Statistical Physics and Biological Information: Hydrophobicity Patterns in Protein Design and Differential Motif Finding in DNA

by

Mehdi Yahyanejad

B.Sc. Sharif University of Technology

M.Sc. University of Toronto

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Physics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2004]
August 2004

Author .....
Department of Physics
August 15, 2004

Certified by .....
Christopher B. Burge
Associate Professor
Thesis Supervisor

Certified by .....
Mehran Kardar
Professor
Thesis Supervisor

Accepted by .....
Thomas J. Greytak
Associate Department Head for Education

Statistical Physics and Biological Information: Hydrophobicity Patterns in Protein Design and Differential Motif Finding in DNA

by

Mehdi Yahyanejad

Submitted to the Department of Physics
on August 15, 2004, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

Abstract

In the past decade, a large amount of biological data has been generated, enabling new quantitative approaches in biology. In this thesis, we focus on two biological questions by using techniques from statistical physics: hydrophobicity patterns in proteins and their impact on the designability of protein structures and regulatory motif finding in DNA sequences.

Proteins fold into specific structures to perform their functions. Hydrophobicity is the main force of folding; protein sequences try to lower the ground state energy of the folded structure by burying hydrophobic monomers in the core. This results in patterns, or correlations, in the hydrophobic profiles of proteins. In this thesis, we study the designability phenomena: the vast majority of proteins adopt only a small number of distinct folded structures. In Chapter 2, we use principal component analysis to characterize the distribution of solvent accessibility profiles in an appropriate high-dimensional vector space and show that the distribution can be approximated with a Gaussian form. We also show that structures with solvent accessibility profiles dissimilar to the rest are more likely to be highly designable, offering an alternative to existing, computationally-intensive methods for identifying highly-designable structures. In Chapter 3, we extend our method to natural proteins. We use Fourier analysis to study the solvent accessibility and hydrophobicity profiles of natural proteins and show that their distribution can be approximated by a multi-variate Gaussian. The method allows us to separate the intrinsic tendencies of sequence and structure profiles from the interactions that correlate them; we conclude that the alpha-helix periodicity in sequence hydrophobicity is dictated by the solvent accessibility of structures. The distinct intrinsic tendencies of sequence and structure profiles are most pronounced at long periods, where sequence hydrophobicity fluctuates less, while solvent accessibility fluctuates more than average. Correlations

between the two profiles can be interpreted as the Boltzmann weight of the solvation energy at room temperature. Chapter 4 shows that correlations in solvent accessibility along protein structures play a key role in the designability phenomenon, for both lattice and natural proteins. Without such correlations, as predicted by the Random Energy Model (REM), all structures will have almost equal values of designability. By using a toy, Ising-based model, we show that changing the correlations moves between a regime with no designability and a regime exhibiting the designability phenomenon, where a few highly designable structures emerge.

Understanding how gene expression is regulated is one of the main goals of molecular cell biology. To reach this goal, the recognition and identification of DNA motifs -- short patterns in biological sequences—is essential. Common examples of motifs include transcription factor binding sites in promoter regions of co-regulated genes and exonic and intronic splicing enhancers. Most motif finder algorithms try to find a functionally relevant (specific) motif in a set of sequences that share a functional property by simply looking for over-represented patterns. They are liable to be misled by other, functionally irrelevant (non-specific) patterns that are over-represented across the genome. To overcome this problem, a "negative" set can be used that is not likely to include the functional motif but may have non-specific patterns. In Chapter 5, We develop an analytical framework for differential motif finding which expands the classical Gibbs motif finder. Both the cases of one and multiple motif occurrences per sequence are developed. In our method, motifs that have strong matches in the negative sequence set are suppressed. As a result, motifs that are differentially enriched in the "positive set" as compared to the "negative set" are found. We show that our method outperforms the classical Gibbs sampler in finding differentially-enriched motifs.

Thesis Supervisor: Christopher B. Burge
Title: Associate Professor

Thesis Supervisor: Mehran Kardar
Title: Professor

Acknowledgments

I would like to thank my advisors, Professor Chris B. Burge and Professor Mehran Kardar, for their continues support and guidance. Chris's support was critical for my entry to biological sciences, and he allowed me to freely explore my new interests in this field. Mehran also has been supportive of me in exploring new opportunities and in giving me advice ranging from physics problems to career paths.

I would also like to thank all my colleagues who have contributed to this thesis either directly or through enlightening discussions:
Dr. Chao Tang (NEC Research Institute),
Dr. Reza Ejtehadi (University of British Columbia),
Professor Ned Wingreen (NEC Research Institute, now at Princeton University),
Professor Leonid Mirny,
(Late) Professor Toyoichi Tanaka,
Dr. Victor Spirin,
Dr. Eldon Emberly, (Rockefeller University)
Gene Yeo,
Uwe Ohler,
Brad Friedman,
and other members of the Burge lab.

Through the years of my PhD, I enjoyed living in Boston. This city has been a source of inspiration for me. Many of my opinions were shaped in this city from what I learned here. This city was also a place where I made many friends who were a source of inspirational conversations as well as entertainment. Among the ones who lived in Boston, I would like to mention: Peyman K., Behrang N., Casey H., Hazhir R., Basak B., Payman K., Gabor C., Parisa F., Michael O., Yoav B., Julia S., Brice S., Ali T., Ali N., Azadeh S., Jasmine C., Ali M., Adel A., Roya B., Fardad H., Chris J., Tom W., Tanya L., Farid G., Salma K., Navid S., Anya O., Selis O., Berker E., Farzan P., Cansu T., Mohammad M., Sohil P., Mammad H., Reza S., and Maryam M. Also, I would like to thank Michelle P. for her continous encouragement as well

as her editing of this manuscript. And great thanks to my family who tolerated my absence and lack of visits for these years in the hope of my success in career and life.

Financial support is acknowledged from the following sources:

Functional Genomics Innovation Award (C.B. Burge and P. Sharp)

MIT Teaching Assistantship

NSF grant no. DMR-940034 (M. Kardar)

KITP Graduate Fellowship

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Abundant biological data has been generated and collected in recent years, including genome sequences for many different organisms, data from gene expression experiments, and the discovery and classification of new protein folds. This abundance of biological data is allowing new quantitative approaches for modeling biological mechanisms; techniques from physics are contributing to these approaches. In my thesis, I focus on the areas of protein folding and motif finding, where physically inspired methods can contribute.

The protein-folding problem has a long history in biological physics, and techniques of statistical physics have been applied to solve different aspects of the problem. I focus on the designability phenomenon: certain protein structures have a far greater number of amino acid sequences folding to them than the average. I explore how geometrical constraints resulting from folding create correlations in solvent accessibility along the protein chain, and how those correlations result in the emergence of highly designable structures. I also study how solvation forces influence the distribution of hydrophobic monomers along the protein chain and offer a method to untangle correlations between solvent accessibility and hydrophobicity to allow the study of the intrinsic correlations within each one.

A motif is a pattern appearing in a DNA sequence that corresponds to a specific function. Finding these motifs has become an important subject in recent years because of their impact on modeling gene regulatory networks. Statistical methods,

some inspired by physics, have been applied. These include Gibbs Sampling, which is inspired by the Metropolis algorithm, often used in statistical physics. However, current motif finders are prone to be misled by strong but functionally irrelevant motifs that overshadow weaker, relevant motifs. To make weaker motifs more pronounced, a second set of sequences-a negative set-can be used. The negative set is chosen not to contain the weak functional motif, though it can still contain the stronger, non-functional motif. By extending the Gibbs sampling algorithm, we offer a discriminative-motif-finding method in which relevant motifs are found by comparing the "positive" and "negative" sequence sets.

1.1 Protein Folding

The transcription of DNA to RNA, and from RNA to proteins constitutes the central dogma of molecular biology. In this picture, the DNA sequence, or genome, completely directs the actions in the cell through the encoding of proteins. Proteins then fold to three-dimensional structures to perform their functions. Proteins are involved in many functions such as enzymatic catalysis, transport and storage, signaling and recognition, and mechanical engines. Much genetic information is being generated by sequencing the genomes of different organisms, and this is resulting in the discovery of many novel genes with as yet unknown functions. To discover the function of these genes, we need to know the structure and function of the proteins to which they are translated.

1.1.1 How proteins are made

DNA sequences with four letter types, or nucleotides, are transcribed to mRNA, which is then linearly translated to proteins with 20 different types of residues. Each triplet of nucleotides on the DNA, called a codon, is mapped to one amino acid residue, or monomer, and is added to a growing polypeptide chain, forming a linear heteropolymer. This chain usually has a length between 30 and 450 residues. Amino acids share amino and carboxyl groups, which are joined by a carbon atom referred to

as C-alpha. The term "acid" is used because the carboxyl group (-COOH) gives up a proton to form the bond. Different side chains, which define the specific properties of the amino acids, are attached to the C-alpha atom. Neighboring amino acids are linked together by a covalent peptide bond between the nitrogen of the amino acid group and the carboxyl carbon atom.

1.1.2 Folding

Each polypeptide chain folds to a specific, fairly compact, three-dimensional structure, called its native state, to perform its function. The information within the sequence uniquely determines the folded structure [1]. It has been shown that the native structure is the global minimum of the free energy for the folding protein. For a sequence with a random composition, the energy landscape is very rough with many local minima, which prevents speedy folding to the global minimum. The fact that protein sequences find their unique ground state in a quick way suggests that evolution has selected the sequences that fold quickly to their unique ground states and have more stable structures.

A number of forces play a role in folding protein chains. The solvation force is considered to be the main force of folding. It results from different affinities of different amino acids toward the solvent. Amino acids can be divided into two major groups, hydrophobic and polar. Hydrophobic amino acids have greasy side chains, which dislike the solvent, while polar amino acids do not. Folding proteins tend to bury their hydrophobic monomers in their core to reduce their contact with the solvent. In the folded protein, there are also interactions between nearby amino acids, such as hydrogen bonding and Van der Waals interactions. Oppositely charged amino-acid side chains can also form salt bridges to further stabilize the protein fold. Cystein residues can also create disulphide bonds, which are useful for proteins that are too small to have a hydrophobic core or function at a high temperature and require a higher degree of stability.

1.1.3 Protein structure

Protein structures are studied using X-ray diffraction and NMR imaging methods. These studies have helped to identify the building blocks of proteins. The simplest components of proteins are known as secondary structures, which are short stretches of amino acids with specific three-dimensional characteristics. The commonly used secondary structures are alpha-helices and beta-sheets. An alpha helix is a right-handed polypeptide coil. The neighboring amino acid residues form hydrogen bonds, which stabilize the helix. A helix can have between 4 to 40 residues [45, 76], and the periodicity of the helix turn is about 3.6 residues. Beta-sheets, on the other hand, consist of pairs of chains (beta-strands) lying side-by-side and are stabilized by hydrogen bonds between neighboring atoms on adjacent chains. The amino acids alternate on either side of the beta-sheet. Other protein segments, such as loops, are sometimes characterized as secondary structures. Secondary structures come together to form what is called a tertiary structure. The tertiary structure, which refers to the three-dimensional structure, can include one or more domains. Each domain often performs a separate function for the protein, and usually has its own, separate, hydrophobic core.

1.1.4 Observed proteins and their classification

Many different protein structures share structural similarities. Proteins sharing the same fold have secondary structures in the same arrangement with the same topological connections. Belonging to the same fold can indicate a common evolutionary origin. It is important to classify protein folds to further understand the evolutionary and functional relationships between different proteins. One broad classification is into five groups of folds including proteins with mainly alpha helices, with mainly beta-sheets, with alternating alpha and beta, with alpha and beta separated along the chain, or with no secondary structures [53].

There are a number of finer structural classification methods, ranging from manual to fully automated methods. Structural classification methods begin by finding the

pair-wise structural alignment between two structures. The two structures are overlapped as best as possible, and the distance between similarly located residues along the backbone is measured. The coordinates of 3D protein structures are recorded in the PDB database [3], and classification databases include FSSP, CATH, and SCOP. FSSP (Families of Structurally Similar Proteins) is obtained by a fully automated method, using the DALI program to generate alignment of single protein chains [37]. CATH (Class Architecture, Topology, Homologous superfamily) is a semi-automated database for protein domains [71]. SCOP (Structural Classification of Proteins) is mostly constructed manually [67] and is considered one of the most accurate protein classifications.

1.1.5 Designability

The concept of designability is a result of the fact that many different sequences can share the same fold. In fact, the number of observed folds is somewhere between 1000 and 2000, depending on the classification method, while the number of observed sequences is much larger by up to three or four orders of magnitude. Among the observed folds, the number of sequences folding to each varies widely. Most sequences fold to a small subset of the observed folds, which are called highly designable folds [71, 37]. Understanding why some folds are highly designable is important for a number of reasons. Not only is it interesting to know why these folds have been selected by evolution, this understanding can also assist in the design of new stable folds. It can also be asked what the role of different forces, such as the solvation force, is in the emergence of highly designable structures. Another important question is how geometrical constraints affect folding and whether they contribute to making some structures more desirable for sequences to fold to.

1.1.6 Modeling Designability

To understand designability, lattice models are often used to make the problem computationally tractable. In lattice models, a simplified polymer with its residues lo-

cated on a grid represents the protein [13, 49, 50]. Interactions are simplified as well. The energy of the system is generally approximated to include only short-range, pair-wise interactions between different monomers. Because hydrophobicity is the main force of folding, the pair-wise interactions can further be simplified to include only hydrophobic forces [13, 17, 20]. In such a model, called the solvation model, there is an energetic advantage for putting hydrophobic residues in the core.

To model the designability problem using lattice proteins, the mapping of all feasible sequences to their corresponding folded structures needs to be determined. It has been observed that the ground state structures of most protein sequences are compact. The target structure space in simulations is thus chosen to be the space of compact structures. Often all the compact structures are generated first. Then the energies of folding to each of those structures are calculated for each sequence. The structure or structures with the lowest energy determine the ground state of the sequence. Sequences with multiple ground states are thrown out, because they are considered non-physical. In nature, proteins need to have a stable unique structure to perform their function. In most lattice simulations, only a small number of sequences have unique ground states. These are called viable sequences. The designability of a structure is defined as the number of viable sequences that have that structure as their unique ground state.

Using lattice protein models, the emergence of highly designable structures has been observed for 2D and 3D compact lattice protein structures. It has been shown that most lattice protein structures have a low value of designability, while only a small number are highly designable, in contrast to a null model in which the designability follows a Poisson distribution [31, 54, 30]. For example, Li et al [54] performed exhaustive calculations for compact structures on a 3x3x3 grid to find the ground state structure of 2^{27} feasible sequences. A small number of structures emerged as highly designable, while many others had either zero or one sequences folding to them. Other studies have explored the designability problem for longer, 3D lattice protein structures [11] and off-lattice proteins[63]. The dependence of the designability phenomenon on the form of the interactions has been investigated as well. Ejtehadi et al

showed that replacing the HP Hamiltonian, which has pair-wise interactions between hydrophbobic and polar amino acids, by a solvation Hamiltonian, which only includes interaction with the solvent, produces a similar designability distribution. The effect of the number of residue letter types on the designability histograms has also been considered. An earlier study suggested that a two-letter model can introduce artifacts which do not exist in models with higher numbers of amino-acid types [8, 7]. A recent study by Li [57] shows that using a 20-letter MJ interaction matrix generates results similar to the HP model for lattice protein structures.

Since solvation forces are the main forces behind folding, it is important to consider their impact on designability in detail. In a purely hydrophobic model, the Hamiltonian can be written as $H = \sum s_i h_i$, where s_i is the degree of solvent exposure of site i along the polypeptide, and h_i is the degree of hydrophobicity of the monomer in position i. In a two-letter model, each of these vectors would be a binary vector. Li et al showed that the solvation Hamiltonian can be written as the Hamming distance between the sequence and structure binary vectors [55]. A lower distance indicates a lower folding energy. In this simple picture, any binary vector can be a feasible sequence, while structures come from a limited number of available binary vectors in the space. A sequence will have a unique ground state if it has only one nearest neighbor structure. In this picture, the Voronoi volume around a structure is proportional to its designability. Li et al showed that highly designable structures typically have a lower number of structure neighbors in the structure space.

The designability of a structure is related to other physical features of the structure. A sequence folding to its ground state can be excited to a different structure, given an amount of energy greater than the "gap energy." The stability of a structure is characterized by the average gap over all sequences that fold to the structure. It has been shown that the designability of a structure correlates well with its degree of stability [61]. It has also been observed that highly designable structures tend to have more surface-to-core transitions [56, 17, 83].

1.1.7 Summary of the conducted research

In this thesis, we focus on how correlations along the polypeptide chain are responsible for creating highly designable structures. Such correlations exist in natural proteins as well. We investigate the correlations in the hydrophobicity profile and the solvent accessibility profile and the interaction between the two. The profiles are given by the values of hydrophobicity or solvent accessibility of each amino acid along the chain. As a result of folding, the geometry of the compact structure creates correlations in the solvent accessibility profile along the protein chain. These correlations are the key for generating the designability phenomenon. We demonstrate that if instead of 2D compact structures, we use binary vectors that have the same correlations along their chain, a designability distribution similar to the one observed in lattice proteins can be generated. As a result, correlations can be used to estimate designability without the need for extensive computational simulations. We also show that reducing the correlations in the binary vectors makes designability more uniform among the structures.

Modeling correlations along the chain can also be useful for understanding natural proteins. Fourier analysis of correlations is a suitable tool for analyzing natural proteins. While proteins have different lengths, Fourier transforms map profiles into an interval of $[0, \pi]$ in frequency space where they can be compared. It has previously been observed that there are correlations both in the hydrophobicity and solvent accessibility profiles of proteins. There have been several studies of one of the two profiles, independent of the other. We know, however, that the two profiles are correlated. The correlation between them exists because a folding protein tries to minimize its solvation energy, which is the inner product of the solvent accessibility profile and the hydrophobicity profile. In our work, we introduce a general framework to model not only the correlations within each profile, but also the correlations among them. We show that the power spectrum of these profiles can be approximated by a multivariate Gaussian distribution. This helps us to untangle the correlations of the two profiles. For example, the model indicates that alpha-helix periodicity

in the hydrophobicity profile is induced by solvent accessibility profile, rather than the reverse. The multivariate Gaussian should also help in predicting the solvent accessibility profile from the hydrophobicity profile, for example for protein sequences with unknown structures.

In the next few sections, we provide a detailed description of Chapters 2-4 of the thesis.

1.1.8 Chap. 2

In previous work, it was shown that protein structures can be mapped to a vector space [56, 8]. In this vector space, distance between structures can be defined as a measure of dissimilarity. It was shown that the protein structures with high designability are the ones that are far away from other structures. We take this idea further and characterize the distribution of the whole vector space of structures to find the regions with low density, in the hope that these regions contain the highly designable structures. The distribution of these vectors can be seen as a cloud of points in a many-dimensional space. After projecting the cloud of points along its eigenvectors, it can be seen that the cloud can be approximated by a Gaussian distribution along each eigenvector. This eigenvalue decomposition method allows us to approximate the whole distribution with a multivariate Gaussian distribution. Using the distribution, it is easy to find the areas of space with low density. Regions of space with a low density of structures are good targets for finding highly designable structures. Being in a low-density region is not a sufficient condition for having high designability, however. A nearby neighbor can prevent a structure from having a high designability by "stealing away" some of the sequences.

We quantify the degree to which structures in low-density regions have high designability. We calculate designability by running extensive simulations and then compare the set of highly designable structures emerging from simulations with the ones predicted to be in the low-density region. We find that 70 percent of predicted structures are indeed highly designable structures, indicating that such an estimator can avoid the need for computationally extensive simulations.

1.1.9 Chap. 3

Correlations along the protein chain have been the subject of many studies [40, 42, 88, 95, 46]. It has been shown that protein sequences can be differentiated from random sequences by the correlations within them. Correlations along the protein chain can be studied for various properties, including hydrophobicity, charge, mutation rate, and solvent accessibility. Correlation patterns can yield insight into the function and underlying interactions. For example, Eisenberg et al [15, 16] observed that alpha helices have a periodicity of 3.6 in their hydrophobicity profile, which is the same as their helical period. This demonstrated that alpha helices tend to have a hydrophobic moment, on one side hydrophobic and the other polar, which contributes to lowering the ground state energy of the folded protein by exposing one side of the helix to solvent while burying the other. Fourier analysis of hydrophobicity profiles also showed that there are long-range correlations along the chain [40]. Solvent accessibility profiles are less studied. Studies of solvent accessibility/surface exposure in lattice and off-lattice proteins indicate that there are long-range correlations along the polypeptide chain created as a result of folding [46]. In a folded protein, a monomer is more likely to be in the core if its neighbors are in the core, creating a positive correlation along the solvent accessibility profile.

Since the hydrophobicity and solvent accessibility profiles are not truly independent, a proper study of the correlation within each needs to incorporate the interactions between the two. We Fourier transform hydrophobicity and solvent accessibility profiles of 1461 protein structures that are selected from a representative set of protein chains in the FSSP database after the multi-domain chains are removed. By Fourier transforming, we are projecting the cloud with 1461 points onto different values of the periodicity, or wave vectors. At each periodicity q, there is a set of 1461 Fourier components of hydrophobicity and solvent accessibility profiles h_q, s_q. We show that these points can be approximated with a multivariate Gaussian distribution. In this Gaussian distribution, we include separate terms for variance in solvent accessibility and hydrophobicity as well as a term incorporating the interaction between the two.

By fitting the data to our proposed Gaussian, we are able to estimate both parameters that we call intrinsic correlations in solvent accessibility and hydrophobicity and the correlations among them. For example, it can be seen that the intrinsic hydrophobicity profile does not have any peak at the alpha-helix periodicity, while there is an alpha-helix peak in the intrinsic solvent accessibility. This indicates that alpha-helix periodicity is induced as a result of existing periodicity in the solvent accessibility. We also show that the interaction term between the solvent accessibility and hydrophobicity profile is similar to a Boltzmann factor with a temperature close to room temperature. Previous studies have argued that such an observation is the result of evolutionary events .

Modeling the distribution of solvent and hydrophobicity profiles by a Gaussian distribution can be helpful in predicting the solvent accessibility profile from the sequence. Solvent accessibility prediction is often one step in secondary structure prediction. Using the Gaussian distribution, we can obtain the conditional probability for the solvent accessibility profile given the hydrophobicity profile.

1.1.10 Chap. 4

In this Chapter, we focus on how changing the correlations along the protein can affect the emergence of designability. Kussel and Shakhnovich used an analytical method to demonstrate that when only additive forces are present, all structures should have the same designability provided that the assumptions of the Random Energy Model (REM) are upheld [48]. It has further been shown that additive forces and solvation forces can be easily mapped onto one another [20, 19]. Moreover, it has been observed that for 2D lattice proteins, the solvation model can still generate highly designable structures [56]. It has been suggested that REM might not be applicable to these 2D examples, because REM often does not perform well in 2D [75]. However, simulations on 4x3x3 lattice proteins have shown that in 3D, solvation forces can still create designability phenomenon [11].

In our work, we show that the emergence of the designability phenomenon can be explained in part by the breakdown of the REM assumption of statistical indepen-

dence between states. We generate sets of artificial structures with different degrees of correlation in the solvent accessibility profile. We then calculate the designability distribution for each of these sets. The artificial sets with a low value of correlations showed a very different distribution of designability from the sets with high value of correlations. For low correlations, designability is almost uniform for different structures, similar to what was predicted by REM. But for sets with higher correlations, the designability is highly non-uniform, and some structures have a high degree of designability. We show that as we increase the correlations within the structures, the correlation between the structures increases as well. This results in the breakdown of the energy independence assumption of REM, and as a result, its prediction.

This work shows that solvation forces can give rise to the designability phenomenon when the correlations along the structures are high enough. Moreover, it suggests that solvation forces alone can be sufficient to describe the emergence of highly designable protein structures in lattice models. Lastly, we explore the degree of correlations in natural proteins. From our analysis, it appears that correlations in solvent accessibility along the chain of natural proteins are strong enough to break the REM assumption when they are compared to the results from lattice proteins.

1.1.11 Future directions

Our work has explored the correlations in the solvent accessibility of lattice protein structures and their connection to designability. These methods can be extended to natural proteins. An interesting future direction is to use our method to estimate the designability of natural protein structures from their solvent accessibility profiles. The result of such an estimation could be compared with, for example, the number of superfamilies of sequences that fold to each structure.

Our method for untangling correlations between solvent accessibility and hydrophobicity could also be used to untangle correlations among other characteristics of a protein chain, such as conservation profile, charge profile, and monomer volume profile. For each protein, a conservation profile can be built by aligning homologous proteins and measuring the frequency of mutation at each site. It is likely that such

a profile is correlated with the solvent accessibility and hydrophobicity profiles, since it is known that hydrophobic monomers inside the protein core are more conserved. These profiles could be modeled by a multivariate distribution similar to the one we used for solvent accessibility and hydrophobicity. Such work could provide valuable information about the interaction among different characteristics of monomers.

1.2 Differential Motif Finding

1.2.1 Introduction to Gene Regulation

Motif finding in biology has become an important topic because of the amount of quantitative biological data being generated in recent years. Motif finding has applications in finding functional elements in protein structures or functional elements in non-coding DNA, which regulate the level of gene expression. Understanding gene regulation is central for understanding how diverse cell types are created from the same genome. It also helps in understanding the nature of diseases that result from malfunctioning regulatory mechanisms.

Regulation is carried out on different levels. One of the most important levels of regulation is the transcriptional level, where a gene on the DNA is read by RNA polymerase and the corresponding mRNA sequence is generated. The magnitude of transcription depends on the attachment of specific proteins, called transcription factors, to a region preceding the start site of the gene, known as the cis-regulatory region. Each protein generally has a specific binding site in the cis-regulatory region called the regulatory factor binding site (RFBS). A binding site can regulate the gene expression for a specific function by itself or in combination with other binding sites [60].

1.2.2 What is Motif finding?

The binding sites corresponding to a specific transcription factor can be represented by a motif model. The motif can be represented by the consensus sequence after

aligning all the binding site sequences and choosing a consensus nucleotide letter to represent the best match in each column. This consensus representation is useful for easy comparison of motifs, and when the motif is fairly sharp, a single letter is highly likely in each position. When dealing with motifs that are less sharp, a matrix can be used to represent the motif. In this representation, the number of each letter in each column of aligned sequences is counted and is recorded in the matrix. Matrix representation is more useful for representing degenerate motifs. Also, it is fairly easy to use motif matrices to find other binding sites that could be present in the cis-regulatory region of other genes[87, 86].

