpha}} - \boldsymbol{e})_*(\beta I - Q)), \tag{2.15}$$

where $diag(\cdot)$ denotes a diagonal matrix and $(\boldsymbol{\alpha})_*$ denotes the step function, *i*.e., $(\alpha_i)_* = 1$ if $\alpha_i > 0$; and $(\alpha_i)_* = 0$ if $\alpha_i \leq 0$.

The main step of Newton method is to solve iteratively the system of linear equations

$$-\bigtriangledown L(\bar{\boldsymbol{\alpha}}^i) + \partial^2 L(\bar{\boldsymbol{\alpha}}^i)(\bar{\boldsymbol{\alpha}}^{i+1} - \bar{\boldsymbol{\alpha}}^i) = \mathbf{0}, \tag{2.16}$$

for the unknown vector $\bar{\boldsymbol{\alpha}}^{i+1}$.

We present below the algorithm, whose convergence was proved in (Fung and Mangasarian, 2003). We denote $\partial^2 L(\bar{\boldsymbol{\alpha}}^i)^{-1}$ as the inverse of the Hessian $\partial^2 L(\bar{\boldsymbol{\alpha}}^i)$.

Start with any $\bar{\boldsymbol{\alpha}}^0 \in \mathbb{R}^{\bar{m}}$. For $i = 0, 1...$:

(i) Stop if $\bigtriangledown L(\bar{\boldsymbol{\alpha}}^i - \partial^2 L(\bar{\boldsymbol{\alpha}}^i)^{-1} \bigtriangledown L(\bar{\boldsymbol{\alpha}}^i)) = 0$.

(ii) $\bar{\boldsymbol{\alpha}}^{i+1} = \bar{\boldsymbol{\alpha}}^{i} - \lambda_i \partial^2 L(\bar{\boldsymbol{\alpha}}^{i})^{-1} \bigtriangledown L(\bar{\boldsymbol{\alpha}}^{i}) = \bar{\boldsymbol{\alpha}}^{i} + \lambda_i \boldsymbol{d}_i$, where $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \cdots\}$ is the Armijo stepsize (Nocedal and Wright, 1999) such that

$$L(\bar{\boldsymbol{\alpha}}^{i}) - L(\bar{\boldsymbol{\alpha}}^{i} + \lambda_i \boldsymbol{d}^{i}) \geq -\delta \lambda_i \bigtriangledown L(\bar{\boldsymbol{\alpha}}^{i}) \cdot \boldsymbol{d}^{i}, \tag{2.17}$$

for some $\delta \in (0, \frac{1}{2})$, and $\boldsymbol{d}^{i}$ is the Newton direction

$$\boldsymbol{d}^{i} = \bar{\boldsymbol{\alpha}}^{i+1} - \bar{\boldsymbol{\alpha}}^{i} = -\partial^2 L(\bar{\boldsymbol{\alpha}}^{i})^{-1} \bigtriangledown L(\bar{\boldsymbol{\alpha}}^{i}), \tag{2.18}$$

obtained by solving (Equation 2.16).

(iii) $i = i + 1$. Go to (**i**) .

2.3 Computational Experiments

2.3.1 Determination of count vector by alpha shape.

Since protein molecules are formed by thousands of atoms, their shapes are complex. In this study, we use the count vector $\boldsymbol{c}$ of pairwise contact interactions derived from the edge simplexes of the alpha shape of a protein structure, where only nearest neighbor atoms in physical contacts are identified. The advantages of this approach are elaborated in (Li et al., 2003). We refer to references (Edelsbrunner, 1993; Liang et al., 1998) for further theoretical and computational details.

2.3.2 Relationship between number of contacts and length of protein.

We found that there is a relationship between the number of total contacts of a protein and the length of the protein. A linear regression on the relationship between the number of total contacts and the length of the protein gives the following equation,

$$N_{contacts} = 3.090 \cdot L_{protein} - 76.182, \tag{2.19}$$

where $N_{contacts}$ is the number of contacts for a protein, and $L_{protein}$ is the number of the protein residues. To eliminate the influence of the length of protein, we normalize the number of contacts for each type of pair-wise contact of a protein using Equation 2.19.

2.3.3 Generating sequence decoys by threading.

We followed Maiorov and Crippen (Maiorov and Crippen, 1992) and used gapless threading to generate a large number of decoys for a simplified test of protein design. We threaded the sequence of a larger protein through the structure of a smaller protein, and obtained sequence decoys by mounting a fragment of the native sequence from the large protein to the full structure of the small protein. We therefore had a set of sequence decoys $(\boldsymbol{s}_N, \boldsymbol{a}_D)$ for each native protein $(\boldsymbol{s}_N, \boldsymbol{a}_N)$ (Fig 2). Because all native contacts were retained, such sequence decoys are quite challenging. This is unlike folding decoys generated by gapless threading (Hu et al., 2004).

2.3.4 Dataset.

We used a list of 1,515 protein chains compiled from the PISCES server (Wang and Dunbrack, 2003). Protein chains in this data set have pairwise sequence identity $< 20\%$, With its structural resolution by

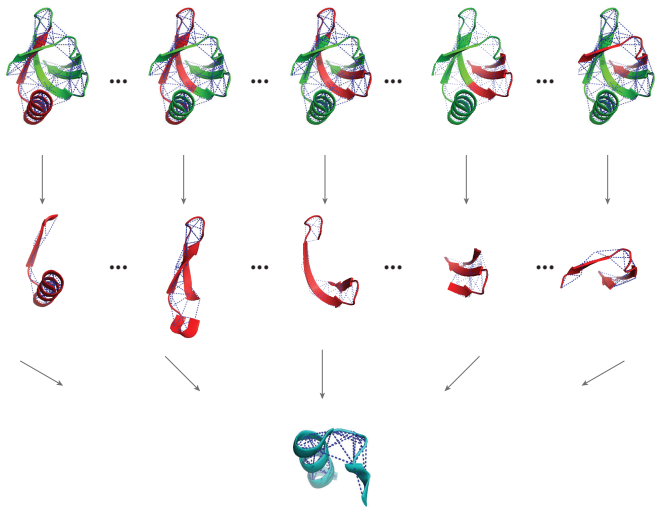

Figure 2: Decoy generation by gapless threading. Sequence decoys can be generated by threading the sequence of a larger protein to the structure of an unrelated smaller protein.

crystallography and has a resolution $\leq$ 1.6 Å, and the R-factor $\leq$ 0.25. We removed incomplete proteins (i.e. those with missing residues), and proteins with uncertain residues (those denoted as ASX, GLX, XLE, and XAA). We further removed proteins with less than 46 and more than 500 amino acids. In addition, we removed protein chains with more than 30% extensive inter-chain contacts. The remaining set of 1,228 proteins are then randomly divided into two sets. One set includes 800 proteins and the other one includes 428 proteins. Using the sequence threading method, we generated 36,823,837 non-protein decoys, together with 800 native proteins as the training set, and 11,144,381 decoy non-proteins with 428 native proteins as the test set.

2.3.5 Selection of matrix A for iterative training.

We used only a subset of the 36 million decoys and native structures so they will fit into the computer memory during training. These structures formed the data matrix A, which will be used to construct the kernel matrix $K(A,\bar{A})$. We used a heuristic iterative approach to construct the matrix A and $\bar{A}$ during each iteration.

Initially, we randomly selected 10 decoys from the set of decoys $\mathscr{D}_j$ for each of the j-th native protein. We have then $m \approx 8,000$ decoys for the 800 native proteins. We further chose only 1 decoy from the selected 10 decoys for each native protein j. These 800 decoys were combined with the 800 native proteins to form the initial matrix A. The contact vectors of a subset of 480 native proteins (60% of the original 800 proteins) and 320 decoys (40% of the 800 selected decoys) were then randomly chosen to form $\bar{A}$. An initial fitness function $H(\boldsymbol{c})$ was then obtained using A and $\bar{A}$. The fitness values of all 36 million decoys and the 800 native proteins were then evaluated using $H(\boldsymbol{c})$. We futher used two iterative strategies to improve upon the fitness function $H(\boldsymbol{c})$.

[**Strategy 1**] In the i-th iteration, we selected the subset of misclassified decoys from $\mathscr{D}_j$ associated with the j-th native protein and sorted them by their fitness value in descending order, so the missclassified decoys with least violation, namely, the negative but smallest absolute values in $H(\boldsymbol{c})$, would be on the top of the list. If there were less than 10 misclassified decoys, we added top decoys that were misclassified in the previous iteration for this native protein, if they exist, such that each native protein would have 10 decoys.

A new version of the matrix A was then constructed using these 8,000 decoys and the corresponding 800 native proteins. To obtain the updated $\bar{A}$, From these 8,800 contact vectors, we randomly selected 480 native proteins (60%) and 3,200 unpaired decoy non-proteins (40%) to form $\bar{A}$.

The iterative training process was then repeated until there was no improvement in the classification of the 36 million decoys and the 800 native proteins from the training set. Typically, the number of iterations was about 10. In subsequent studies, we experimented with different percentage of selected decoys, ranging from 10% to 100% to examine the effect of the size of $\bar{A}$ on the effectiveness of the fitness function $H(\boldsymbol{c})$.

[**Strategy 2**] In the *i*-th iteration, we selected the top 10 correctly classified decoys sorted by their fitness value in ascending order for each native protein, namely, those correctly classified decoy with positive but smallest absolute values are selected. These contact vectors of 8,000 selected decoys were combined with the 800 native proteins to form the new data matrix A.

To construct $\bar{A}$, we first selected the most challenging native proteins by taking the top 80 correctly classified native proteins (10%) sorted by their fitness value in descending order, namely, those that were negative but with the smallest absolute values in $H(\boldsymbol{c})$. We then randomly took 400 native proteins (50%) from the rest of the native protein set, so altogether we had 480 native proteins (60%). Similarly, we selected the top one decoy that was most challenging from the 10 chosen decoys in A for each native protein, namely, the top decoy that is correctly classified with positive but smallest value of $H(\boldsymbol{c})$. We then randomly selected 3 decoys for each native protein from the remaining decoys in A to obtain 3,200 decoy non-proteins (40%). The matrix $\bar{A}$ was then constructed from the selected 480 native proteins and 3,200 decoy non-proteins. The iterative training process was repeated until there was no improvement in classification of the 36 million decoys and 800 native proteins in the training set. Typically, the number of iterations was about 5.

In subsequent studies, we experimented with different choice of challenging native proteins. The selection ranges from the top 10% to 60% most challenging native proteins. The choice of the challenging

decoys was also varied, where we experimented with choosing the top one to the top four most challenging decoys for each native protein, while the number randomly selected decoys varied from three to zero.

2.3.6 Learning parameters.

There were two important parameters: the constant γ in the kernel function $e^{-\gamma||\boldsymbol{c}_i-\boldsymbol{c}||^2}$, and the cost factors C, which was used during training so errors on positive examples were adjusted to outweigh errors on negative examples. Our experimentation showed that $\gamma = 5.0\times10^{-5}$ and $C = 1.0\times10^{4}$ were reasonable choices.

2.3.7 Timing information.

The algorithm was implemented in the C language. It called LAPACK(Anderson et al., 1999) and used LU decomposition to solve the system of linear equations. It also called an SVD routine to determine the 2-norm of a matrix for calculating $\beta = 1.1(1/C + ||DA-\boldsymbol{e}||_2^2)$. Once matrices A and $\bar{A}$ were specified, the fitness function $H(\boldsymbol{c})$ could be derived in about 2 hours and 10 minutes on a 2 Dual Core AMD Opteron(tm) Processors of 1,800 MHz with 4 Gb memory for an A of size $8,800\times210$ and an $\bar{A}$ of size $3,680\times210$. The evaluation of the fitness of 14 million decoys took 2 hours and 10 minutes using 144 CPUs of a Linux cluster (2 Dual Core AMD Opteron(tm) Processors of 1.8 GHz with 2 Gb memory for each node). Because of the large size of the data set, the bottleneck in computation is disk IO.

2.3.8 F_β score.

We use the F_β score to measure the performance of fitness function in classification. The F_β score is defined as

$$F_\beta = (1+\beta^2)\frac{Precision\times Recall}{\beta^2\times Precision + Recall} \tag{2.20}$$

where $Precision = TP/(TP+FP)$ and $Recall = TP/(TP+FN)$. TP is true positive, FP is false positive, and FN is false negative. Since the imbalanced data set between the native proteins and the decoys, we put more weight on small size of native proteins, therefore we set $\beta = 10$. The F_β score is calculated for two different strategies on both training set and test set.

Comparing these two strategies by F_β score, we found the strategy 2 (table TABLE III) has better performance than the strategy 1 (table TABLE II). In strategy 2, We may select the fitness function of preselecting top 1 decoys and top 50% native proteins as best candidate function to classify native proteins and decoys.

2.4 Results

2.4.1 Performance in discrimination.

We used the set of 428 natives proteins and 11,144,381 decoys for testing the designed fitness function. We took the sequence $\boldsymbol{a}$ as the predicted sequence such that $\boldsymbol{c} = f(\boldsymbol{s}_N, \boldsymbol{a})$ has the best fitness value.

Sequence decoys obtained by gapless threading were quite challenging, since all native contacts of the protein structures were maintained, and decoy sequences were from real proteins. In a previous study, we showed that no linear fitness function can be found that would succeed in the challenging task of identifying all 440 native sequences in the training set (Hu et al., 2004). Because we are unaware of any other development of design fitness functions amenable for high-throughput tests, and frequently no distinctions were made between protein folding potential and protein design fitness function, we compared our fitness function with several well-established scoring functions developed for protein folding.

Here we succeeded in obtaining a simplified nonlinear fitness function for protein design that are capable of discriminating 796 of the 800 native sequences (Table TABLE I). It also succeeded in correctly identifying

Method	Number of misclassifications			
	Training set		Test set	
	800/36M	440/14M	428/11M	201/3M
Nonlinear function	4/988	NA	20/218	NA
Tobi *et. al.*	NA	192/39,583	NA	44/53,137
Bastolla *et. al.*	NA	134/47,750	NA	58/29,309
Miyazawa & Jernigan	NA	173/229,549	NA	87/80,716

TABLE I: The number of misclassifications using simplified nonlinear fitness function, optimal linear scoring functions taken as reported in (Tobi et al., 2000; Bastolla et al., 2001), and Miyazawa-Jernigan statistical potential (Miyazawa and Jernigan, 1996) for both native proteins and decoys (separated by "/") in the test set and the training set. The simplified nonlinear function is formed using a basis set of 3,680 (480 native + 3,200 decoy) contact vectors derived using strategy 2.

95% (408 out of 428) of the native sequences in the independent test set. Results for other methods were taken from literature obtained using much smaller and less challenging data set. Overall, the performance of our method was better than results obtained using optimal linear scoring function taken as reported in (Tobi et al., 2000) and in (Bastolla et al., 2001), which succeeded in identifying 78% (157 out of 201) and 71% (143 out of 201) of the test set, respectively. Our results are also better than the Miyazawa-Jernigan statistical potential (Miyazawa and Jernigan, 1996) (success rate 58%, 113 out of 201). This performance is also comparable with a more complex nonlinear fitness function, with $> 5,000$ terms reported in (Hu et al., 2004), which succeeded with a correct rate of 91% (183 out of 201).

2.4.2 Effect of the size of the basis set $\bar{A}$ using strategy 1.

The matrix $\bar{A}$ contains both proteins and decoys from A and its size is important in discrimination of native proteins from decoys. In our fitness function, Gaussian kernels centered around these selected contact vectors were used as basis set to interpolate the global landscape of protein design.

We examined the effects of different sizes of $\bar{A}$ using strategy 1. For a data matrix A consisting of 800 native proteins and 8,000 sequence decoys derived following the procedure described earlier, we tested different choice of $\bar{A}$ on the performance of discrimination. With the data matrix A, we fixed the selection of the 480 native proteins (60%), and experimented with random selection of different number of decoys, ranging from 800 (10%) to 8,000 (100%) to form different $\bar{A}$s.

The results of classifying both the training set of 800 native proteins with 36 million decoys and the test set of 428 native proteins with 11 million decoys are shown in Table TABLE II. When 60% (480) native proteins and 100% (8,000) decoys are included, there were only 5 native proteins misclassified in the test set and 24 native proteins in the training set.

2.4.3 Effect of the size of the pre-selection of dataset using strategy 2

We now then explored the effects of different choices in constructing matrix $\bar{A}$ using strategy 2. We varied our selection of the most challenging native proteins from the top 10% to 60%, and varied selection of the most challenging decoys from the top one to the top four decoys for each native protein, as described earlier. Results are shown in Table TABLE III. We found that the performance of the discrimination of both the training set and test set had change when either native proteins selection rate was changed from 10% to 60%, or decoys selection rate is changed from the top 1 to the top 4. Overall, these results suggest that for the blind test developed here, a fitness function with good discrimination can be achieved with about 480 native proteins and 3,200 decoys, along with 400 pre-selected native proteins and 800 pre-selected top-1 decoys. Our final fitness function used in Table TABLE I was constructed using a basis set of 3,680 contact vectors. We also observed that the average number of iterations was about 5 using strategy 2, which is much faster than strategy 1.

2.5 Discussion

In this study, we have developed a simplified nonlinear kernel function for fitness landscape of protein design using a rectangular kernel and a fast Newton method. The results in a blind test are encouraging. They suggest that for a simplified task of simultaneously designing 428 proteins from a set of 11 million decoys, the search space of protein shape and sequence can be effectively parametrized with just about 3,680 basis set of contact vectors. It is likely that the choice of matrix A is important. We showed that once A is carefully chosen, the overall design landscape is not overly sensitive to the specific choice of the basis set contact vectors for $\bar{A}$.

The native protein list in both training and test set come from the PISCES server, which has the lowest pair-wise identity (20%), finer resolution cutoff (1.6 Å) , and lower R-factor cutoff (0.25). This native dataset is better than previous studies conducted on a (Hu et al., 2004) dataset derived from the WHATIF database, which has looser constraints: pair-wise sequence identity $< 30\%$, resolution cutoff < 2.1 Å, and R-factor cutoff < 2.1. We compared our results with classic studies of Tobi *et. al.* (Tobi et al., 2000), Bastolla *et. al.* (Bastolla et al., 2001) and Miyazawa and Jernigan (Miyazawa and Jernigan, 1996). Although the training set and test set are different, we observed that our simplified nonlinear function can detect 95% (208) native proteins from 11 million decoys and only misclassified 218 decoys as native proteins, which outperformed Tobi *et. al.* (Tobi et al., 2000) (78% correct rate for native proteins, 53,137 misclassification for decoys), Bastolla *et al.* (Bolon and Mayo, 2001) (71% correct rate for native proteins, 29,309 misclassification for decoys), and Miyazawa and Jernigan (Miyazawa and Jernigan, 1996) methods (57% correct rate for native proteins, 80,716 misclassification for decoys) on much smaller blind test set of 201 native proteins and 3 million decoys.

As protein length is linearly correlated with the total number of contacts, we found that length corrections were important for improving the fitness function.