Motif finding methods try to find common patterns among a set of co-regulated genes. Co-regulated genes are genes with similar transcriptional rates in the presence of a specific stimulus. In gene expression experiments, the rate of transcription of many genes is measured. By clustering the data, co-regulated genes are identified[79, 89, 26, 85]. Motif finders can then be used to search for over-represented patterns in the cis-regulatory region of these genes. Predictions from bioinformatics methods can be verified experimentally. For example, if the transcription factor is known, it can be mixed with random DNA sequences. The sequences that bind to the transcription factor can be recovered and aligned to generate the motif model [77, 80].

1.2.3 Methods of motif finding

Most computational motif finders fall into two main classes: enumerative methods and alignment methods. In enumerative methods, all the oligomers of a certain length in the cis-regulatory region of co-regulated genes are counted. The oligomers with high counts as compared to what is expected from a background model are chosen as motif candidates. These oligomers can be clustered to generate motif models [90, 5]. In alignment methods, however, sequences are aligned so as to identify significant local similarities [86, 51, 2]. Alignment methods use a matrix representation of the motif. They begin by choosing one or a few start sites for the binding site on each sequence. They then try to find better sites, ones that increase the likelihood of the sequences with that motif in it. The likelihood function is defined as the chance of

observing the sequence with a certain alignment when a motif model is assumed. There are a number of different methods to maximize the likelihood. We focus on describing the Gibbs Sampler because it will be used later in our work.

1.2.4 Gibbs Sampler

Our goal is to maximize the likelihood of observing the sequence data given a motif model [10, 51, 2]. To achieve this, Gibbs sampling samples from the likelihood function in such a way that more time is spent where the likelihood is the highest [28].

The Gibbs sampler is a Markov chain Monte Carlo method. A Markov chain is a process where each state is only dependent on the preceding state. The system moves from one state to another through a transition matrix. The transition matrix is called ergodic if it fulfills certain conditions: every state is accessible from every other state in a finite number of steps, and there is at least one state for which there is non-zero probability of return in one step. A system evolving under an ergodic transition matrix eventually reaches a stationary state where the likelihood function is properly sampled [28, 59].

Metropolis et al [62] suggested a way to construct a transition matrix given a probability distribution that cannot be easily normalized or directly sampled from. Transitions between states are based on the likelihood. In this method, the ratio of the likelihoods of the two states is compared. If the likelihood of the new state is better, then the move is made, whereas otherwise the move only takes place with a probability equal to the ratio of the likelihoods. This method is widely used in physics in situations where obtaining an analytical normalization for the probability distribution, or partition function, is impossible. In physics, the likelihood function is usually proportional to the Boltzmann weight $exp(-H/kT)$. Since it is often difficult to obtain a closed form for the partition function, it is impossible to directly sample from the probability distribution. The advantage of using the Metropolis algorithm is that it does not require knowledge of the partition function. Hastings introduced the Metropolis algorithm to statistics. He extended the Metropolis algorithm to cases where the moves from one state to another are not symmetric [35].

The Gibbs sampler can be considered as a special case of the Metropolis-Hasting method where the probability distribution is a function of multiple variables. In this method, all the model parameters except one are fixed at each step. A new value for that parameter is sampled from a conditional distribution for that parameter over all the other model parameters. This procedure is then repeated for the other parameters [59].

To use the general Gibbs sampling algorithm for the specific task of motif finding, a likelihood function for observing the set of sequences with instances of motifs in them is needed. These typical model parameters usually include the location of the motif instances, the background composition of the sequences, and a model description of the motif. The model parameters are initially set randomly. For example, this can be done by picking random sites as the start positions for the motif on each sequence. One sequence is then chosen at random. A new start site for the motif instance is sampled based on the scores of each segment against the weight matrix constructed from the motif instances in other sequences. Then another sequence is chosen and the same procedure is repeated there. This is continued until the system converges. Convergence is reached when the weight matrix generated from the motif instances stabilizes. This can be quantified by measuring the information content of the weight matrix [86, 51, 58, 69].

1.2.5 Challenges

Motif finders that try to use only one set of sequences to find the functional motif are prone to generate many false positives [93, 70]. Any extra information regarding the motif, such as the exact size of the motif or the number of instances in each sequence, can help to reduce the number of false positives. Clustering methods used to group co-regulated genes are also error-prone. Some of these genes might not be regulated by the same factor, but could have similar expression profiles as a result of a secondary response. As a result, the binding site might not be found in the cis-regulatory region of the genes. It is thus important to have the flexibility of searching for motifs that do not necessarily have an instance in each sequence.

1.2.6 The differential motif finder

Our goal is to find a motif that is highly present in one set of sequences as compared to another set of sequences. For example, we imagine that based on expression profiles, a group of genes can be separated into two sets: a "positive" set that is expected to contain the motif, and a "negative set" that is not. We define a total likelihood that combines the probability of observing the motif in the first set with a penalty for its appearance in the second set. The total likelihood contains a likelihood function for the first set similar to that developed by Liu et al [58, 69] divided by a term that contains an estimate of the chance of observing the motif in the second set. In this method, sampling is only done over the positive set, and the negative set is treated as a whole block. We use the weight matrix generated from the motif instances in the first set to estimate its presence in the second set. Since we are not actually looking for the positions of the motif in the second set but only the strength of the motif there, we integrate over all of the feasible locations of motif occurrences in the second set. This produces a partition function, which is a function only of the weight matrix and the expected number of motifs in the second set. The weight matrix is known, because it comes from the first set. For the expected number of motifs, we choose an expected number that maximizes the value of the partition function.

Chapter 2

Structure Space of Model Proteins: A Principal Component Analysis

2.1 Introduction

Proteins fold into specific structures to perform their biological function. Despite the huge diversity in their functions, evolutional paths, structural details, and sequences, the vast majority of proteins adopt only a small number ($\sim$ 1000) of folds ("topology"). [25, 12, 6, 71, 92, 33] This observation has intrigued a number of authors and lead to the concept of *designability*. [25, 9, 99, 30, 54] The designability of a structure is defined to be the number of sequences that have that structure as their unique lowest-energy state. [54] It has been shown in various model studies that structures differ drastically in their designability; a small number of highly designable structures emerge with their associated number of sequences much larger than the average. [54, 31, 56, 7, 36, 11, 63] Highly designable structures are also found to possess other protein-like properties, such as thermodynamic stability, [54] fast folding kinetics, [30, 61] and tertiary symmetry. [99, 54, 91] These results suggest that there may be a designability principal behind nature's selection of protein folds; these small number of folds were selected because they are readily designed, stable against mutations, and thermodynamically stable.

Why are some structures more designable than others? How do we identify highly

designable structures? Finkelstein and co-workers argued that certain motifs are easier to stabilize and thus more common because they either have lower (e.g. bending) energies or have unusual energy spectra over random sequences. [25, 24, 23] Govindarajan and Goldstein suggested that the topology of a protein structure should be such that it is kinetically "foldable". [30, 31, 32] More recently, it was noted that an important clue resides in the distribution of structures in a suitably defined structure space, with highly designable structures located in regions of low density. [56, 7] In particular, within a hydrophobic model, Li *et al.* showed that the distribution of structures is very nonuniform, and that the highly designable structures are those that are far away from other structures. [56] However, identifying highly designable structures still remains a tedious task, requiring either full enumeration or sufficiently large sampling of both the structure and the sequence spaces, making studies of large systems prohibitive.

In this paper, we investigate the properties of the structure space of the hydrophobic model of Li *et al.*, starting from a Principal Component Analysis (PCA). We show that while the distribution of the structures is not uniform, it can be approximated as a cloud of points centered on a single peak. The principal directions of this cloud are almost coincident with those obtained by rotation into Fourier space; the coincidence is in fact exact for the subset of cyclic structures. An interesting feature is that the eigenvalues of PCA, describing the extent of the density cloud along the principal axis, vary continuously with the Fourier label q, with a minimum at $q = \pi$ corresponding to alternating patterns. The continuity of the eigenvalues suggests an expansion around $q = \pi$, which leads to an analytical conjecture for the density of structures in the N-dimensional binary space. Assuming the validity of this conjecture in more general models, it provides a means of estimating density, and hence indirectly designability, of structures by simply analyzing their sequences, without the need for extensive enumerations of other possible structures.

The rest of the paper is organized as follows. In Section II we review the hydrophobic model and the designabilities of structures. In Section III we discuss the methods and the results of PCA applied to the structure space, and relate the density

and designability of a structure to its projections onto the principal axes. In Section IV we demonstrate that Fourier transformation provides a very good approximation to PCA, and show that in fact the two procedures are equivalent for the subset of cyclic structures. As a comparison with real structures, in Section V we introduce and study an ensemble of pseudo-structures constructed by a Markovian process. Finally, in Section VI we synthesize the numerical results of PCA analysis, and develop a conjecture for the density of points in structure space.

2.2 The Hydrophobic Model

We start with a brief review of the hydrophobic model of Li *et al.* [56] and the designabilities of structures. Model sequences are composed of two types of amino acids, H and P. Each sequence $\{h_i\}$ (for $i = 1, 2, \cdots, N$) is represented by a binary string or vector, with $h_i = 0$ for a P-mer and $h_i = 1$ for an H-mer. We take the polymer length $N = 36$, for which there are 2^{36} sequences. Each of these sequences can fold into any one of the many compact structures on a 6×6 square lattice (Fig. 2-1). There are $57,337$ such compact structures unrelated by rotation and mirror symmetries. In the hydrophobic model, the only contribution to the energy for a sequence folded into a structure is the burial of the H-mers in the sequence into the core of the structure. So if one represents a structure by a binary string or vector, $\{s_i\}$, for $i = 1, 2, \cdots, 36$, with $s_i = 0$ for the surface sites and $s_i = 1$ for the core sites (Fig. 2-1), the energy is

$$E = -\sum_{i=1}^{N} h_i s_i, \tag{2.1}$$

where h_i is the sequence vector.

The designability of a structure is defined as the number of sequences that have the structure as their unique lowest-energy state. To obtain an estimate for designabilities of structures, we randomly sampled $50,000,000$ sequences and found the unique lowest-energy structure, if any, for each of them. In Fig. 2-2, we plot the histogram of designabilities, *i.e.* number of structures with a given designability. Note

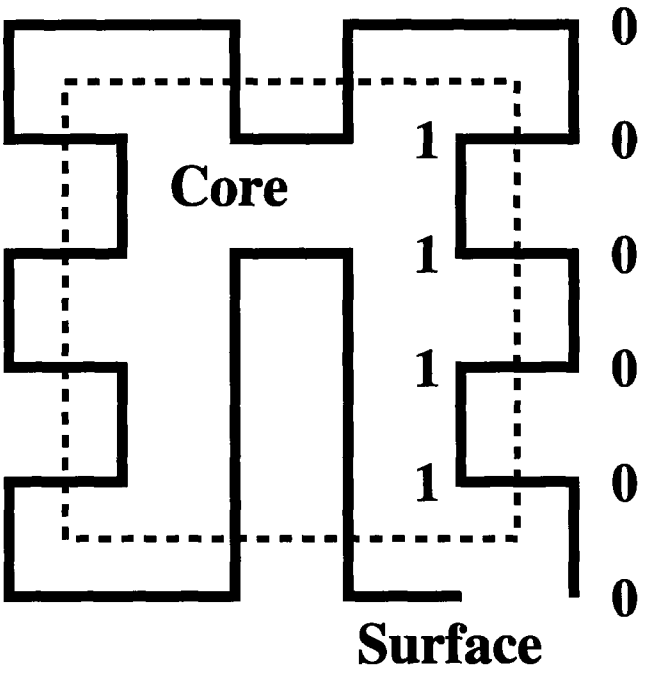


Figure 2-1: A possible compact structure on the 6×6 square lattice. The 16 sites in the core region, enclosed by the dashed lines, are indicated by 1's; the 20 sites on the surface are labeled by 0's. Hence this structure is represented by the string 00110011000011000011001100001111110O. Note that each 'undirected' open geometrical structure can be represented by two 'directed' strings, starting from its two possible ends (except for structures with reverse-labeling symmetry where the two strings are identical). It is also possible for the same string to represent different structures which are folded differently in the core region. For the 6×6 lattice of this study, there are 26929 such 'degenerate' structures, which are by definition non-designable.

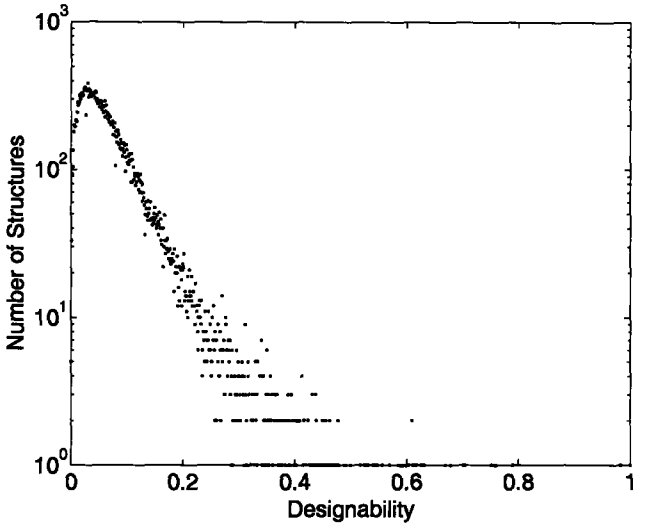


Figure 2-2: Number of structures with a given designability versus relative designability for the 6×6 hydrophobic model. The data is generated by uniformly sampling 5×10^7 strings from the sequence space. The designability of each structure is normalized by the maximum possible designability.

that we have normalized designability so that its maximum value of 2981 is scaled to one. In this paper, we define highly designable structures to be the top one percent of designable structures (structures with nonzero designability), which means 307 structures with a designability larger than 0.47.

In the hydrophobic model, both sequences and structures can be regarded as points in a 36-dimensional binary space, or corners of a hypercube in a Euclidean space of similiar dimension. In this representation, the lowest-energy state of a sequence is simply its nearest structure point. [56] Designabilities can then be obtained by constructing Voronoi polyhedra around all points corresponding to structures in this space; the designability of each structure is then the number of sequence points that fall within the corresponding Voronoi polytope (Fig. 2-3). Structures in the lower density regions have larger Voronoi polytopes and higher designability. Understanding how the structure points are distributed in this 36-dimensional space can thus help us address questions concerning designability. In the next section we examine the distribution of the structure points via the method of PCA.

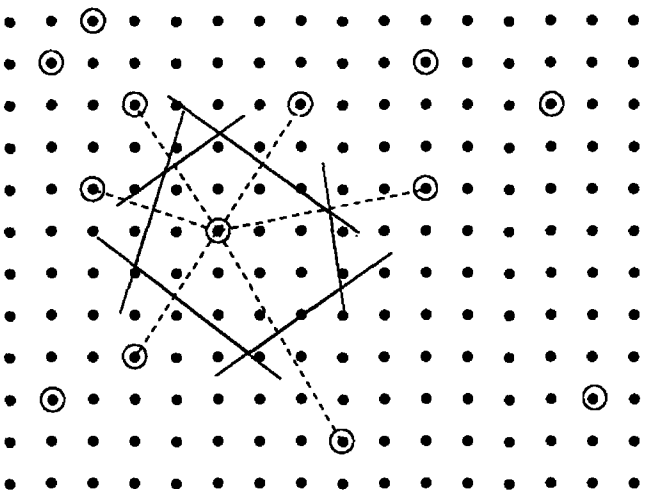


Figure 2-3: Schematic representation of the 36-dimensional space in which sequences and structures are vectors or points. Sequences, represented by dots, are uniformly distributed in this space. Structures, represented by circles, occupy only a sparse subset of the binary points and are distributed non-uniformly. The sequences lying closer to a particular structure than to any other, have that structure as their unique ground state. The designability of a structure is therefore the number of sequences lying entirely within the Voronoi polytope about that structure.

2.3 Principal Component Analysis

First, let us note that while sequences are uniformly distributed in the 36-dimensional hypercube, structures are distributed on a 34-dimensional hyperplane because of the following two geometrical constraints. The first constraint on structure vectors comes from the fact that all compact structures have the same number of core sites, and thus

$$\sum_{i=1}^{36} s_i = 16. \tag{2.2}$$

The second constraint is that since the square lattice is bipartite, and any compact structure traverses an equal number of 'black' and 'white' points,

$$\sum_{i=1}^{36} (-1)^i s_i = 0. \tag{2.3}$$

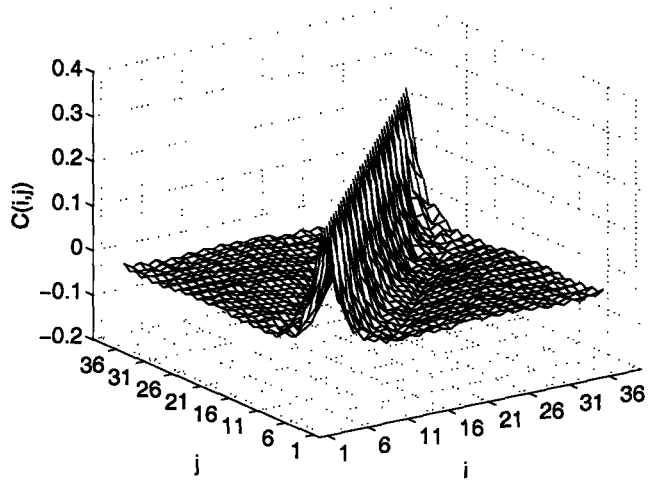


Figure 2-4: Covariance matrix C_{ij} of all compact structures of the 6×6 square.

Next, let us define the covariance matrix of the structure space as

$$C_{i,j} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle , \tag{2.4}$$

where $i, j = 1, 2, \cdots, 36$, and the average is over all the $57,337$ possible $(s_1, s_2, \cdots, s_{36})$ for compact structures. The 36×36 covariance matrix is symmetric $C_{i,j} = C_{j,i}$, and also satisfies the condition $C_{i,j} = C_{37-i,37-j}$. The latter is due to the reverse-labeling degeneracy of the structure ensemble, since if the string $(s_1, s_2, \cdots, s_{36})$ is in this ensemble, then its reverse $(s_{36}, s_{35}, \cdots, s_1)$ is also included. This symmetry implies that if $(v_1, v_2, \cdots, v_{36})$ is an eigenvector of the matrix $C_{i,j}$, then $(v_{36}, v_{35}, \cdots, v_1)$ is also an eigenvector with the same eigenvalue. Therefore, for every eigenvector of $C_{i,j}$ we have either $v_j = v_{37-j}$ or $v_j = -v_{37-j}$.

As depicted in Fig. 2-4, the matrix C_{ij} is peaked along the diagonal and decays off-diagonally with short range correlations. This feature reflects a general property of compact self-avoiding walks; if a monomer is in the core (on the surface) the neighboring monomers along the chain have enhanced probability to be in the core (on the surface). Another characteristic of C_{ij} is that it is almost a function of $|i-j|$ only, *i.e.* $C_{ij} \approx F(|i-j|)$, barring some small end and parity effects. We expect this feature of approximate translational invariance to be generic beyond the 6×6 lattice model studied here. We also looked at the covariance matrix for the subset

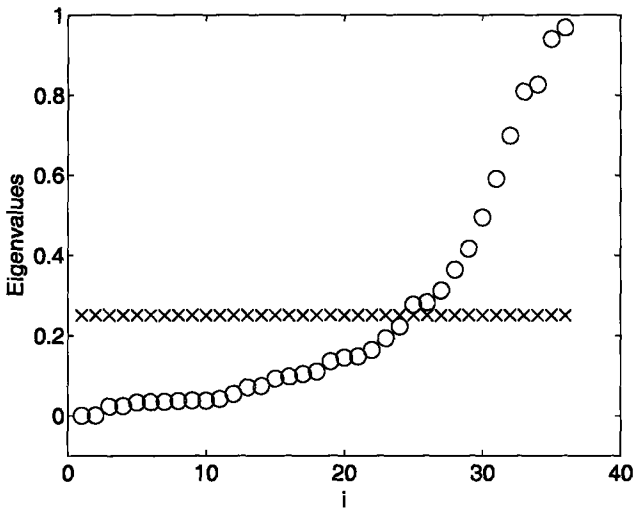


Figure 2-5: Eigenvalues of the covariance matrix for the structure vectors (circles), and for all points in sequence space (crosses).

of highly designable structures. While qualitatively similar, it tends to decay faster off-diagonally than that of all structures. This is attributed to the fact that highly designable structures tend to have more frequent transitions between core and surface sites. [56, 11, 83]

For PCA of structure space, the matrix C_{ij} is diagonalized to obtain its eigenvectors $\{\vec{v}^{(k)}\}$, and the corresponding eigenvalues $\{\lambda_k\}$ for $k = 1, 2, \cdots, 36$, which are shown in Fig. 2-5. The two zero eigenvalues ($\lambda_1 = \lambda_2 = 0$) result from the constraints in Eqs. (2.2) and (2.3), with the corresponding eigenvectors of $v_i^{(1)} = 1$, and $v_i^{(2)} = (-1)^i$ for $i = 1, 2, \cdots, 36$, respectively. The remaining 34 nonzero eigenvalues range smoothly from zero to one, making any further dimensional reduction not obvious. For comparison, the 36 eigenvalues of the uniformly distributed points of sequence space are all the same ($\lambda = 1/4$). (It is easy to show that the covariance matrix for the sequence space is $C_{ij} = \delta_{ij}/4$.) On the other hand, a uniform distribution on the 34-dimensional hyperplane where the structure points reside would result in 34 identical eigenvalues of $360/1377 \approx 0.26$.[1]

[1]This can be seen from the following argument: Since all points in the 34-dimensional hyperplane are equivalent up to parity, the most general form of the covariance matrix is $C_{ij} = x + y(-1)^{i-j} + z\delta_{ij}$. Requiring zero eigenvalues for eigenvectors $(1, 1, 1, 1, ...)$ and $(-1, 1, -1, 1, ...)$ gives the constraints $36x + z = 0$, and $36y + z = 0$, i.e. $x = y = -z/36$. The value of z is then set by $C_{ii} = \langle s_i \rangle (1 - \langle s_i \rangle) = 20/81$, where $\langle s_i \rangle = 16/36$. So we have $x = y = -10/1377$ and $z = 360/1377$. It is then easy to see

Identification of the principal axes and eigenvalues does not necessarily provide information about the distribution of points in space. To examine the latter, we first project each structure vector onto its components along the eigenvectors. Using the rotation matrix R_{ki} that diagonalizes the covariance matrix, the component y_k of the structure vector along principal axis k is obtained as

$$y_k = \sum_{i=1}^{36} (s_i - \langle s_i \rangle) R_{ki}. \tag{2.5}$$

Interestingly, we find that along each of the principal directions, the distribution of components is a bell-shaped function with a single peak at zero. Such distributions can then be well approximated by Gaussians whose variances are the corresponding eigenvalues λ_k, *i.e.*

$$\rho_k(y_k) \approx \frac{1}{\sqrt{2\pi\lambda_k}} e^{-\frac{y_k^2}{2\lambda_k}}. \tag{2.6}$$

In Fig. 2-6 we show the distribution of projections y_k on two principal axes $k = 16$ and $k = 36$, along with the corresponding Gaussian distributions.

Equation (2.6) provides a good characterization of the density of structures in the N dimensional space. Highly designable structures are expected to lie in regions of this space where the density of structures is small, while the number of available sequences is large. Let us consider a structure characterized by a vector $\vec{y}$. If the density of structural points in the vicinity of this point is $\rho_{str}(\vec{y})$, the number of available structures in a volume V around this point is $V\rho_{str}(\vec{y})$. Neglecting various artifacts of discreteness, the volume of the Voronoi polyhedron (see Fig. 3) around this point is given by $V(\vec{y}) \approx 1/\rho_{str}(\vec{y})$. The designability is the number of structures within this volume, and estimated as $\rho_{seq}(\vec{y})/\rho_{str}(\vec{y})$, where $\rho_{seq}(\vec{y})$ is the density of sequences. The sequence density is in fact uniform in the N-dimensional space. The structure density can be approximated as the product of Guassians along the principal

that the 34 nonzero eigenvalues are 360/1377.

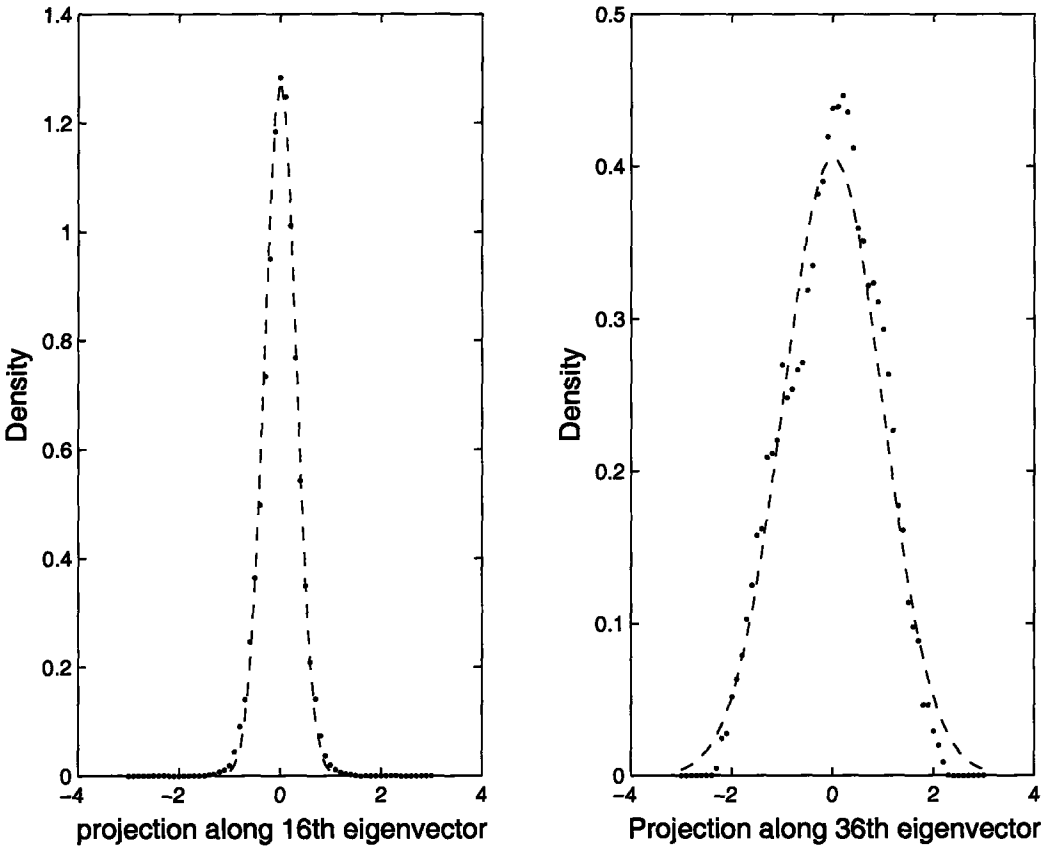


Figure 2-6: Distributions of projections y_k onto principal axes $k = 16$ (a), and $k = 36$ (b), for all 57337 structures (dots). Also plotted are Gaussian forms with variances λ_{16} and λ_{36}, respectively (dashed lines).

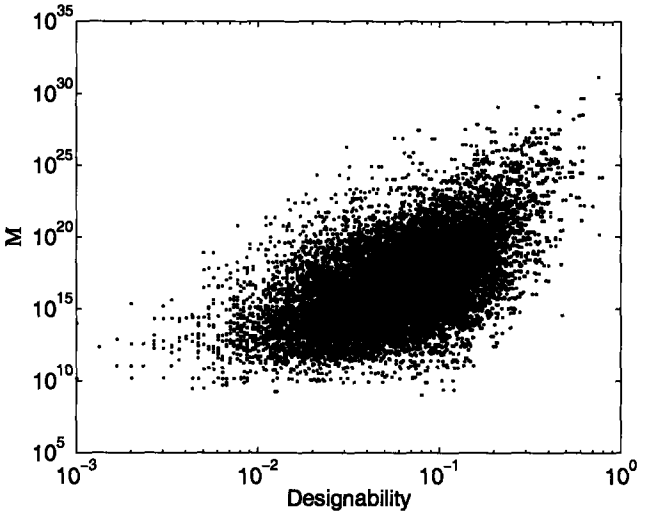


Figure 2-7: The estimate $\mathcal{M}$ (Eq. (2.7)) versus scaled designability for all designable structures on the 6×6 square.

projections, and thus

$$\text{Designability} \approx \frac{\rho_{seq}(\vec{y})}{\rho_{str}(\vec{y})} \propto \prod_{k=3}^{36} \frac{1}{\rho_k(y_k)} \propto \exp\left[\sum_{k=3}^{36} \frac{y_k^2}{2\lambda_k}\right] \equiv \mathcal{M}(\vec{y}). \qquad (2.7)$$

We have neglected various proportionality constants in the above equation, leading to the quantity $\mathcal{M}(\vec{y})$ which is our estimator for designability. In Fig. 2-7, the estimate $\mathcal{M}$ is plotted against the actual designability for all designable structures. There is a reasonably good, but by no means perfect, correlation between the designability and the estimator $\mathcal{M}$. The structures with the top one percent value of $\mathcal{M}$ include 39% of the highly designable structures.

2.4 Fourier Decomposition And Cyclic Structure

In discussing Fig. 2-4, we already noted that the covariance matrix C_{ij} is approximately a function of $|i-j|$, with corrections due to end effects. If this were an exact symmetry, the matrix would be diagonal in the Fourier basis. Even in the presence of the end effects, Fourier decomposition provides a very good approximation to the eigenvectors and eigenvalues of PCA, as demonstrated below. For each structure

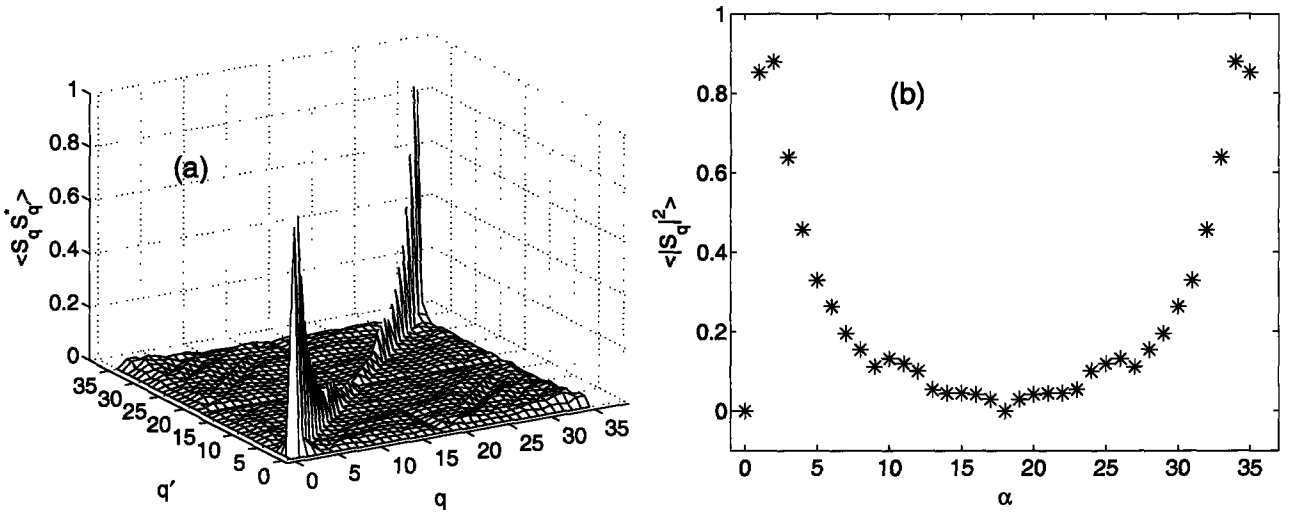


Figure 2-8: (a) The Fourier transformed covariance matrix $\langle S_q S^*_{q'} \rangle$ (Eq. (2.9)); and (b) its diagonal elements $\langle |S_q|^2 \rangle$.

vector $\{s_j\}$, the Fourier components are obtained from

$$S_q = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} e^{iqj} \left(s_j - \langle s_j \rangle \right), \tag{2.8}$$

where $q = 2\pi\alpha/N$, with $\alpha = 0, 1, \cdots, N-1$. The average value of $\langle s_j \rangle$ is subtracted for convenience. With this subtraction, the two constraints in Eqs. (2.2) and (2.3) correspond to two zero modes in Fourier space, as $S_0 = 0$ and $S_\pi = 0$, and since $\{s_j\}$ are real $S^*_q = S_{-q}$.

The covariance matrix in the Fourier space is

$$\langle S_q S^*_{q'} \rangle = \frac{1}{N} \sum_{j,j'=1}^{N} e^{i(qj - q'j')} C_{jj'}, \tag{2.9}$$

and is both real and symmetric (since $C_{jj'} = C_{j'j}$). If $C_{jj'}$ is translationally invariant, *i.e.* $C_{jj'} = F(|j - j'| \bmod N)$, Eq. (2.9) becomes

$$\langle S_q S^*_{q'} \rangle = \delta_{q,q'} \lambda_q, \tag{2.10}$$

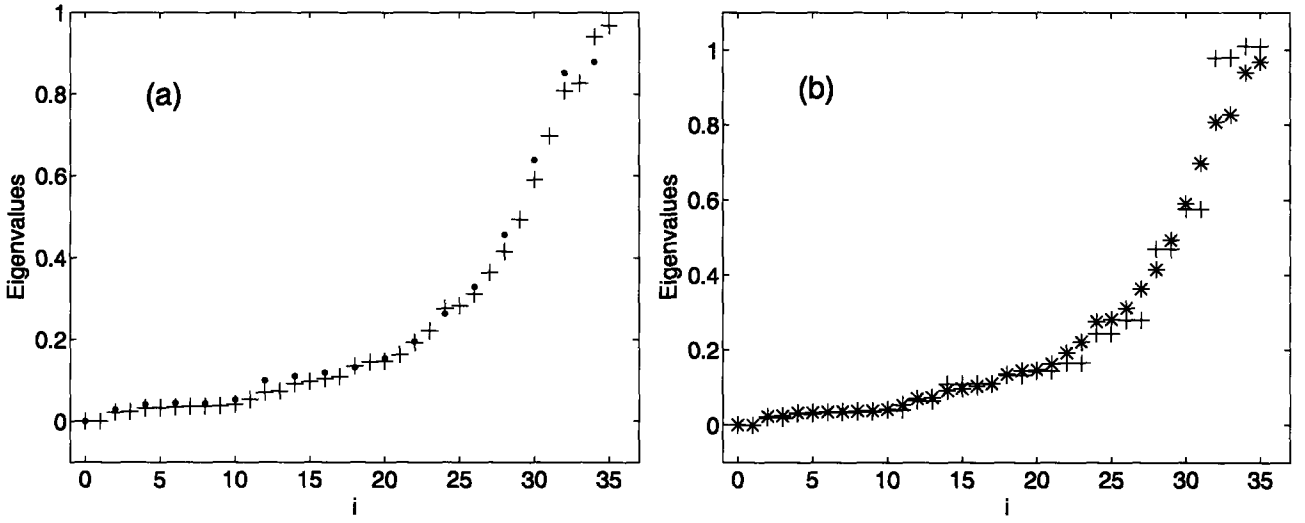


Figure 2-9: (a)The diagonal elements $\langle |S_q|^2 \rangle$ (dots) plotted together with the eigenvalues of the covariance matrix (pluses). (b) Diagonal elements of the Fourier transformed C_{cyclic}, which are also the eigenvalues of the covariance matrix of cyclic structures (pluses). Eigenvalues of the covariance matrix for all structures are indicated by stars.

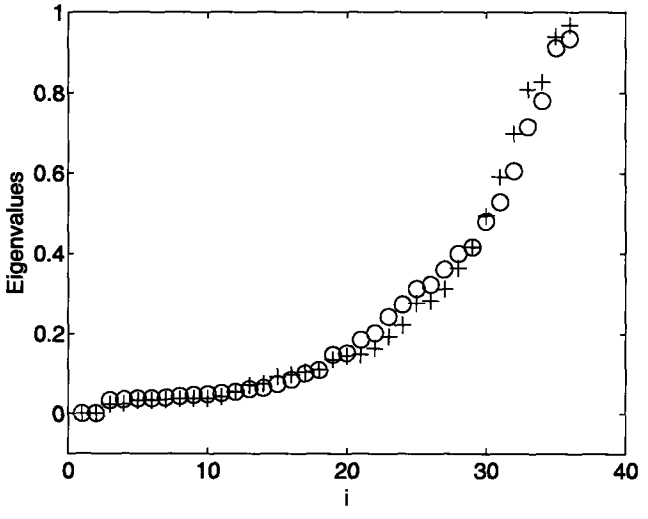


Figure 2-10: Eigenvalues of the covariance matrix for the structures generated by the Markov model (circles), and that of the true structure space (pluses).

where

$$\lambda_q = \sum_{k=0}^{N-1} e^{iqk} F(k) = \left\langle |S_q|^2 \right\rangle , \tag{2.11}$$

are the diagonal elements of the diagonalized matrix in Eq. (2.10), and hence the eigenvalues of $C_{jj'}$. Note that because the matrix is real–symmetric, its eigenvalues appear in pairs, *i.e.*

$$\lambda_q = \lambda_{-q} . \tag{2.12}$$

Since our covariance matrix is not fully translationally invariant, $\left\langle S_q S_{q'}^* \right\rangle$ is not diagonal. However, as shown in Fig. 2-8a, its off diagonal elements are very small. As required by symmetry, the diagonal elements form pairs of identical values. These diagonal elements, plotted versus the index α in Fig. 2-8b, should provide a good approximation to the eigenvalues obtained by PCA. This is corroborated in Fig. 2-9a, where we compare $\langle |S_q|^2 \rangle$ with the true eigenvalues of the covariance matrix $C_{jj'}$.

Finally, we note that the end effects that mar the translational invariance of the covariance matrix are absent in the subset of *cyclic structures.* Any structure whose two ends are neighboring points on the lattice can be made cyclic by adding the missing bond. Any one of the $N = 36$ bonds on the resulting closed loop can be broken to generate an element of the original set of structures, and the corresponding structure strings are cyclic permutations of each other. Thus, the covariance matrix $C_{cyclic}(j, j')$ of the set of all cyclic structures is translationally invariant. In our model of 6×6 compact polymers, there are a total of 36×276 cyclic structures. The Fourier transform of their covaraince matrix is diagonal as expected, with diagonal elements depicted in Fig. 2-9b. The corresponding Fourier eigenvalues are quite close to the eigenvalues of the full matrix obtained in the PCA (Fig. 2-9b). Thus the end effects do not significantly modify the correlations, and this is specially true for the smallest eigenvalues which make the largest contributions to the density in Eq. 2.7.

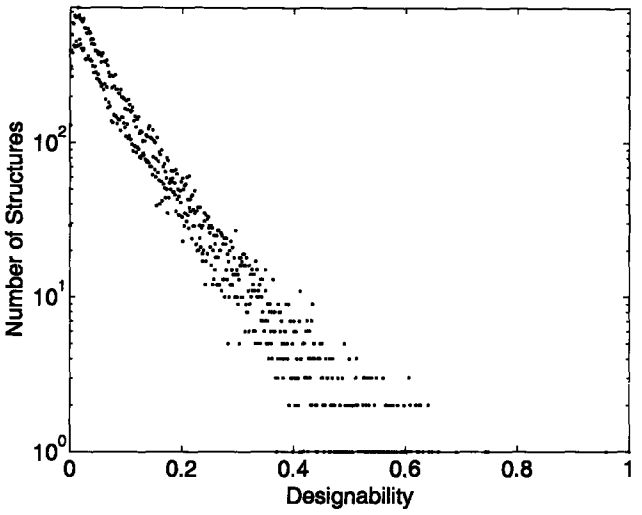


Figure 2-11: Number of pseudo-structures with a given designability versus designability for the pseudo-structure strings randomly generated using the Markov model. The data is generated by uniformly sampling 5×10^7 binary sequence strings. The designability of each pseudo-structure is normalized by the maximum possible designability.

2.5 A Markovian Ensemble of Pseudo–Structures

The geometry of the lattice, and the requirement of compactness constrain the allowed structure strings of zeros and ones in a non-trivial fashion. In our estimation of designabilities so far, we have focused on the covariance matrix which carries information only about two point correlations along these strings. In principle, higher point correlations may also be important, and we may ask to what extent the covariance matrix contains the information about the structures' designabilities? As a preliminary test, we performed a comparative study with an artificial set of strings, not corresponding to real structures, but constructed to have a covariance matrix similar to true structures on the 6×6 lattice.

Specifically, we generated a set of random strings $\{\vec{t}\}$, of zeros and ones of length 36, using a third order Markov process as follows. For each string, the first element t_1 is generated with probability $P(t_1 = 1) = \langle s_1 \rangle$, where $\langle s_1 \rangle$ is the fraction of the true structure strings with $s_1 = 1$. The second element t_2 is generated according to a transition probability $P(t_1 \rightarrow t_2)$ which is taken to be the "conditional probability" $P(s_2|s_1)$ extracted from the true structure strings. The third point t_3 is

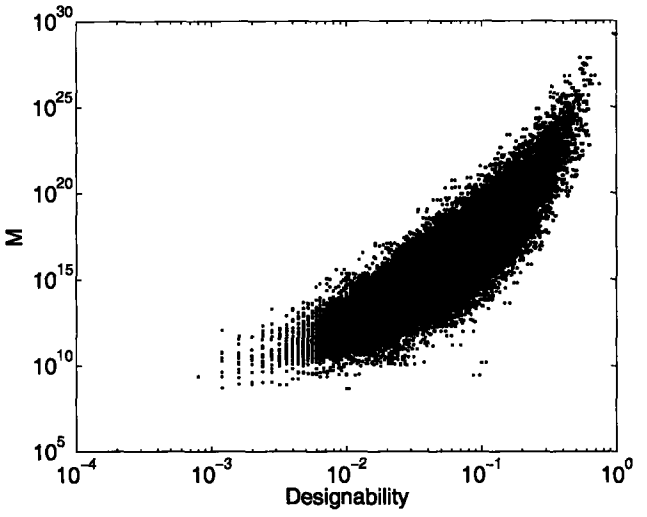


Figure 2-12: The quantity $\mathcal{M}$ versus designability for all designable pseudo-structures generated by the Markov model.

generated according to a transition probability $P(t_1 t_2 \rightarrow t_3)$ which is the "conditional probability" $P(s_3|s_1 s_2)$ extracted from the true structure strings. All the remaining points t_j, $j = 4, 5, \cdots, 36$, are generated according to the transition probabilities $P(t_{j-3} t_{j-2} t_{j-1} \rightarrow t_j)$ equal to the true "conditional probabilities" $P(s_j|s_{j-3} s_{j-2} s_{j-1})$ of actual structures. Sequences that do not satisfy the global constrains of Eqs. (2.2) and (2.3) are thrown out. For every Markov string generated, we also put its reverse in the pool, unless the string is its own reverse.

The above Markovian ensemble has a covariance matrix, and corresponding eigenvalues, very similar to those of the true structures, as shown in Fig. 2-10. We then calculated the designabilities of these "pseudo-structures" using Eq. (2.1) by uniformly sampling 5×10^7 random binary sequences. The histogram of the designabilities (Fig. 2-11) is qualitatively similar to that of the true structures (Fig. 2-2). Next we constructed the designability estimator $\mathcal{M}$ (Eq. (2.7)) for the pseudo-structures, using the eigenvalues and eigenvectors of their covariance matrix. The quantity $\mathcal{M}$ is plotted versus designability in Fig. 2-12 for all the artificial pseudo-structures with non-zero designability. The pseudo-structures with the top one percent value of $\mathcal{M}$ include 60% of the highly designable psuedo-structures.

These results suggest that a considerable amount of information about the des-

ignability is indeed contained in the two point correlations of the string. The designability estimator, Eq. (2.7) in fact does a somewhat better job in the case of pseudo-structures generated according to short-range Markov rules.

2.6 Conclusions

One of the most intriguing properties of compact structures, which emerged from early extensive enumeration studies,[54] is that designabilites range over quite a broad distribution of values. Such a large variation in designability is a consequence of a non-uniform distribution of structure vectors, with highly designable structures typically found in regions of low density.[56] However, our study of 6×6 lattice structures using PCA indicates that the non-uniform density actually has a rather simple form that can be well approximated by a single multi-variable Gaussian, as in Eq. (2.7). Since this method of estimating structure designability is based only on the overall distribution of structures, it can be a useful tool in cases where there is not enough computational power to enumerate the whole structure space and calculate the designability. To obtain an accurate enough covariance matrix requires only a uniform sampling of the structure space.

We can also attempt to use the numerical results as a stepping stone to a more analytical approach for calculating the density of structures. *First*, we note that the covariance matrix for all structures is rather similar to that of the subset of cyclic structures, and that for the latter PCA is equivalent to Fourier decomposition. *Second*, we observe that the multi-variable Gaussian approximation to structure density in Eq. (2.7) is most sensitive to the eigenvalues that are close to zero. In terms of Fourier components, these are eigenvalues corresponding to values of q close to π, and related to the constraint of Eq. (2.3). There is also a zero eigenvalue for $q = 0$, related to condition (2.2). However, the latter global constraint appears not to have any local counterpart, as there is a discontinuity in the eigenvalues close to zero. *Third*, the continuity of the eigenvalues as $q \to \pi$, along with the symmetry of Eq. (2.12), suggests as expansion of the form $\lambda_q = K(q-\pi)^2 + \mathcal{O}\left((q-\pi)^4\right)$. Indeed the numerical

results indicate that the important (smaller) eigenvalues can well be approximated by $K(q-\pi)^2$, with $K \approx 0.04$.[2]

With this approximation, the designability estimate of Eq. (2.7) becomes

$$\mathcal{M}(\{\vec{s}\}) \approx \exp\left[\sum_q \frac{|S_q|^2}{2K(q-\pi)^2}\right] = \exp\left[\frac{1}{2K}\sum_{i,j=1}^{N}(-1)^i s_i J_N(|i-j|)(-1)^j s_j\right]. \tag{2.13}$$

The first form in the above equation expresses the estimate in terms of the Fourier modes of the structure string, while the second term is directly in terms of the elements $\{s_i\}$. The function $J_N(|i-j|)$ is the discrete Fourier transform of $1/q^2$, which for large N behaves as $|i-j|$. Equation (2.13) is thus equivalent to the Boltzmann weight of a set of unit charges on a discrete line of N points marked by parity. The charges on the sublattice of the same parity attract each other with a potential $J_N(r)$, while those on different sublattices repel. Such an interaction gives a larger weight (and hence designability) to configurations in which the charges alternate between the core and surface sites, as observed empirically.[56, 11, 83]

In would be revealing to see how much of the above results, developed on the basis of a lattice hydrophobic model, can be applied to real protein structures. One could use the exposure level of residues to the solvent in building up the structure vectors. Current methods deal with structure strings of a fixed length, equal to the dimension of the structure space. Since real proteins have different lengths, there is a need for a scaling method to handle them all together. Our study shows that the two point correlations of structure vectors are approximately translationally invariant, and can be captured by Fourier analysis. This suggests the possibility of casting the density of points in structure space in universal functional forms dependent only of a few parameters encoding the properties of the underlying polymers. If so, it would be possible to provide good estimates for designability with polymers of varying length, without the need for extensive numerical computations.

[2]If not forced to go through zero for $q = \pi$, a somewhat better fit can be obtained with $\lambda_q = 0.03 + 0.04(q-\pi)^2 + \mathcal{O}\left((q-\pi)^4\right)$. Fourier transforming back to real space, the additional constant leads to screened Coulomb interactions for $J_N(s)$.

Chapter 3

Untangling influences of hydrophobicity on protein sequences and structures

3.1 Introduction

How the sequence of amino acids determines the structure and function of the folded protein remains a challenging problem. It is known that hydrophobicity is an important determinant of the folded state; hydrophobic monomers tend to be in the core, and polar monomers on the surface [44, 52, 16, 66]. Several studies have examined the correlations in the hydrophobicity of amino-acids along the protein chain [40, 74, 88, 95], which are useful in secondary structure prediction [18], and in the design of good folding sequences [96]. Naturally, sequence correlations arise from the locations of the amino-acids in the folded protein structure, and are best interpreted in conjunction with solvent accessibility profiles (which indicate how exposed a particular amino-acid is to water in a specific structure). For example, Eisenberg *et al* [16] note that for secondary structures lying at the protein surface, which have a strong periodicity in their solvent accessibility, hydrophobicity profiles also exhibit the period of the corresponding α helix or β strand. Constraints from forming compact

structures induce strong correlations in the solvent accessibility profile [34, 88, 98, 47], which should in turn induce similar correlations in the hydrophobicity profiles. It is desirable to quantify and separate the resulting correlations in protein sequences and structures.

In this paper, we aim for a unified treatment of hydrophobicity and solvent accessibility profiles, and the interactions between them. The *sequence* of each protein is represented by a profile $\{h_i\}$, where h_i is a standard measure of the hydrophobicity of the i-th amino-acid along the backbone [4]. Its *structure* has a profile $\{s_i\}$ for $i = 1,\ 2,\ \cdots, N$, where s_i is a measure of the exposure of the amino-acid to water in the folded structure [52]. While we do not expect perfect correlations between these profiles, we can inquire about the statistical nature of these correlations, and in particular whether they are diminished or enhanced at different periods. To this end, we employ the method of Fourier transforms and examine the statistics of the resulting amplitudes $\{\tilde{h}_q, \tilde{s}_q\}$, and power spectra $\{|\tilde{h}_q|^2, |\tilde{s}_q|^2\}$, for a database of 1461 non-homologous proteins. In a sense, this can be regarded as extending the work of Eisenberg *et al* [16] who explore correlations between hydrophobicity and solvent accessibility independent of specific locations along the backbone. Of course, the use of Fourier analysis is by no means new, and has for example been employed to study hydrophobicity profiles [16, 78, 40, 42]. However, we are not aware of its use as a means of correlating sequence and structure profiles.

Our results suggest that the hydrophobicity and solvent accessibility profiles are well approximated by a joint Gaussian probability distribution. This allows us to obtain the *intrinsic* correlations in the hydrophobicity profile, as distinct from correlations induced by solvent accessibility. For example, the α-helix periodicity in hydrophobicity profiles is shown to be induced by the corresponding periodicity in the solvent accessibility profiles. We also find that at long wavelengths the two profiles have different intrinsic characteristics: solvent accessibility profiles are positively correlated while hydrophobicity profiles are anti-correlated. Interestingly, the coupling between the two profiles is independent of wave-number, and hence can be interpreted as the Boltzmann weight of the solvation energy. The corresponding temperature is

close to room temperature, consistent with the "mean" temperature estimated in previous work from the frequencies of occurrence of amino acid residues in the core and on the surface [64, 23].

3.2 Methods and Results

For our protein data set, we selected 2200 representative chains from the Dali/FSSP database. Any two protein chains in this set have more than 25 percent structural dissimilarity. We removed all the multi-domain chains by using the CATH domain definition database, leaving 1461 protein chains [37, 72, 66]. The hydrophobicity profiles, $\{h_i\}$, were generated from the sequence of amino-acids using the experimentally measured scale of Fachere and Pliska [22] (in units of kcal/mol). We used the *relative solvent accessibility* reported by NACCESS [39] to generate solvent accessibility profiles $\{s_i\}$. (The relative solvent accessibility is the ratio of the solvent accessibility of a residue to the solvent accessibility of that residue in an extended tripeptide ALA-X-ALA for each amino acid type X.) We then computed the corresponding Fourier components as

$$\tilde{s}_q = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} e^{iqj} \left(s_j - \frac{\sum_{j=1} s_j}{N} \right), \tag{3.1}$$

where $q = 2\pi\alpha/N$, with $\alpha = 0, 1, \cdots, N-1$, and similarly for $\tilde{h}_q$. (The average values were subtracted to remove the DC component in the Fourier transform.)

Our results for the power spectra of solvent accessibility and hydrophobicity profiles are indicated respectively in Figs. 3-1(a) and 3-1(b) (q is related to the periodicity λ through $\lambda = \frac{2\pi}{q}$). A prominent feature of both plots is the peak at the α-helix periodicity $\lambda = 3.6$ [16]. Its presence in the solvent accessibility spectrum indicates that solvation energy plays a role in the spatial arrangement of α helices– they tend to lie at the surface, exposed to the solvent on one side and buried on the other side [16].