We developed two strategies to search for improving fitness landscapes. Strategy 1 mostly uses misclassified decoys in the next iteration of construction of matrix A. On average, 10 iteration is necessary to arrive at a good fitness function, which had excellent performance of only 5 misclassifications for the training data set. The misclassification rate in the test set is comparable to other fitness functions (Tobi et al., 2000; Bastolla et al., 2001; Miyazawa and Jernigan, 1996). Strategy 2 selected the most challenging decoys by the fitness value landscape in the matrix A for the next iteration. We pre-selected certain percentage of the

number of native proteins and certain number of decoys before generating the basis set matrix $\bar{A}$. Overall, strategy 2 performs better than strategy 1, not only in reducing both native proteins and decoys misclassifications in the blind test set, but also in speeding up the search process in deriving the final fitness function with the number of iterations reduced from 10 to 5. With Strategy 2, the updated fitness landscape is only adjusted by challenging decoys, it can identify the most challenging decoys and native proteins, leading to improvement in the fitness landscape in the next iteration.

Our final fitness landscape can correctly classify most of the native proteins, except 4 proteins (1ft5 chain A, 1gk9 chain A, 2p0s chain A, 2qud chain A) in the training set and 20 proteins in the test set (TABLE IV). Among misclassified proteins, 4 of which have $> 20\%$ contacts due to inter chain interactions. In addition, 14 misclassified proteins have metal ions and organic compounds bound that provide additional conformation stability. The covalent bonds between these organic compounds, metal ions and the rest of the protein are not reflected in the protein description. It is likely that substantial or strong contacts with other protein chains, DNA, or co-factors alter the conformation of the protein. The conformations of these proteins may be different upon removal of these contacts. Altogether, 21 of the 24 misclassified proteins have explanations, and the fitness function truly failed only for 3 proteins.

The representation of protein structures will likely have important effects on the success of protein design. The approach of reduced nonlinear function is general and applicable when alternative representations of protein structures are used, *e.g.*, adding solvation terms, including higher-order interactions.

2.6 Conclusion

We showed that a simplified nonlinear fitness function for protein design can be can be obtained using a simplified nonlinear kernel function via a finite Newton method. We used a rectangular kernel with a basis set of native proteins and decoys chosen *a priori*.

We succeeded in predicting 408 out of the 428 (95%) native proteins and misclassified only 218 out of 11 million decoys in a large blind test set. It should be noted that the test sets used were different, as other method were based on relatively small (201 native proteins and 3 million decoys) blind test sets. Our result outperformed statical linear scoring function (87 out of the 201 misclassifications, 57% correct rate) and optimized linear function (between 44 and 58 misclassifications out of the 201, 78% and 71% correct rate). The performance was also comparable with results obtained from a far more complex nonlinear fitness function with $> 5,000$ terms (18 misclassifications, 91% correct rate). Our results further suggest that for the task of global sequence design of 428 selected proteins, the search space of protein shape and sequence can be effectively parametrized with just about 3,680 carefully chosen basis set of native proteins and non-native protein decoys.

The rectangle kernel matrix with a finite Newton method works well in constructing fitness landscapes. In addition, we showed that the overall landscape is not overly sensitive to the specific choice of dataset.

Overall, our strategy of reduced kernel can be generalized to constructing other types of fitness functions.

Select native proteins rate	Select decoys rate	Iteration	Number of misclassifications			
			Training set		Test set	
			800/36M	F_β	428/11M	F_β
60%	0%	4	21/1,374	0.958	26/387	0.931
60%	10%	5	14/922	0.972	24/216	0.940
60%	20%	6	16/902	0.969	28/250	0.930
60%	30%	6	10/1,037	0.975	29/304	0.926
60%	40%	10	16/812	0.970	27/199	0.933
60%	50%	10	13/1,112	0.971	25/269	0.936
60%	60%	12	15/802	0.972	27/237	0.932
60%	70%	9	13/947	0.973	24/256	0.939
60%	80%	8	11/1,078	0.973	28/278	0.929
60%	90%	9	12/690	0.977	27/170	0.934
60%	100%	5	5/2,681	0.962	24/609	0.931

TABLE II: Effects of the size of basis set $\bar{A}$ on performance of discrimination using strategy 1. The number of misclassifications of both native proteins and decoys (separated by "/") in both training set and test set are listed.

Pre-select native proteins top	Pre-select decoys top	Iteration	Number of misclassification			
			Training set		Test set	
			800/36M	F_β	428/11M	F_β
0%	1	6	8/1,010	0.978	25/212	0.938
10%	1	5	5/997	0.982	24/242	0.939
20%	1	6	9/625	0.981	26/174	0.936
30%	1	6	9/689	0.980	24/211	0.940
40%	1	6	8/869	0.980	25/218	0.937
50%	1	5	4/988	0.983	20/218	0.949
60%	1	5	6/1,039	0.980	24/280	0.938
10%	1	5	5/997	0.982	24/242	0.939
10%	2	5	6/1,270	0.977	22/372	0.941
10%	3	7	9/934	0.978	22/247	0.944
10%	4	5	5/1,071	0.981	24/210	0.944

TABLE III: Test results using different size both for the pre-selected native proteins, which changes from 10% to 60% while fixing the pre-selected decoys top 1, and the pre-selected decoys changes from the top 1 to the top 4 while fixing pre-selected native proteins 10% using strategy 2. Misclassifications in two tests using different numbers of native proteins and decoys are listed (see text for details).

Molecular name		Classification	Ligand(s)	PDBID	Chain	Fitness value
Catalase	○	Oxidoreductase	1 HEM and 3 SO_4	1gwe	A	0.1085
Streptavidin	○	Biotin Binding Protein	1 BTN and 2 GOL	2f01	A	0.1407
Acutohaemonlysin	○	Toxin	2 IPA	1mc2	A	0.1728
Endonuclease I	○	Hydrolase	1 Mg and 2 Cl	2pu3	A	0.1900
cytochrome c, putative	○	Electron Transport	2 SO_4, 1 Na and 2 HEM	2czs	A	0.2664
Cytochrome F	○	Electron transport protein	1 HEME C	1e2w	A	0.6023
Bowman-Birk type trypsin inhibitor		Hydrolase Inhibitor	None	2fj8	A	0.8463
Uncharacterized protein with erredoxin-like fold	○	Structural Genomics,Unknown Function	1 Unkown ligand	3e8o	A	1.1592
General secretion pathway protein G	○	Protein Transport	1 Zn	1t92	A	1.3175
ARF GTPase-activating protein git1	△	Signaling Protein	None	2w6a	A	1.6581
Cystatin B	△	protein binding	None	2oct	A	1.8043
SNAP-25A		Transport protein	None	1n7s	D	1.9074
Lin2189 protein	○	Structural Genomics, Unknown Function	2 GOL	3b49	A	2.0142
Fibritin		Chaperone	None	2ibl	A	2.1211
Oxalate oxidase 1	○	Oxidoreductase	1 Mn, 1 GLV	2et1	A	2.9975
Alpha-2-macroglobulin receptor-associated protein	○	Lipid Transport/endocytosis/chaperone	2 Ca, 1 Na and 3 MPD	2fcw	B	3.5660
Recombination endonuclease VII	○	plasma protein	1 Zn and 7 SO_4	1e7l	A	3.7397
Hypothetical protein YDCE	○ △	Isomerase	1 BEZ	1gyx	A	4.2697
Syntaxin 1a	△	Transport protein	None	1n7s	B	5.0204
Bacteriophage t4 short tail fibre	○	Structural protein	1 CIT ,2 SO_4 and 1 Zn	1ocy	A	8.0264

TABLE IV: 20 native proteins in the test set are misclassified using strategy 2, The number of ligands bound to the protein are listed. The molecules are sorted by the fitness value. 14 of them (marked by ○) have ligand(s) bound to the protein. 4 of them (marked by △) have $> 20\%$ contacts due to inter chain interactions. The covalent bonds between these organic compounds, metal ions and the rest of the protein and inter chain interaction provide additional stability beyond intra-residue interactions of the descriptors.

CHAPTER 3

CHROMATIN STRUCTURE

3.1 Introduction

Interactions among different parts of chromosomes are a fundamental component of any physical model of gene and genome regulation. It is well-known that genes are widely dispersed linearly along chromatin. Gene regulation is not only controlled by linear proximity, but also by long-distance interactions (Gibcus and Dekker, 2013). Recent studies suggested that functional elements that are far away from each other on a linear scale cooperate to regulate gene expression by engaging in long-range chromatin looping interactions (van Heyningen and Hill, 2008). However, how these distal functional elements are assembled inside the nucleus is still unknown. Understanding the spatial organization of chromatin is a key to gain understanding of the mechanism of gene activities, nuclear functions, and maintenance of epigenetic of cells (Fraser and Bickmore, 2007). Developing a generic approach to determine the spatial organization of chromatin is essential for identification of long-range relationships between genes and their distant regulatory elements.

Recently, the development of chromosome conformation capture (3C) technologies (Dekker et al., 2002; Hagège et al., 2007; Abou El Hassan and Bremner, 2009) give us capability to study the three-dimensional structure of chromosome. and its high-throughput modifications (Simonis et al., 2006; Zhao et al., 2006; Würtele and Chartrand, 2006; Lomvardas et al., 2006; Ling et al., 2006; Bantignies et al., 2011; Dostie et al., 2006; Umbarger et al., 2011; Lieberman-Aiden et al., 2009; Duan et al., 2010; Tanizawa et al., 2010; Kalhor et al., 2012; Horike et al., 2005; Fullwood et al., 2009; Tiwari and Baylin, 2009;

Schoenfelder et al., 2010). Dekker et al. use 3C methods (Dekker et al., 2002): formaldehyde cross-linking, fragment DNA, ligation proximity loci, and with next generation sequencing technologies equipped us to map interactions (Duan et al., 2010; Lieberman-Aiden et al., 2009). 3C-based assays are much powerful in that it can discover chromatin loops on chromosome conformation.

Output of 3C-based methods could be used to estimate the overall 3D folding of chromatin. This idea is based on the hypothesis that the interaction frequency of a pair of loci, is inversely related to the average spatial distance between them.

Recently, several new approaches have been developed that remove biases in the 3C-assay contact maps (Baù et al., 2010; Yaffe and Tanay, 2011; Hu et al., 2012) by using a more deterministic approach for 3D modeling of genomes and genomic domains (Duan et al., 2010; Fraser et al., 2009; Jhunjhunwala et al., 2008; Baù et al., 2010). All these approaches have in common that, to the largest possible satisfy the experimental interaction data, they developed diverse experiments (3C, 4C, 5C, HiC) (Duan et al., 2010; Fraser et al., 2009; Jhunjhunwala et al., 2008) and computation to build 3D chromatin structure (Baù et al., 2010). One important caveat of all these methods is that they ignore the importance of non-specific physical interactions. In eukaryotic cells, chromatin is contained within the nuclues and most of polymer theory does not consider this confinement effect (Lieberman-Aiden et al., 2009; Bohn et al., 2007; Barbieri et al., 2012). Consequently the non-specific interactions arising from confinement effect are omitted. Another important caveat of all models is inefficient sampling and expensive calculations based on simplifying assumptions such as fractal structure (Lieberman-Aiden et al., 2009), random loop connections (Heermann et al., 2012), or extensive simulation based on the 3C-assay data by assuming harmonic force

between loci (Baù et al., 2010; Duan et al., 2010). All these approaches are time consuming, and do not satisfy all the experimental constraints that would reflect real chromatin structures.

To overcome such biases and limitations, we developed a two steps approach that removes the non-significant interactions from 5C experiments and then construct a physical model that satisfy the constraints from significant interactions. In the 5C experiments (Baù et al., 2010) on a 500 kilo base (kb) α-globin gene domain located near the left telomere of human chromosome 16 , Baù et al. used HindIII, a type II site-specific deoxyribonuclease restriction enzyme which cleaves palindromic DNA sequence AAGCTT, to cleave chromatin segment. They designed 30 forward and 25 reverse primers for paired-end sequencing at the end of the HindIII sites to detect long range interactions between two targeted sets of genomic loci on both GM12878 and K562 cells. An interaction can only happen between a reverse and forward primer. We applied this approach to determine the higher order spatial organization and we developed differential activation mechanism of a 500 kilo base (kb) α-globin gene domain based on the frequency matrix data captured by the 5C experiments.

3.2 Physical model

We modeled chromatin fiber chain as a self-avoiding polymer consisting of beads that represent fiber and its persistence length property.

Bystricky et al. did measurement on the chromatin persistence length, found that chromatin persistence length $= 170 - 220$ nm, mass density $\approx 110 - 150$ bp/nm (Bystricky et al., 2004).

There are also several studies suggesting that the persistence length of a chromatin is much lower than 170 nm (Langowski and Heermann, 2007; Dekker, 2008). In our model, we used 30 nm chromatin fiber diameter and 150 nm persistence length as our model parameters.

A canonical 30 nm chromatin fiber has a mass density of 11 nm/kb (Dekker, 2008). Each 30 nm bead contains 2727 bp DNA. As a result, we represented the chromatin fiber as a self-avoiding walk polymer chain that consists of beads and each bead has a diameter of 30 nm.

3.2.1 Obtaining significant interactions

We pursued a multistage data cleaning procedure. Our first step was to remove the biases from the 5C data, and second step was removing random interactions from the data using a reference state.

3.2.1.1 Removing bias: short segments

The length of fragment is too long or too short will cause biase (Naumova et al., 2012). In this study, we eliminate HindIII segments which are less than 2727 bp long, which is equivalent to a 30 nm diameter sphere. Therefore, we removed 11 forward HindIII segments and 2 reverse HindIII segments from Baù et al. 5C data (Baù et al., 2010). 19 forward HindIII segments and 23 reverse HindIII segments were kept.

3.2.1.2 Removing bias: proximity effects

To overcome proximity effects, we ignoreed the interactions between neighboring segments. For instance, forward primer segment 1 has interaction frequency 5823 with reverse primer segment 2 in GM12878 cell, and 13686 in K562 cell. These interaction frequencies were discarded in both GM12878 and K562 cells to avoid bias. 5 proximity events were removed (F1-R2, F10-R10, F11-R11, F24-R23, F28-R24, F30-R25).

3.2.1.3 Removing bias: random model

3.2.1.3.1 From 5C data to polymer chain

For 500 kb α-globin gene domain chromatin fiber, we divided the chromatin fiber into different length segments separated by 42 HindIII primer sites. Each primer site was represented by a bead and the distances

between beads were approximately set to the persistence length. In some case, HindIII segments were smaller than persistence length. In that case, the distance between beads were the size of that HindIII segments. There were also cases where a HindIII segment was larger than the persistence length. In that case, we created virtual primer sites such that divide the HindIII segment into small pieces by persistence length. In the circos diagram, the inner thinest blue ring with bars represents the final division, the long black bar denotes real primer sites and the short grey bar stands for virtual sites (Figure. 7 B). Therefore there are a total of 54 sites along the chromatin fiber, meaning that the polymer chain is represented by 54 beads. As a result, we grew a self-avoiding polymer within a sphere equivalent to the cell nucleus to create our reference state.

3.2.1.3.2 Nucleus size

To estimate the diameter of the sphere that α-globin gene domain chromatin fiber would occupy, we proportioned it to that of the whole human genome chromosome. The lymphoblast cell size varies from 10 to 20 μm (Rozenberg, 2002). A nucleus occupies about 10% of the total cell volume in a eucaryotic cell (Alberts et al., 2007). the nucleus size from $\frac{4}{3}\pi(\frac{D}{2})^3 \times 10\% = \frac{4}{3}\pi(\frac{d}{2})^3$, where D is the diameter of cell which ranged from 10 to 20 μm, therefore nucleus diameter of 4.64 to 9.28 μm. Also, the human entire genome has 6 billion pair of bases in diploid cells, (Chromosome size source: National Center for Biotechnology Information. Human Reference Sequence from Build 33 of the Human Genome released April 14, 2003). Therefore for 500 kb α-globin domain, we have $\frac{\frac{4}{3}\pi(\frac{d}{2})^3}{6{,}000{,}000{,}000} = \frac{\frac{4}{3}\pi(\frac{d_\alpha}{2})^3}{500{,}000}$, where d_α is the diameter of the sphere that 500 kb α-globin gene domain chromatin fiber diameter would occupy. Finally d_α ranged from 203 to 405 nm, here we adopt $d_\alpha = 330$ nm.

3.2.1.3.3 Refrence state

We used a three-dimensional self-avoiding walk polymer model to represent a chromatin conformation. A length-n chromatin conformation was represented by a connected chain $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$, where the ith node of the conformation is located at the site $x_i = (x_{i1}, x_{i2}, x_{i3})$ in the three-dimensional space and each chain was connected by varies length of segment.

For α-globin gene domain chromatin conformations, the nodes position should satisfy some constraints. In our C-SAC model, these nodes are limited inside the 330 nm sphere with self-avoiding, distance, and loop constraints. At the beginning, the starting node randomly grows inside the 330 nm sphere. Second, Euclidean distance between neighbouring nodes x_i and x_{i+1} must be same as the physical model derived. Third, the direction of the vector $x_{i+1} - x_i$ should be uniformly distributed in the three-dimension space. In this study, we used the Yershova (Yershova et al., 2010) method to generate 640 uniform deterministic sample nodes over the rotation group SO(3) in three-dimension space. To mimic the rod segment, we interpolated a certain number beads of diameter 30 nm between x_i and x_{i+1}, for example, if the length of x_{i+3} and x_{i+4} is 150 nm, then there would be four beads of diameter 30 nm between these 2 nodes. Fourth, we enforce the self-avoiding constraint. Any candidate node was not permitted to be closer than 30 nm with the partial chain represented by 30 nm beads (Figure 3). Fifth, we restricted that none of the nodes can go outside of the 330 nm sphere which is α-globin gene domain volume size.

3.2.1.3.4 Sequential Importance Sampling for reference state

In reality, exhaustive enumeration of all three-dimensional self-avoiding walk polymers to discover geometrically complexity and interesting features is computationally impossible, especially for long chain polymers, because the number of possible self-avoiding walk polymers increases exponentially with chain length.

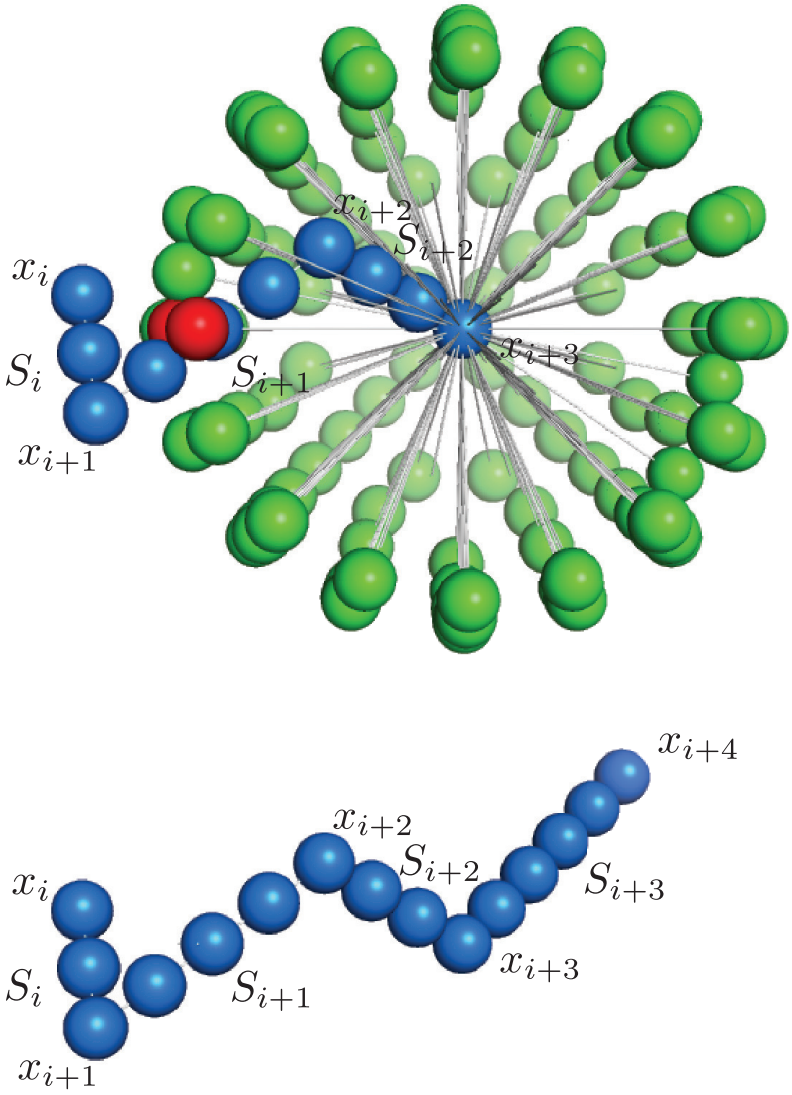

Figure 3

Figure 3 *(previous page)*: Illustration of the bead rod self-avoiding walking model. We have one partial chain including consecutive connected segments S_i, S_{i+1}, S_{i+2}, and corresponding nodes $x_i, x_{i+1}, x_{i+2}, x_{i+3}$. The segment S_i, S_{i+1}, S_{i+2} length is 70, 140 and 90 individually. Each segment is evenly interpolated with some certain 30 nm balls, S_i is inserted one 30 nm ball between x_i and x_{i+1}, S_{i+1} is inserted three 30 nm balls between x_{i+1} and x_{i+2}, while S_{i+2} is inserted two 30 nm balls between x_{i+2} and x_{i+3}. The next segment S_{i+3} connected by node x_{i+3} is 150 nm. Given the location of x_{i+3}, for example, we have uniformly distributed 160 candidate nodes on a sphere of radius 150 nm satisfying the segment distance 150 nm restriction. We randomly pick one candidate node which has no conflicts with previous nodes, including the interpolated balls. Then we continue interpolation of four 30 nm balls on segment S_{i+3}. For clarity, we only show a diagram of partial segments from S_i to S_{i+3}, ignoring the previous segments between S_1 and S_i. The blue balls represent one partial chain, each ball diameter is 30 nm. The first three balls include x_i and x_{i+1} nodes represent segment S_i, S_{i+1} and S_{i+2} are represented by five and four 30 nm balls respectively. The green and red balls represent candidate nodes with a distance 150 nm with node x_{i+3}.

(previous page continue): For clarity in this cartoon, we do not show all 160 candidate nodes, we clip one plane of the sphere of radius 150 nm, take a inside view of sphere to show the relationship between the candidate nodes and previous segments represented by 30 nm balls. The grey lines indicate the segments between candidate nodes and node x_{i+3} which have length 150 nm. There are two red 30 nm balls have collided with the middle interpolated ball of the segment S_{i+1}. In our bead rod self-avoiding walking model, we do not consider these red balls conflicting with previous partial chain, therefore we random pickup one node x_{i+4} among the 158 candidate nodes as the segment S_{i+3} other end, then interpolate 30 nm balls on the segment S_{i+3}. We iteratively grow the next segment from the newly partial chain end by above method. The chain including all nodes and segments is inside one pre-defined sphere as well. In our calculation for the α-globin gene domain, we uniformly generate 640 candidate nodes on certain sphere depending on the growing segment length to randomly pick one to grow the chain. These random growing chain are always located inside one 330 nm sphere.

Monte Carlo methods are often used to generate samples from all possible conformations to gain estimation of feature statistics from these generated samples. When chain length becomes large, however, the direct generation of self-avoiding walk polymers with rejecting method from the uniform distribution of all possible self-avoiding walk polymers becomes difficult. In two dimensional lattice space, for instance, the success rate σ_N of generating self-avoiding walk polymers decreases exponentially along the number of monomer N, $\sigma_N \approx Z_N/(4 \times 3^{N-1})$. Z_N is the normalized constant, which is the total number of different self-avoiding walk polymers with N monomers. For $N = 20$, this rate is approximately $\sigma \approx 21.6\%$, while for $N = 48$, this number decreases to 0.79% (Liu, 2008). The success rate σ_N will become very low in the three-dimensional bead rod self-avoiding walk polymer model. To conquer this attrition problem, Rosenbluth and Rosenbluth introduced biased sampling method to correct the bias samples by weighting them, and satisfy the uniform distribution of the polymers when growing one chain (Rosenbluth and Rosenbluth, 1955). The essential idea of the method is to iteratively grow one more monomer for a t-polymer chain after $t-1$ successive steps with self-avoiding walk, until reaching the desired length $t = n$. The position of the tth monomer is influenced by the current $t-1$ polymer chain conformation. If there are n_t candidate positions (i.e. unoccupied placements) for the tth monomer, we randomly selected any one of the n_t sites. Nevertheless, there is a limitation of the resulting samples that is biased to more compact conformations and does not strictly conform to the uniform distribution. We assigned a weight to adjust for this bias. The weight was recursively set as $w_t = n_t w_{t-1}$ in the Rosenbluth chain growth method. Then we could obtain any statistics from these weighted samples.

Liu and Chen (Liu and Chen, 1995) extended the Rosenbluth biased sample growth method (Rosenbluth and Rosenbluth, 1955) by setting up a general framework of Sequential Importance Sampling (SIS)

method. More flexible and effective algorithms can be established from this framework. The SIS chain growth strategy (Liu and Chen, 1995; Liu, 2008) is being wildly used, and it is successful when applied to proteins structure studies (Liang et al., 2002; Lin et al., 2008). In the context of growing polymer chain, SIS can be formulated as follows. Let $(x_1,\ldots,x_t)$ be the position of the t monomers in a chain of length t. Let $\pi_1(x_1),\pi_2(x_1,x_2),\ldots,\pi_t(x_1,\ldots,x_t)$ be a sequence of target distributions with $\pi(x_1,\ldots,x_n)=\pi_n(x_1,\ldots,x_n)$ being the final target distribution from which we wish to draw an inference from. Let $g_{t+1}(x_{t+1}|x_1,\ldots,x_t)$ be a sequence of trial distributions which indicates the growing of the polymer chain. Then we have:

The configurations of successfully generated polymers ensemble $\{(x_1^{(j)},\ldots,x_n^{(j)})\}_{j=1}^m$ and their associated weights $\{w_n^{(j)}\}_{j=1}^m$ can be used to estimate any properties of the polymer chains, such as radius of gyration, compactness, and local environment. The objective inference $\mu_h=E_\pi[h(x_1,\ldots,x_n)]$ is estimated with

$$\hat{\mu}_h=\frac{\sum_{j=1}^m w_n^{(j)}\cdot h(x_1^{(j)},\ldots,x_n^{(j)})}{\sum_{j=1}^m w_n^{(j)}}, \tag{3.1}$$

for any integrable function h of interests.

The Rosenbluth method (Rosenbluth and Rosenbluth, 1955) is a special case of SIS. Its target distributions $\pi_t(x1,\ldots,x_t)$ is the uniform distribution of all self-avoiding walking polymer chains of length t. Its sampling distribution $g_{t+1}(x_{t+1}|x_1,\ldots,x_t)$ is the uniform distribution among all $n_{t+1}(x_1,\ldots,x_t)$ unoccupied neighbouring sites of the last monomer x_t, and the weight function is

$$w(x_1,\ldots,x_t,x_{t+1})=w(x_1,\ldots,x_t)\cdot n_{t+1}(x_1,\ldots,x_t).$$

Algorithm 1 Sequential Important Sampling algorithm for random model

1: Draw $x_1^{(j)}, j = 1, \ldots, m$ from $g_1(x_1)$

2: Set the incremental weight $w_1^{(j)} = \pi_1(x_1^{(j)})/g_1(x_1^{(j)})$

3: **for** $t = 1 \to n-1$ **do**

4: **for** $j = 1 \to m$ **do**

// Sampling for the $(t+1)$th monomer for the jth sample

5: Draw position $x_{t+1}^{(j)}$ from $g_{t+1}(x_{t+1}|x_1^{(j)}, \ldots, x_t^{(j)})$

// Compute the incremental weight

6: $$u_{t+1}^{(j)} \leftarrow \frac{\pi_{t+1}(x_1^{(j),\ldots,x_{t+1}^{(j)}})}{\pi_t(x_1^{(j)}, \ldots, x_t^{(j)}) \cdot g_{t+1}(x_{t+1}^{(j)}|x_1^{(j)}, \ldots, x_t^{(j)})}$$

7: $w_{t+1}^{(j)} \leftarrow u_{t+1}^{(j)} \cdot w_t^j$

8: **end for**

9: Resampling

10: **end for**

When there is no candidate positions for the $(t+1)$th monomer to grow, the number of unoccupied neighbouring sites $n_{t+1} = 0$. In this case, we cannot continue to grow the chain one monomer by one monomer. We then discard the whole partial chain and regrow one new polymer chain. In the case of the Rosenbluth method, no resampling is used. We generate 100,000 random bead rod self-avoiding walking polymer chains as one ensemble.

3.3 Significant interactions for GM12878 and K562 cells

3.3.1 Remove non-specific interaction from 5C data

We remove the non-specific interactions from 5C data by using the random ensemble including 100,000 random polymer chains generated by using sequential importance sampling interaction frequencies among primer sites in 5C data both for GM12878 and K562 cells are mapped to the node-node interaction in physical model. We define the propensity $prop_{ij}$ of node i and node j to be in spacial contact, and this is modeled as an odds ratio. We have

$$prop_{ij} = \frac{q_{ij}}{q_{ij}^R},, \tag{3.2}$$

where the probability $q_{ij} = \frac{n_{ij}}{n}$ is the probability of node i and node j to be in contact in real chromatin fibers as measured in the 5C experiment. n_{ij} is the interaction frequency between nodes i and j among all experimental tests, and n is the total number of interaction frequencies in a cell type. The random probability $q_{ij}^R = \frac{\sum_k w_k I(i,j)}{\sum_i \sum_j \sum_k w_k I(i,j)}$ is the probability of node i and node j to be in contact in the reference state. w_k is the weight of kth chain in the ensemble, and $I(i,j)$ is a indicator function when node i and j has contact is 1 in distance threshold $d_c = 84$ nm, otherwise 0.

3.3.2 *p*-value

To test if the interaction between node i and node j is significant, we compare how many times experimental q_{ij} less than bootstrapped q_{ij}^R by bootstrapping 1000 times of 100,000 random chains with replacement. Where $q_{ij_m}^R = \frac{\sum_{k'} w_{k'} I(i,j)}{\sum_i \sum_j \sum_{k'} w_{k'} I(i,j)}$, and $w_{k'}$ is the weight of k'th random chain from the mth bootstrapped 100,000 samples with replacement. The probability p_{ij} of an interaction is:

$$p_{ij} = \frac{\sum_{m=1}^{M} I(q_{ij} < q_{ij_m}^R)}{M}, \tag{3.3}$$

where M = 1000, and $I(\cdot)$ is a indicator function when condition is satisfied is 1, otherwise 0. Then False Discovery Rate (FDR) control is employed to correct multiple comparisons in these multiple hypothesis test. Considering the genomic distance between primer sites has effect in final interaction frequency, we employed FDR control on the groups where the genomic separation between nodes are K, where $K = j - i$ is constant. For each constant K, sort p_{ij} ascendantly to get new p-value set $\{p_{ij}^{(m)}\}$, such that $p_{ij}^{(1)} \leqslant p_{ij}^{(2)} \leqslant \cdots \leqslant p_{ij}^{(m)}$ are ordered, where m is the total number of the set $\{K | K = j - i, i$ and j is the node index of the α-globin gene domain chain of physical model $\}$. Then we use Hochberg adjustment method (Hochberg, 1988) to adjust p-value $p_{ij}^{(m)}$,

$$\tilde{p}_{ij}^{(l)} = \begin{cases} p_{ij}^{(m)} & \text{for } l = m, \\ \min(\tilde{p}_{ij}^{(l+1)}, \frac{m}{l} p_{ij}^{(l)}) & \text{for } l = m-1, \ldots, 1. \end{cases} \tag{3.4}$$

When the entire family of tests is considered, and the adjusted p-value is less than a significance level $\alpha = 5\%$, we reject the null hypothesis. After the FDR is used for different $K = 1 \ldots 53$ nodes in the

separation set, we obtain 132 and 83 significant interactions for GM12878 and K562 individually out of total 426 interactions.

3.3.3 Mapping significant interactions to distance

We map the minimum of the propensity corresponding to the distance threshold $d_c = 84$ nm, the maximum of the propensity correspond to the collision distance $\mu = 30$ nm, and assume the propensity $prop_{ij}$ versus spatial distance between i and j follows half Gaussian distribution,

$$\frac{prop_{ij}}{\max prop_{ij}} = \exp \frac{-(d_{ij}-\mu)^2}{2\sigma^2}, \text{and } d_{ij} > \mu. \tag{3.5}$$

where $\sigma = \frac{d_c-\mu}{\sqrt{2\log \frac{\max prop_{ij}}{\min prop_{ij}}}}$. Therefore from the Equation 3.5, we have the entry d_{ij} of the distance matrix D between node i and node j, which has significant interaction as,

$$d_{ij} = \mu + \sqrt{2\sigma^2 \log \frac{\max prop_{ij}}{prop_{ij}}}, \text{and } prop_{ij} > 0. \tag{3.6}$$

3.4 Growth model for α-globin gene domain

3.4.1 SIS algorithm to reconstruct 3D conformation of α-globin gene domain

In this study, we propose Sequential Importance Sampling algorithm for reconstructing the 3D conformation of the α-globin gene domain based on the physical model and distance matrix. Besides generating 3D structures, we can also obtain physical properties of the α-globin gene domain ensembles.

Without loss of generality, one can get conformations by minimizing an error function measuring deviation in distance from the distance matrix constraints.

$$\mathscr{E}(\boldsymbol{x}_n^{(k)}) = \frac{\sum\limits_{(i,j)\in P_{\boldsymbol{x}_n^{(k)}}} \left| \| x_i - x_j \| - d_{i,j} \right|}{\sum\limits_{(i,j)\in P_{\boldsymbol{x}_n^{(k)}}} d_{i,j}}, \tag{3.7}$$

where $P_{\boldsymbol{x}_n^{(k)}} = \{(i,j)|$ significant contact node pair i and j exist in the kth conformation $\boldsymbol{x}_n^{(k)}\}$ contains significant contact node pairs of kth conformation $\boldsymbol{x}_n^{(k)}$ and significant contact pair i and j has corresponding distance is d_{ij} in which the distance constraints are partial and incomplete in the 5C data. Our objective is to generate a set of conformations satisfying all distance constraints and following certain target distribution $\pi(\boldsymbol{x}_n^{(k)})$,

$$\pi(\boldsymbol{x}_n^{(k)}) = \exp(-\mathscr{E}(\boldsymbol{x}_n^{(k)})) \tag{3.8}$$

Let $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ be a vector for the partial chain node positions from node 1 to node t. The joint trial distribution for a partial chain $\boldsymbol{x}_t$ is

$$g_t(\boldsymbol{x}_t) = g_1(\boldsymbol{x}_1) g_2(x_2|\boldsymbol{x}_1) \ldots g_t(x_t|\boldsymbol{x}_{t-1}).$$

where $g_t(x_t|\boldsymbol{x}_{t-1})$ is trial distribution in that given the previous nodes positions condition $\{\boldsymbol{x}_{t-1}|x_1, \ldots, x_{t-1}\}$, the possible positions x_t with different probabilities for the node t may retain.

Following the principle of importance sampling (Marshall, 1956; Liang et al., 2002; Liu, 2008), We assign a weight which is given by

$$w(\boldsymbol{x}_n^{(k)}) = \pi(\boldsymbol{x}_n^{(k)}) / g_n(\boldsymbol{x}_n^{(k)})$$

to each conformation sample $\boldsymbol{x}^{(k)}$, $k = 1, \ldots, m$, where $g_n(\boldsymbol{x}^{(k)})$ is the full chain trial distribution. Therefore we can estimate the expected mean value of the physical properties of the chromatin.

$$E_\pi(h(\boldsymbol{x}_n)) \simeq \frac{\sum_{k=1}^{m} w(\boldsymbol{x}_n^{(k)}) \cdot h(\boldsymbol{x}_n^{(k)})}{\sum_{k=1}^{m} w(\boldsymbol{x}_n^{(k)})},$$

To maintain the samples diversity, we generate sample conformations by adopting the Fearnhead et al. framework (Fearnhead and Clifford, 2003). In Algorithm 2, we set m_t as the number of samples in the tth iteration and $m_{max} = \max(m_t)$.

3.4.2 Priority score $\beta_t^{(l)}$

Constructing a high quality priority scores $\beta_t^{(l)}$ is a crucial step in algorithm 2, which works as the trial distribution $g_t(x_t|\boldsymbol{x}_{t-1})$ to guide the growth of the partial chains $\boldsymbol{x}_{t-1}$ towards more profitable regions and satisfies the target distribution, such that the full chain will eventually obey all the distance constrains. We developed a priority score consisting of three components: growth potential from collision constraints, growth potential from distance constraints and growth potential from distance loop consideration. The last two potential components of the priority score incorporate the distance information of future nodes.

3.4.3 Growth potential from collision constraint

Due to 30 nm chromatin fiber in the model, this growth potential function penalizes the violation of the lower bound constraint 30 nm,

$$f_1(x_t) = \sum_{B_{\boldsymbol{x}_{t-1}}} h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0), \tag{3.9}$$

Algorithm 2 Sequential Importance Sampling algorithm for growing model

Set $m_1 = 1$, $w_1^{(1)} = 1.0$ and place the first residue at fixed $x_1^{(1)}$

for $t = 2 \rightarrow n$ **do**

    $L_t = 0$

    // L_t: number of length t chains that can be obtained from samples obtained at step $t-1$.

    for sample $j = 1 \rightarrow m_{t-1}$ **do**

        Find all of the valid sites $x_t^{(i,j)}$, $i = 1, \cdots, l_t^{(j)}$ for placing x_t next to partial chain $\boldsymbol{x}_{t-1}^{(j)}$

        // $l_t^{(j)}$ = number of available sites to place x_t next to partial chain $\boldsymbol{x}_{t-1}^{(j)}$.

        Generate $l_t^{(j)}$ number of t-length chain $\widetilde{\boldsymbol{x}}_t^{(L_t+i)} = (\boldsymbol{x}_{t-1}^{(j)}, x_t^{(i,j)})$

        $\widetilde{w}_t^{(L_t+i)} = w_{t-1}^{(j)}$

        // Temporary weights for uniform distribution.