We would like to untangle correlations between the two profiles, so as to determine their intrinsic tendencies, by finding a joint probability distribution $P(\{s_i\}, \{h_i\})$. Clearly this cannot be decomposed as a product of contributions from different sites i, as neighboring components such as s_i and s_{i+1}, are highly correlated. We antici-

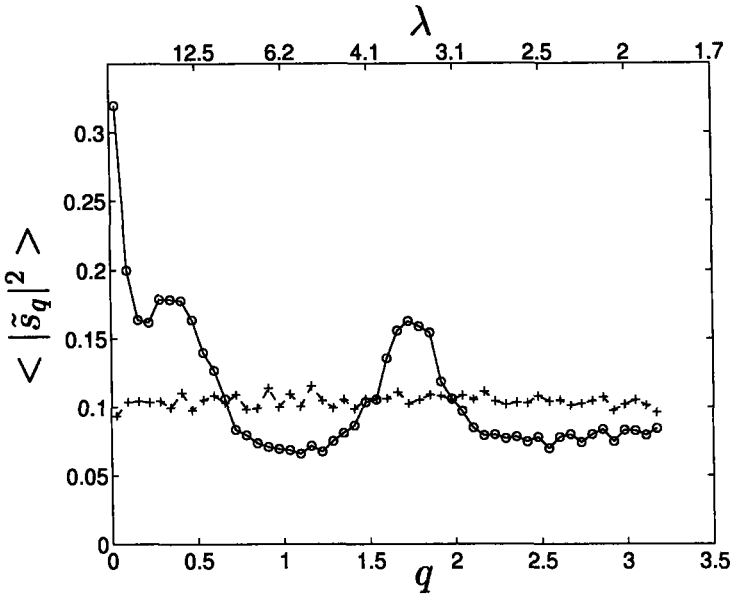


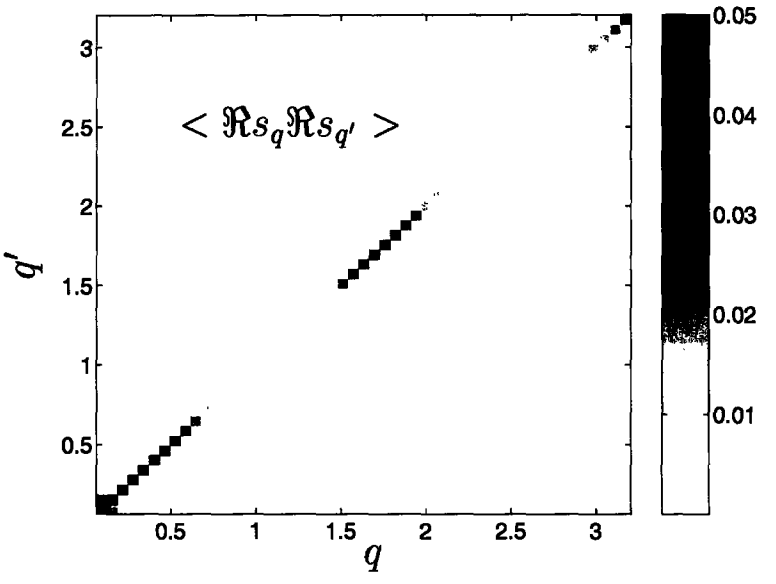
Figure 3-1: Power spectra from averaging over 1461 proteins, for (a) solvent accessibility (there are no units for $|\tilde{s}_q|^2$, since it based on accessibility relative to solution); (b) hydrophobicity (the units of $|\tilde{h}_q|^2$ are (k cal/mole)^2). The plus signs in each case are obtained from random permutation of the sequences.

pate that the Fourier components for different q are independently distributed (i.e. $P(\{\tilde{h}_q, \tilde{s}_q\}) = \prod_q p(\tilde{h}_q, \tilde{s}_q)$) for the following reasons: **(i)** For the subgroup of cyclic proteins [94] the index i is arbitrary, and the counting can start from any site. The invariance under relabeling then implies that the probability can only depend on $i-j$, and hence is separable into independent Fourier components. This exact result does not hold for open proteins because of end effects, but should be approximately valid for long sequences when such effects are small. **(ii)** Numerical analysis of a lattice model of proteins in Ref. [98] confirms the exact decomposition into Fourier modes for cyclic structures, and its robustness even for open structures of only $N = 36$ monomers. To test this hypothesis, we examine all possible covariances involving $\{\tilde{h}_q, \tilde{s}_q\}$ for different q. Note that the Fourier amplitudes are complex (i.e. $\tilde{s}_q = \Re s_q + i\Im s_q$, and similarly for $\tilde{s}_q$), and hence there are 4×4 covariance plots, such as in Fig. 3-2(a)) for the covariance of $\Re s_q$ with itself. In all cases we find that the off-diagonal terms are small; the only exceptions are at small q where we expect end effects to be most pronounced.

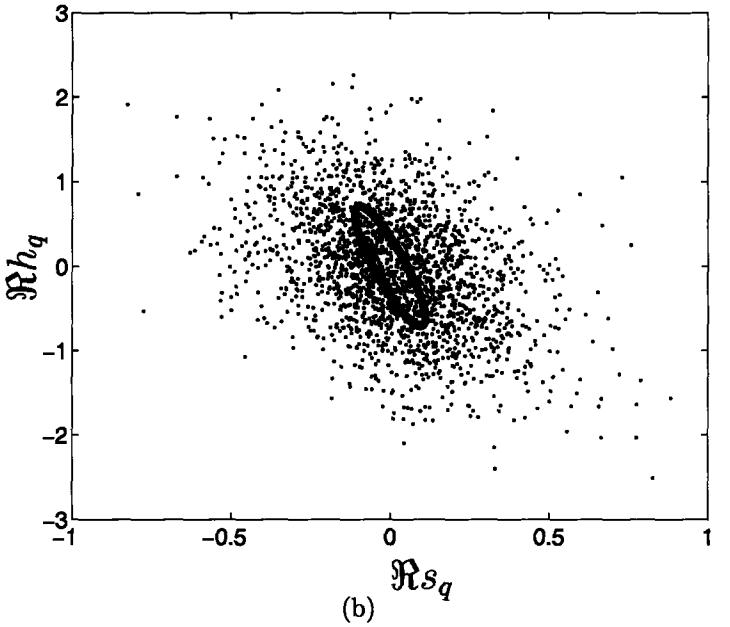
One can make a similar case for the independence of the real and imaginary components at a given q. (For cyclic structures the phase is arbitrary.) The real (imaginary) components are, however, correlated as illustrated by the scatter plot of $(\Re h_q, \Re s_q)$ for $q = 0.9$ in Fig. 3-2(b). We made similar scatter plots for different values of q in the interval 0 to π, with similar results which were well fitted by Gaussian forms. Based on these results, we describe the joint probability distribution in Fourier space by the multivariate Gaussian form

$$\begin{aligned} P(\{\tilde{h}_q, \tilde{s}_q\}) &= \prod_q \exp\left[-\frac{(\Re s_q)^2}{2A_q} - \frac{\Re s_q \Re h_q}{B_q} - \frac{(\Re h_q)^2}{2C_q}\right] \\ &\times \exp\left[-\frac{(\Im s_q)^2}{2A'_q} - \frac{\Im s_q \Im h_q}{B'_q} - \frac{(\Im h_q)^2}{2C'_q}\right], \end{aligned} \tag{3.2}$$

with the parameters plotted in Fig. 3-3. If the probabilities depend only on the separation $i - j$ between sites, the real and imaginary Fourier amplitudes should follow the same distribution. In our fits we allowed the corresponding parameters to



(a)



(b)

Figure 3-2: (a) Covariance of $\Re s_q$ with $\Re s_{q'}$. There is very little correlation between off-diagonal terms. (b) Scatter plot of $\Re h_q$ versus $\Re s_q$ for $q = 0.90$, and the half-width half-maximum locus of a Gaussian fit (solid line).

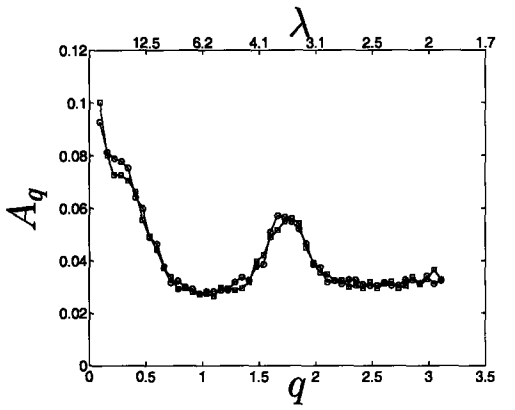
be different to obtain a measure of the accuracy of the model and the fitting procedure. As indicated in Fig. 3-3 the resulting values of real and imaginary amplitude are quite close, differing by less than 5%.

We interpret $\{A_q\}$ and $\{C_q\}$ as measures of *intrinsic* tendencies of hydrophobicity and surface exposure profiles, while $\{B_q\}$ is inversely proportional to the strength of the interactions that correlate them. In the absence of any such interactions, $\{A_q\}$ and $\{C_q\}$ would be the same as the power spectra in Fig. 3-1. With this in mind, let us now examine these plots in more detail. They are related to the original variables $< |s_q|^2 >$ and $< |h_q|^2 >$ through:

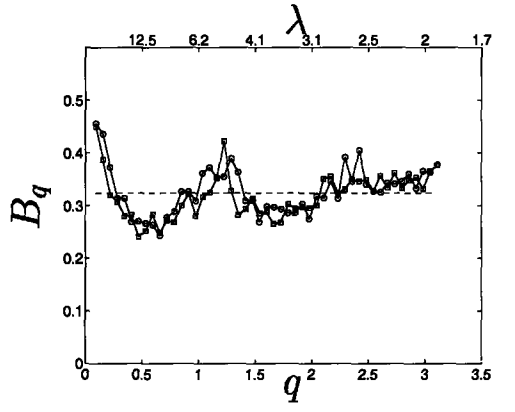
$$\begin{aligned} < |s_q|^2 > &= \frac{A_q}{1 - \frac{A_q C_q}{B_q^2}} + \frac{A'_q}{1 - \frac{A'_q C'_q}{B'^2_q}} \\ < |h_q|^2 > &= \frac{C_q}{1 - \frac{A_q C_q}{B_q^2}} + \frac{C'_q}{1 - \frac{A'_q C'_q}{B'^2_q}} \end{aligned} \tag{3.3}$$

In the absence of interactions, or $1/B_q \to 0$, these equations reduce to $< |\tilde{s}_q|^2 >= A_q + A'_q$ and $< |\tilde{h}_q|^2 = C_q + C'_q$. As shown in Fig. 3-3, the average value of B_q for the data is $\simeq 0.32$. As a result, the α-helix peak in $< |\tilde{s}_q|^2 >$ is 42% larger than the peak in A_q, and also an α-helix peak is induced in $< |\tilde{h}_q|^2 >$ as result of the peak in A_q even though there is no peak in C_q. With the current value of $B_q \simeq 0.32$, the peak in A_q is magnified by 42% in $< |s_q|^2 >$, and a peak is induced in $< |h_q|^2 >$ because of existing peak in A_q.

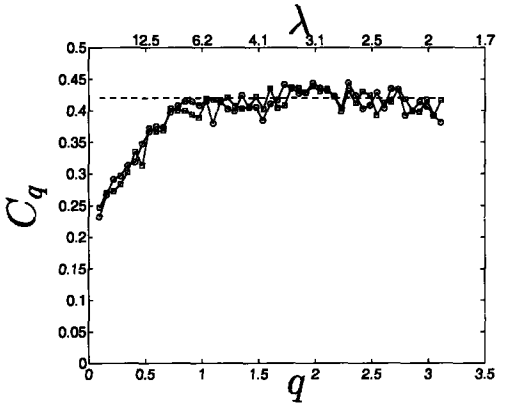
The prevalence of α-helices in structures is reflected in the peak at $\lambda = 3.6$ in Fig. 3-3(a). As a check, we repeated the analysis for 493 proteins in our database that are classified as mainly β by CATH [72]. The α-helix peak disappears completely for this subset, and a weaker peak corresponding to β strands at $\lambda = 2.2$ (which was not visible in Fig. 3-3(a)) emerges as a weak peak. This may indicate that the formation and arrangement of β strands is less influenced by hydrophobic forces. The other prominent feature of Fig. 3-3(a) is the increase in A_q as $q \to 0$. We believe this reflects the fact that at a coarse level the protein is a *compact polymer*; it is well known that polymer statistics leads to long-range correlations in the statistics of segments



(a)



(b)



(c)

Figure 3-3: *Intrinsic* variances of solvent accessibility and hydrophobicity profiles are described by A_q and C_q respectively, while B_q is related to the interaction that correlates them. The square and circle symbols correspond to the parameters of the imaginary and real components, respectively. These figures are calculated for our set of 1461 proteins. Dashed lines indicate respectively the average value of B_q [in (b)], and the asymptotic behavior of C_q [in (c)].

in the interior of a compact structure [34]. While the precise manner in which this could lead to correlations as in Fig. 3-3(a) has not been worked out, we note that similar effects have been observed before in studies of protein-like structures in three dimensions [47], and compact lattice polymers in two dimensions [98].

The α-helix peak, which is prominent in the hydrophobicity power spectrum of Fig. 3-1(b) is absent from Fig. 3-3(c). Thus, the observed periodicity in sequence data is not an intrinsic feature of the amino-acid profiles, but dictated by the required folding of structures. If the sequence of amino-acids were totally random, we would expect a distribution $P(\{h_i\}) = \prod_i p_a(h_i)$, where $p_a(h_i)$ indicates the frequency of a particular base. The corresponding distribution in Fourier space would also be independent of q. The observed $\{C_q\}$ are indeed constant (approximately 0.42 ± 0.02), at large q. The value of $C_q + C_q^*$ is different from the average indicated in Fig. 3-1(b), with the assumption that the amino-acids are distributed randomly. This difference is due to the interaction term in equation 3.2.

Reduced values of C_q are observed as $q \to 0$, corresponding to large periodicities, as seen in Fig. 3-3(c). A similar feature is also present in the power spectrum in Fig. 3-1(a), as noted before by Irback *et al.* [41] who suggest that anti-correlations can be advantageous for removing the degeneracies of ground state for folding sequences. More recent studies also indicate that long stretches of hydrophobic monomers, which could be a source of long range positive correlations, are avoided [81]. Further investigations of this issue would be helpful.

Finally, we note that the interaction terms $\{B_q\}$ in Fig. 3-3(b) which correlate sequence and structure profiles (at different periodicities) are approximately constant. As $\sum_q \tilde{h}_q \tilde{s_q}^* = \sum_i h_i s_i$, these terms can be regarded as arising from the Boltzmann weight $\exp[-E/(k_B T)]$ of a solvation energy $E = \sum_i h_i s_i$ at some temperature T. Using $\overline{B_q} \approx 0.32 \pm 0.03$ kcal/mol, we can extract a corresponding temperature of $T = (2\overline{B_q})/k_B = 323 \pm 30$°K. Interestingly, this fictitious T is around room temperature, i.e. in the range of temperatures where most proteins fold and function. This indicates that an important factor in correlating sequence hydrophobicity and structural solvent accessibility is indeed the free energy of solvation. This conclusion is also consistent

with the analysis done by Miller [64], which estimated differences in the free energies of amino-acids between the surface and the core of the proteins by counting their relative frequencies in the different locations. Finkelstein*et al* [23] provides a more thorough discussion on why we expect this fictitious temperature to be near room temperature.

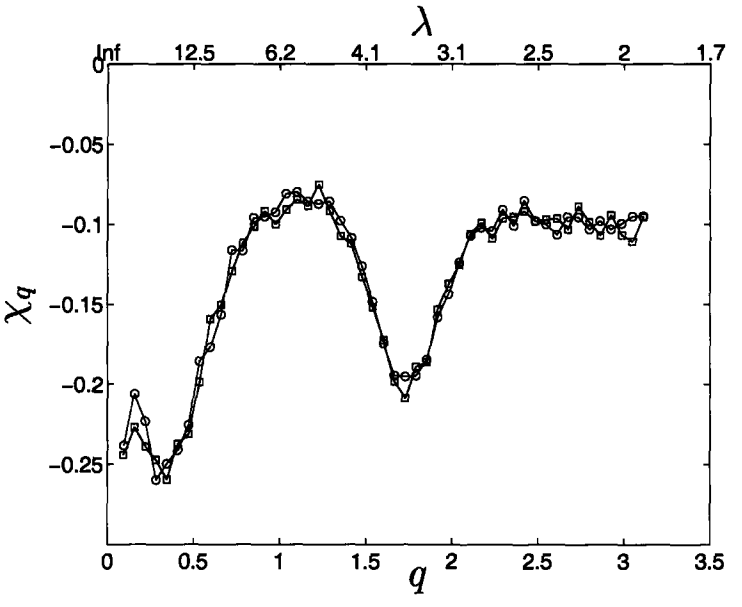


Figure 3-4: The susceptibility χ_q is negative since the more hydrophobic monomers tend to be in less solvent exposed sites. The circle and square symbols correspond to real and imaginary components, respectively.

In principle, the Gaussian distribution in Eq. 3.2 can be used as a tool for predicting structures, at least as far as their surface exposure profile is concerned. Given a specific sequence, we can calculate the hydrophobicity profile $\{h_i\}$, and the corresponding $\{\tilde{h}_q\}$. The conditional probability for surface exposure profiles is then given by

$$\begin{aligned} P(\{\tilde{s}_q|\tilde{h}_q\}) &= \prod_q p(\tilde{s}_q|\tilde{h}_q) \propto \\ &\prod_q \exp\left[-\frac{(\Re s_q - \chi_q \Re h_q)^2}{2\sigma_q^2} - \frac{(\Im s_q - \chi'_q \Im h_q)^2}{2{\sigma'_q}^2}\right]. \end{aligned} \tag{3.4}$$

Thus $\tilde{s}_q$ is Gaussian distributed with a mean value of $\chi_q \tilde{h}_q$ and a variance σ_q^2 with 'susceptibility' χ_q and 'noise' σ_q easily related to (A_q, B_q, C_q). χ_q is plotted in Fig. 3-4. The corresponding distribution of $\{s_i\}$ in real space is then obtained by Fourier transformation.

3.3 Conclusions

We investigated correlations between protein sequences and structures due to hydrophobic forces by application of Fourier transforms to profiles of hydrophobicity and solvent accessibility. Each Fourier component is separately well approximated by a Gaussian distribution; their joint distribution is described by a product of multivariate Gaussians at different periodicities. This approach enables us to separate the intrinsic tendencies of the profiles from the interactions that couple them. We thus find that α-helix periodicity is an intrinsic feature of structures and not sequences, and that at long periods the structural profiles are more correlated than average, while the sequences are less correlated. A quite satisfying outcome is that the correlations between the two profiles can be explained by the Boltzmann weight of the solvation energy at room temperatures.

Our joint distribution can be used in applications such as predicting solvent accessibility from hydrophobicity profiles [68], or protein interaction sites [27]. Incorporating the impact of correlations within solvent accessibilities is likely to improve predictions. The distribution can also be used in analytical approaches to protein folding, wherever there is a need for taking into account the complexities of structure and sequence space.

Chapter 4

Could solvation forces give rise to designability in proteins?

4.1 Introduction

Protein structure classes are populated by very different numbers of observed proteins. The structures representing classes with a high number of protein folds are called highly designable protein structures [53, 71, 25]. A fundamental question yet to be answered is whether this is due to the natural evolution, geometrical constraints in folding, or the nature of interaction forces.

Model protein simulations have shown that most sequences fold to only a few highly designable structures [31, 54]. In these studies, based on ground-state search, the Hamiltonian is often simplified to include only short-range pairwise interactions, $H(\{\sigma_i\},\{r_i\}) = \sum_{i<j}^{N} U(\sigma_i,\sigma_j)\Delta_{ij}$, where Δ_{ij} (the (i,j) element of contact matrix, $\mathbf{\Delta}$) is unity if monomers i and j are in contact but not adjacent along the chain, and otherwise is zero. $U(\sigma_i,\sigma_j)$ is the contact energy between monomer types σ_i and σ_j.

The pair contact potentials have been evaluated for all possible pairs of the 20 amino acids based on the frequency of occurrence of pair contacts in native structures in the Protein Data Base [65]. Eigenvalue decomposition of the 20×20 interaction matrix shows [55] that U can be approximated as $U(\sigma_i,\sigma_j) = -(\sigma_i + \sigma_j + \lambda\sigma_i\sigma_j + E_c)$, with only 22 independent energy parameters, including 20 σ's corresponding to

different amino acid types, a mixing parameter λ, and a residue-independent contact energy E_c. It also shows that σ's can be divided to hydrophobic and polar residues. When considering completely compact structures, E_c can be set to zero as its only effect is a constant shift in the energy spectrum of sequences that fold to a given structure. Since $\lambda\sigma_i\sigma_j$ is small compare the additive terms, we set the λ to zero to study the influence of the additive terms. Setting $\lambda = 0$ reduces the pair contact model to a solvation model [20]:

$$H(\{\sigma_i\}, \Delta) = -\frac{1}{2}\sum_{i,j}^{N}(\sigma_i + \sigma_j)\Delta_{i,j} = -\sum_{i}\sigma_i b_i \tag{4.1}$$

where $b_i = \sum_j^N \Delta_{i,j}$. $\mathbf{b}$ is called the contact vector [82, 43], and b_i represents the degree to which residue i is buried in the protein structure.

Recently, Kussell and Shakhnovich [48] have demonstrated with an elegant analytical technique that when only additive forces are present, the designability of all structures will be the same given that the assumptions of the Random Energy Model are upheld. Their result was consistent with the findings of Ejtehadi *et al* [20] for lattice structures on a $3 \times 3 \times 3$ cube. However, there is evidence of non-uniform designability for longer chains, in both two and three dimensions [56, 11].

A principal assumption of REM is the statistical independence of the energies of states (protein structures) over disorder (sequences) [73]. We examine this independence assumption in a general form when solvation forces are the dominant forces. To control correlation between structures we introduce a method of using the Ising model to construct pseudo-structures and show that by breaking the statistical independence between the states, highly designable structures emerge. By comparing our results from Ising pseudo-structures with both lattice and natural protein structures, we conclude that solvation forces alone can be sufficient to result in the emergence of highly designable protein structures.

4.2 Methods & Results

To rederive the REM prediction for equal designability of all structures within the solvation model, we start with the approximation of Ref. [48] for the designability of a given structure Δ,

$$\mathcal{D}(\Delta) = \sum_{E,C} p(E,C) n(E,\Delta,C), \tag{4.2}$$

where $n(E,\Delta,C)$ is the energy spectrum of structure Δ for all sequences with fixed composition C (the fraction of number of hydrophobic residues in the chain).

$p(E,C)$ is the probability for a sequence with composition C to fold to a structure Δ with folding energy E. We coarse grained the structure-dependent contact vector $\mathbf{b}$ to a binary vector with zero elements for surface sites and 1's for buried sites. For a structure with N_c core sites and N_s surface sites ($N_c + N_s = N$),

$$n(E,\Delta,C) = \begin{cases} \binom{N_c}{-E}\binom{N_s}{C-(-E)}, & \text{if } -E < C; \\ 0, & \text{if } -E > C; \end{cases} \tag{4.3}$$

Given that N_c is almost the same for all completely compact structures with a given length N, $n(E,\Delta,C) \approx n(E,C)$ with no dependence on structure Δ. As a result, $\mathcal{D}(\Delta)$ is independent of Δ, and the designability is the same for all structures, visible as a sharp peak in a histogram plot of the designability of all structures. In practice, very narrow and sharp Gaussian curves are consistent with this approximation (see Fig. 4-1).

To investigate the breakdown of REM, we examine the statistical dependence of energies by calculating the covariance between two arbitrary structures α and β [73], $\langle \delta E^\alpha, \delta E^\beta \rangle_{\boldsymbol{\sigma}} \equiv \langle E^\alpha E^\beta \rangle_{\boldsymbol{\sigma}} - \langle E^\alpha \rangle_{\boldsymbol{\sigma}} \langle E^\beta \rangle_{\boldsymbol{\sigma}}$, where
$E^\alpha = -\sum_{i=1}^{N} \sigma_i b_i^\alpha$ and the average is over all feasible sequences, $\boldsymbol{\sigma}$. We then obtain $\langle \delta E^\alpha, \delta E^\beta \rangle_{\boldsymbol{\sigma}} \equiv B^2 Q_{\alpha,\beta}$ where $B^2 = \langle \boldsymbol{\sigma}^2 \rangle - \langle \boldsymbol{\sigma} \rangle^2$ and $Q_{\alpha\beta} = \sum_i b_i^\alpha b_i^\beta$. Defining $P(Q) = \sum_{\alpha,\beta} \delta(Q - Q_{\alpha\beta})$, we obtain

$$\langle Q \rangle = \langle Q_{\alpha,\beta} \rangle_{\alpha,\beta} = \sum_i \langle b_i \rangle^2 \tag{4.4}$$

The maximum correlation happens between a structure with itself: $Q_{max} = \sum_i \langle b_i^2 \rangle$. To measure the degree to which a system is correlated, we calculate $\langle \gamma \rangle = \langle Q/Q_{max} \rangle = \frac{\sum_i \langle b_i \rangle^2}{\sum_i \langle b_i^2 \rangle}$. $\langle b_i \rangle$ and $\langle b_i^2 \rangle$ do not change significantly as a function of i, except for i's near the ends of the proteins. This observation can further simplify the equation to $\langle \gamma \rangle = \frac{\langle b \rangle^2}{\langle b^2 \rangle}$. For example, for a structure with M core sites ($b = 1$) and $N - M$ surface sites ($b = 0$), we have $\langle b \rangle = \langle b^2 \rangle = \frac{M}{N}$, yielding $\langle \gamma \rangle = \frac{M}{N}$.