        $L_t = L_t + l_t^{(j)}$

        if $L_t \leq m_{max}$ **then**

            Let $m_t = L_t$ and $\{(\boldsymbol{x}_t^{(j)}, w_t^{(j)})\}_{j=1}^{m_t} = \{(\widetilde{\boldsymbol{x}}_t^{(l)}, \widetilde{w}_t^{(l)})\}_{l=1}^{L_t}$

        else

            Let $m_t = m_{\max}$

            for $l = 1 \rightarrow L_t$ **do**

Algorithm 2 Sequential Importance Sampling algorithm for growing model (continued)

Assign a priority score $\beta_t^{(l)}$ for chain $\widetilde{\boldsymbol{x}}_t^{(l)}$ according to the constraints

Find constant c such that $\sum_{l=1}^{L_t} \min\{c\beta_t^{(l)}, 1\} = m_{\max}$

// by binary search.

end for

Draw r from uniform distribution $\mathscr{U}[0,1)$

for $j = 1 \to m_{\max}$ **do**

Find integer J_j such that $\sum_{l=1}^{J_j-1} \min\{c\beta_t^{(l)}, 1\} < r_j \leq \sum_{l=1}^{J_j} \min\{c\beta_t^{(l)}, 1\}$

Select sample $\boldsymbol{x}_t^{(j)} = \widetilde{\boldsymbol{x}}_t^{(J_j)}$

Set weight $w_t^{(j)} = \widetilde{w}_t^{(J_j)} \cdot (\gamma_t^{(J_j)} / \beta_t^{(J_j)})$

end for

end if

end for

end for

for $j = 1 \to m_n$ **do**

Calculate importance weight $w(\boldsymbol{x}_n^{(j)}) \propto w_n^{(j)} \pi(\boldsymbol{x}_n^{(j)})$

end for

where h_1 is the loss function to measure the violation of constraint of x_t with its previous partial chain $B_{\boldsymbol{x}_{t-1}}$, which not only include the previous nodes $\{x_1,\ldots,x_{t-1}\}$, but also the middle balls interpolating each segment between two end nodes x_i and x_{i+1}, where i from 1 to $t-1$, according to our bead rod self-avoiding walking polymer model. And

$$h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = I(||x_t - \tilde{x}_i|| < r_0), \text{any } \tilde{x}_i \in B_{\boldsymbol{x}_{t-1}},$$

where $I(\cdot)$ is a indicator function, such that $h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = 0$, when $||x_t - \tilde{x}_i|| \geqslant r_0$, and $h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = 1$, when $||x_t - \tilde{x}_i|| \leqslant r_0$, here $r_0 = 30$ nm. The more violation of the 30 nm lower bound distance constraints, the bigger the value of $f_1(x_t)$ will be.

3.4.4 Growth potential from distance constraints

Given a partial chain $\boldsymbol{x}_{t-1}$, if the position of a current node t ($x_t \notin \boldsymbol{x}_{t-1}$) is strongly constrained, the node t should be restricted in a small limited spatial region. We generate a number of uniformly distributed candidate nodes x_t which have distance $d_{t-1,t}$ with node x_{t-1} in three dimensional space. We have growth potential from distance constraints $f_2(x_t)$ to encourage x_t to satisfy the distance constrains.

$$f_2(x_t) = h_2((\parallel x_{i_1} - x_t \parallel, \ldots, \parallel x_{i_K} - x_t \parallel), (d_{i_1,t}, \ldots, d_{i_K,t})), \tag{3.10}$$

where $i_k \in P_t = \{i \mid \text{node } i \in \{ \text{ node of partial chain } \boldsymbol{x}_{t-1}\}, \text{ node } i \text{ has significant contact with node } t\}$, K is the number of nodes that have existing significant contacts with node x_t in the partial chain $\boldsymbol{x}_{t-1}$, and $d_{i_k,t}$ is the distance matrix D element between significant contact node pair i_k and t. h_2 is the loss function to measure the similarity between node positions and distant constraints.

$$h_2((\| x_{i_1} - x_t \|, \ldots, \| x_{i_K} - x_t \|), (d_{i_1,t}, \ldots, d_{i_K,t})) = \frac{\sum_{i_k \in P_t} |\| x_{i_k} - x_t \| - d_{i_k,t}|}{\sum_{i_k \in P_t} d_{i_k,t}},$$

3.4.5 Growth potential from loop constraints

For the node growth inside the loops, especially for some node t without any significant interaction pairs with the partial chain $\boldsymbol{x}_{t-1}$, we have to impose distance constrains to enforce the node t to follow the triangle distance inequality rule. We propose the potential function from the loop constrains $f_3(x_t)$ to punish the node t from going far away,

$$f_3(x_t) = h_3(x_t, O_t), \tag{3.11}$$

where $O_t = \{(t_{i_k}, t_{j_k}) \mid \text{significant interaction node pair } t_{i_k} \text{ and } t_{j_k} \text{ exist and } t_{i_k} < t < t_{j_k}\}$, and as before, h_3 is a loss function to measure the triangle inequality,

$$h_3(x_t, O_t) = \sum_{(t_{i_k}, t_{j_k}) \in O_t} I(\Big| \| x_t - x_{t_{i_k}} \| - d_{t_{j_k}, t_{i_k}} \Big| > \sum_{l=t}^{t_{j_k}-1} d_{l,l+1}), \tag{3.12}$$

which punishes the violation of triangle inequality. $d_{l,l+1}$ is the length of segment between node l and $l+1$, l from t up to $t_{j_k} - 1$, and $I(\cdot)$ is a indicator function such that when the distance of node t between the node t_{i_k} greater than the sum of the rest segment length between the node t and the node t_{j_k}, which means the position x_t will not satisfy the distance constraint, then value is 1, otherwise the value is 0.

3.4.6 Combined priority score

The combined priority score $\beta_t^{(l)}$ for chain $\tilde{\boldsymbol{x}}_t^{(l)}$ is set as

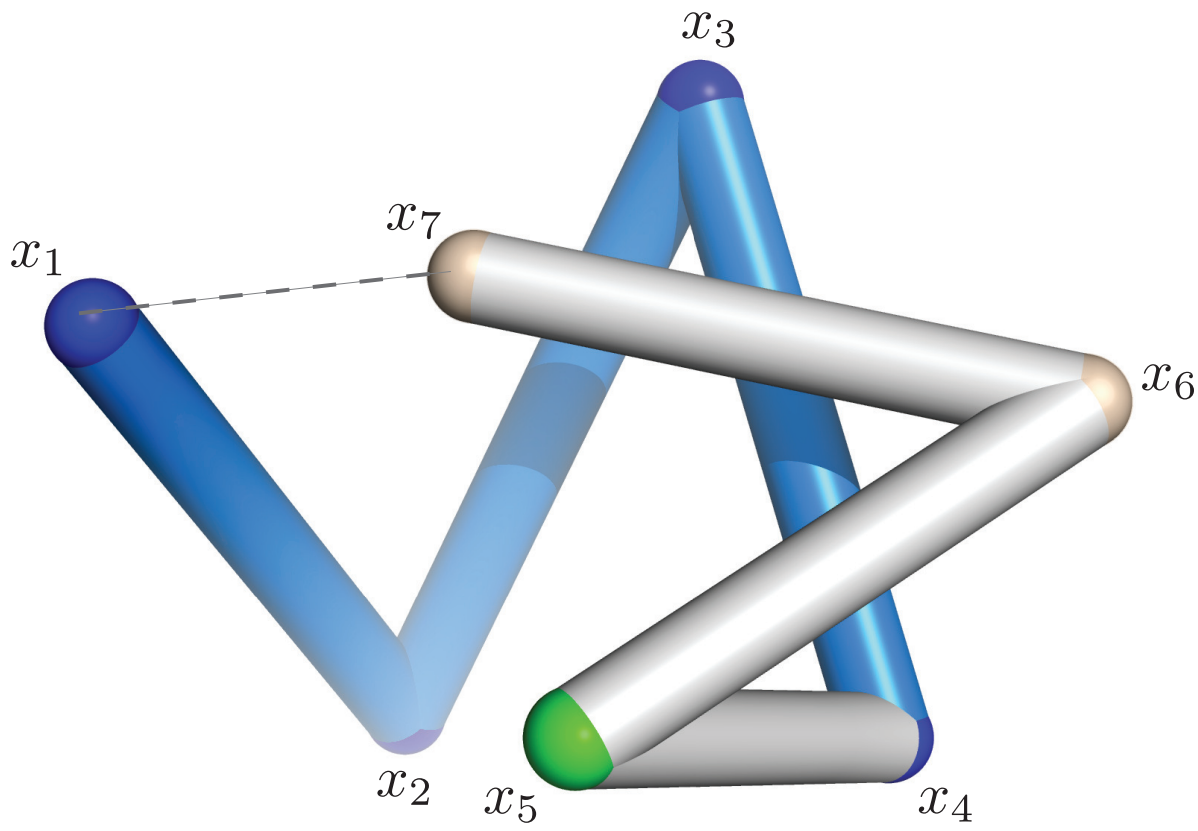

Figure 4: Growth potential from loop constraints. For simplicity, in the cartoon, there is only one loop constraint connected between node 1 and node 7 (dashed line) and the distance is $d_{1,7}$. The partial chain is $\boldsymbol{x}_4$ represented by blue, and node colored by dark blue. The current growing node 5 (green) and the future nodes 6 and 7 (orange). The grey line indicates the rod segments that will potentially be placed. The loop constraint applies on node 5 is the deviation of $\| x_5 - x_1 \|$ and $d_{1,7}$ should less than the summation of length of segments S_5 and S_6. Especially, when growing to the candidate node 6, the partial chain $\boldsymbol{x}_5$ is already determined, then the loop constraint is exactly the triangle inequality rule, $|\| x_6 - x_1 \| - d_{1,7}|$ should less than $d_{6,7}$.

$$\beta_t^{(l)} = \exp\left[-\frac{\rho_1 f_1(\tilde{x_t}^{(l)}) + \rho_2 f_2(\tilde{x_t}^{(l)}) + \rho_3 f_3(\tilde{x_t}^{(l)})}{\tau_t}\right] \tag{3.13}$$

where ρ_1, ρ_2, and ρ_3 are coefficients of the three growth potential functions (growth potential from collision constrains, growth function from distance constrains and growth potential from loop constrains), τ_t is a temperature like variable. In this study, we set $\rho_1 = \rho_2 = \rho_3 = \tau_t = 1$.

3.4.7 Target score $\gamma_t^{(l)}$

The target score $\gamma_t^{(l)}$ for chain $\tilde{\boldsymbol{x}}_t^{(l)}$ is set as

$$\gamma_t^{(l)} = \exp\left[-\frac{\rho_1 f_1(\tilde{x_t}^{(l)}) + \rho_2 f_2(\tilde{x_t}^{(l)})}{\tau_t'}\right], \tag{3.14}$$

where ρ_1 and ρ_2 are coefficients of the two growth potential functions, and same as those of the priority score $\beta_t^{(l)}$. τ_t' is a temperature like variable, and $\tau_t' = \frac{1}{2}\tau_t$. It is different from the random model, where we assume the uniform distribution, while we try to rebuild the 3D structure conforming to the experimental observation, i.e. distance constrains, and self-avoiding walk in growth model.

Figure 5 is one of example to show how we grow the chromatin polymer chain one step by one step according to the potential functions Equation 3.9, Equation 3.10, Equation 3.11 and target score Equation 3.14.

3.5 Calculation Details

3.5.1 Root mean square deviation (*RMSD*) of distance

$$RMSD(\boldsymbol{c}_m, \boldsymbol{c}_n) = \sqrt{\frac{\sum_{k=1}^{l} ||c_m^k - c_n^k||^2}{l}}, \tag{3.15}$$

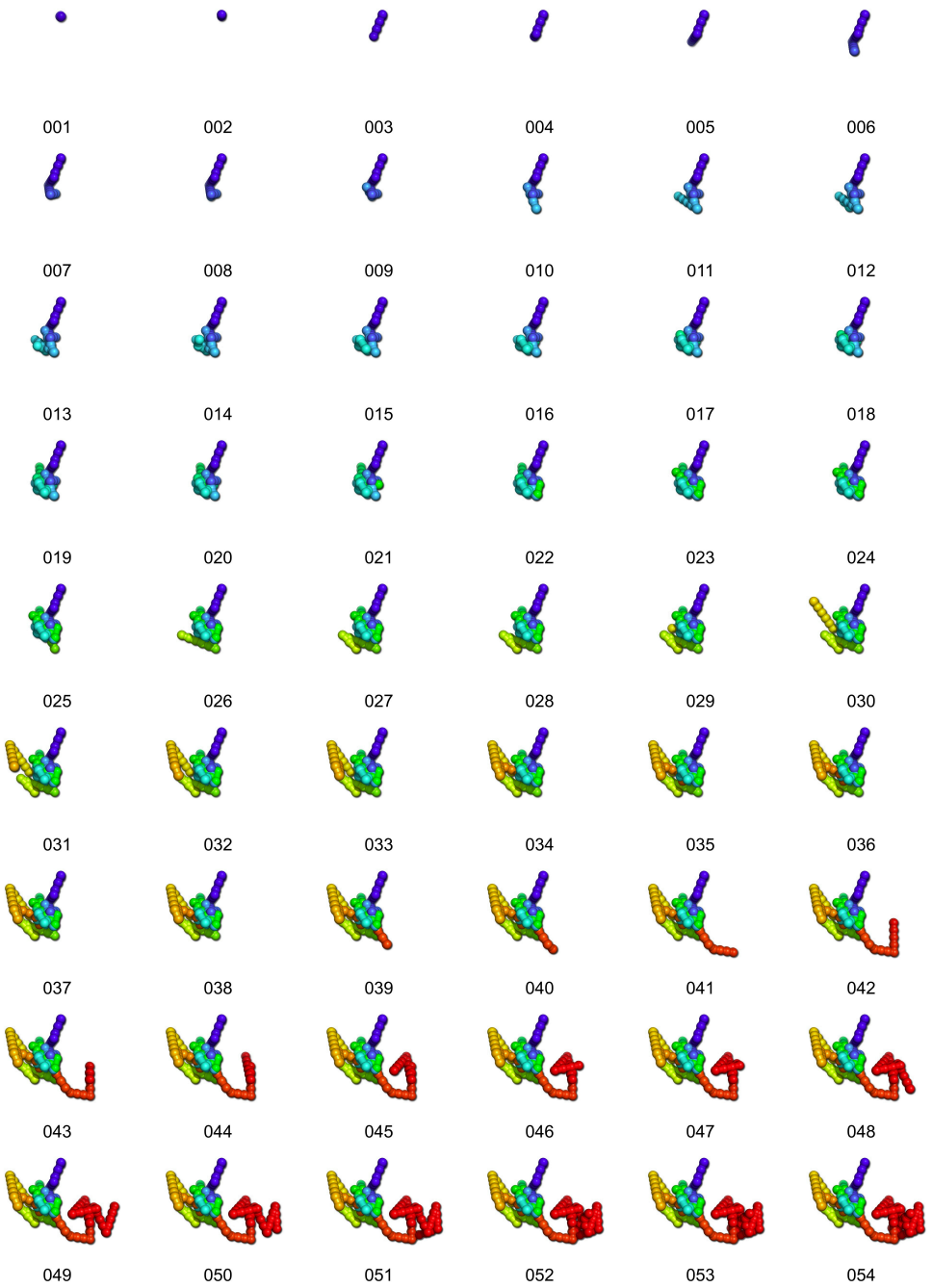
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054

Figure 5

Figure 5 *(previous page)*: Illustration of growing α-globin gene domain. We grow the α-globin gene domain under the potential functions Equation 3.9, Equation 3.10, Equation 3.11 and target score Equation 3.14. This figure shows the example how we grow the polymer chain. Our model contains 54 sites for the α-globin gene domain. Between two consecutive nodes, various number nodes are filled such that the bead-string polymer chain is formed. In each growth step, we randomly generate the next site, assign its growth value, then select it according to its probability. We color the chain from blue to red to clearly visualize how the α-globin gene polymer chain grows from the beginning to the end. Each step is numbered serially to show the growth procedure.

where $\boldsymbol{c}_m = \{dist_{ij} | ij \text{ pair is the significant interactions}\}$ is the all significant contacts distance collection of mth predicted conformation, l is the total number of significant interactions and c_m^k is the kth distance element in the set $\boldsymbol{c}_m$.

3.5.2 Probability q_{ij}^{pred} of interactions between i and j in predicted model

$$q_{ij}^{pred} = \frac{\sum_{k'} w_{k'} I(i,j)}{\sum_i \sum_j \sum_{k'} w_{k'} I(i,j)}, \tag{3.16}$$

where $I(i,j)$ is indicator function, if the distance between monomer i and j in the k'th conformation less than the threshold d_c, $I(i,j) = 1$, otherwise 0. k' is range from 1 to 10,000 conformations for the GM12878 cell or the K562 cell.

3.5.3 Propensity $Prop_{ij}^{pred}$ of interaction between i and j in predicted model

$$Prop_{ij}^{pred} = \frac{q_{ij}^{pred}}{q_{ij}^{R}} = \frac{\frac{\sum_{k'} w_{k'} I(i,j)}{\sum_i \sum_j \sum_{k'} w_{k'} I(i,j)}}{\frac{\sum_k w_k I(i,j)}{\sum_i \sum_j \sum_k w_k I(i,j)}}, \tag{3.17}$$

where $I(i,j)$ is indicator function, if the distance between monomer i and j in the k'th conformation less than the threshold d_c, $I(i,j) = 1$, otherwise 0. k is range from 1 to 100,000 conformations for the random model we generated. k' is range from 1 to 10,000 predicted conformations for the GM12878 cell or the K562 cell.

3.5.4 Percentage P_{ij}^{pred} of interacting between i and j in predicted model

$$P_{ij}^{pred} = \frac{\sum_{k'} w_{k'} I(i,j)}{\sum_{k'} w_{k'}} \tag{3.18}$$

where P_{ij}^{pred} is to measure the frequency of the interaction between monomers i and j in the whole predicted ensemble of conformations.

3.5.5 Contact index CI_i of monomer i in the predicted model

$$CI_i = \sum_{j \in S} P_{ij}^{pred} = \sum_{j \in S} \frac{\sum_{k'} w_{k'} I(i,j)}{w_{k'}}, \tag{3.19}$$

where CI_i is the contact index of monomer i to measure the monomer i contact propensity in the predicted model. High CI_i indicates higher propensity with other monomers. S is a set of monomers $\{j | j$ is any one of the monomer of the physical model and $j \neq i\}$ in the physical model.

3.5.6 Alpha shape

Alpha shape is a subset of delaunay triangulation, which can represent the geometry more accurately and can capture contact interactions (Edelsbrunner, 1987). In addition, no fictitious contacts will be introduced between two monomers when there is a third intervening monomer (Li et al., 2003).

In this study, we treat the segment ends as 30 nm sphere, and the probe sphere diameter is 54 nm (d_c = 84 nm = 54 nm + 30 nm).

3.5.7 Probability of local monomer interactions by alpha shape

For discovering local monomer interaction environment, we calculate three types of connections by alpha shape to detect interactions between i and j. (1) Triplet connection, in which monomer k direct connect both monomer i and j when monomer i and j has connection: (2) k-connection, in which monomer k only connect with monomer i or j when monomer i and j has connection: (3) Intercept connection, in which monomer k connect both monomer i and j when monomer i and j do not connect each other.

3.5.8 Triplet connection

$$L_{i-j}^{k_{triplet}} = \frac{\sum_{l \in E_{i-j}^{triplet}, k \in S_{i-j}^{triplet}} w_l}{\sum_k \sum_{l \in E_{i-j}^{triplet}, k \in S_{i-j}^{triplet}} w_l} \tag{3.20}$$

The probability $L_{i-j}^{k_{triplet}}$ describes the frequency with which some k monomer connect two monomers i and j, when monomer i and j has connection. $S_{i-j}^{triplet} = \{k | i \leftrightarrow j, i \leftrightarrow k, j \leftrightarrow k\}$, and $E_{i-j}^{triplet} = \{\text{chain}|\ \text{satisfy}\ S_{i-j}^{triplet}\}$

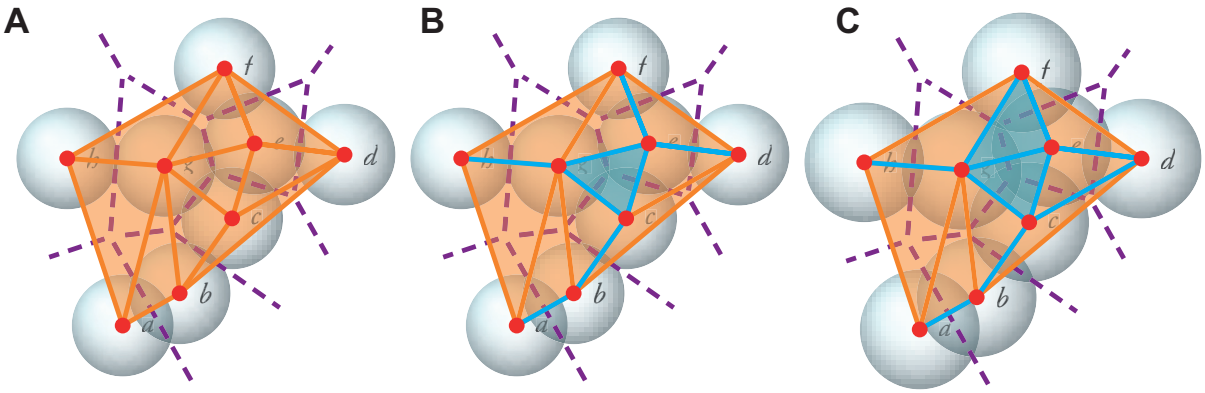

Figure 6: Alpha shape diagram. This is a simple two dimensional alpha shape representation. (A) In the two dimensional space, there are total 8 spheres with the same radii r, marked with lowercase letter a to h, the center colored as red points and the sphere is colored light blue. Violet dashed line split the region in this voronoi diagram, with each region contain only one center of the sphere. Each violet dashed line vertically evenly divided the segment connected by two centers of the spheres. Any point of each dashed line has same distance with the centers which it divide. The dual formation is the delaunay triangulation represented by an orange color in this figure. (B) Alpha shape is a subset of delaunay triangulation colored by cyan, containing segment $ab, bc, ce, de, fe, ge, cg, gh$ (tow spheres are overlapped) and one triangle ceg (three spheres are overlapped). (C) With the radius of spheres increasing, the alpha shape has more additional segments fg and triangular efg. If the radius of spheres expand to infinity, all spheres are overlapped, the alpha shape will become the delaunay triangulation (orange segment and triangle area in (A)). If the radius of spheres decrease close to zero, no any spheres overlapping, then there is no alpha shape.

3.5.9 k connection

3.5.9.1 k connection with i

$$L_{i-j}^{k_{-i}} = \frac{\sum_{l \in E_{i-j}^{-i}, k \in S_{i-j}^{-i}} w_l}{\sum_k \sum_{l \in E_{i-j}^{-i}, k \in S_{i-j}^{-i}} w_l} \tag{3.21}$$

Where the probability $L_{i-j}^{k_{-i}}$ detects how often k monomer connect only monomer i , when monomer i and j has connection. $S_{i-j}^{-i} = \{k | i \leftrightarrow j, k \leftrightarrow i, k \nleftrightarrow j\}$ and $E_{i-j}^{-i} = \{\text{chain}|\text{ satisfy } S_{i-j}^{-i}\}$.

3.5.9.2 k connection with j

$$L_{i-j}^{k_{j-}} = \frac{\sum_{l \in E_{i-j}^{j-}, k \in S_{i-j}^{j-}} w_l}{\sum_k \sum_{l \in E_{i-j}^{j-}, k \in S_{i-j}^{j-}} w_l} \tag{3.22}$$

Where the probability $L_{i-j}^{k_{j-}}$ detects how often k monomer connect only monomer j , when monomer i and j has connection. $S_{i-j}^{j-} = \{k | i \leftrightarrow j, i \nleftrightarrow k, j \leftrightarrow k\}$ and $E_{i-j}^{j-} = \{\text{chain}|\text{ satisfy } S_{i-j}^{j-}\}$

3.5.10 intercept connection

$$L_{i-j}^{k_{intercept}} = \frac{\sum_{l \in E_{i-j}^{intercept}, k \in S_{i-j}^{intercept}} w_l}{\sum_k \sum_{l \in E_{i-j}^{intercept}, k \in S_{i-j}^{intercept}} w_l} \tag{3.23}$$

Where the probability $L_{i-j}^{k_{intercept}}$ detect how often k monomer intercept monomer i and j , when monomer i and j has no connection. $S_{i-j}^{intercept} = \{k | i \nleftrightarrow j, i \leftrightarrow k, k \leftrightarrow j\} E_{i-j}^{intercept} = \{\text{chain}|\text{ satisfy } S_{i-j}^{intercept}\}$

3.5.11 Density-based algorithm

In this study, we adapted a density-based clustering method (Ester et al., 1996) to cluster 10,000 conformations for both GM12878 and K562 cells separately.

In the clustering conformations for the GM12878 and K562 cell, we measure the RMSD (Equation 3.15) for each pair conformation, then use traditional density-based algorithms to cluster the GM12878 and K562 cells conformations. If the conformation has at least 5 neighbouring conformations for which RMSD is less than 34 nm, we treat this conformation as one cluster of core conformations, otherwise this will be a border conformation or a noise conformation.

We also use density-based algorithms to cluster which monomer have more neighbouring monomers surrounding. We measure the percentage P_{ij}^{pred} of interactions in the predicted conformations (Equation 3.16), the higher value of P_{ij}^{pred}, the more likely it is to be connected between monomer i and j in the predicted conformations, therefore we query neighbour monomers by the condition of greater than some certain probability of interactions in the predicted conformations P_{ij}^{pred}.

3.6 Results

Our approach for generating 3D structure of α globin gene domain as an interacting, self-avoiding polymer chain from a derivation of the 5C interaction frequency matrix includes three steps with several substeps: (1) data translation into monomer interactions and obtaining the significant interactions by removing random interactions from 5C data, (2) translating the significant interactions into distance constraints and model building by sequential importance sampling, (3) ensemble analysis of the candidate 3D structures and proposing a mechanism for the activation of α-globin gene in K562 cells.

The following sections describe the results of each of these key steps in our approach to 3D structure determination of α-globin gene domain located in human chromosome 16.

3.6.1 From 5C data to polymer chain and obtaining the significant 5C interactions

3.6.1.1 5C analysis of α-globin gene domain

5C, described previously (Dostie et al., 2006; Umbarger et al., 2011), uses highly multiplexed ligation-mediated amplification to detect sets of 3C ligation products. We used the 5C data that was obtained by Baú et al. (Baù et al., 2010) specifically for the 500-kb α-globin gene domain of human chromosome 16 using HindIII restriction enzyme. The position of each ligation product is determined by designing forward and reverse primer sequences, where interactions are only possible between reverse and forward primer sites. According to their experiments, in total, there are 30 forward primers and 25 reverse primers, which were capable of detecting 750 unique pairwise chromatin interactions (Baù et al., 2010).

Due to the distribution of HindIII restriction sites and non-alternating design of the primers, the resulting interaction matrix is partial and biased. It is a very challenging task to predict the full 3D structure of α-globin gene domain from this partial and biased data. Our aim is to construct a general framework to build up a pipeline to get as much information as possible from partial 5C interaction frequencies and utilize it to discover known long range interactions, as well as to uncover interactions that cannot be discovered by 5C assay due to the design of the experiments.

3.6.1.2 Physical model

We build up a physical model as our basis to study the α-globin gene domain structure based on the 5C data. In our model, a canonical 30 nm chromatin fiber diameter, the mass density of 11 nm/kb (Bystricky et al., 2004; Dekker, 2008), and the persistence length of 150 nm (Wedemann and Langowski, 2002) are used. We discard short HindIII segments, which are less than 2727 bp which is equivalent to a 30 nm sphere. To correct proximity events, we ignored the interactions between consequent segments. We mapped

the segments that are longer than persistence length on to several monomers. By doing this, the 500 kb α-globin gene domain is split into 53 consecutive segments that have varying lengths and 54 distinctive sites. Then the 5C contact frequencies are mapped on the monomers where the primers of each HindIII segments are located. In total, we end up having 54 primer sites and each primer site corresponds to a monomer in our polymer chain. Figure 7A shows the location of genes on α-globin gene domain along with the primer locations and their corresponding numbering.

3.6.1.3 Random model and removal of non-significant interactions

In order to remove the expected interactions from the experimental data, we created a population of random C-SAC chains, as described in detail before (Gürsoy et al., 2014). Based on our physical model, we sequentially grow each monomer (primer site) randomly as a self-avoiding walk, until we reach the total length of 500 kb in a confined space of nucleus. A population of 100,000 properly weighted random chains is generated and the p-value of each 5C interaction frequency is calculated by bootstrapping. After the FDR adjustment, 132 out of 425 and 83 out of 367 interactions remained significant for GM12878 and K562 cells, respectively. In Figure 7B, you can see the circos diagram of α-globin gene domain with the interactions that are obtained from 5C experiment for both GM12878 and K562 cell lines. Figure 7C shows the interactions after removing the background from 5C experiment. We, then, adjusted the frequencies of the significant interactions by calculating their propensity, $prop(i, j)$ = observed/expected.

Notably, novel long-range interactions were identified after cleaning the data from random interactions. For example, in both cell lines, while α-globin gene almost equally interacts with a locus called LUC7L, the interaction between α-globin gene and distant upstream regulatory element HS48 is only significant in K562 cell line (Figure 7D). Similarly, in both cell lines, α-globin gene interacts with upstream regulatory

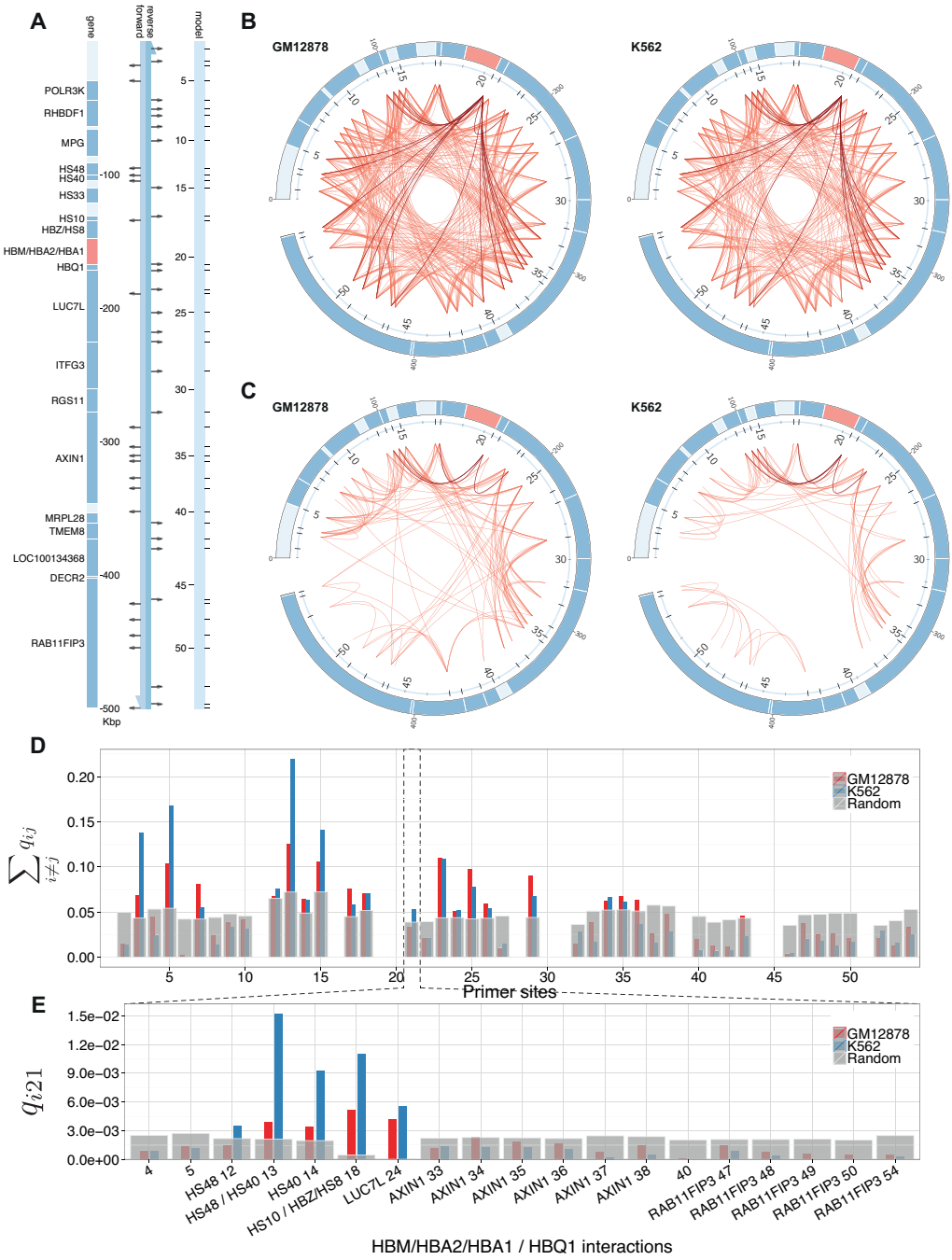
A
gene
forward
reverse
model
POLR3K
RHBDF1
MPG
HS48
HS40
HS33
HS10
HBZ/HS8
HBM/HBA2/HBA1
HBQ1
LUC7L
ITFG3
RGS11
AXIN1
MRPL28
TMEM8
LOC100134368
DECR2
RAB11FIP3
-100
-200
-300
-400
-500
Kbp
B
GM12878
K562
C
GM12878
K562
D
GM12878
K562
Random
Primer sites
E
GM12878
K562
Random
4
5
HS48 12
HS48 / HS40 13
HS40 14
HS10 / HBZ/HS8 18
LUC7L 24
AXIN1 33
AXIN1 34
AXIN1 35
AXIN1 36
AXIN1 37
AXIN1 38
40
RAB11FIP3 47
RAB11FIP3 48
RAB11FIP3 49
RAB11FIP3 50
RAB11FIP3 54
HBM/HBA2/HBA1 / HBQ1 interactions

Figure 7

Figure 7 *(previous page)*: Mapping of primer sites onto C-SAC model chain and removal of non-specific interactions.

(A) Linear diagram of the 500 Kbp α-globin gene domain. The first vertical line depicts the coordinates of the genomic regions. The α-globin gene is highlighted in red, and other genes are in dark blue. The second vertical line represents 5C HindIII segments (Baù et al., 2010). The forward strand is in light blue, and the reverse strand is in dark blue. Primer sites are marked by arrows. The third vertical line represents the position of the each primer site on model chain with their corresponding node index.

(B) Circos diagram of 500 Kbp α-globin gene domain with all interactions obtained from unprocessed 5C frequency data (Baù et al., 2010). The rings are the circos version of the first and third vertical line of (A). Red curves represent spatial interactions between HindIII primer sites. Dark red curves represent interactions between the α-globin gene and the rest of the HindIII primer sites.

(C) Significant interactions of the α-globin gene domain for the GM12878 and K562 cells, respectively. Out of 750 original 5C interactions, there are 132 significant interactions in the GM12878 cell and 83 significant interactions in the K562 cell after removal of non-specific interactions.

(D) The overall probability of a primer site to interact with other primer sites for the GM12878 cell (red) and for the K562 cell (blue) are shown. The overall probability of random interactions are shown in grey.

(E) The probability of interactions of between α-globin gene and another primer site. The probability of significant interactions in the GM12878 cell (red) and the K562 cell (blue) are shown. The probability of random interactions are in grey.

elements HS40 and HS48 (Figure 7D). However, the probability of these interactions is more than 3 fold higher in active K562 cells than in inactive GM12878 cells. Similar results were experimentally observed from an independent 3C study earlier (Dekker et al., 2002). It is also remarkable that, the total number of significant interactions observed for α-globin gene is greater than the total number of interactions in the random population in active K562 cells, though it is lower than the total number of interactions in the random population in inactive GM12878 cell line (Figure 7E).

3.6.2 From significant interactions to spatial distance constraints and chain growth

After calculating the propensity for the significant interactions, we assume that the relationship between propensity and the spatial distance between two primer sites follows a half-Gaussian distribution. We calculated the spatial distances between monomers and constructed our distance constraints (See SI). We employ the Sequential Importance Sampling (Liang et al., 2002; Lin et al., 2008; Liu, 2008) approach to grow the polymer chain for α-globin gene domain by using the distance constraints we estimated from 5C interaction frequencies. We generated 10,000 conformations for the GM12878 and K562 cell individually by properly weighting them. The highest weighted chains in both populations are the ones that satisfy the most distance constraints.

3.6.3 3D structure of α-globin gene domain in GM12878 and K562 cells

Our Importance Sampling Algorithm generated 10,000 3D models for α-globin gene domain by searching for a spatial arrangement of all primer sites that minimized the violation of the imposed distance constraints for each cell type. Although, any chain that violates the excluded-volume effect, had been discarded regardless. The highest weighted chains for both GM12878 and K562 cells represent the 5C frequency data best and are shown in Figure 8A. The main difference between two cell lines is the α-globin gene is a

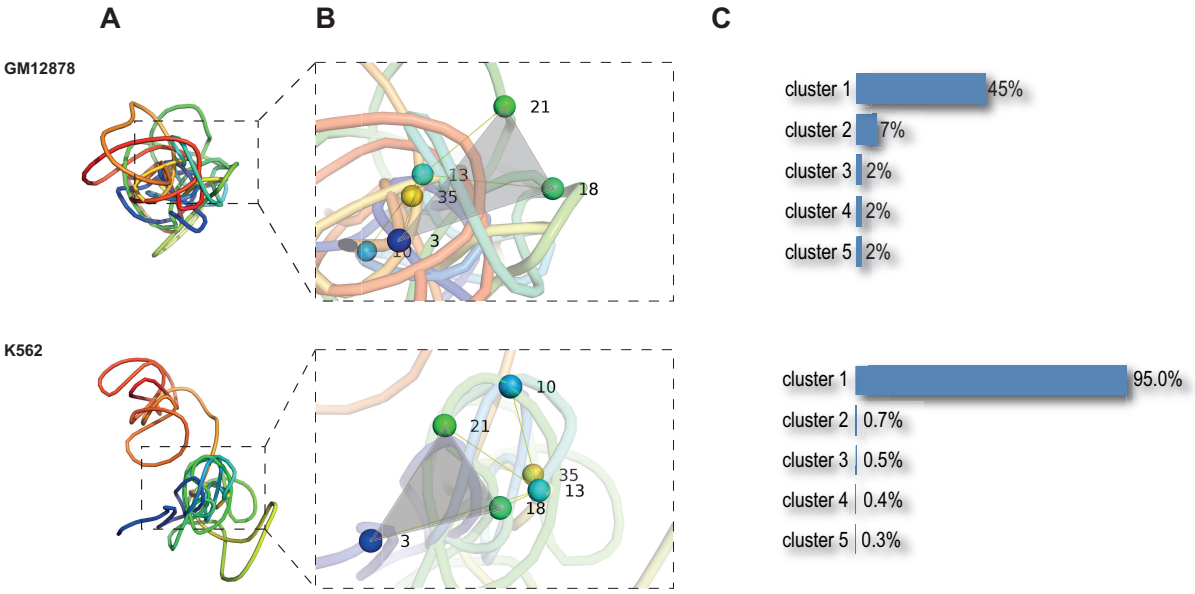

Figure 8: Representative three-dimensional chromatin structures of the α-globin gene domain in the GM12878 and the K562 cells and clustering.