To be able to control the correlations within our structures, we use pseudo structures generated using an one-dimensional Ising model, with 1's resembling the core sites and 0's the surface sites. The geometrical constraints in the protein structures are reduced to site correlations in a one-dimensional chain. We only constrain the size and the magnetization of the system to have the same core to surface ratio of the geometrical structures. For example, in the case of 6×6 square lattice structures, we set the string length N to 36 and set the number of 1's in each string to $M = 16$. The interaction among the monomers is defined as $E = -J \sum_{i=1}^{N-1} b_i b_{i+1}$, where J is the interaction constant (or reciprocal temperature). This choice of interaction creates a positive correlation along the structure; a residue that is on the surface (core) will tend to have its neighbor on the surface (core) as well. Positive correlations in contact vector or solvent accessibility of lattice proteins and off-lattice proteins have been observed in previous work [98, 46]. Running a Monte Carlo simulation of the Ising model, we generate a set of strings $\mathbf{b}(t)$, which is a function of time. The final pseudo-structure set is created by sampling from $\mathbf{b}(t)$. The sampling rate is taken to be larger than the relaxation time of the simulations to ensure that the space of structures is sampled uniformly. We generate sets of Ising pseudo-structures with different values of $J = 0, 1, 3, 6$. Only the set generated by setting $J = 0$ has the same properties as a set generated by randomly putting $M = 16$ ones in $N = 36$ slots of each structure vector. The value $J = 3$ is chosen because the average energy of sampled structures, $\langle \sum_i b_i b_{i+1} \rangle$, is the same as 6×6 lattice structures. The measured results of $\langle \gamma \rangle$ using simulation are shown in Table 4.1. For all cases, the values are very close to the predicted values of $\langle \gamma \rangle \simeq \frac{M}{N} = \frac{16}{36} = 0.44$.

The designability histograms for the Ising sets with $J = 0$ and 3 and the set

data set	$\langle\gamma\rangle$	var(γ)	$\langle\mathcal{D}\rangle$	max($\mathcal{D}$)	std($\mathcal{D}$)	$\langle E\rangle$
Ising J=0	0.44	0.088	455	646	44	-6.86
Ising J=1	0.44	0.089	453	750	65	-9.00
Ising J=3	0.44	0.15	374	5263	459	-12.30
Ising J=6	0.44	0.28	324	4618	478	-14.69
2D 6x6	0.44	0.14	209	2938	185	-12.54

Table 4.1: The values of average and variance of γ and the designability $\mathcal{D}$ for different Ising structure sets and 6×6 square lattice structures. The values reported for variance of γ are from simulation, these values differ less than one percent with the values obtained from Eq.4.5 or Eq.4.6.

from the 6×6 square lattice are plotted in Fig. 4-1. Even though $\langle\gamma\rangle$ is the same for all cases, they show different designability characteristics. In the case of $J = 0$, the designability has a Gaussian shape, while for $J = 3$, the designability curve drops almost exponentially, similar to the designability histogram of the 6×6 square lattice. (R-squares of the fits are reported in Table 4.1.) In the case of $J = 3$, similar to the 6×6 square lattice, the highly designable structures are far more designable than other structures. This is distinctly different from the REM prediction for the designability plot and suggests that the REM assumption might have been violated. Since $\langle\gamma\rangle$ is the same for both the $J = 0$ and $J = 3$ case, the difference should be due to the different variance of γ.

The difference in the low-designability part of the histogram is due to the lack of long-range geometrical correlations in our Ising set. A model which takes into account longer range correlations can reproduce the designability plot of lattice proteins more accurately [98]. Nonetheless, a simple nearest-neighbor Ising model is sufficient to examine how the designability histogram depends on correlations among the structures. Our simple Ising pseudo-structures, generated by tuning only one correlation parameter J, highlights this dependency. Two clearly different phases are visible: in the low correlation regime, the REM prediction is obtained, and in the high correlation regime it is not.

Even though the average value of γ ($= Q/Q_{max}$) depends on the number of core sites and the length of the chain, it can be seen from Fig. 4-2 that its variance does not. Calculating $Var(\gamma) = Var(Q)/Q_{max}^2$ shows that the variance of *interstructural*

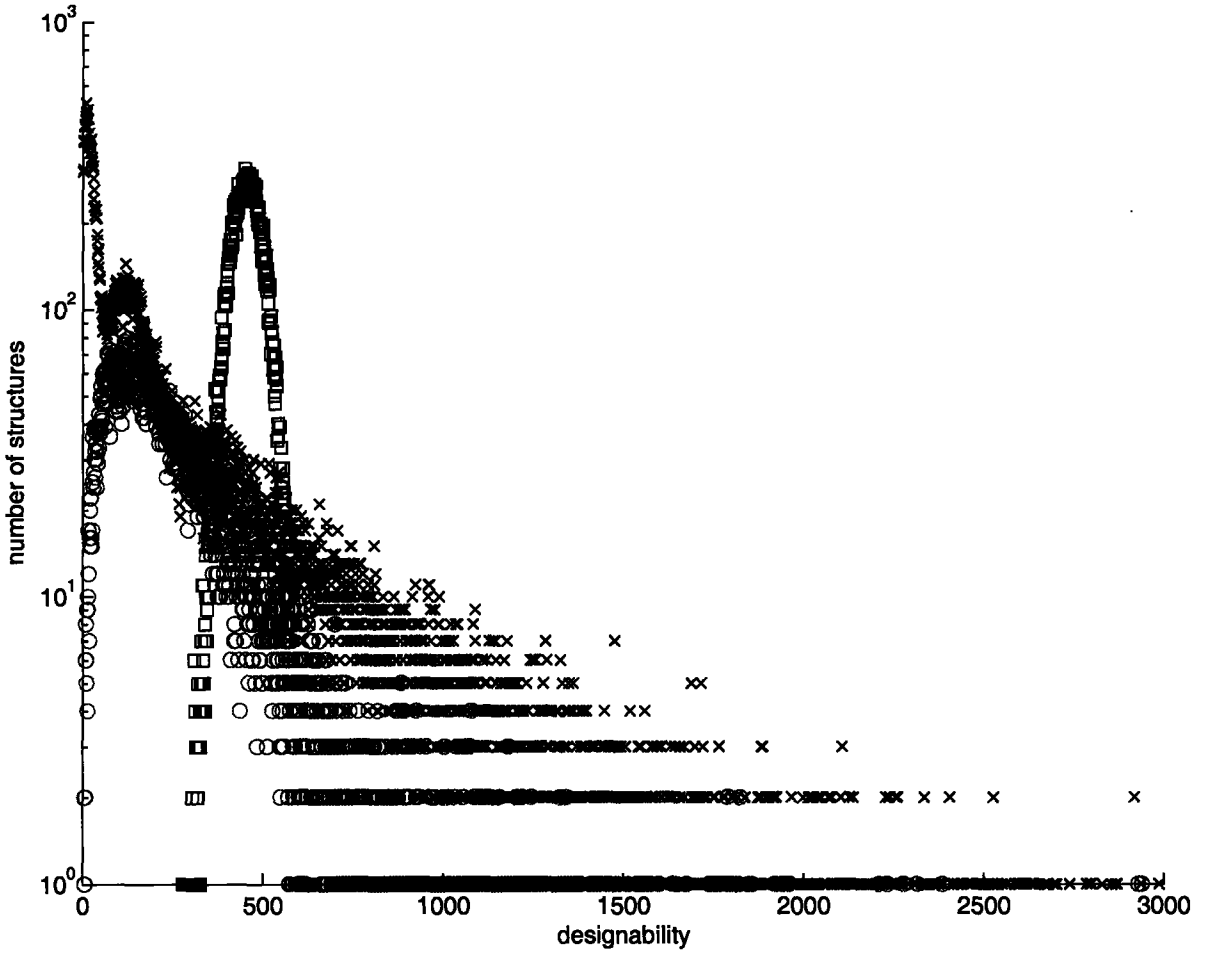


Figure 4-1: Designability of structures for different sets of data is plotted. For each set 50 million sequences were sampled.

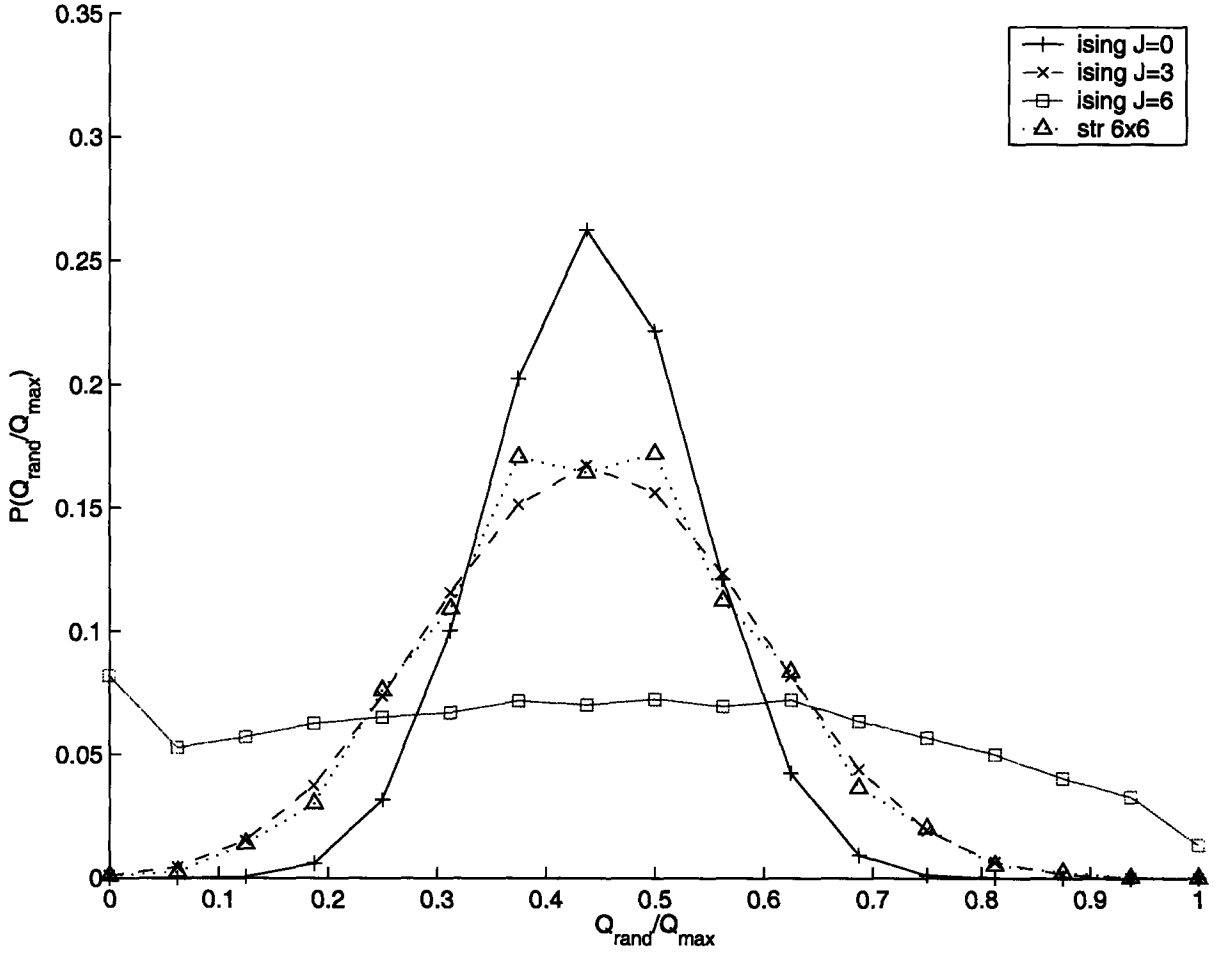


Figure 4-2: Histogram of $\frac{Q}{Q_{max}}$ for different sets of structures.

correlations can be related to *intrastructural* correlations C_{ij}:

$$\begin{aligned} Var(Q) &= Var(\sum_i b_i^\alpha b_i^\beta) \\ &= \sum_{i,j} \langle b_i^\alpha b_j^\alpha \rangle_\alpha \langle b_i^\beta b_j^\beta \rangle_\beta - \sum_i \langle b_i^\alpha \rangle_\alpha^2 \langle b_i^\beta \rangle_\beta^2 \\ &= \sum_{i,j} C_{ij}^2 + 2 \sum_{i,j} C_{ij} \langle b_i^\alpha \rangle_\alpha^2 \langle b_j^\beta \rangle_\beta^2 \\ &\simeq \sum_{i,j} C_{ij}^2, \end{aligned} \tag{4.5}$$

where $C_{ij} = \langle b_i^\alpha b_j^\alpha \rangle_\alpha - \langle b_i^\alpha \rangle_\alpha^2$. The last step was done using the approximate translational invariance of $\langle b_i^\alpha \rangle_\alpha = const$ as well as assuming that the average solvent accessibility is the same for all protein structures with length N. This assumption is exact for compact lattice proteins and is a reasonable approximation for natural proteins. The result is $Var(\gamma) \simeq \sum_{ij} C_{ij}^2 / M^2$. A comparison to the actual value of the variance of γ for the data sets reported in Table 4.1 shows that these approximations lead to a result with less than 1% error. The above equation shows how correlations within structures, C_{ij}, which are controlled by J, can affect the correlations between structures by increasing the variance of γ. Increasing the variance of γ results in more structures in the system with a considerable correlation, breaking the independence assumption of REM and resulting in the emergence of power-law designability plots.

Table 4.1 shows that $J = 3$ and 6×6 structures have nearly equal energy averages. It can be seen from Fig. 4-2 that they also have similar γ distributions. Even though there are only a few common structures between these two sets ($\sim$ 6%), they have very similar designability distributions (Fig. 4-1). This indicates how most of the information regarding the overall shape of the designability plot is embodied in the correlations among monomers, and how most of those correlations can be approximated by a simple nearest neighbor Ising-type interaction. We also observed that the designability of each structure from the Ising set anti-correlates with its Ising energy (data not shown). Analogously to 2D-lattice structures, this means that those structures which frequently switch between surface and core sites are more designable

[56, 48].

Since 2D geometrical constrants were not considerd in constructing the Ising pseudo-structures, we believe the result of Ising pseudo-structures can be compared with 3D protein structures with a similar correlations along their chain. The geometrical dimension and constraint control the correlations, number of structures, and the ratio of the core to surface residues. In the case of compact $3 \times 3 \times 3$ structures, the designability histogram is fairly sharp and conforms to the REM prediction. It has $\langle\gamma\rangle = \frac{7}{27}$ and the simplicity of the geometry allows the calculation of designability by combinatorial methods [19]. However, increasing the length of the chain changes the story. Recently, Cejtin *et al* [11], in a very huge enumeration study for all compact structures in a $4 \times 3 \times 3$ cube, reported an exponentially decaying designability, in contradiction with the REM prediction. We have seen that an Ising data set with $J = 1$ was able to reproduce the same designability distribution. Here, because the number of structures is larger, a lower value of γ ($= 12/36$) can violate the REM independent energy assumption. We observed a Gaussian distribution of designability for pseudo-structures of this length, corresponding to very small values of J.

It is widely believed that the solvation energy plays an important role in the stability of protein structures [14, 38]. Based on the above, we try to estimate whether REM is applicable to natural proteins if a solvation (additive) potential is the dominant force in folding.

In the case of natural proteins, it is difficult to calculate the covariance matrix since proteins have different lengths. We convert Eqn. 4.5 to its equivalent in Fourier space, which can be easily applied to proteins with different lengths.

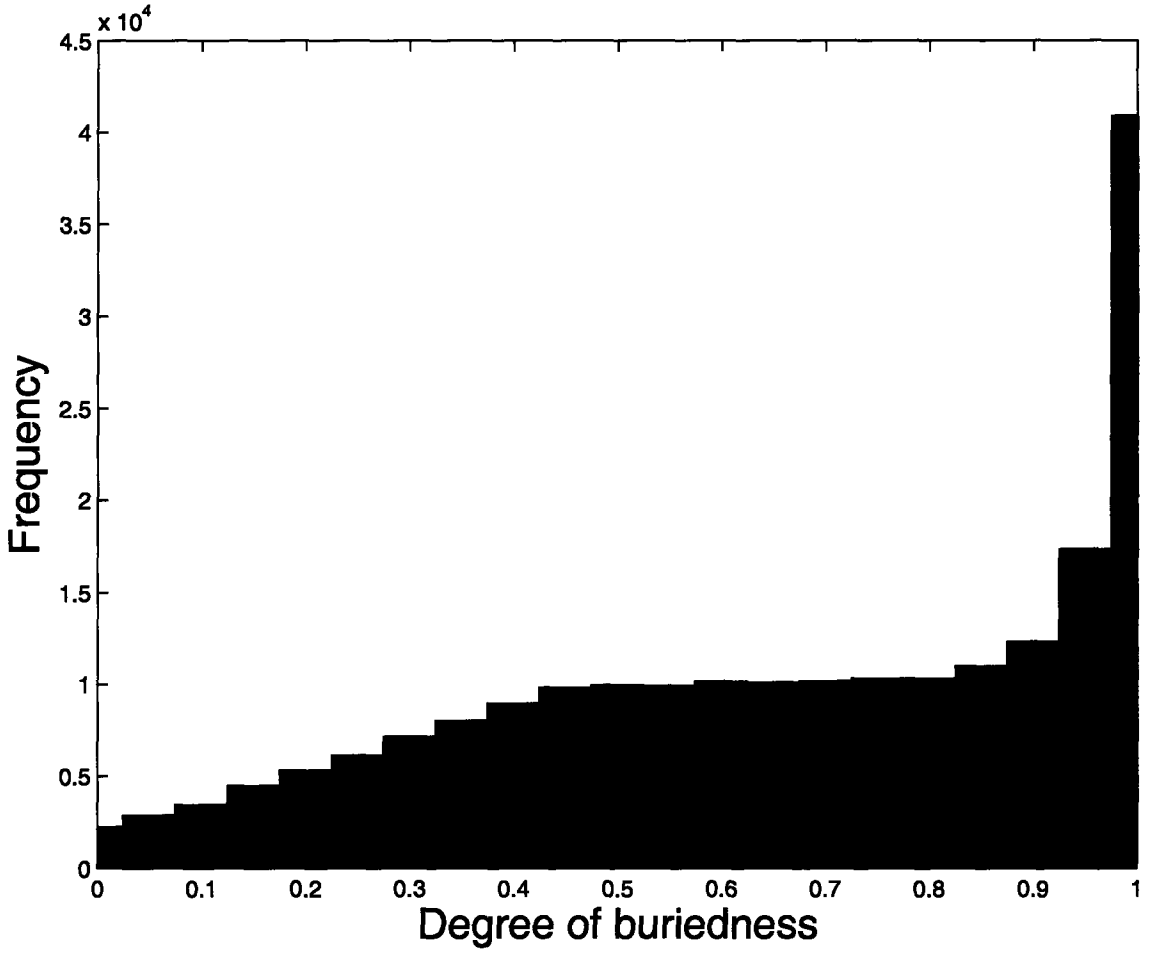


Figure 4-3: Histogram of degree of buriedness for all residues in 1461 protein structures taken from representative set of FSSP database.

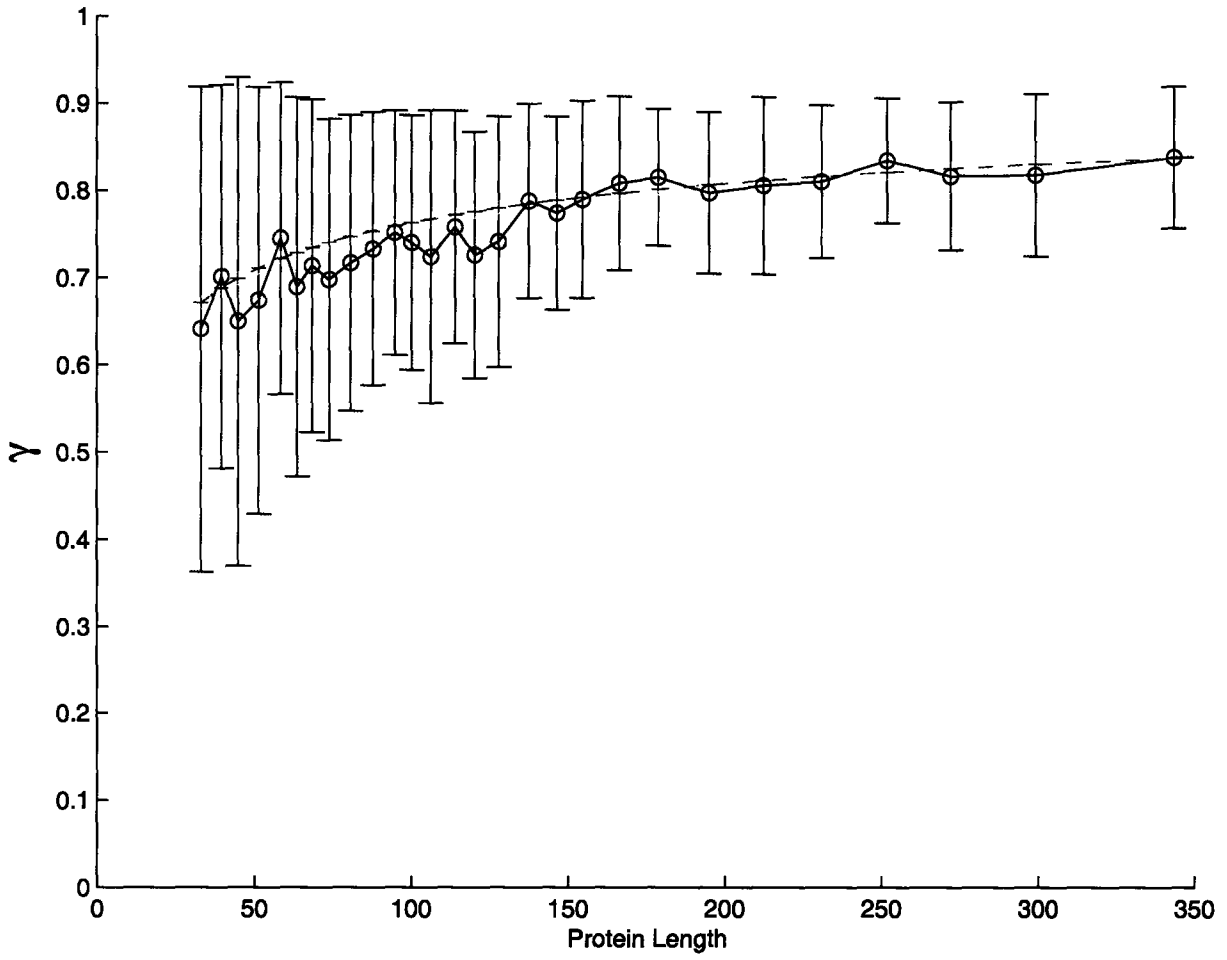


Figure 4-4: Circles represent the values of γ for natural proteins as a function of length. Dashed line is a fit based on a theoretical prediction.

$$
\begin{aligned}
Var(Q) &= \sum_{i,j} C_{i,j}^2 \\
&= \sum_{q,q'} |C_{q,q'}|^2 \text{ (Parseval Theorem)} \\
&\simeq \sum_{q,q'} |C_{q,q}\delta_{q,q'}|^2 \\
&= \sum_{q} C_{q,q}^2 \simeq \sum_{q} |b_q|^4
\end{aligned}
\tag{4.6}
$$

The approximation was made because off-diagonal terms for $C_{qq'}$ are small compared to the diagonal terms[98].

To estimate the mean and variance of γ for natural proteins, we selected 2200 representative chains from the Dali/FSSP database. Any two protein chains in this set have more than 25 percent structural dissimilarity. We removed all the multi-domain chains by using the CATH domain definition database, leaving 1461 protein chains [37, 72, 29]. We used the *relative solvent accessibility* reported by NACCESS [39] to generate solvent accessibility profiles $\{s_i\}$. The relative solvent accessibility is the ratio of the solvent accessibility of a residue in the protein's native structure to the solvent accessibility of that residue in an extended tripeptide ALA-X-ALA, for each amino acid type X. We converted solvent accessibilities to our vector of buriedness b, through $b_i = 1 - s_i$. The histogram of b_i for all residues in all studied proteins is shown in Fig. 4-3. In order to calculate $\langle\gamma\rangle = \frac{\langle b_i\rangle^2}{\langle {b_i}^2\rangle}$, we calculated $\langle b_i\rangle^2$ and $\langle {b_i}^2\rangle$ for each protein separately. Since γ depends on the protein length N, we calculated γ for different length intervals separately (Fig. 4-4). Error bars are based on two type of variations: 1)the variation in calculation of $\langle\gamma\rangle$ because of the different estimates in different proteins. 2) the variation resulting from Eqn. 4.6 because of correlations within the solvent accessibility profiles of proteins.

If we assume a compact protein is simply a sphere with radius R filled with monomers with radius a, then $R/a \simeq N^{1/3}$. Then γ can be estimated as $\frac{M}{N} = \frac{((R/a)-w)^3}{(R/a)^3}$, where w is a free parameter representing the width of the effective surface

layer. The best fit to the data in Fig. 4-4 is achieved by $w = 0.4$. Even though we do not have an accurate value for the critical γ separating the phases in which REM works or does not, the values of $\langle\gamma\rangle$ in Fig. 4-4 are too high to assume energy independence in the proteins studied. This suggests that REM is not able to predict the designability distributions observed in real proteins.

4.3 Conclusion

In summary, using a finite length one dimensional Ising model we have produced pseudostructures that reproduce the overall designability characteristics of lattice proteins. Our Ising interaction parameter J has been used as a correlation control parameter. Considering an additive (solvation) potential, we have calculated and compared the designability of the pseudo-structures with compact lattice structures. We have shown that increasing intra-structural correlations can create inter-structural correlations, which causes the energy independence assumption of REM to break down, both in two and three dimensions. This breakdown allows the emergence of an exponentially decaying designability–similar to what has been observed in natural proteins– even when the design interaction potential is additive (or solvation). This shows that the correlations in the solvent accessibility profiles of protein native structures are responsible for generating the observed designability behavior. Based on our model, because of the high correlation of protein structures, in contradiction with the REM prediction, solvation forces could be sufficient to describe the observation of highly designable protein structures.

Chapter 5

Finding Differential Motifs

5.1 Introduction

Understanding how gene expression is regulated is one of the main goals of molecular cell biology. In recent years, considerable computational effort has been devoted to detecting motifs– short patterns in biological sequences that correspond to functionally important elements in DNA sequences. Common examples of motifs include transcription factor binding sites in the promoter regions of co-regulated genes and exonic or intronic splicing enhancers.

Motif-finding methods rely on finding over-represented patterns in DNA sequence data; these patterns are likely to be functionally important. Common methods include the maximization of a likelihood function for motifs represented by weight matrices [86, 51, 2, 79] , exact word-counting [90, 84] or a combination of the two. Since these methods try to find a functionally relevant (specific) motif in a set of sequences that share a functional property by simply looking for over-represented patterns, they are liable to being misled by other, functionally irrelevant (non-specific) patterns that are over-represented across the genome. To avoid this problem, users are forced to find not just the strongest pattern, but as many patterns as possible, and then filter them based on the magnitude of enrichment of the motif in the set of sequences that share that functional property versus another set that does not [89]. This is not necessarily effective, because the non-specific motifs can overlap with the specific

ones and prevent identification of the weaker but functionally relevant ones.

For example, when looking for exonic splicing enhancers (ESE's), a set of exons with weak splice sites are used, since it is expected that they contain ESE's to compensate for the weak splicing site signals. Classical motif finder will return functionally irrelevant patterns that result from the biased codon usage in exons. To overcome this problem, another "negative" set, including exons with strong splice sites believed not to contain ESE's, is used to measure the enrichment. Such sets were used to find ESE's using exact word-counting methods [21]. Tompa has offered a more general method which allows mismatches in words. These methods share the common problems of word-based methods, which are limited to short sequence segments because they slow down exponentially with increasing motif length. Also, this type of motif finders requires a post clustering of the over-represented words to obtain the motifs. This post processing requires ad hoc assumptions to define the boundary of clusters.

Matrix-based methods can offer many advantages. For example, they are more sensitive to degenerate motifs. It has also been shown that they are related to the binding potentials in protein-DNA interaction. At least two matrix-based methods have been developed to improve motif-finding sensitivity. Workman and Stormo have offered a neural network based algorithm to suppress motifs that are both in the positive set and the negative set [97]. Wareham *et. al.* have suggested a heuristic method to improve the sensitivity of motif finding that avoids low-complexity sequences that might appear in the background. We offer a method which is an extension of the classical Gibbs sampler. This method incorporates the effect of a negative sequence set while it is searching for motifs in a positive set. This method is similar to that of Workman and Stormo method in using the concept of a partition function to measure the existence of the motif in the negative set. However, our method is not based on neural networks and can be incorporated in many motif finders that are weight-matrix based.

In this chapter, we develop the analytical framework for a differential motif finding, expanding on the classical Gibbs motif finder. In our method, the classical Gibbs sampling motif finder is modified to suppress motifs that have strong matches in the

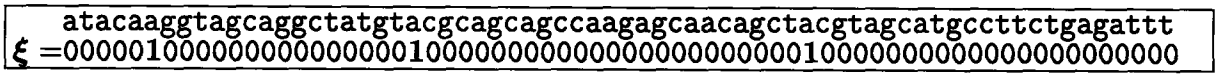


Figure 5-1: A representation of the vector $\boldsymbol{\xi}$.

negative sequence set. Our method can be applied to cases in which there is only one occurrence per sequence as well as when there are variable numbers of instances. We examine the basic assumptions in this method. We also show that this method outperforms the classical Gibbs sampler in finding differentially-enriched motifs.

5.2 Methods

5.2.1 Notation

In the methods section, first we introduce our notation and describe the formalism of the classical Gibbs Sampling algorithm using our notation. Next, we explain our formulation of a differential Gibbs Sampler.

In our notation, a sequence is represented by a binary matrix. For example, the sequence of ACGCCT is represented by:

$$\mathbf{u} = \begin{pmatrix} A: & 1 & 0 & 0 & 0 & 0 & 0 \\ C: & 0 & 1 & 0 & 1 & 1 & 0 \\ G: & 0 & 0 & 1 & 0 & 0 & 0 \\ T: & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.1}$$

The matrix elements are identified by u^{mn}, where m is the index of the columns and n the index of the rows. In a long sequence S, we represent a subsequence, or a motif occurrence, with width w using $\boldsymbol{u}_i$, where i is the start location of the subsequence in the sequence S. Again, the elements in this subsequence are identified using superscripts such as u_i^{mn}. In this representation, the first motif occurrence in the sequence of fig.5-1 can be represented by $\boldsymbol{u}_i$, and the second letter of that subsequence, a G, can be identified by $u_6^{2,3}$.

Motifs are also represented with weight matrices. Θ^{mn} is the logarithm of the

probability of having letter n in position m of the motif. In this model different columns in the weight matrix contribute independently to the motif. The matching score of the sequence segment starting at position i to the motif Θ is obtained by $s_i = \sum_{m=1}^{w} \sum_{n=1}^{4} (\Theta^{mn} - \Theta_b^n) u_i^{mn}$, where e^{s_i} is the probability of observing sequence u_i under the motif model Θ and background model Θ_b. Note the difference between the two. Θ is a two dimensional matrix with length w and width 4, while Θ_b is a column vector with width 4. We also define $s_{b,i}$ to be the score of position i under the background model Θ_b.

The version of the Bernoulli Gibbs sampler allowing multiple sites per sequence is often called the Bernoulli Gibbs Sampler[69] . In our notation, the probability of observing the set of sequences is written as $P(S, \boldsymbol{\xi}|\Theta, \Theta_b, q)$. S represents all the sequences in the sequence set concatenated together with a total number of bases N. $\boldsymbol{\xi}$ represents the locations of the motif in the set of sequences. If there is a motif present starting at position i, ξ_i is set to 1, otherwise 0. Θ is the motif model and Θ_b is the model for the background. q is the prior probability of having a motif starting in any given position. In the Bernoulli Gibbs Sampler, P is defined as:

$$P(S, \boldsymbol{\xi}|\Theta, \Theta_b, q) = \left(\prod_{i=1}^{L} e^{s_{b,i}} \right) \prod_{i=1}^{L} e^{s_i \xi_i + \xi_i \log(q) + (1-\xi_i)\log(1-q)} \tag{5.2}$$

Since $\prod_{i=1}^{L} e^{s_{b,i}}$ is independent of Θ and $\boldsymbol{\xi}$, it can considered as a constant C, which will not play any role in the rest of our discussions. Since the goal is to maximize $P(S, \boldsymbol{\xi}|\Theta, \Theta_b)$, we need to find update rules for $\boldsymbol{\xi}$, Θ and q that result in increasing P. Sites along the sequence are randomly chosen and ξ_i is updated based on how it affects the total value of P. The updated value of ξ_i can be obtained as follows:

$$\frac{p(\xi_i = 1)}{p(\xi_i = 0)} = \frac{e^{s_i + \log(q)}}{e^{\log(1-q)}} \rightarrow p(\xi_i = 1) = \frac{q e^{s_i}}{q e^{s_i} + (1 - q)} \tag{5.3}$$

Using the above equation, each position in the sequence can be sampled to be considered as a candidate for the motif model. After each update of $\boldsymbol{\xi}$, Θ and q need

to be re-estimated. To obtain Θ and q, we have to maximize the likelihood for Θ and q. The $[\sum_{n=1}^{4} \exp(\Theta^{mn}) - 1]$ needs to be added to the log likelihood with the Lagrange multiplier λ to ensure that Θ^{mn} remains normalized.

$$\frac{\partial \log P}{\partial \Theta^{mn}} = \sum_{i}^{N} \xi_i \frac{\partial s_i}{\partial \Theta^{mn}} - \lambda e^{\Theta^{mn}} = \sum_{i}^{L} \xi_i u_i^{mn} - \lambda e^{\Theta^{mn}} \tag{5.4}$$

The derivation of the last step uses the relation $\frac{\partial s_i}{\partial \Theta^{mn}} = \frac{\partial \sum_{mn} \Theta^{mn} u_i^{mn}}{\partial \Theta^{mn}} = u^{mn}$. By setting Eq. 5.4 to zero, Θ can be obtained estimated:

$$\Theta^{mn} = \log \frac{1}{\lambda} \sum_{i=1}^{N} \xi_i u_i^{mn} \tag{5.5}$$

$$\lambda = \sum_{n=1}^{4} \sum_{i=1}^{L} \xi_i u_i^{mn} = \sum_{i=1}^{L} \xi_i \tag{5.6}$$

We define $g = \sum_{i=1}^{L} \xi_i$, which shows the number of motifs in the set. q can be estimate similarly by maximizing the likelihood with respect to q:

$$\frac{\partial \log P}{\partial q} = \frac{\sum_{i=1}^{L} \xi_i}{q} - \frac{1 - \xi_i}{1 - q} \equiv 0 \rightarrow q = \frac{\sum_i \xi_i}{N} \tag{5.7}$$

The steps in the Bernoulli Gibbs sampler are summarized in (Fig. 5-2).

1. Choose random non-overlapping positions, ξ_i, in the sequence based on the initial value of q. The expected number of motif occurrences is Nq.

2. Build $\Theta^{mn} = \log(\frac{\sum_{i=1}^{L} \xi_i u_i^{mn}}{g})$, where u_i's are sequence segments length w starting in location i, and g is the number of motif occurrences in the sequence.

3. Choose a random position i.

4. Score the sequence segment; $s_i = \sum_{m=1}^{w} \sum_{n=1}^{4} (\Theta^{mn} - \Theta_b^{mn}) u^{mn}$.

5. Sample position i with the probability $p(\xi_i = 1) = \frac{q}{(1-q)e^{-s_i}+q}$.

6. Update $q = \frac{1}{N} \sum_i \xi_i$.

7. Go to step 2 and repeat until convergence.

8. Θ is the output motif.

Figure 5-2: Steps in the Bernoulli Gibbs Sampler

5.2.2 Formulation of a Differential Gibbs Sampler

To remove the non-functional motifs or low-complexity patterns that can hinder motif finding process, a second set of sequences can be used that is equally likely to have such non-functional elements but is not expected to have the functional motif. Here, we develop a Gibbs sampling method to enable us to use a second set of sequences to improve the search in the first sequence set by avoiding the non-functional motifs that are present in both sets.

In this case, we represent the two sets of sequences, the positive and negative sets, with S_1 and S_2. The locations of motifs are shown with $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ in the first and the second sequence set respectively. The expected number of motifs in the two sets can be different as well. q_1 is used to represent the expected probability of having a motif in any position in the first set, and q_2 is used for the second set. To represent the motif, we use a weight matrix Θ which is common between the two sets. Θ will be determined using the motif occurrences in the first set only. We are looking for a motif that has a high likelihood of appearing in the first set, and at the same time has a lesser likelihood of appearing in the second set. To find the over-represented motif Θ, we choose to maximize $\frac{P(S_1,\boldsymbol{\xi}|\Theta,q_1)}{Z(S_2|\Theta,q_2)}$. To measure the chance of observing the motif in the second sequence set, we use the partition function $Z(S_2|\Theta,q_2) = \sum_{\{\boldsymbol{\eta}\}} P(S_2,\boldsymbol{\eta}|\Theta)$. To obtain Z, we need to sum over all the 2^{L_2} combinations of $\{\boldsymbol{\eta}\}$:

$$
\begin{aligned}
Z(S_2|\Theta,q_2) &= \sum_{\{\boldsymbol{\eta}\}} P(S_2,\boldsymbol{\eta}|\Theta,q_2) && (5.8)\\
&= \prod_{i=1}^{L_2} \sum_{\xi_i=0,1} e^{s_i\xi_i+\xi_i\log q_2+(1-\xi_i)\log(1-q_2)} && (5.9)\\
&= \prod_{i=1}^{L_2} \left(e^{s_i+\log q_2} + e^{\log(1-q_2)}\right) && (5.10)
\end{aligned}
$$

There is no dependency on η left in $Z(S_2|\Theta,q_2)$. This means that there is no need for sampling in the second set. The whole set is treated as a block. $Z = Z(S_2|\Theta,q_2)$ is used as an overall measure of the existence of the motif Θ in set two. We will designate the sequence of set one with u and sequence of the set two with v. $g_1 = \sum_{i=1}^{L} \xi_i$ also

represents the number of motif occurrences in set one. To proceed, we make another assumption. Since Θ has to represent the motif coming from the motif instances u_i in set one, so instead of estimating Θ, we set it to

$$\Theta^{mn} = \log(\frac{1}{g_1}\sum_{i=1}^{g_1}\xi_i^{mn}u_i^{mn}) \tag{5.11}$$

q_2 can be estimated by maximizing the partition function $Z(S_2|\Theta, q_2)$. For easier calculation, we maximize $\log Z = \sum_{i=1}^{L_2}\log(q_2e^{s_i} + (1-q_2))$ as a function of q_2:

$$\frac{\partial \log Z}{\partial q_2} = \sum_{i=1}^{L_2}\frac{e^{s_i}-1}{q_2e^{s_i}+(1-q_2)} = 0 \tag{5.12}$$

,which yields:

$$\sum_{i=1}^{L_2} f_i(q_2) = L_2q_2, \tag{5.13}$$

where,

$$f_i(q_2) = \frac{q_2}{(1-q_2)e^{-s_i}+q_2} \tag{5.14}$$

where f_i is the chance of having an instance of the motif Θ in position i. f_i is similar to the updating rule of site, eqn. 5.3, in the classical Bernoulli Gibbs sampler discussed in the previous chapter. This is the equation of self-consistency for q_2. On the left side of the equation, there is a sum over the chance that each site is a start position for a motif, and the right hand side is the estimate of the total number of motif occurrences. Given a Θ, all the s_i's can be computed for the sequence set and by solving Eqn. 5.12, q_2 can be obtained. For a given weight matrix this equation can be solved. The value of resulted q is expect to be small near zero. It is important to note that Z as a function q is a concave function:

$$\frac{\partial \log Z}{\partial q_2} = \sum_{i=1}^{L_2}\frac{-(e^{s_i}-1)^2}{(q_2e^{s_i}+(1-q_2))^2} < 0 \tag{5.15}$$

Also, for $q_2 = 0$, Z is equal to zero. For $q_2 = 1$, Z is equal to $\sum_i^{L_2} s_i$, which is strongly negative in most practical cases. This means there are only two times of plots for Z

as a function of q_2, which are both demonstrated in Fig. 5-11(a) and Fig. 5-11(b) of the next chapter.

Now that we know how to calculate Z, the only thing left is to obtain the update rule for sampling the sites in the first set or $p(\xi_i = 1)$. For this purpose, we need to calculate the change in the overall likelihood if the site i is flipped from the unoccupied state, $\xi_i = 0$, to an occupied state, $\xi_i = 1$:

$$\frac{p(\xi_i = 1)}{p(\xi_i = 0)} = \frac{\frac{P(S_1|\Theta(\xi_i=1),q_1)}{Z(S_2|\Theta(\xi_i=1))}}{\frac{P(S_1|\Theta(\xi_i=0),q_1)}{Z(S_2|\Theta(\xi=0))}} \tag{5.16}$$

$$\log\frac{p(\xi_i = 1)}{p(\xi_i = 0)} = \log(P(\xi_i = 1)) - \log(P(\xi_i = 0)) - \log(Z_2(\xi_i = 1)) + \log(Z_2(\xi_i = 0)) \tag{5.17}$$

The term $\log(P(\xi_i = 1)) - \log(P(\xi_i = 0))$ is identical to its version in classical Bernoulli Gibbs sampler, eqn.5.3:

$$\log(P(\xi_i = 1)) - \log(P(\xi_i = 0)) = s_i + \log(q_1) - \log(1 - q_1) \tag{5.18}$$

To estimate the term $\log(Z(\xi_i = 1)) - \log(Z(\xi_i = 0))$, we need to use the following approximation:

$$\log(Z(\xi_i = 1)) - \log(Z(\xi_i = 0)) \approx \frac{\partial \log Z}{\partial \xi_i}\Delta\xi_i \tag{5.19}$$

$$= \left[\frac{\partial \log Z}{\partial q_2}\frac{\partial q_2}{\partial \xi_i} + \sum_{mn}\frac{\partial \log Z}{\partial \Theta^{mn}}\frac{\partial \Theta^{mn}}{\partial \xi_i}\right] \times (\Delta\xi_i) \tag{5.20}$$

The change of Z is a result of the change of q_2 and Θ, which are dependent on the change in ξ_i . Since we already set the q_2 to maximize Z, the term $\frac{\partial \log Z}{\partial q_2}$ is equal to zero. $\frac{\partial \log Z}{\partial \Theta^{mn}}$ can be calculated using Eqn. 5.8:

$$\frac{\partial \log Z}{\partial \Theta^{mn}} = \sum_{i=1}^{L_2}\frac{\partial \log Z}{\partial s_i}\frac{\partial s_i}{\partial \Theta^{mn}} = \sum_{i=1}^{L_2}\frac{q_2}{q_2 + (1 - q_2)e^{-s_i}}v_i^{mn} = \sum_{i=1}^{L_2} f_i v_i^{mn} \tag{5.21}$$

$\frac{\partial \Theta^{mn}}{\partial \xi_i}$ can be calculated using Eqn. 5.11:

$$\Theta^{mn} = \log(\frac{1}{g_1}\sum_{i=1}^{L_1}\xi_i u_i^{mn}) \rightarrow \frac{\partial \Theta^{mn}}{\partial \xi_i} = \frac{u_i^{mn}}{\sum_j \xi_j u_j} - \frac{1}{g_1} \tag{5.22}$$

Putting together all the pieces yields:

$$\begin{aligned}\log\frac{p(\xi_i=1)}{p(\xi_i=0)} &= s_i + \log(q_1) - \log(1-q_1) - \sum_{mn}\sum_{k=1}^{L_2} f_k v_k^{mn}(\frac{u_i^{mn}}{\sum_j \xi_j u_j^{mn}} - \frac{1}{g_1}) \quad (5.23)\\ &= s_i + \log(\frac{q_1}{(1-q_1)}) - \sum_{mn}(\frac{\sum_{k=1}^{L_2} f_k v_k^{mn}}{\sum_{j=1}^{L_1}\xi_j u_j^{mn}})u_i^{mn} + \frac{\sum_{i=1}^{L_2} f_i \sum_{mn} v_i^{mn}}{g_1}\end{aligned}$$

Since the length of the motif is w, $\sum_{m=1}^{w}\sum_{n=1}^{4} v_i^{mn} = w$. This further simplifies our calculation:

$$\log\frac{p(\xi_i=1)}{p(\xi_i=0)} = s_i + \log\left(\frac{q_1}{1-q_1}\right) - \sum_{mn}(\frac{\sum_k f_k v_k^{mn}}{\sum_j \xi_j u_j^{mn}})u_i^{mn} + \frac{L_2 q_2}{g_1}w \tag{5.24}$$

This can be rewritten as:

$$\log\frac{p(\xi_i=1)}{p(\xi_i=0)} = \log\left(\frac{q_1}{1-q_1}\right) + \sum_{m=1}^{w}\sum_{n=1}^{4}\left[\Theta^{mn} - \Theta_b^{mn} - (\frac{\sum_j f_j v_j^{mn}}{\sum_j \xi_j u_j^{mn}}) + \frac{L_2 q_2}{g_1}\right] u_i^{mn} \tag{5.25}$$

To simplify this result, we define Θ' to be:

$$\Theta'^{mn} = \Theta^{mn} - \frac{\sum_j f_j v_j^{mn}}{\sum_j \xi_j u_j^{mn}} + \frac{L_2 q_2}{g_1} \tag{5.26}$$

Then, we will have

$$\frac{p(\xi_i=1)}{p(\xi_i=0)} = \frac{q_1}{1-q_1}e^{s'_i} \tag{5.27}$$

Now the first set gets updated as before but the scores, s'_i, are computed using Θ' instead of Θ. Θ' is an effective weight matrix that includes the contributions from the second sequence set. The steps for the Differential Motif Finder are summarized in Fig. 5-3.

1. Choose random non-overlapping positions, ξ_i's , in sequence 1 based on the expected number of motif occurrences $L_1 q_1$.

2. Build $\Theta^{mn} = \log(\frac{\sum_{i=1}^{L} \xi_i u_i^{mn}}{g_1})$, where u_i's are sequence segments of length w in set one.

3. Using Θ, score all the sites in sequence set two estimate q_2 and then compute
$\Theta' = \Theta - \frac{\sum_j f_j v_j^{mn}}{\sum_j \xi_j u_j^{mn}} + \frac{L_2 q_2}{g_1}$, where $f_j = \frac{q_2}{(1-q_2)e^{-s_i}+q_2}$. v_i are sequence segments in set two.

4. Choose a random position i in sequence set 1.

5. Score the sequence segment $s'_i = \sum_{m=1}^{w} \sum_{n=1}^{4} (\Theta'^{mn} - \Theta_b^{mn}) u^{mn}$.

6. Sample position i with probability $p(\xi_i = 1) = \frac{q_1}{(1-q_1)e^{-s'i}+q_1}$.

7. Update $q_1 = \frac{1}{L_1} \sum_i \xi_i$.

8. Go to step 2 and repeat until convergence.

9. Θ is the output motif.

Figure 5-3: Steps for Differential Bernoulli Gibbs Sampler

5.2.3 The one motif occurrence per sequence case

Instead of having a variable number of occurrences of a motif in each sequence, we can limit the number to exactly one occurrence per sequence in both data sets. In this case, the sequences in the data set are not concatenated. So, there are N_1 and N_2 sequences in the first and second data set respectively. Sequences can have different length. The length of each sequence in the first set is shown by L_i and in the second set by L'_i. A_i represents the location of the start of the motif in the sequence i. A'_i is used for the locations of the motifs in the second set. The score of a motif occurrence in sequence i starting in position j is written as $s_{i,j}$. The P and Z for a set of sequences are defined as:

$$P(S_1|\boldsymbol{A},\Theta) = \prod_{i=1}^{N_1} e^{s_{i,A_i}} \tag{5.28}$$

$$Z(S_2|\Theta) = \sum_{\{A'_i\}} P(S_2,\boldsymbol{A}|\Theta) = \sum_{\{A_i\}} \prod_{i=1}^{N_2} e^{s_{i,A_i}} = \prod_{i=1}^{N_2} \sum_{A'_i=1}^{L'_i} e^{s_{i,A_i}} \tag{5.29}$$

$$\log Z = \sum_{i=1}^{N_2} \log \left(\sum_{A'_i=1}^{L'_i} e^{s_{i,A'_i}} \right) \tag{5.30}$$

We maximize a term similar to what we had in previous sections, $\frac{P(S_1,A|\Theta(A))}{Z(S_2|\Theta(A))}$. In this case, A refers to the position of the occurrences in the first set. We calculate the change in this overall likelihood when the motif occurrence in sequence i is moved from location $A_i = a$ to a new position $A_i = b$.

$$\log \left(\frac{p(A_i=b)}{p(A_i=a)} \right) = \log \frac{\frac{P(A_i=b)}{Z(A_i=b)}}{\frac{P(A_i=a)}{Z(A_i=a)}} = s_{i,b} - s_{i,a} - \sum_{mn} \frac{\partial \log Z}{\partial \Theta^{mn}} \frac{\partial \Theta^{mn}}{\partial A_i} \tag{5.31}$$

$\frac{\partial \log Z}{\partial \Theta^{mn}}$ can be computed easily:

$$\frac{\partial \log Z}{\partial \Theta^{mn}} = \sum_{i=1}^{N_2} \sum_{A'_i=1}^{L'_i} \left[\left(\frac{e^{s_{i,A_i}}}{\sum_{j=1}^{L'_i} e^{s_{i,j}}} \right) \frac{\partial s_{i,A_i}}{\partial \Theta^{mn}} \right] \tag{5.32}$$

$$= \sum_{i=1}^{N_2} \sum_{A'_i=1}^{L'_i} \left(\frac{e^{s_{i,A_i}}}{\sum_{j=1}^{L'_i} e^{s_{i,j}}} \right) v_{i,j}^{mn} \tag{5.33}$$

$$= \sum_{i=1}^{N_2} \sum_{A'_i=1}^{L'_i} f_{i,j} v_{i,j}^{mn}, \tag{5.34}$$

where $f_{i,j} = \left(\frac{e^{s_{i,j}}}{\sum_{k=1}^{L'_i} e^{s_{i,k}}} \right)$. For $\frac{\partial \Theta^{mn}}{\partial A_i}$, we have:

$$\frac{\partial \Theta^{mn}}{\partial A_i} = \frac{\partial \log(\frac{1}{N_1} \sum_j u_j^{mn})}{\partial A_i} = \frac{u_{i,b}^{mn} - u_{i,a}^{mn}}{\sum_{j=1}^{N_1} u_{j,A_j}^{mn}} \tag{5.35}$$

In this case, we define Θ' as

$$\Theta'^{mn} = \Theta^{mn} - \frac{\sum_{i=1}^{N_2} \sum_{j=1}^{L'_i} f_{i,j} v_{i,j}^{mn}}{\sum_{i=1}^{N_1} u_{i,A_i}^{mn}} \tag{5.36}$$

This time, each sequence in set one is scored using Θ'. A new site is then sampled in the sequence based on the score of each site. Then, that new site will replace the only motif occurrence in the motif model. Then, the second set of sequences is scanned using Θ, which is generated using the motif occurrences in the first set. This produces the correction term coming from the second set that can help to generate Θ'. These steps are summarized in Fig. 5-5. They can be compared to the steps of classical Gibbs sampler version for one occurrence motif per sequence in Fig. 5-4.

1. Choose one random position A_i in each of the sequences.
2. Build $\Theta = \log(\frac{1}{N_1}\sum_{i=1}^{N_1}\xi_i u_{i,A_i}^{mn})$, where $u_{i,j}$'s are the sequence segments length w in the sequence i starting in location j.
3. Choose one sequence randomly, i.
4. Score all the segments on the sequence using $s_i = \sum_{m=1}^{w}\sum_{n=1}^{4}(\Theta^{mn} - \Theta_b^{mn})u^{mn}$.
5. Sample a site, A_i, in sequence i with the probability $p(A_i = 1) \propto e^{s_i}$
6. Go to step 2 and repeat until convergence.
7. Θ is the output motif.

Figure 5-4: Steps for classical Gibbs sampler with One Occurrence Per Sequence

1. Choose one random position A_i in each of the sequences.
2. Build $\Theta = \log(\frac{1}{N_1}\sum_{i=1}^{N_1}\xi_i u_{i,A_i}^{mn})$, where $u_{i,j}$'s are the sequence segments length w in the set one.
3. Using Θ, score all the sites in the sequence set two and compute $\Theta' = \Theta - \frac{\sum_{i=1}^{N_2}\sum_{j=1}^{L_i} f_{i,j} v_{i,j}^{mn}}{\sum_{i=1}^{N_1} u_{i,A_i}^{mn}}$, where $f_{i,j} = \left(\frac{e^{s_{i,j}}}{\sum_{k=1}^{L_i} e^{s_{i,k}}}\right)$ and $v_{i,j}$'s are the sequence segments in the set two.
4. Choose one sequence randomly, i.
5. Score all the segments on the sequence using $s'_i = \sum_{m=1}^{w}\sum_{n=1}^{4}(\Theta'^{mn} - \Theta_b^{mn})u^{mn}$.
6. Sample a site, A_i, in sequence i with the probability $p(A_i = 1) \propto e^{s'_i}$
7. Go to step 2 and repeat until convergence.
8. Θ is the output motif.