(A) The genomic coordinates change from blue at the beginning to red at the end of the α-globin gene domain. The α-globin gene domain has only one domain and is more compact in the GM12878 cell. In the K562 cell, the α-globin gene domain has two distinct domains, which forms more extended structure.

compact conformation and there is only one chromatin domain formed in GM12878 cell, while it is more in open conformation and is formed by two distinct domains in K562 cells.

To assess the difference between cell lines in detail, we performed a structural alignment between structures based on similarly positioned primer sites (triangle in Figure 8A). While the primer sites 13 (HS40/48), 18 and 35 are clustered together in both cell lines, it seems like primer site 3 is in direct interaction with this

Figure 8 *(previous page)*: (B) Close up views of selected primer sites (3, 10/(MPG), 13/(HS48 / HS40), 18/(HS10 / HBZ/HS8), 21/(HBM/HBA2/HBA1 / HBQ1), 35/(AXIN1)) and their orientation in the α-globin gene domain for both GM12878 and K562 cells. The orientation of 3, 18, 21 are superimposed in both cell lines for clear comparison. There is a significant rearrangement of the position of the genes in the α-globin gene domain. The primer sites 10, 13, 35 are located to the left side of primer sites 18 and 21, which are close to the primer site 3 in the GM12878 cell, while the same group is located on the right side of primer sites 18 and 21, which is far away from the site 3 in the K562 cell. Primer site 3 and 13 are much closer to each other in the GM12878 cell than in the K562 cell.

(C) The ensemble of 10,000 modeled three-dimensional structures of the α-globin gene domain in both the GM12878 cell and the K562 cell form clusters. There are 5 main structured clusters for the GM12878 cell excluding the ones that contain less than 1.5% of the population. 45% of the population forms the largest cluster. There are 5 main structured clusters for the K562 cell excluding the ones that contain less than 0.3% of the population. 95% of the population are contained in the largest cluster. The conformation of α-globin gene domain is less fluctuating in the K562 cell. In contrast, there is a lot of diversity in the GM12878 cell conformations.

cluster only in GM12878 cell. Furthermore, primer 10 is in direct interaction with primer 13 (HS40/48) and primer 21 (α-globin gene) only in K562 cell. The primer site 3 is very close to primer site 13 (HS40/48) in GM12878 cell, while it is far away from 13 in K562 cell. Primer 13 (HS40/48) is upstream regulatory element and the interaction between 13 and 21 (α-globin gene) is necessary for the activation of the cell (Vernimmen et al., 2009). Although primer 13 is in direct interaction with primer 21 in both cell lines, it interacts with primer 3 only in inactive GM12878 cell. This extra interaction of primer 13 might be competing with the interaction between primer 13 (HS40/48) and 21 (α-globin gene), which is necessary for the activation of the cell in the active GM12878 cell.

10,000 model chains are clustered according to their structural similarities. The model chromatin chains for GM12878 cell are clustered in total of 487 different conformations (Figure 8). The only most populated cluster contained 45% of the entire population. Strikingly, models obtained for K562 cells formed more static sets of solutions, with a total of 92 clusters, including the top cluster spanning the 95% of the population.

3.6.4 Models reproduce known long-range interactions

We determined whether the 3D models reflected the known long-range interactions. We calculated all interactions from the spatial distances between primer sites from a weighted ensemble of 10,000 model chains for both GM12878 and K562 cells. There are a total of 460 significant interactions for the GM12878 cell. 120 of 460 significant interactions exist in the 5C frequency matrix. Only 12 interaction from the 5C measurements cannot be satisfied simultaneously with other 5C constraints (Figure 10A). There are a total 426 significant interactions in the K562 cell. 80 of them exist in the 5C interaction matrix and only 3 interactions from 5C measurements cannot be captured by our model (Figure 10A). The weighted distance

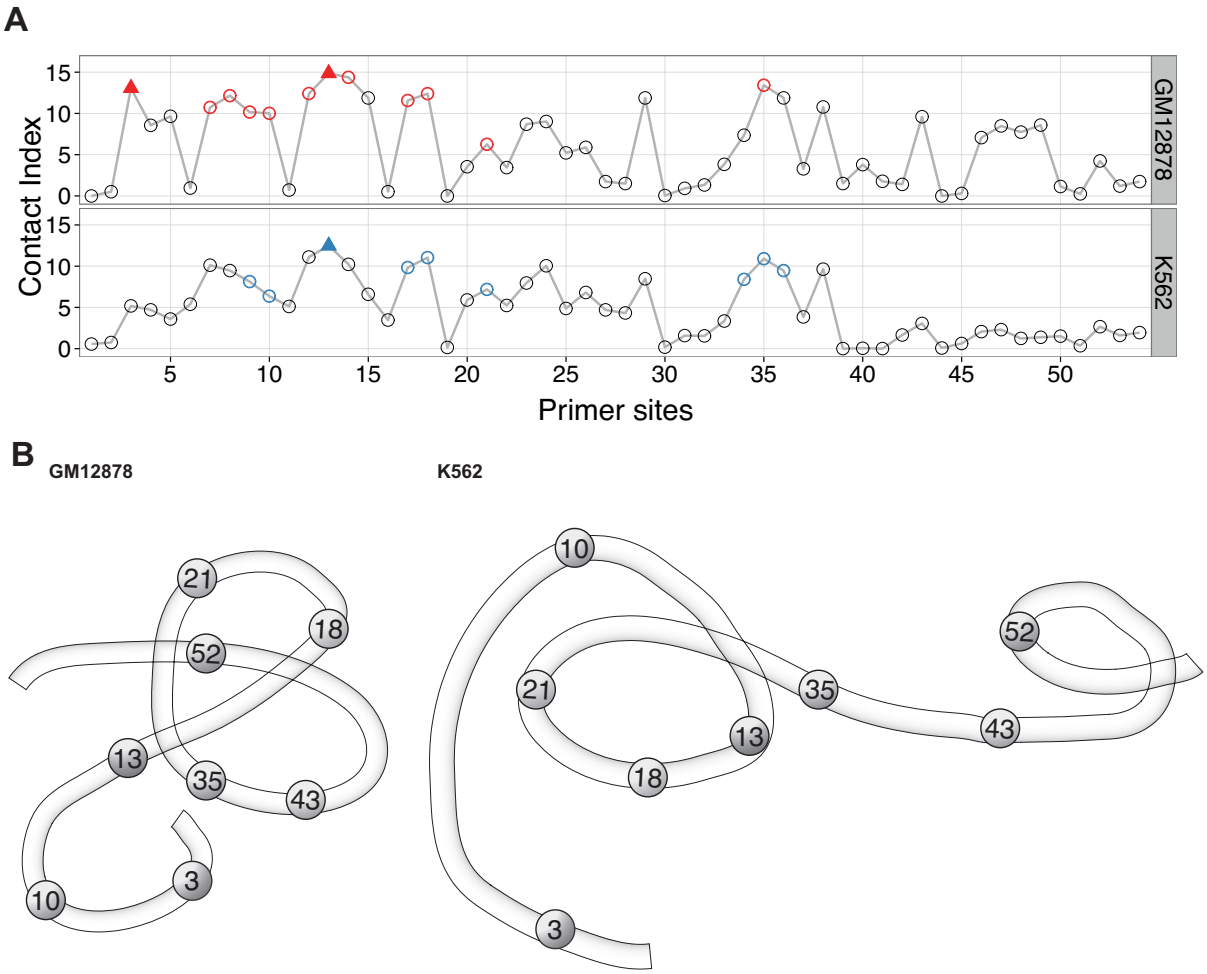
A
Contact Index
15
10
5
0
GM12878
K562
5
10
15
20
25
30
35
40
45
50
Primer sites
B
GM12878
K562
21
18
52
13
35
43
10
3
10
21
35
52
13
43
18
3

Figure 9

Figure 9 *(previous page)*: Spatially clustered primer sites and proposed model conformations of the α-globin gene domain for both GM12878 and K562 cells.

(A). Contact index, namely, is the propensity of a site to make interactions with other primer sites, for both GM12878 and K562 cells. The condition to form a spatial cluster is (1) at least 60% of the chains will have the specific $i-j$ interaction and (2) each i will have at least 7 spatial neighbors. The clustered primer sites are colored in red and blue for the GM12878 and K562 cells, respectively. Solid triangles are core sites satisfying both conditions, the colored circles are boundary sites that are not in the core but are directly connected to the core sites. One tight cluster exists for both the GM12878 and K562 cells. The primer sites 3 and 13 are the core primer sites surrounded by boundary sites 7, 8, 9, 10, 12, 14, 17, 18, 21 and 35 in the GM12878 cell. In the K562 cell, the only core primer site 13 is spatially surrounded by boundary sites 9, 10, 17, 18, 21, 34, 35, and 36.

(B). Model conformation of the α-globin gene domain in the GM12878 and K562 cells. We constructed a three dimensional model for the α-globin gene domain based on the averaged pairwise distances in predicted conformations, and their spatial clustering. We use one continuous tube chain to represent the 30 nm chromatin fiber of the α-globin gene domain, circles where on the tube are clustered primer sites. It is obvious that the three-dimensional conformation is more extended in the K562 cell than in the GM12878 cell.

matrices and the circos diagrams of the interactions in the model chains (Figure 10B-C) also suggest a two domain structure for K562 cell and one single domain structure for GM12878 cell. This is consistent with the circos diagrams of the significant 5C interactions from Figure 7C. We also predicted 340 and 346 new interactions for GM12878 and K562 cells, respectively.

3.6.5 Validation by ChIA-PET data

We checked the results from an independent method, ChIA-PET (Li et al., 2010), to validate a particular aspect of our 3D models for α-globin gene domain. According to ChIA-PET study on K562 cell, there are 13 CTCF-mediated interactions in the α-globin gene domain(Figure 11A). 10 among these 13 interactions exist in our 10,000 populations. Among these 10 correctly predicted CTCF-mediated interactions, only 2 of them exist in the 5C interaction matrix, 8 of them are our newly predicted interactions, which also shows the predictive power of our model. Only 3 ChIA-PET measured CTCF-mediated interactions are undetected. We further investigated why our model could not catch these 8 ChIA-PET interactions. Among them, the 5C frequency is 0 between primer 8 and 40, meaning there is no spatial proximity between these primers according to 5C study, which contradicts with ChIA-PET. There is no primer designed for site 11 in the 5C study, hence there is no spatial constraint on this particular site. Therefore it is not possible to detect an interaction between site 11 and 40. We also missed the interaction between site 9 and 40. However, according to Broad Institute CTCF enrichment data (Bernstein et al., 2005; Bernstein et al., 2006; Mikkelsen et al., 2007), there is no CTCF binding site on site 9. That's why, it is not clear whether site 9 can have significant CTCF-mediated interactions with any other site.

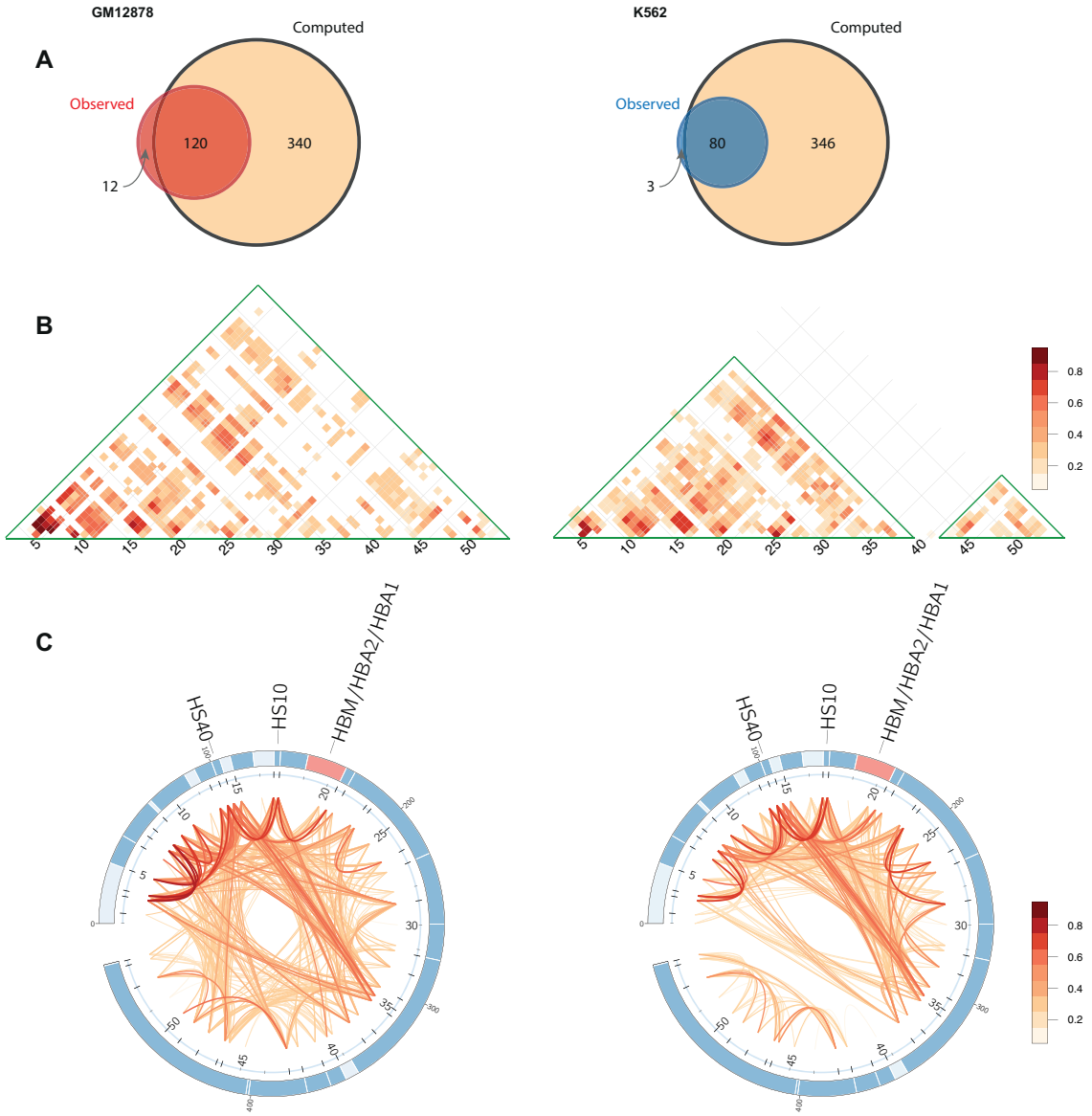
GM12878
K562
A
Computed
Observed
120
340
12
80
346
3
B
C
HS40
HS10
HBM/HBA2/HBA1

Figure 10

Figure 10 *(previous page)*: Spatial interactions in the α-globin gene domain including both directly measured 5C contacts and newly predicted contacts from spatial constraints of the C-SAC models.

(A) Significant interactions among the primer sites in the α-globin gene domain for both GM12878 and K562 cells. All interactions are calculated from the spatial distances between primer sites from weighted ensemble of 10,000 reconstructed structures of α-globin gene domain. There are a total of 460 significant interactions for the GM12878 cell, 120 (91%) of which exist in the 5C measurements, and 340 are newly predicted interactions. 12 (9%) interactions from the 5C measurements cannot be satisfied simultaneously with other 5C constraints. There are a total of 426 significant interactions for the K562 cell, 80 (96%) of which exist in the 5C measurements, 346 are newly predicted interactions. 3 (4%) interactions from the 5C measurement cannot be satisfied simultaneously with other 5C constraints.

(B) Heat map of significant interaction contributions among primer sites in the constructed α-globin gene domain for both the GM12878 and the K562 cells. The contribution of $i-j$ interactions in percentage is color coded. After non-specific interactions are removed based on our random C-SAC model, the formation of domains can be clearly seen. In the GM12878 cell, α-globin gene domain exist as a single domain, while in the K562 cell, it has two separate domains.

(C) Circos diagram to illustrate the interactions shown in the heat maps. Darker color in the curved links indicate interactions observed in majority of constructed 3D models. The percentage of contribution are color coded.

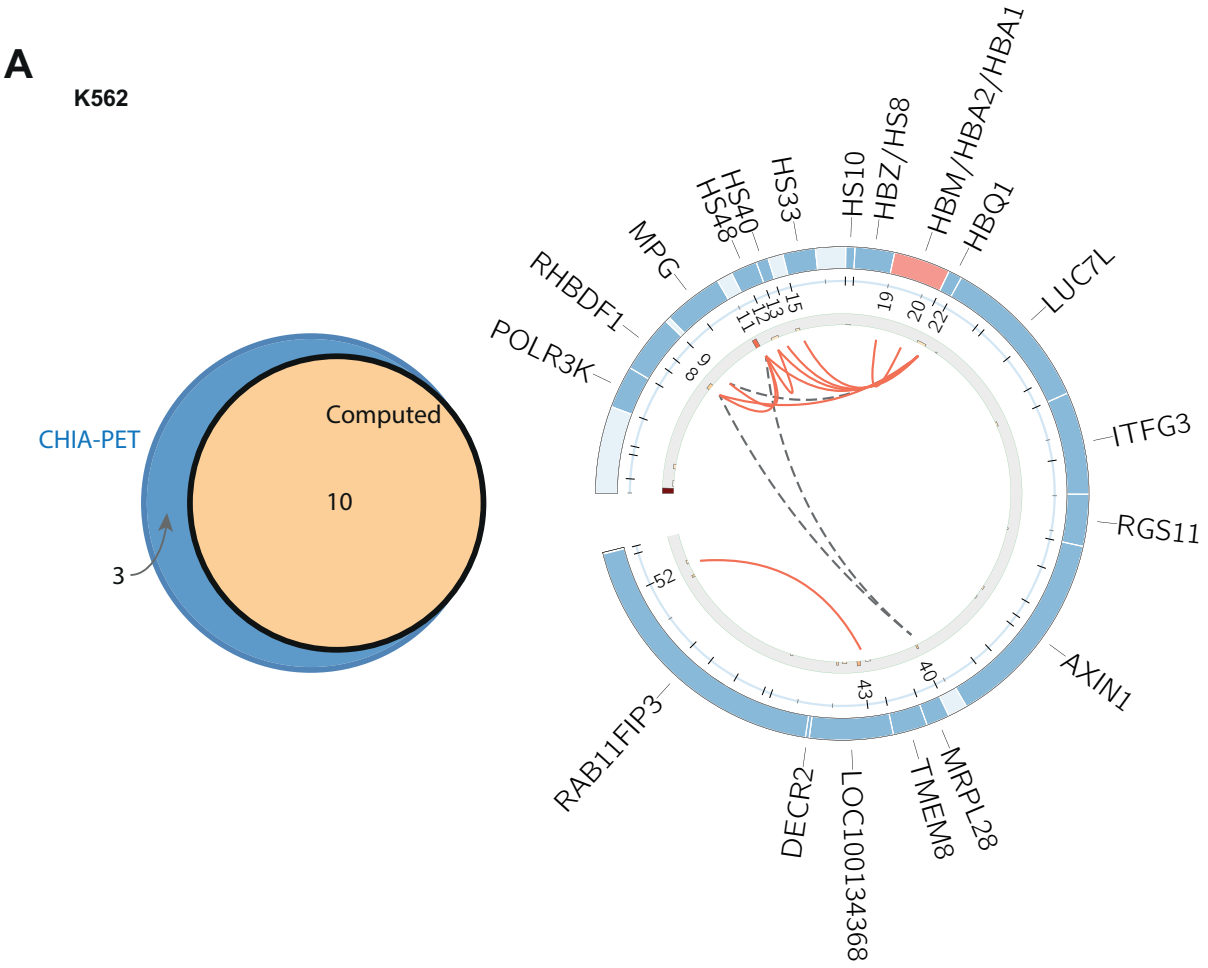
A
K562
CHIA-PET
Computed
10
3
HBM/HBA2/HBA1
HBZ/HS8
HS10
HS33
HS40
HS48
MPG
RHBDF1
POLR3K
HBQ1
LUC7L
ITFG3
RGS11
AXIN1
MRPL28
TMEM8
LOC100134368
DECR2
RAB11FIP3

Figure 11

Figure 11 *(previous page)*: Comparison between predicted CTCF mediated interactions based on 5C measurements and independently detected interactions based on ChIA-PET technique (Li et al., 2010). The Venn diagram on the left shows that the coverage of predicted interactions accounting for 10 of the 13 ChIA-PET measured CTCF interactions. The circos diagram shows the CTCF interactions detected by ChIA-PET technique in the α-globin gene domain of the K562 cell, the inner grey ring is based on the Broad institute CTCF enrichment data. The red solid curves between primer sites are ChIA-PET detected CTCF mediated interactions. The grey dashed curves are miss-predicted interactions.

Our predicted significant interactions recover 10 (77%) of ChIA-PET detected CTCF mediated interactions. Only 2 of them are also observed in the 5C measurements. That is, 8 interactions confirmed by ChIA-PET and are predicted from our constructed three-dimensional C-SAC model. Only 3 measured CTCF interactions are undetected. Among them, the 5C frequency is 0 between primer site 8 and 40, meaning there should be no spatial interactions between them according to 5C study, which contradicts with ChIA-PET data. There is no primer in the 5C study for site 11, hence there is no spatial constraint on this primer site from 5C measurements. Therefore it is not possible to detect interaction between 11 and 40. For primer site 9, the Broad institute CTCF enrichment data shows that there is no CTCF site on this primer site. Therefore it is not clear whether site 9 can have significant CTCF mediated interactions.

3.6.6 Proposed mechanism for the activation of α-globin gene domain

The major difference between the cell lines GM12878 and K562 is the expression level of the α-globin gene (Baù et al., 2010). K562 cell express high level of α-globin gene compared to GM12878 cells. Previous 3C studies suggested that the differential activation of the gene is regulated by the spatial proximity between the α-globin gene and its upstream functional elements HS48, HS46 and HS40 (Baù et al., 2010). After obtaining a population of 10,000 weighted model structures, we calculated how the primers are clustered in 3D space based on their spatial distances for both GM12878 and K562 cells (Figure 9). The corresponding spatial clusters are very similar in both cell lines. However, there exist extra primer sites in the spatial cluster of GM12878. The main difference between the clusters of two cell lines is the spatial location of primer site 3. In GM12878 cells, primer site 3 is in the core of the spatial cluster along with the primer site 13 (HS40/48), while it is far away from primer site 13 in K562 cells.

We further propose a mechanism for the differential activation of α-globin gene domain in different cell lines according to spatial distances between genes and regulatory elements from the population of model structures.

In GM12878 cells, interaction between primer site 3 and 13 (HS40/48) competes with the interaction between primer site 13 (HS40/48) and 21 (α-globin gene). In contrast, in K562 cells, primer sites 13 and 21 are in direct interaction without competing with any other interactions and primer site 3 is far away from 13. This agrees with Vernimmen et al. observation in which kicking HS40 element will dramatically decrease the active α-globin gene expression level (Vernimmen et al., 2009). This could mean, in the compact GM12878 cell, HS40 (primer site 13) tightly interacts with the primer site 3 and other sites, so that it is not easily accessed by other proteins and cannot be expressed easily. For extended K562 cell conformation,

HS40 (primer site 13) is easily accessed by other proteins, which may express the α-globin gene in a high level. Another main difference is that the α-globin gene domain can be represented by a main structure in K562 cells, while its conformation is fluctuating in GM12878 cells as was previously described clustering analysis. Diverse conformational space in GM12878 model chains and a unique conformation for K562 cell could reflect the difference between the activation mechanism of α-globin gene. Since the active K562 cell is in a unique conformational state, it might be easier for the cell to maintain the necessary long-range interactions for the gene activation. The fluctuating nature of GM12878 may result in instability to hold the interactions between the α-globin gene and its distant regulatory elements.

3.7 Discussion

Some research groups have proposed theoretical polymer models to calculate the scaling property of chromatin folding. Bohn et al. developed a loop model for random attraction with random probability and explained leveling off scaling property at large genomic distance (Bohn et al., 2007). Lieberman-Aiden et al. developed a fractal model to show that chromatin structure follow a scale property without knots (Lieberman-Aiden et al., 2009). Barbieri et al. developed a strings and binders switch model which can recapture the scaling property of chromatin folding (Barbieri et al., 2012). Since the adventure of high throughput 3C-assay approach which can detect the frequency of the spatial proximity of loci on the chromatin, various chromatin models were proposed to study the chromatin structure. Most research groups try to inverse the frequency of loci spatial proximity to distance, and apply the Monte Carlo approach to study chromatin structure from small domain scale (Baù et al., 2010) to genome wide scale (Duan et al., 2010; Tanizawa et al., 2010). For the theoretical polymer model, they only discuss the general scaling property of chromatin folding, and they did not consider the scaling property under confinement

volume. For other models using 3C-assay frequency matrix, all studies have ignored the random non-physical interactions under the confined volume, although some research groups have tried to correction the bias of the 3C-assay experiments (Yaffe and Tanay, 2011; Hu et al., 2012). Our C-SAC model considers these important factors, randomly generates conformations in certain confined volume, then we can remove the random non-physical interaction to determine real spatial proximity interactions.

Baú et al. also proposed one model to study α-globin gene domain for both GM12878 and K562 cells (Baù et al., 2010). Their model proportions the primer site lengths to the different size of ball, balls have spring interaction if the 5C frequency between the corresponding primer segments, then apply Monte Carlo simulation to get clustered conformations. Baú's model may not catch up the right interactions between longer segments and smaller segments and may not be efficient. Our C-SAC model not only removes the non-physical interactions but also can quickly generate more accurate conformations so that we can study the mechanism of gene regulation through the generated structures.

Vernimmen et al. found that α-globin has a strong correlation with upstream segment HS40 by deleting and ectopically reinserting HS40 element in the α-globin gene domain (Vernimmen et al., 2009). Our C-SAC growing model also reveals this relation between HS40 and α-globin gene. Comparing with Baú's model, we not only propose α-globin gene domain in the K562 cell structure, but also in the GM12878 cell structure and we propose a potential mechanism for α-globin gene express is active in the K562 cell and silence in the GM12878 cell.

Our C-SAC model is robust and efficient to rebuild chromatin structures. The C-SAC random model may properly generate random conformation under the confined volume so that the real spatial proximity will stand out. The C-SAC growing model can efficiently and quickly produce the chromatin conformation

following to the 3C-assay experiment, and discover new spatial proximity which 3C-assay can not report. The C-SAC model is also flexibility as long as we can put more constraints (CTCF frequency matrix, CHIA-PET frequency matrix, etc.) on the guidance function, then we may get more accurate picture of the target chromatin structures. The C-SAC is also dependent on the chromatin physical properties: persistence length, mass density, confined volume and predefined connection distance threshold. The C-SAC model gave a way to enlighten us on the mechanism of gene regulation.

The C-SAC model cannot only apply to small chromatin domains, but also on genome wide scale due to its efficiency. In the future, we plan to apply our C-SAC model on the multiple chromosome chains to study genome wide properties.

3.8 Conclusion

The 5C experiments for capturing the pairwise DNA interactions reveals chromosomal architecture and gene regulation in a detailed way. The correct biological interpretation of the outcome of the experiments relies on extensive three-dimensional modeling of the chromatin as the experiments are limited to restriction enzyme sites. We have presented an algorithm for the analysis of the 5C interaction frequencies and reconstruction of the full three-dimensional structure of the chromatin and identified main chromosomal players that are involved in gene activation. Analysis of the 5C interaction frequencies that were corrected by our structural random model demonstrates how to eliminate the background noise and provides reproducible insights into chromosomal architecture.

We showed that given the correct experimental constraints, our algorithm is capable of fully satisfying the experimental data and the differential activation of α-globin gene in different cell lines can be captured by structural details obtained from this model. Using this approach, we were able to predict the existing

CTCF-mediated interactions that were not captured by 5C (Baù et al., 2010), but validated by an independent study ChIA-PET (Li et al., 2010). The data also showed that all the known long-range interactions by 5C can be represented in our ensemble model chromatin chains. We also applied global clustering for the analysis of the reconstructed chromatin structures to decipher the structural differences between chromosomal architecture of active and inactive cell lines. This approach demonstrated one cluster representing a static conformational state of chromosomal architecture in active cell lines, and a fluctuating conformational landscape of chromosomal architecture in inactive cell lines. The more detailed, local clustering of the primer locations according to their spatial positioning also revealed a structural difference between cell lines, where the accessibility of regulatory elements were blocked by parts of the chromatin in inactive cell lines.

The approach we propose here is general and can facilitate the determination of chromatin structure by use of any type of 3C-based data, aiming at the prediction of chromosomal structures at a genome level. Although our approach can generate a large ensemble of chromatin chains that satisfy almost all experimental constraints, there still exists uncertainty in the physical parameters used in the current C-SAC model, including persistence length, chromatin fiber diameter, and mass density. These issues will likely be resolved when chromosomal properties are better understood and the C-SAC algorithm is further developed.

APPENDICES

Open-Access License

No Permission Required

PLOS applies the Creative Commons Attribution (CC BY) license to all works we publish (read the human-readable summary or the full license legal code). Under the CC BY license, authors retain ownership of the copyright for their article, but authors allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in PLOS journals, so long as the original authors and source are cited. **No permission is required from the authors or the publishers.**

In most cases, appropriate attribution can be provided by simply citing the original article (e.g., Kaltenbach LS et al. (2007) Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration. *PLOS Genet* 3(5): e82. doi:10.1371/journal.pgen.0030082). If the item you plan to reuse is not part of a published article (e.g., a featured issue image), then please indicate the originator of the work, and the volume, issue, and date of the journal in which the item appeared. For any reuse or redistribution of a work, you must also make clear the license terms under which the work was published.

This broad license was developed to facilitate open access to, and free use of, original works of all types. Applying this standard license to your own work will ensure your right to make your work freely and openly available. Learn more about open access. For queries about the license, please contact us.

CITED LITERATURE

[Abou El Hassan and Bremner, 2009]Abou El Hassan, M. and Bremner, R.: A rapid simple approach to quantify chromosome conformation capture. Nucleic acids research, 37(5):e35, April 2009.

[Alberts et al. , 2007]Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P.: Molecular Biology of the Cell. Garland Science, 5 edition, November 2007.

[Anderson et al. , 1999]Anderson, E., Bai, Z., and Bischof, C.: LAPACK users' guide. Society for Industrial Mathematics, 1999.

[Avery et al. , 1944]Avery, O. T., Macleod, C. M., and McCarty, M.: Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. The Journal of experimental medicine, 79(2):137–158, feb 1944.

[Bantignies et al. , 2011]Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G.: Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell, 144(2):214–226, January 2011.

[Barbieri et al. , 2012]Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., and Nicodemi, M.: Complexity of chromatin folding is captured by the strings and binders switch model. Proceedings of the National Academy of Sciences of the United States of America, 109(40):16173–16178, October 2012.

[Bastolla et al. , 2001]Bastolla, U., Farwer, J., Knapp, E. W., and Vendruscolo, M.: How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins, 44(2):79–96, aug 2001.

[Bastolla et al. , 2000]Bastolla, U., Vendruscolo, M., and Knapp, E. W.: A statistical mechanical method to optimize energy functions for protein folding. Proceedings of the National Academy of Sciences of the United States of America, 97(8):3977–3981, apr 2000.

[Baù et al. , 2010]Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J., and Marti-Renom, M. A.: The three-dimensional folding of the -globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology, 18(1):107–114, December 2010.

[Becker et al. , 2006]Becker, O. M., Dhanoa, D. S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., Nudelman, R., Kauffman, M., and Noiman, S.: An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. Journal of medicinal chemistry, 49(11):3116–3135, June 2006.

[Bernstein et al. , 2005]Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., and Lander, E. S.: Genomic maps and comparative analysis of histone modifications in human and mouse. Cell, 120(2):169–181, January 2005.

[Bernstein et al. , 2006]Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S.: A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell, 125(2):315–326, April 2006.

[Bohn and Heermann, 2010]Bohn, M. and Heermann, D. W.: Diffusion-driven looping provides a consistent framework for chromatin organization. PloS one, 5(8):e12218, 2010.

[Bohn et al. , 2007]Bohn, M., Heermann, D. W., and van Driel, R.: Random loop model for long polymers. Physical review E, Statistical, nonlinear, and soft matter physics, 76(5 Pt 1):051805, November 2007.

[Bolon and Mayo, 2001]Bolon, D. N. and Mayo, S. L.: Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences of the United States of America, 98(25):14274–14279, dec 2001.

[Burges, 1998]Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data mining and knowledge discovery, 2(2):121–167, 1998.

[Bystricky et al. , 2004]Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M.: Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proceedings of the National Academy of Sciences of the United States of America, 101(47):16495–16500, November 2004.

[Crick, 1958]Crick, F. H.: On protein synthesis. Symposia of the Society for Experimental Biology, 12:138–163, 1958.

[Crick, 1970]Crick, F.: Central Dogma of Molecular Biology. Nature, 227(5258):561–563, August 1970.

[Dahiyat and Mayo, 1997]Dahiyat, B. I. and Mayo, S. L.: De novo protein design: fully automated sequence selection. Science (New York, NY), 278(5335):82–87, oct 1997.

[Dekker, 2008]Dekker, J.: Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. The Journal of biological chemistry, 283(50):34532–34540, December 2008.

[Dekker et al. , 2002]Dekker, J., Rippe, K., Dekker, M., and Kleckner, N.: Capturing chromosome conformation. Science (New York, NY), 295(5558):1306–1311, February 2002.

[Desmet et al. , 1992]Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I.: The dead-end elimination theorem and its use in protein side-chain positioning. Nature, 356(6369):539–542, apr 1992.

[Deutsch and Kurosky, 1996]Deutsch, J. and Kurosky, T.: New Algorithm for Protein Design. Physical review letters, 76(2):323–326, jan 1996.

[Dostie et al. , 2006]Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J.: Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research, 16(10):1299–1309, October 2006.

[Drexler, 1981]Drexler, K. E.: Molecular engineering: An approach to the development of general capabilities for molecular manipulation. Proceedings of the National Academy of Sciences of the United States of America, 78(9):5275–5278, sep 1981.

[Duan et al. , 2010]Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S.: A three-dimensional model of the yeast genome. Nature, 465(7296):363–367, 2010.

[Edelsbrunner, 1987]Edelsbrunner, H.: Algorithms in combinatorial geometry. Springer Verlag, 1987.

[Edelsbrunner, 1993]Edelsbrunner, H.: The union of balls and its dual shape. In Proceedings of the ninth annual symposium on Computational geometry, pages 218–231, New York, NY, USA, 1993.

[Ester et al. , 1996]Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, pages 226–231. Kdd, 1996.

[Fearnhead and Clifford, 2003]Fearnhead, P. and Clifford, P.: On-line inference for hidden Markov models via particle filters. Journal of the Royal Statistical Society Series B-Statistical Methodology, 65(4):887–899, November 2003.

[Fraser et al. , 2009]Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., and Dostie, J.: Chromatin conformation signatures of cellular differentiation. Genome biology, 10(4):R37, 2009.

[Fraser and Bickmore, 2007]Fraser, P. and Bickmore, W.: Nuclear organization of the genome and the potential for gene regulation. Nature, 447(7143):413–417, May 2007.

[Fullwood et al. , 2009]Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E., and Ruan, Y.: An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462(7269):58–64, November 2009.

[Fung and Mangasarian, 2003]Fung, G. and Mangasarian, O.: Finite Newton method for Lagrangian support vector machine classification. Neurocomputing, 55:39–55, 2003.

[Futreal et al. , 2004]Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R.: A census of human cancer genes. Nature Reviews Cancer, 4(3):177–183, March 2004.

[Gibcus and Dekker, 2013]Gibcus, J. H. and Dekker, J.: The hierarchy of the 3D genome. Molecular cell, 49(5):773–782, March 2013.

[Gürsoy et al. , 2014]Gürsoy, G., Xu, Y., Kenter, A. L., and Liang, J.: Spatial Confinement is a major Determinant of Folding Landscape of Human Chromosomes and Genetic Programming of Cell. manuscript, 2014.

[Hagège et al. , 2007]Hagège, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forné, T.: Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nature protocols, 2(7):1722–1733, 2007.

[Hawking, 1996]Hawking, S. W.: The Illustrated a Brief History of Time. Random House Digital, Inc., 1996.

[Heermann et al. , 2012]Heermann, D. W., Jerabek, H., Liu, L., and Li, Y.: A model for the 3D chromatin architecture of pro and eukaryotes. Methods (San Diego, Calif.), 58(3):307–314, November 2012.

[Hershey and Chase, 1952]Hershey, A. D. and Chase, M.: INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. The Journal of general physiology, 36(1):39–56, September 1952.

[Hill et al. , 2000]Hill, R. B., Raleigh, D. P., Lombardi, A., and DeGrado, W. F.: De novo design of helical bundles as models for understanding protein folding and function. Accounts of chemical research, 33(11):745–754, nov 2000.

[Hochberg, 1988]Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75(4):800–802, 1988.