Figure 5-5: Steps for Differential Motif Sampling with One Occurrence Per Sequence

```
atgcgccgtagcagttgatagtcaaatgctcatctactac
atgtgacgtagcacaaatgagcaaggttgtgcgccgcttg
ctctcacgttgcagtgctagagcctacctgtctgttaccc
cgagtacgtagcaattcaagtagatcgggacttctcgcgt
acggcacgcagcacggattaaatagcctgagtcttatggt
agactacgcagcataatccgcctagcaatctcggaacgga
gaactacgttgcaacgtacgcggggctacaaagtattact
taggaacgtagcacagcttgaatcatagcctttatttctt
gcgggacgtagcacgagcaacagcttgccttctgagattt
atacaaggtagcaggctatgtgccaagacgactaccctca
```

Figure 5-6: A set of artificially generated sequences with a strong motif in position 6 of all of them.

5.3 Results and Discussions

5.3.1 Gaining intuition for the differential Gibbs sampler

To have an intuition for this new method, we need to investigate how different variables such as $P(S_1, \boldsymbol{\xi}|\Theta, q_1)$ or $Z(S2|\Theta, q_2)$ compare to one another, and how they change through the process of motif finding. We generate a set of 20 sequences each 20 bases long. A strong motif with a relative information content of 12 bits is planted in position 6 of each sequence. We choose a fixed position to help visualization of the motif, and the motif finder will not use this information. The consensus sequence for the motif is ACGTAGCA.

We pick 20 random positions in these sequences, and using the 8-mers starting in those positions, we build frequency matrix as well as a weight matrix from these segments. An example of such log-odd probability weight matrix, $\Theta - \Theta_b$, is shown in Figure 5-7.

Using the weight matrix Θ, we can obtain the score for each site using the equation $s = \sum_{m=1}^{w} \sum_{n=1}^{4} (\Theta^{mn} - \Theta_b^n) u^{mn}$. To study the relationship between the scores and the weight matrices, two weight matrices are chosen, one with low information content similar to the background distribution, and the second with a high information content

$$\Theta - \Theta_b = \begin{pmatrix} -0.66 & 0.04 & -0.08 & 0.47 & -0.08 & -0.08 & -0.08 & -0.08 \\ 0.37 & 0.26 & -0.15 & -0.56 & -0.56 & 0.37 & 0.37 & 0.00 \\ 0.10 & 0.23 & -0.18 & -0.18 & -0.03 & 0.10 & -0.60 & -0.18 \\ -0.03 & -1.15 & 0.38 & -0.03 & 0.50 & -0.74 & 0.12 & 0.26 \end{pmatrix}$$

Figure 5-7: Weight Matrix

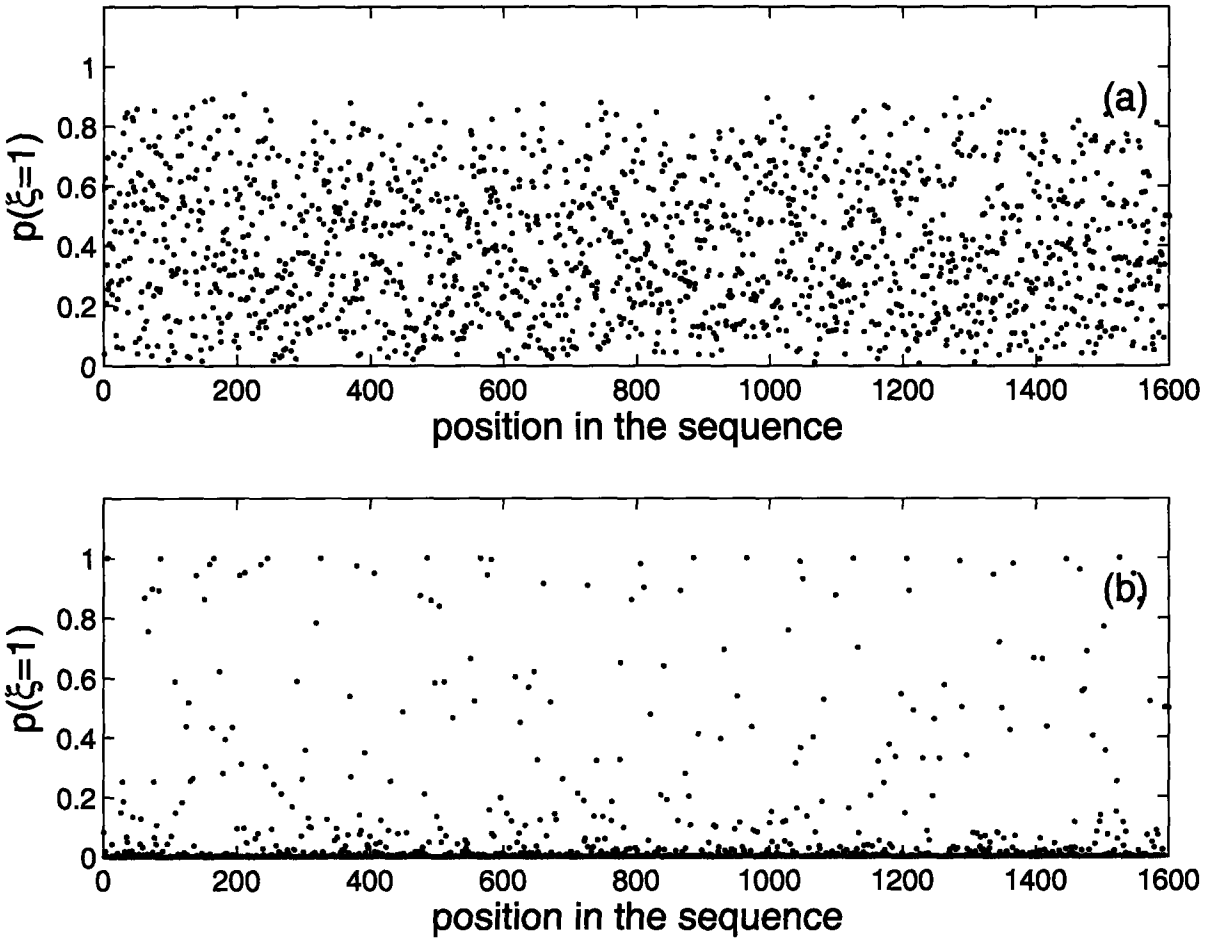


Figure 5-8: The value of score, s_i, across the artificial set. (a)Using the weight matrix generated by random locations. (b) Using the weight matrix generated from the occurrences of the planted motif.

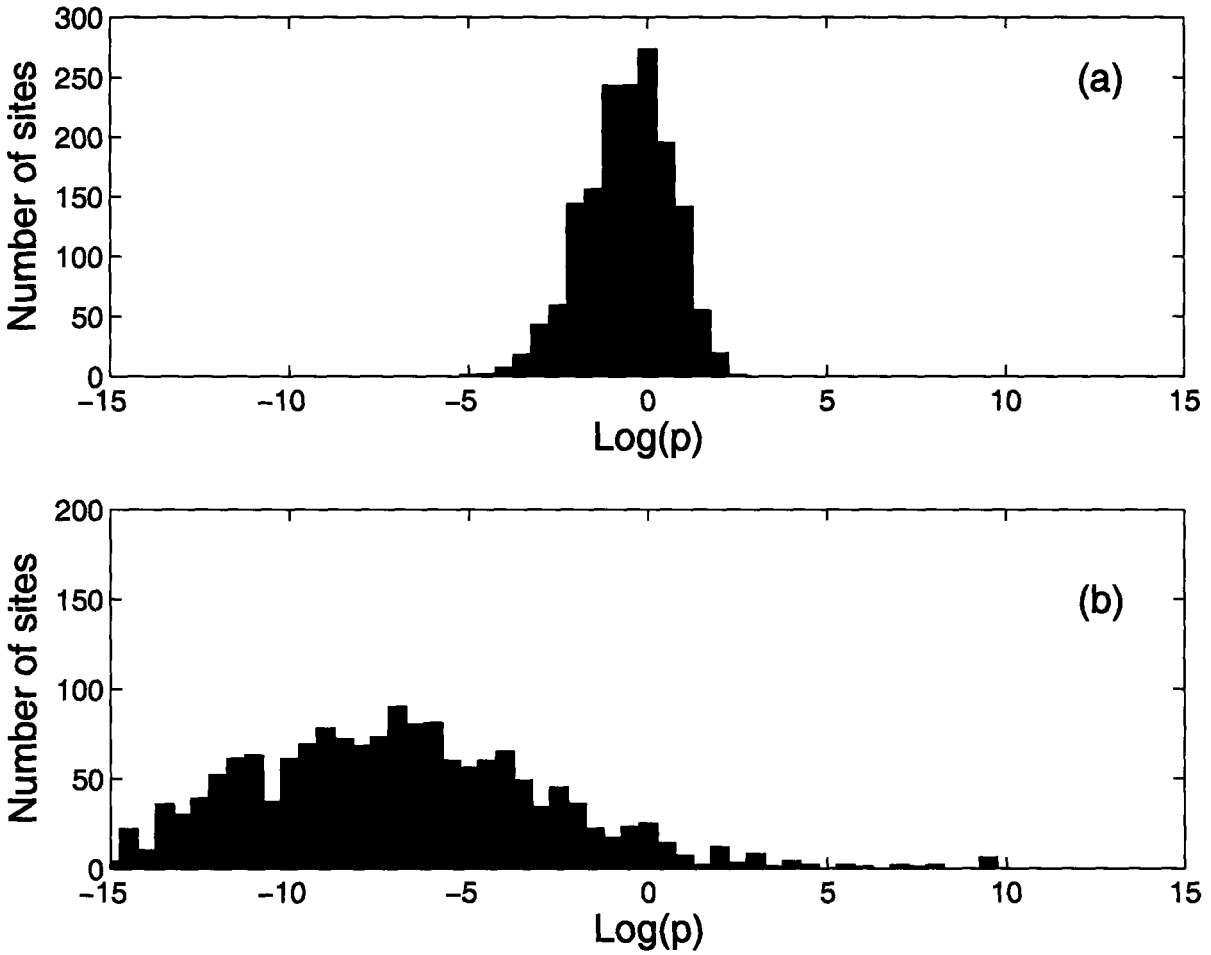


Figure 5-9: The histogram of scores across the artificial set. (a)Using the weight matrix generated by random locations. (b) Using the weight matrix generated from the occurrences of the planted motif.

resembling the planted motif in the sequence set. Their profile across The histogram of the scores, s_i's, can also be seen in Fig. 5-9. The scores for the weak weight matrix are mostly clustered around zero, while for the strong weight matrix, most of the scores are highly negative except a few which belong to the sites of the planted motif. This shows that strong weight matrices (with more information content) push the scores of most of the sites toward negative.

If we are expecting to see 20 occurrences of our motif in the sequence set 1600-base long, then we can set $q = \frac{20}{1600}$. Using this value of q, probability of sampling sites, $p(\xi_i = 1) = \frac{q}{q+(1-q)e^{-s_i}}$ can be calculated (Fig.5-8). It can be seen that for the weak weight matrix, there more sites that are potential choices for sampling, but for the

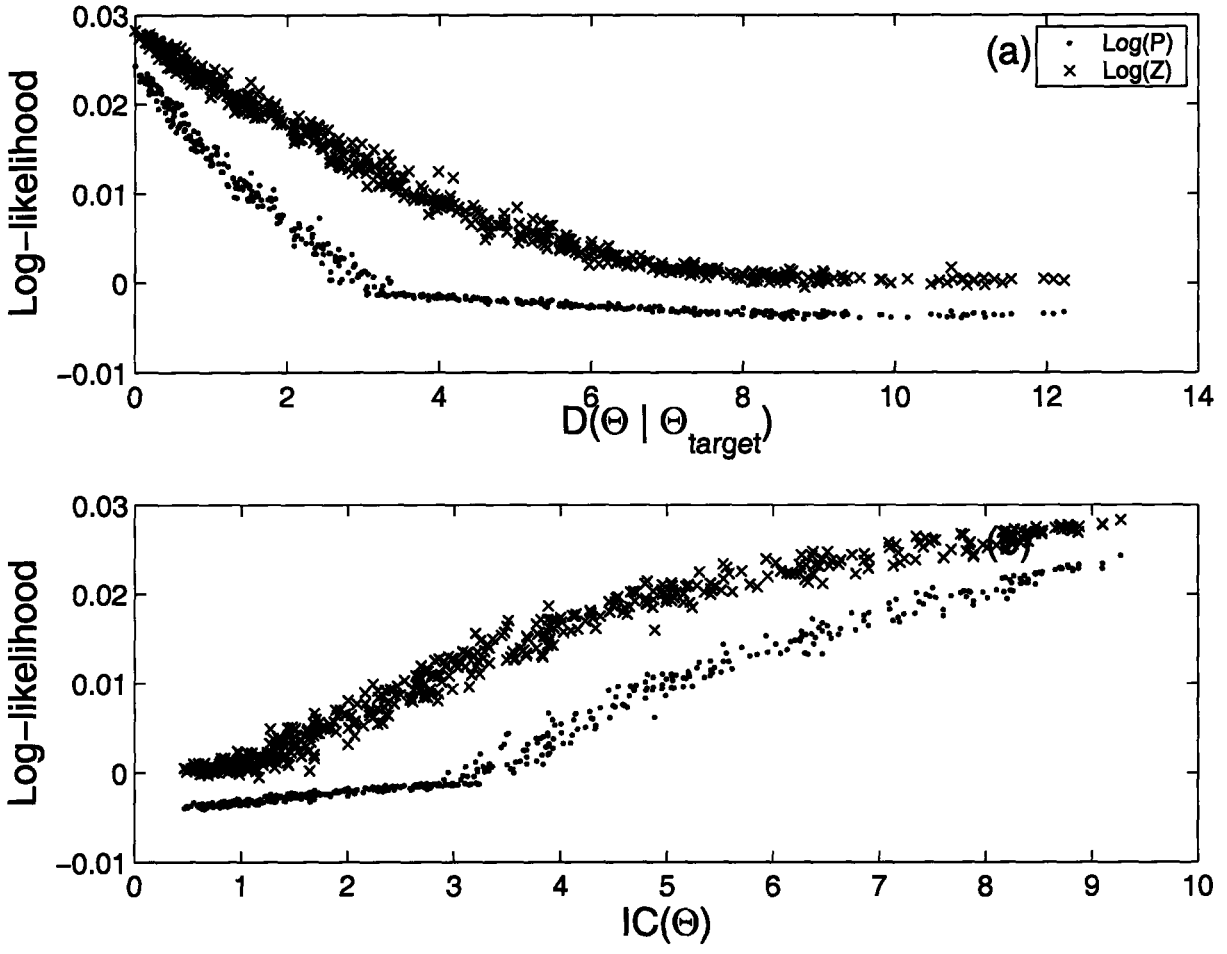


Figure 5-10: (a)$\frac{1}{N}\log(Z)$ versus the distance of random weight matrix from the planted motif, $D(\Theta|\Theta_{planted})$. (b)$\frac{1}{N}\log(Z)$ vs information content of the random weight matrix

stronger weight matrix, only the sites that match well with the weight matrix have a high probability and the rest of the sites have low scores.

In the model that was discussed in the previous section, we used a partition function Z to estimate the degree of presence of a certain motif Θ in the sequence set. We examine how Z changes when Θ becomes more similar to the planted motif. We generated a number of weaker weight matrices, Θ, by adding pseudo-counts to the planted motif $\Theta_{planted}$. Then we calculated $\log(Z(\Theta))$ as a function of the dissimilarity of Θ with $\Theta_{planted}$ (Fig. 5-10). To measure dissimilarity, we used a symmetric

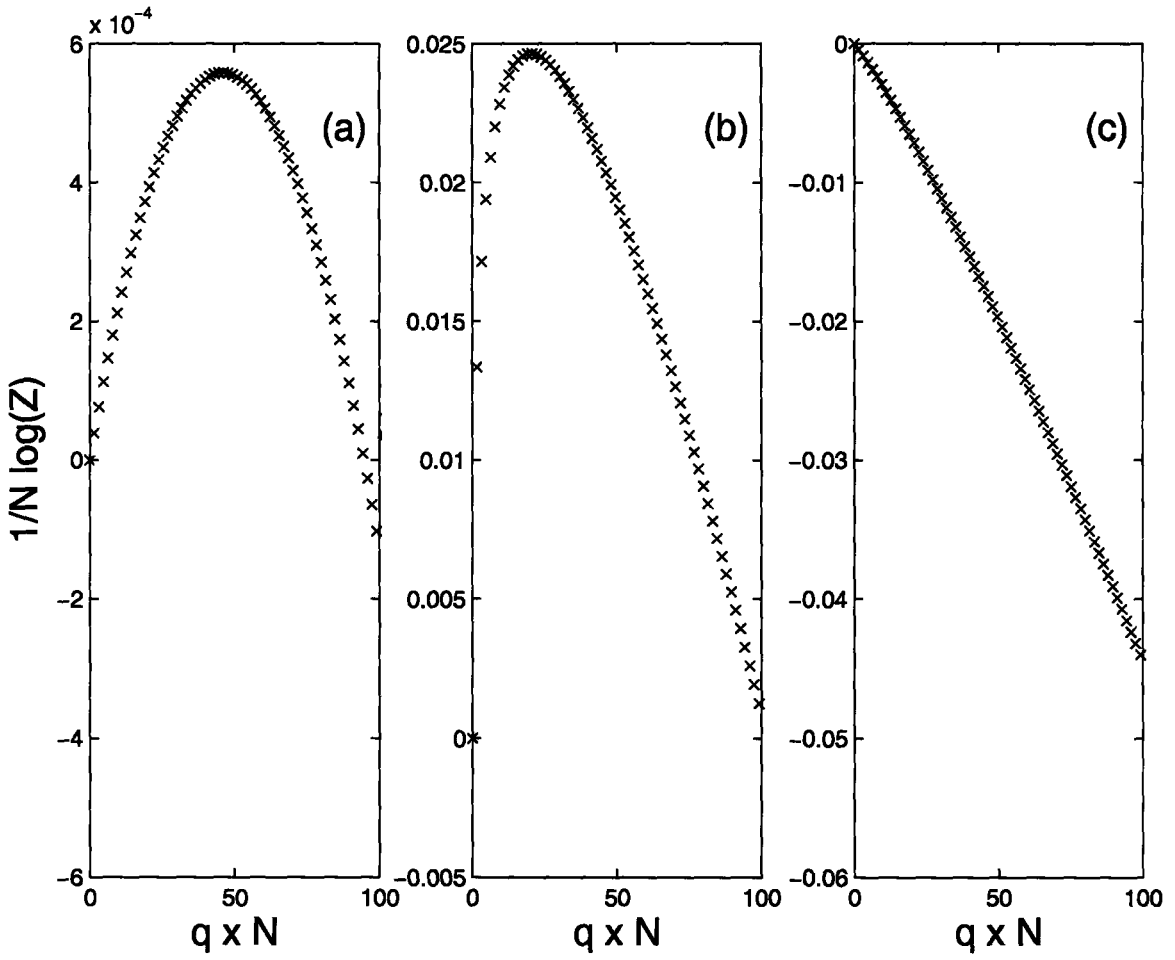


Figure 5-11: $\frac{1}{N}\log(Z)$ as a function of $q \times N$ for (a)Using the weight matrix generated by random locations and (b) Using the weight matrix generated from the occurrences of the planted motif. (c) Using a strong weight matrix, which does not have any occurrences in the sequence set. It can be seen that $\frac{1}{N}\log(Z)$ is substantially lower for the low-information content weight matrix compare to the weight matrix generated using the planted motif. Also, it can be seen that a weaker motif maximizes $\frac{1}{N}\log(Z)$ in a higher value of $q \times N$. For the weight matrix without any match in the sequence set, the maximum of $\frac{1}{N}\log(Z)$ is at $q = 0$.

Kullback-Leibler divergence formula:

$$D(F_1, F_2) = \sum_{mn} F_1^{mn} \log(\frac{F_1^{mn}}{F_2^{mn}}) + F_2^{mn} \log(\frac{F_2^{mn}}{F_1^{mn}}), \tag{5.37}$$

where F^{mn} is the frequency matrix defined as $F^{mn} = \frac{1}{g}\sum_{i=1}^{g} u_i^{mn}$. g is the number of motif occurrences.

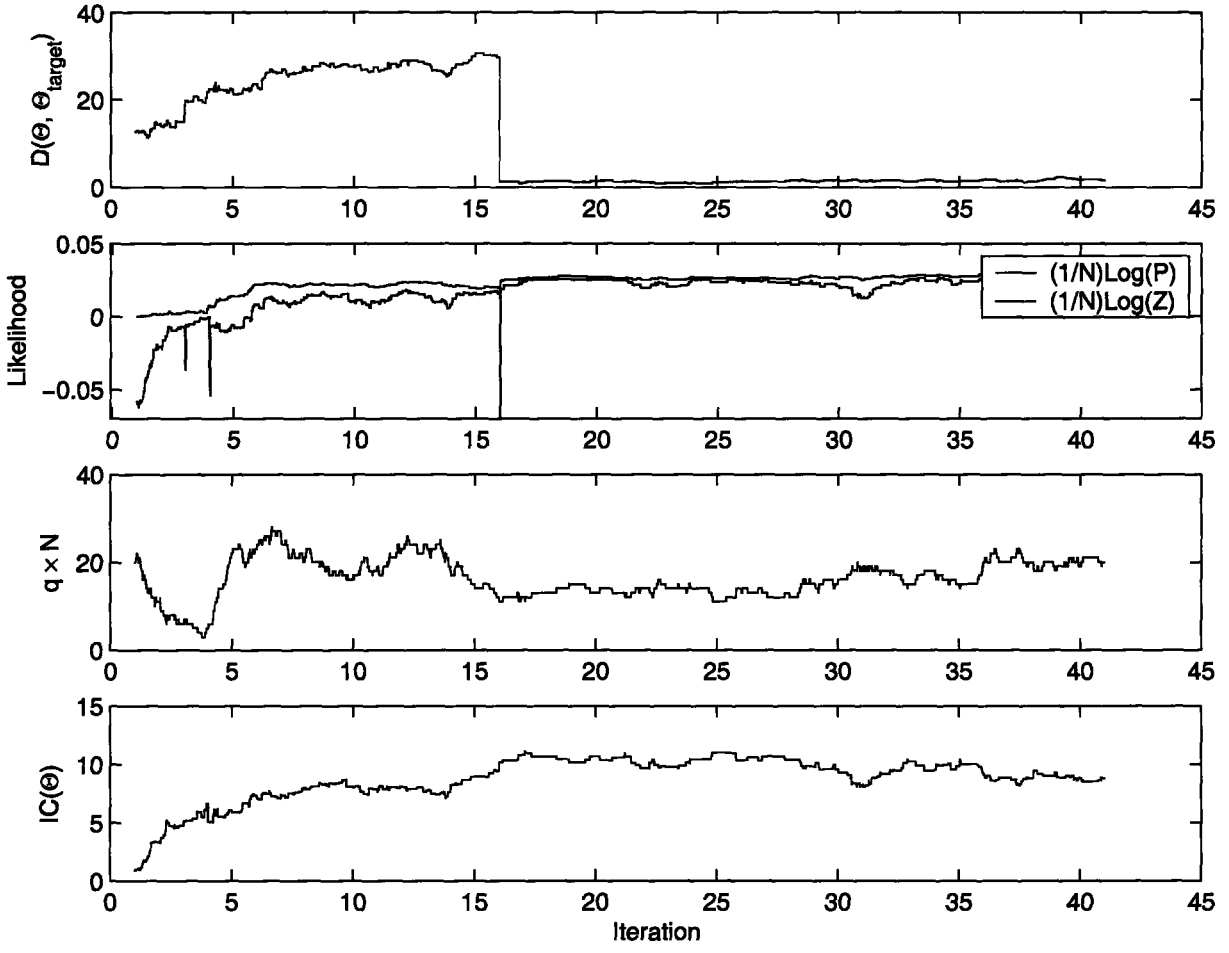


Figure 5-12: Running classical Bernoulli Gibbs sampler on a 1600-base-long sequence, which has 20 occurrences of a motif is planted in it.(a) The change of D as a function of iteration. Each unit on the x-axis represents sampling all the sites along the sequence. It can be seen that at iteration 16, D suddenly decreases. This means that the planted motif is found. The sudden decrease in fact is many iterations that is not visible due to the low resolution of the figure. (b)The change in $\log(P)$ and $log(Z)$ as a function of iteration. (c) the number of motif occurrences as function of iteration. It can be seen that Bernoulli Gibbs sampler does not quite converges even when it finds the planted motif. (d) Information content for the weight matrix as the sampler evolves.