[Hoffman et al. , 2012]Hoffman, R., Benz, Jr, E. J., Silberstein, L. E., Heslop, H., Weitz, J., and Anastasi, J.: Hematology. Basic Principles and Practice, Expert Consult Premium Edition - Enhanced Online Features. Elsevier Health Sciences, November 2012.

[Horike et al. , 2005]Horike, S.-i., Cai, S., Miyano, M., Cheng, J.-F., and Kohwi-Shigematsu, T.: Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nature genetics, 37(1):31–40, January 2005.

[Hu et al. , 2004]Hu, C., Li, X., and Liang, J.: Developing optimal non-linear scoring function for protein design. Bioinformatics (Oxford, England), 20(17):3080–3098, November 2004.

[Hu et al. , 2012]Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S.: HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics (Oxford, England), 28(23):3131–3133, December 2012.

[Huang et al. , 2010]Huang, S.-Y., Grinter, S. Z., and Zou, X.: Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Physical chemistry chemical physics : PCCP, 12(40):12899–12908, oct 2010.

[Huang and Zou, 2006]Huang, S.-Y. and Zou, X.: An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. Journal of computational chemistry, 27(15):1866–1875, nov 2006.

[Huang and Zou, 2008]Huang, S.-Y. and Zou, X.: An iterative knowledge-based scoring function for protein-protein recognition. Proteins, 72(2):557–579, aug 2008.

[International Human Genome Sequencing Consortium, 2004]International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature, 431(7011):931–945, October 2004.

[Jacak et al. , 2012]Jacak, R., Leaver-Fay, A., and Kuhlman, B.: Computational protein design with explicit consideration of surface hydrophobic patches. Proteins, 80(3):825–838, March 2012.

[Jagielska et al. , 2008]Jagielska, A., Wroblewska, L., and Skolnick, J.: Protein model refinement using an optimized physics-based all-atom force field. Proceedings of the National Academy of Sciences of the United States of America, 105(24):8268–8273, June 2008.

[Jhunjhunwala et al. , 2008]Jhunjhunwala, S., van Zelm, M. C., Peak, M. M., Cutchin, S., Riblet, R., van Dongen, J. J. M., Grosveld, F. G., Knoch, T. A., and Murre, C.: The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. Cell, 133(2):265–279, April 2008.

[Jiang et al. , 2008]Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D.: De novo computational design of retro-aldol enzymes. Science (New York, NY), 319(5868):1387–1391, mar 2008.

[Joachimiak et al. , 2006]Joachimiak, L. A., Kortemme, T., Stoddard, B. L., and Baker, D.: Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. Journal of molecular biology, 361(1):195–208, aug 2006.

[Kalhor et al. , 2012]Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L.: Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature biotechnology, 30(1):90–98, January 2012.

[Kleinberg, 1999]Kleinberg, J. M.: Efficient algorithms for protein sequence design and the analysis of certain evolutionary fitness landscapes. Journal of computational biology : a journal of computational molecular cell biology, 6(3-4):387–404, 1999.

[Langowski and Heermann, 2007]Langowski, J. and Heermann, D. W.: Computational modeling of the chromatin fiber. Seminars in cell & developmental biology, 18(5):659–667, October 2007.

[Lazar et al. , 2006]Lazar, G. A., Dang, W., Karki, S., Vafa, O., Peng, J. S., Hyun, L., Chan, C., Chung, H. S., Eivazi, A., Yoder, S. C., Vielmetter, J., Carmichael, D. F., Hayes, R. J., and Dahiyat, B. I.: Engineered antibody Fc variants with enhanced effector function. Proceedings of the National Academy of Sciences of the United States of America, 103(11):4005–4010, mar 2006.

[Lee and Mangasarian, 2001]Lee, Y.-J. and Mangasarian, O. L.: RSVM: Reduced support vector machines. In Proceedings of the First SIAM International Conference on Data Mining, pages 1–17, 2001.

[Li et al. , 2010]Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H. S., Tennakoon, C., Wei, C.-L., Ruan, Y., and Sung, W.-K.: ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome biology, 11(2):R22, 2010.

[Li et al. , 1996]Li, H., Helling, R., Tang, C., and Wingreen, N.: Emergence of preferred structures in a simple model of protein folding. Science (New York, NY), 273(5275):666–669, aug 1996.

[Li et al. , 2003]Li, X., Hu, C., and Liang, J.: Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins, 53(4):792–805, December 2003.

[Li et al. , 2013]Li, Z., Yang, Y., Zhan, J., Dai, L., and Zhou, Y.: Energy functions in de novo protein design: current challenges and future prospects. Annual review of biophysics, 42:315–335, 2013.

[Liang et al. , 1998]Liang, J., Edelsbrunner, H., Fu, P., and Sudhakar, P.: Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. Proteins: Structure Function and Genetics, 1998.

[Liang et al. , 2002]Liang, J., Zhang, J., and Chen, R.: Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. The Journal of chemical physics, 117(7):3511, 2002.

[Liang and Grishin, 2004]Liang, S. and Grishin, N. V.: Effective scoring function for protein sequence design. Proteins, 54(2):271–281, February 2004.

[Lieberman-Aiden et al. , 2009]Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY), 326(5950):289–293, October 2009.

[Lin et al. , 2008]Lin, M., Lu, H.-M., Chen, R., and Liang, J.: Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. The Journal of chemical physics, 129(9):094101, September 2008.

[Ling et al. , 2006]Ling, J. Q., Li, T., Hu, J. F., Vu, T. H., Chen, H. L., Qiu, X. W., Cherry, A. M., and Hoffman, A. R.: CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. Science (New York, NY), 312(5771):269–272, April 2006.

[Liu, 2008]Liu, J. S.: Monte Carlo Strategies in Scientific Computing. Springer, January 2008.

[Liu and Chen, 1995]Liu, J. S. and Chen, R.: Blind Deconvolution via Sequential Imputations. Journal of the American Statistical Association, 90(430):567–576, June 1995.

[Lomvardas et al. , 2006]Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J., and Axel, R.: Interchromosomal interactions and olfactory receptor choice. Cell, 126(2):403–413, July 2006.

[Lu and Skolnick, 2001]Lu, H. and Skolnick, J.: A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins, 44(3):223–232, aug 2001.

[Maiorov and Crippen, 1992]Maiorov, V. N. and Crippen, G. M.: Contact potential that recognizes the correct folding of globular proteins. Journal of molecular biology, 227(3):876–888, oct 1992.

[Májek and Elber, 2009]Májek, P. and Elber, R.: A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. Proteins, 76(4):822–836, sep 2009.

[Mangasarian, 1994]Mangasarian, O.: Nonlinear programming. Society for Industrial Mathematics, 1994.

[Marshall, 1956]Marshall, A. W.: The use of multistage sampling schemes in Monte Carlo computations. In Symposium on Monte Carlo methods, University of Florida, 1954, pages 123–140. John Wiley and Sons, Inc., New York, 1956.

[Mikkelsen et al. , 2007]Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 448(7153):553–560, August 2007.

[Mintseris and Weng, 2003]Mintseris, J. and Weng, Z.: Atomic contact vectors in protein-protein recognition. Proteins, 53(3):629–639, November 2003.

[Miyazawa and Jernigan, 1985]Miyazawa, S. and Jernigan, R. L.: Estimation of Effective Interresidue Contact Energies From Protein Crystal-Structures - Quasi-Chemical Approximation. Macromolecules, 18(3):534–552, 1985.

[Miyazawa and Jernigan, 1996]Miyazawa, S. and Jernigan, R. L.: Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of molecular biology, 256(3):623–644, mar 1996.

[Naumova et al. , 2012]Naumova, N., Smith, E. M., Zhan, Y., and Dekker, J.: Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods (San Diego, Calif.), 58(3):192–203, November 2012.

[Nocedal and Wright, 1999]Nocedal, J. and Wright, S. J.: Numerical optimization. Springer Verlag, 1999.

[PABO, 1983]PABO, C.: Molecular technology: Designing proteins and peptides. Nature, 301(5897):200–200, 1983.

[Pokala and Handel, 2005]Pokala, N. and Handel, T. M.: Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. Journal of molecular biology, 347(1):203–227, March 2005.

[Qian et al. , 2007]Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J., and Baker, D.: High-resolution structure prediction and the crystallographic phase problem. Nature, 450(7167):259–264, November 2007.

[Ravikant and Elber, 2010]Ravikant, D. V. S. and Elber, R.: PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins, 78(2):400–419, feb 2010.

[Rippe, 2001]Rippe, K.: Making contacts on a nucleic acid polymer. Trends in biochemical sciences, 26(12):733–740, December 2001.

[Rosenbluth and Rosenbluth, 1955]Rosenbluth, M. N. and Rosenbluth, A. W.: Monte Carlo calculation of the average extension of molecular chains. The Journal of chemical physics, 23:356, 1955.

[Röthlisberger et al. , 2008]Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, y., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D.: Kemp elimination catalysts by computational enzyme design. Nature, 453(7192):190–195, may 2008.

[Rozenberg, 2002]Rozenberg, G.: Microscopic Haematology: A Practical Guide for the Laboratory, 2nd edition. CRC Press, 2 edition, December 2002.

[Sachs et al. , 1995]Sachs, R. K., van den Engh, G., Trask, B., Yokota, H., and Hearst, J. E.: A random-walk/giant-loop model for interphase chromosomes. Proceedings of the National Academy of Sciences of the United States of America, 92(7):2710–2714, March 1995.

[Schoenfelder et al. , 2010]Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J. A., Umlauf, D., Dimitrova, D. S., Eskiw, C. H., Luo, Y., Wei, C.-L., Ruan, Y., Bieker, J. J., and Fraser, P.: Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nature genetics, 42(1):53–61, January 2010.

[Schölkopf, 2002]Schölkopf, B.: Learning with kernels: support vector machines, regularization, optimization, and beyond. The MIT Press, may 2002.

[Shakhnovich and Gutin, 1993]Shakhnovich, E. I. and Gutin, A. M.: Engineering of stable and fast-folding sequences of model proteins. Proceedings of the National Academy of Sciences of the United States of America, 90(15):7195–7199, aug 1993.

[Shifman et al. , 2006]Shifman, J. M., Choi, M. H., Mihalas, S., Mayo, S. L., and Kennedy, M. B.: Ca2+/calmodulin-dependent protein kinase II (CaMKII) is activated by calmodulin with two bound calciums. Proceedings of the National Academy of Sciences of the United States of America, 103(38):13968–13973, sep 2006.

[Siegel et al. , 2010]Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D.: Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science (New York, NY), 329(5989):309–313, jul 2010.

[Simonis et al. , 2006]Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W.: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature genetics, 38(11):1348–1354, November 2006.

[Tanaka and Scheraga, 1976]Tanaka, S. and Scheraga, H. A.: Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules, 9(6):945–950, nov 1976.

[Tanizawa et al. , 2010]Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K.-i.: Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic acids research, 38(22):8164–8177, December 2010.

[Tiwari and Baylin, 2009]Tiwari, V. K. and Baylin, S. B.: Combined 3C-ChIP-cloning (6C) assay: a tool to unravel protein-mediated genome architecture. Cold Spring Harbor protocols, 2009(3):pdb.prot5168, March 2009.

[Tobi et al. , 2000]Tobi, D., Shafran, G., Linial, N., and Elber, R.: On the design and analysis of protein folding potentials. Proteins, 40(1):71–85, jul 2000.

[Umbarger et al. , 2011]Umbarger, M. A., Toro, E., Wright, M. A., Porreca, G. J., Baù, D., Hong, S.-H., Fero, M. J., Zhu, L. J., Marti-Renom, M. A., McAdams, H. H., Shapiro, L., Dekker, J., and Church, G. M.: The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. Molecular cell, 44(2):252–264, October 2011.

[van Heyningen and Hill, 2008]van Heyningen, V. and Hill, R. E.: Long-Range Control of Gene Expression. Academic Press, March 2008.

[van Holde and Zlatanova, 1996]van Holde, K. and Zlatanova, J.: What determines the folding of the chromatin fiber? Proceedings of the National Academy of Sciences of the United States of America, 93(20):10548–10555, oct 1996.

[van Holde and Zlatanova, 2007]van Holde, K. and Zlatanova, J.: Chromatin fiber structure: Where is the problem now? Seminars in cell & developmental biology, 18(5):651–658, oct 2007.

[Vapnik and Chervonenkis, 1974]Vapnik, V. N. and Chervonenkis, A. J.: Theory of Pattern Recognition. (1974), 1974.

[Vapnik, 1999]Vapnik, V.: The Nature of Statistical Learning Theory (Information Science and Statistics). Springer, 2nd edition, nov 1999.

[Vendruscolo et al. , 2000]Vendruscolo, M., Najmanovich, R., and Domany, E.: Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins, 38(2):134–148, feb 2000.

[Vendruscolo and Domany, 1998]Vendruscolo, M. and Domany, E.: Pairwise contact potentials are unsuitable for protein folding. The Journal of chemical physics, 109(24):11101, 1998.

[Vernimmen et al. , 2009]Vernimmen, D., Marques-Kranc, F., Sharpe, J. A., Sloane-Stanley, J. A., Wood, W. G., Wallace, H. A. C., Smith, A. J. H., and Higgs, D. R.: Chromosome looping at the human α-globin locus is mediated via the major upstream regulatory element (HS-40). Blood, 114(19):4253–4260, November 2009.

[Wagner et al. , 2004]Wagner, M., Meller, J. a., and Elber, R.: Large-scale linear programming techniques for the design of protein folding potentials. Mathematical programming, 101(2), jul 2004.

[Wang and Dunbrack, 2003]Wang, G. and Dunbrack, R. L.: PISCES: a protein sequence culling server. Bioinformatics (Oxford, England), 19(12):1589–1591, aug 2003.

[Watson and Crick, 1953]Watson, J. and Crick, F.: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171(4356):737–738, apr 1953.

[Wedemann and Langowski, 2002]Wedemann, G. and Langowski, J.: Computer simulation of the 30-nanometer chromatin fiber. Biophysical journal, 82(6):2847–2859, June 2002.

[Widom, 1989]Widom, J.: Toward a unified model of chromatin folding. Annual review of biophysics and biophysical chemistry, 18:365–395, 1989.

[Würtele and Chartrand, 2006]Würtele, H. and Chartrand, P.: Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 14(5):477–495, July 2006.

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

[Yershova et al. , 2010]Yershova, A., Jain, S., Lavalle, S. M., and Mitchell, J. C.: Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. The International Journal of Robotics Research, 29(7):801–812, June 2010.

[Yue and Dill, 1992]Yue, K. and Dill, K. A.: Inverse protein folding problem: designing polymer sequences. Proceedings of the National Academy of Sciences of the United States of America, 89(9):4163–4167, may 1992.

[Zhao et al. , 2006]Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R.: Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature genetics, 38(11):1341–1347, November 2006.

[Zhu et al. , 2008]Zhu, J., Fan, H., Periole, X., Honig, B., and Mark, A. E.: Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins, 72(4):1171–1188, September 2008.

VITA

Yun Xu

EDUCATION

Doctor of Philosophy in Bioinformatics Aug 2005 - May 2014

University of Illinois at Chicago

Master of Science in Mathematics Sep 2001 - Jun 2004

Fudan University

Bachelor of Science Sep 1991 - Jul 1995

Shanghai Fisheries University (Shanghai Ocean University)

EXPERIENCE

Research Projects 2005 - Present

Computational Biology Lab, University of Illinois at Chicago, Chicago IL, USA

- ◇ Formulated a simplified global nonlinear function for fitness landscape of protein design.
  - – Developed a global nonlinear fitness landscape via a finite Newton method and a rectangular kernel with a basis set of native proteins and decoys chosen a priori,
  - – Successfully predicted 95% native proteins from a large blind test set including 428 native proteins and 11 million decoys,
- ◇ Formulated a coiled-coil model for β-barrel membrane protein structure.
  - – Developed a coiled-coil model based on differential geometry for β-barrel membrane protein structure prediction,
  - – improved the accuracy of ab initio structure prediction from 44% to 73%, comparing to the 7% by random prediction
- ◇ Developed a sequential Monte Carlo sampling method for generating 3D chromatin fiber conformations for studying of long range gene regulation.

– Working on the development of sequential Monte Carlo sampling method for constructing 3D chromatin fiber conformations by statistical mechanical models, and to study the chromatin basis of long range gene regulation.

HONORS

Outstanding Graduate Thesis of Shanghai 2005

Shanghai, China

Excellent Graduate Student Jun 2004

Fudan Unviersity, Shanghai, China

New Century Xie Xide Scholarship (First Rank Graduate Scholarship) Oct 2003

Fudan Unviersity, Shanghai, China

PUBLICATIONS

[1] Yun Xu, Gamze Gürsoy, Amy L Kenter, and Jie Liang. Chromatin structure modelling on α-globin gene domain. *Manuscript*, 2014.

[2] Yun Xu, Changyu Hu, Yang Dai, and Jie Liang. On simplified global nonlinear function for fitness landscape: A case study of inverse protein folding. *Manuscript*, 2014.

[3] Gamze Güsoy, Yun Xu, Amy L Kenter, and Jie Liang. Spatial confinment is a major determinant of folding landscape of human chromosomes. *Manuscript*, 2014.

[4] Hammad Naveed, Yun Xu, Ronald Jackups, and Jie Liang. Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. *Journal of the American Chemical Society*, 134(3):1775–1781, January 2012.

[5] Fu-Yao Ren, Yun Xu, Wei-Yuan Qiu, and Jin-Rong Liang. Universality of stretched Gaussian asymptotic diffusion behavior on biased heterogeneous fractal structure in external force fields. *Chaos, Solitons & Fractals*, 24(1):273–278, 2005.

[6] Yun Xu, Fu-Yao Ren, Jin-Rong Liang, and Wei-Yuan Qiu. Stretched Gaussian asymptotic behavior for fractional Fokker–Planck equation on fractal structure in external force fields. *Chaos*, 20:581, May 2004.

[7] Wei-Yuan Qiu, Fu-Yao Ren, Yun Xu, and Jin-Rong Liang. Stretched Gaussian Asymptotic Behavior for Fractional Giona–Roman Equation on Biased Heterogeneous Fractal Structure in External Force Fields. *Nonlinear Dynamics*, 38(1-4):285–294, 2004.

[8] Fu-Yao Ren, Wei-Yuan Qiu, Yun Xu, and Jin-Rong Liang. Answer to an open problem proposed by E Barkai and J Klafter. *Journal of Physics. A. Mathematical and General*, 37(42):9919, 2004.

[9] Fu-Yao Ren, Jin-Rong Liang, Wei-Yuan Qiu, and Yun Xu. Universality of stretched Gaussian asymptotic behaviour for the fractional Fokker–Planck equation in external force fields. *Journal of Physics. A. Mathematical and General*, 36(27):7533, 2003.

[10] Fu-Yao Ren, Jin-Rong Liang, Wei-Yuan Qiu, and Yun Xu. Fractional Fokker–Planck equation on heterogeneous fractal structures in external force fields and its solutions. *Physica A: Statistical Mechanics and its Applications*, 326(3):430–440, 2003.

[11] Fu-Yao Ren, Jin-Rong Liang, Wei-Yuan Qiu, Xiao-Tian Wang, Yun Xu, and R R Nigmatullin. An anomalous diffusion model in an external force fields on fractals. *Physics Letters A*, 312(3):187–197, 2003.