We see that $\log(Z)$ increases as Θ becomes more similar to the planted $\Theta_{planted}$ (Fig. 5-10). When $D(\Theta|\Theta_{planted})$ is zero when Θ and $\Theta_{planted}$ are identical. As it can be seen from the figure $\log(Z)$ is in its maximum in when $D(\Theta|\Theta_{planted}) = 0$. In Fig. 5-10(b), the log likelihoods are plotted as a function of information content of

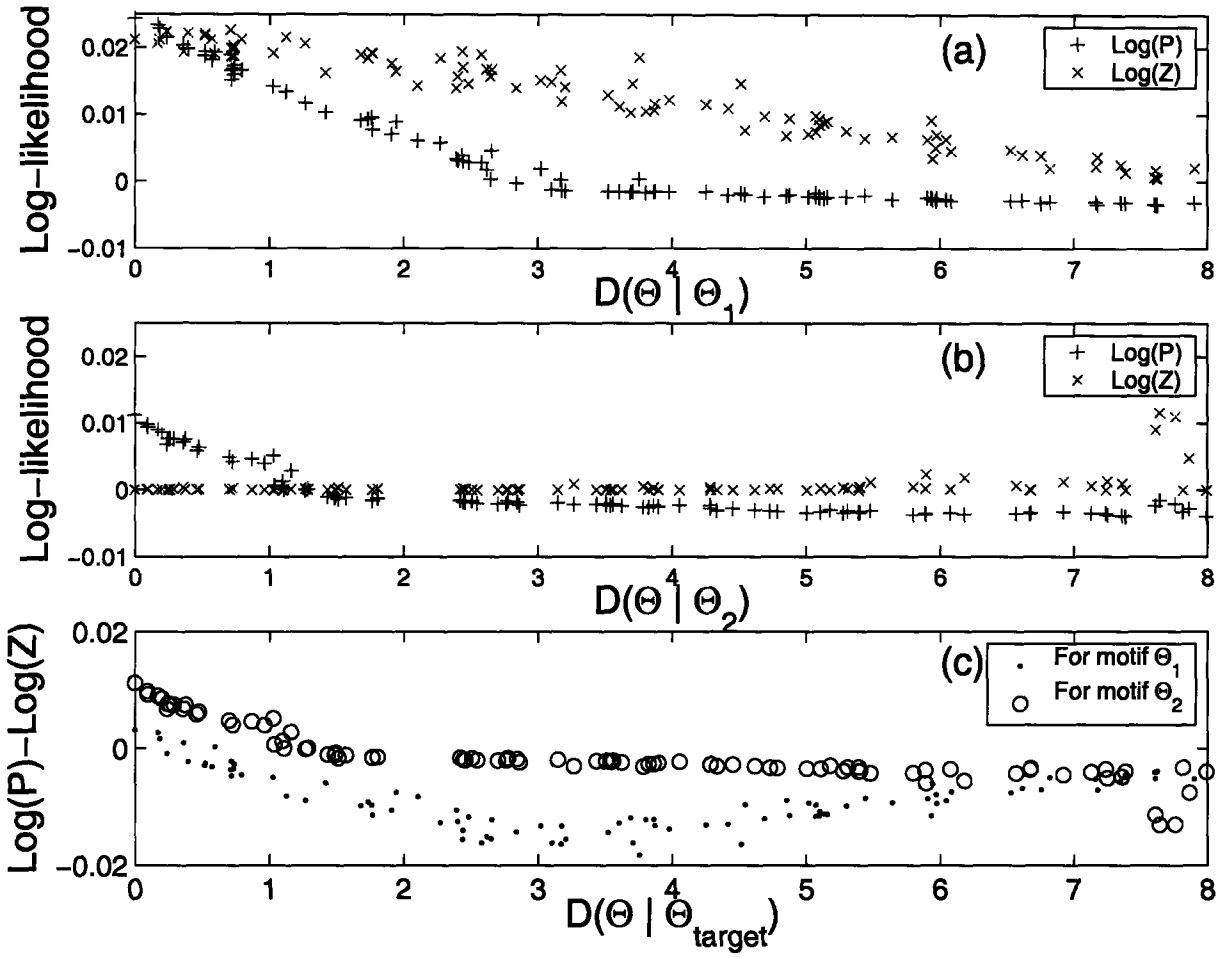


Figure 5-13: The value of $P^*(S_1|\Theta, q_1)$ and $Z(S_2|\Theta)$ is calculated for two sets of sequences. In the first set of sequences, there is two planted motifs of Θ_1 and Θ_2. The second sequence set only contains Θ_1. Random Θ's are generated, and for each Θ, the values of P and Z are calculated. (a) When Θ approaches Θ_1 in the left of the plot, $\log(P^*)$ and $\log(Z)$ increases. This is due to the fact that Θ_1 exists in both sequence sets and is detected both by P and Z. (b) As Θ approaches Θ_2, $\log(P^*)$ increases while $\log(Z)$ remains constant. This is because Θ_2 is present in the first set and is detected by $P^*(S_1|\Theta, q_1)$, while it is not present in the second set and does not contribute to $Z(S_2|\Theta)$. (c) The plot for $\log(P^*(S_1|\Theta, q_1)) - \log(Z(S_2|\Theta))$ as Θ approaches to Θ_1 or Θ_2. This difference is bigger for Θ_2. This means maximizing this difference can lead to Θ_2, which is differentially enriched.

Θ. The information content is calculated using:

$$IC(F) = \sum_{mn} F^{mn} \log(\frac{F^{mn}}{F_b^{mn}}), \tag{5.38}$$

Calculating q^*

To calculate $Z(\Theta)$, we need to maximize $Z(\Theta, q)$ for q. The value of $\log(Z)$ is plotted as a function of expected number of motif occurrences, $q \times N$ in figure 5-11. As it was shown in Eq. 5.15, the curve is concave, and the value of $\log(Z(q = 0)) = 0$. In Fig. 5-11(a), a weight matrix is used that was generated using random starting sites along the sequence. This weight matrix has a very low information content $IC(\Theta) = 0.7$bits. In Fig. 5-11(b), the motif which was planted in the sequence was used. In this case, the value of $\log(Z)$ is higher, and the maximum occurs at $q^* \times N = 20.4$, which matches well with the fact that 20 occurrences of the motif were planted in the sequences. In Fig. 5-11(c), a strong motif with an information content of 7.6 bits was used that did not have any occurrences planted in the sequence. For such a weight matrix, all the scores along the sequence, s_i's, are negative, and as a result all the terms such as $\log(qe^{s_i} + (1 - q))$ will be less than zero for any value of $q > 0$. This results in a negative $\log(Z) = \sum_i \log(qe^{s_i} + (1 - q))$ for any value of $q > 0$. This means for a strong weight matrix with no occurrences in the data set, the maximum value of $\log(Z)$ is at $q = 0$.

Figure 5-12 shows running classical Bernoulli Gibbs sampler on the set of sequences which has one planted motif, we can study the behaviour of different variables in the system. The original expected number of occurrences is set to 20, and 20 random locations are chosen for the occurrences of the motif. The unit for the x-axis is one full iteration through all the sites in the sequence. The order of picking sites for updating is random to avoid artifacts. By looking at Fig. 5-12(a), we see that the motif is found near 16 iteration when suddenly the distance between the motif found with the planted motif drops to zero. The drop looks sudden because of limited resolution in the plot. As can be seen in Fig. 5-12(c), the number of occurrences in the set fluctuates as a function of iteration. We also see in Fig. 5-12(d) that information content increases in a fairly monotonic manner.

Differential Likelihood

To study the behavior of differential Gibbs sampler, we need to study how $P(S_1, \boldsymbol{\xi}|\Theta, q_1)$ and $Z(S_2|\Theta)$ change as a result of the change in Θ. It is important to show that for the motifs that are differentially enriched $\frac{P(S_1,\boldsymbol{\xi}|\Theta,q_1)}{Z(S_2|\Theta)}$ is higher compare to motifs that are equally present in both sets of sequences. We choose two sets of sequences. In the first set, we plant two motifs, Θ_1 and Θ_2, each with 20 occurrences. In the second set, we only plant 20 occurrences of the motif Θ_1. In this example, we set Θ_1 to be a stronger motif with 9 bits of information, while Θ_2 has only 7.7 bits of information content. A differential motif finder is expected to find the motif which is differentially represented represented in the two sets, even if it is not the strongest. We generate a number of random weight matrices and measure the $\log(P^*(\Theta))$ in the first set and the $\log(Z(\Theta))$ in the second set. P^* is the maximum of value of $P(S_1, \boldsymbol{\xi}|\Theta, q_1)$ for a fixed Θ but variable q_1 and ξ. It is the highest value P can take under current Θ.

These values are plotted as function of the dissimilarity of Θ with Θ_1 or Θ_2 (Fig. 5-13). From Fig. 5-13(a), it can be seen that $\log(P)$ and $\log(Z_1)$ approach one another when $D(\Theta, \Theta_1) \rightarrow 0$ since Θ_1 is present in the two sequence sets. However, in Fig. 5-13(b), we see that $\log(Z(\Theta))$ is near zero for the Θ's that are similar to Θ_2. This is due to the fact that Θ_2 is not present in the second set. In Fig. 5-13(c), we see that the value of $\log(P(\Theta)) - \log(Z(\Theta))$ is bigger when the dissimilarity of Θ is measured from Θ_2. This means for a motif finder that maximizes $\log(P) - \log(Z)$ will converge to Θ_2.

5.3.2 Using Test Sequence Sets

To test our differential motif finder, we use a number of artificial examples. The motif finder is expected to find differentially-enriched motifs with a higher likelihood. Throughout the search, a second set of sequences is used to penalize motifs that are present in the second data set. We compare the performance of the differential motif finder with that of the classical Gibbs sampler in finding differentially-enriched motifs.

We generate two sets of sequences. One is a positive set containing the functional motif and a non-functional motif. The other is a negative set containing only the non-functional motif. The two sequence sets are generated with 20 sequences each. Each sequence is 60-bases long. The background model to generate the sequences is set to [A:32% C:18% G:18% T:32%]. The first sequence set contains two non-overlapping motifs F1 and F2, while the second set only contains one of the motifs, F1. Motifs are 10 bases long. To generate the motifs, a random matrix is generated where each of its elements is choosen using a uniform distribution from a range of $[0, 1]$. A motif occurrence is produced by sampling from the probability distribution given by this matrix.

```
>set 1:  the positive set
gagtccggcgagaaagtattgtttacttcaagtagatcggtaattctagattatctaaag
actatcagcgaccatagcaatctcgctgcggaaagtagagaagctcggattaaatagcat
actttctgggctcaactgacttctaattgaccccggtctacaaagtattactataatcct
tgatacggagaaaaaattgcattctcatagtttaggacattacctcagcttgaatcatat
taaagcggaacgacacaatagtagacttgatctcgaaagacgactacactaataggtttt
aaagtcggatacaaaggacttttcccttggtccagtttagtttaattttcgagccaaagt
cgaggcggggacaatgaagctaagtcttgaattagaatagcaacagcattcttttggttt
cccaactgacatccaaacatttaattttgatgaagtttaaattctagagaaaattttcaa
aattacggagataatttggtaatacattgctatagtatattcgcggtctgacttattttg
gaagtcggcgacaaaggttatccccttggcatcgctacttataaactatacgattttgg
ctgaacggagacaatattaatcatacttctaaaagtagggcaaagtatttttttccaagac
cacatcggcgtacaagatccaaggtcttggcatcgtcccagaactcctgcgaagcaaaga
ggggacggagcgcatacacgagaacctgggtaaagggaattgatttatatcattaactgt
tcggcctgataccatcaactgcaacgttggttccgttaagctacaagacatagagttaag
aacatcggatcccatgtaattactaatgctaaaagcgcaaggggacttcaactacaaata
atttacggtgccacagaattaattatttcgtgaaggtaattaagatagccagtgcagtcc
atttacggaggtaaatatatacggagttcgcctagttgctcgataacagaacttcagtgt
tatcacggatagcattagtataagccttcggtaaggtcgagatgataatattaatctact
cttcgcgggcagaaagaagtaaagtctgcttaatgatatcgctgagtttttcaatatatt
agcttcggggcaaaaaagtctctgcctcggattagaaatgtaaaacagtctcgggttagg

>set 2:  the negative set
aaaatcagaccacaacacgatcctgtgcaaaagcaatatctgagagtacgcctggtcagc
taagtcggagagcataaattgagtacaattaaataggctttgtacagtgggccttacata
tctatcggcgcttcatccttgggtattaagaattcataacacttaatcttcatccagatg
taccacggacagactaattcatgatttcaaacacgttgtgataccgactatactgaatac
aattacggagacaaaaatcttaccaccaaagttttacaatttccacagagggctccttat
gctcccggtgaccaaattaacctgtgaacgtttaattcaacgtagggcagcatctttaag
gccttctgcgagcaacacttcattccagtttattaagagtgtgaggcgctgacatcttcg
catcccggaaacaatccttgacggtaaaacaaacatgcggccttcgtaagaagcttagca
aaagacggcgccccactaggatgccgtataaggggtgttgatgttgagtaacctgtctgg
tcagtcggataccatcaatgtttacgtctcattcataaactagtaggtcctaacgttcgc
tattccggcgcgcatggaatcacacttggataacagcaaagcgtaatacttgttggggaa
atattcggcctaaattaatcgtttgatcatttaacggcgctactaaaatttcttgaacag
taactcggcgacaaacgaaatcaaacgttgtaaagatcatatagaaccccctatagccaa
tatctcggagaccatgtatagtgtctaacccggcaatgggccgagtgatacaacgtcaac
agacacggaaaccaaatcaagtgaaagctatctatcagctaaatggcggtatacatgaat
tttcgcagataccaattccactacgtagtgatcggttttaaccatacctacagaatacta
aatagcggagataaaagaccggggagaatgcctgaaaggtttactagatattcgttatca
atatacggcgaccaaggctcacctttttttcaaaggaccctatctagcggatcaagtttat
taagtcggagcgaaaaaatttacccttaaaaattgcaagaagattcagttagatgtcaag
attaacggagcccatttaatccaaatactaattgttatggcgagtttttttttaaattat
```

Figure 5-14: Two sequence sets are generated. The first has two motifs A and B, one strong and one weak. The second sequence set only has the strong motif.

Since such a weight matrix has very low information content, we raise it to the power of 5 to generate sharp motifs. Since all the elements are less than one, raising the elements of the matrix to any power will reduce the elements closer to zero much more than those close to one. This creates a frequency matrix that is sharper compared to the original one. In practice, the original random matrix has about 2-4 bits of information content, but after raising it to the power of 5, the information content of the matrix increases to 8-12 bits, which is desired for our test. Instead of this method, people sometimes use a consensus sequence and place variations of that consensus sequence as the motif occurrences. We find that this is not a good method for testing weight-matrix based motif finders, since these motif finders are not well suited for handling compensatory mutations. Their performance can better be examined by using motifs obtained from weight matrices.

We first test the differential Bernoulli Gibbs sampler(DBGS). This sampler can accept multiple motif occurrences in the sequences of the first and second sets. In our test set, there is only one motif occurrence per sequence, but the differential Bernoulli Gibbs sampler does not use this information to improve its search results. We implement the code for the differential Bernoulli Gibbs sampler in MATLAB. We present a test result for running DBGS on 100 test data sets that contained different F1 and F2 motifs. In each test, DBGS was run for 3 cycles. Each cycle includes full iteration through all the sites in the first sequence set. Since the total length of the sequence in the first set is 1200, each cycle includes 1200 updates. Sites in the first sequence set are picked in a randomly permuted order and are updated based on the updating rule of DBGS5-3. Based on the theory, the change in Z, the partition function of the sequence set, has to be computed after updating a site in the first sequence set. This change is proportional to $\frac{\sum_j f_j v_j^{mn}}{\sum_j \xi_j u_j^{mn}} + \frac{L_2 q_2}{g_1}$ 5-3. To calculate this quantity, q_2 has to be calculated first. To calculate q_2, Z needs to be maximized as a function of q_2. This can be done by using the Newton method for equation 5.12. An initial value for q_2 can be set and then by iteration of equation 5.12, q_2 converges to the value that maximizes Z. More complex numerical methods can be used as well to find the value of q_2 that maximizes Z. In practice, since calculating the change in

Z is computationally intensive, it is only done every 10 updates of the sites the first sequence set.

The output weight matrix of the program is examined by comparing it to the originally planted F1 and F2 motifs. To test whether it matches any of these motifs, the distance between the output weight matrix and the planted motifs is measured. We use

$$D(F_{planted}, F_{output}) = \sum_{mn} F^{mn}_{output} \log(\frac{F^{mn}_{output}}{F^{mn}_{planted}}) + F^{mn}_{planted} \log(\frac{F^{mn}_{planted}}{F^{mn}_{output}}), \qquad (5.39)$$

If the distance between to motifs is less than 5, they are declared to be the same motif. Output motifs of the program are either marked as a match to F1, F2, or none. In figure 5-15, each point is one test run with its own F1 and F2. The information content of each of these planted motifs is measured and makes up the x and y axis. The points are shown with different markers depending on how the output motif matches the planted motifs. If the output motifs matches F1, the points are shown as blue triangles, and if it matches F2 as red squares. Output weight matrices that do not match either of these two are shown with green points. As can be seen, all the motifs that match one of the planted motifs match F2. To better understand this figure, the line of $x = y$ is plotted. Any point in the lower right side of this line corresponds to a data set that has a weaker F2 motif compared to F1. For the points in the upper left the situation is opposite. Without a differential method, we would expect the output motifs in the lower right to match F1 and the output motifs in the upper left to match F2, since motif finders tend to find the motif with the highest information content. However, as it can be seen in figure 5-15, the output motifs in the lower right still match F2.

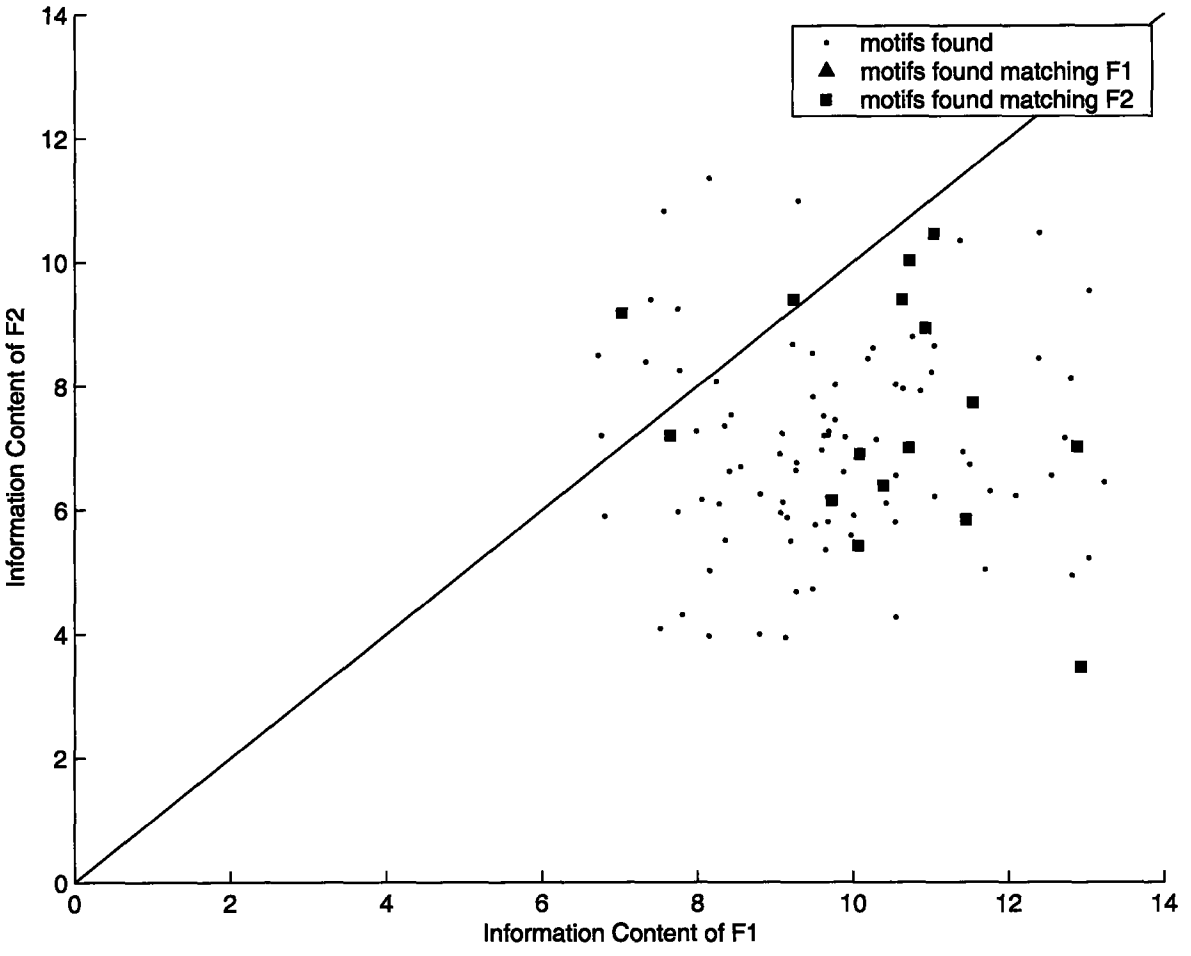


Figure 5-15: Results for the differential Bernoulli Gibbs sampler with variable number of motif occurrences in the sequence. The motif F1 is present in two sequence sets, but the motif F2 is only present in the first sequence set. Each point represents the weight matrix that the motif finder converged to for a different test set. The points marked by red squares and blue triangles are the output motifs of the program that matched the original, planted F2 and F1 motifs respectively. The green dots are the output motifs that did not match any of the planted motifs. Note that none of the output motifs matched the planted motif F1.

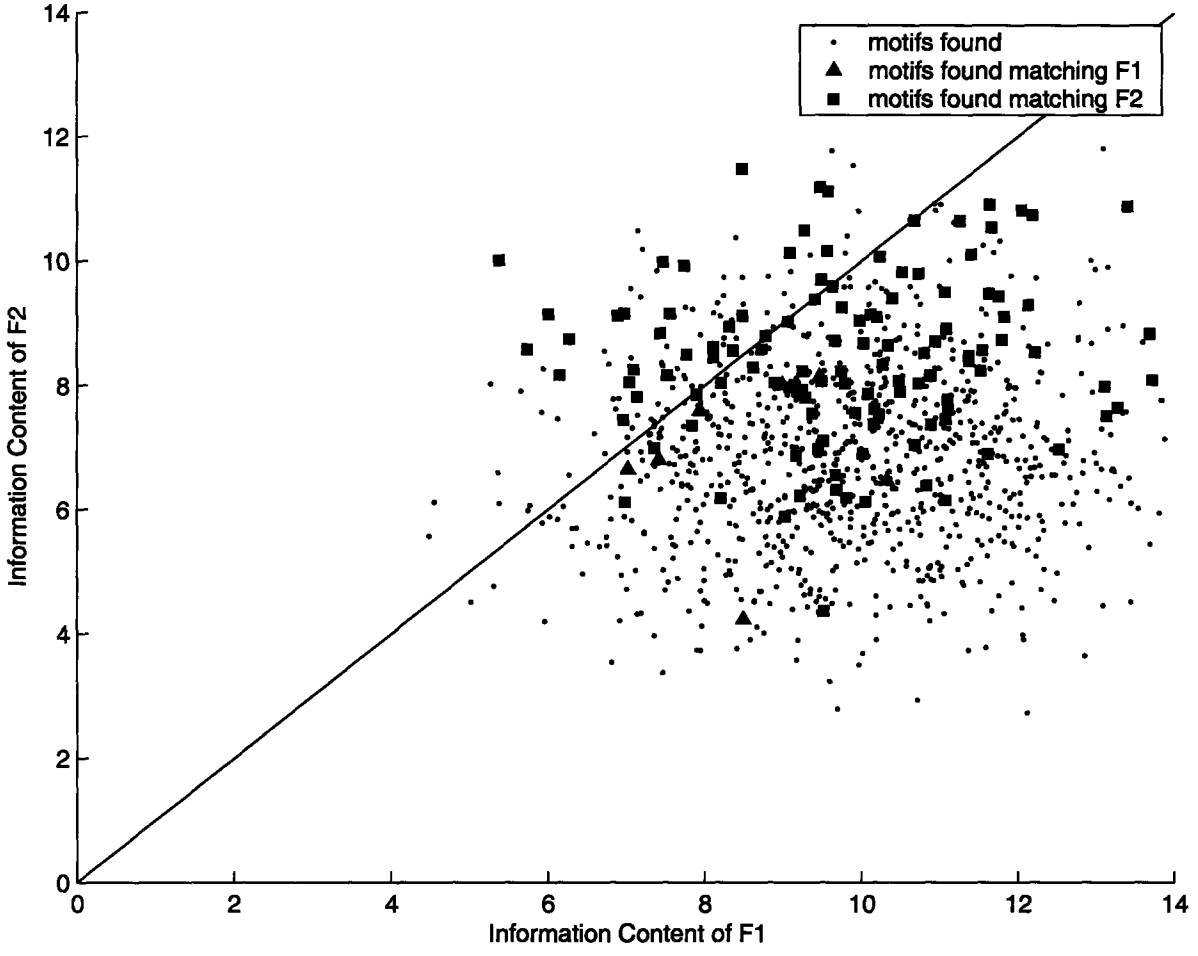


Figure 5-16: Results for the differential Bernoulli Gibbs sampler with only one motif occurrence per sequence. The motif F1 is present in two sequence sets, but the motif F2 is only present in the first sequence set. The points marked by red squares and blue triangles are the output motifs of the program that matched the original, planted F2 and F1 motifs respectively. The green dots are output motifs that did not match any of the planted motifs.

To test the differential Gibbs sampler with one motif occurrence per sequence, the same test sets were used as for the differential Bernoulli Gibbs sampler. In this case, one sequence chosen at random from the first set is scored using the weight matrix. Then, a site is sampled and the weight matrix is updated. This weight matrix is used on the second set to estimate the contribution from the second set to the weight matrix, which is $\frac{\sum_{i=1}^{N_2}\sum_{j=1}^{L_i} f_{i,j} v_{i,j}^{mn}}{\sum_{i=1}^{N_1} u_{i,A_i}^{mn}}$. In this method, there is no need to calculate q_2 since the number of motif occurrences in the second set is set to one in each sequence. The results for this test are shown in figure 5-16. In this implementation to speed up the program, instead of calculating the full partition function for the second set, we used sampling to estimate Z and to produce $\frac{\sum_{i=1}^{N_2}\sum_{j=1}^{L_i} f_{i,j} v_{i,j}^{mn}}{\sum_{i=1}^{N_1} u_{i,A_i}^{mn}}$. Instead of summing over all the sites, one site in the sequence is sampled based on the probability of $f_{i,j}$. Because of the speed up, more data sets were used in this case. The code for this method was implemented both in C and MATLAB. The results from the MATLAB version are shown in the plot. Note that most of the data sets are generated so as to have sets in which F2 has a lower information content, since it is not surprising if the motif finder finds the stronger motif. The purpose of the differential Gibbs sampler is to use the information in the second set to find the motif that are potentially weaker than other motifs but are present differentially. As can be seen from the figure, the motif finder tends to find F2 in the great majority of cases. In a few cases, it finds F1. This could be because in these cases the two motifs shared some similarity, and as a result the output motif was a mixture of the two motifs.

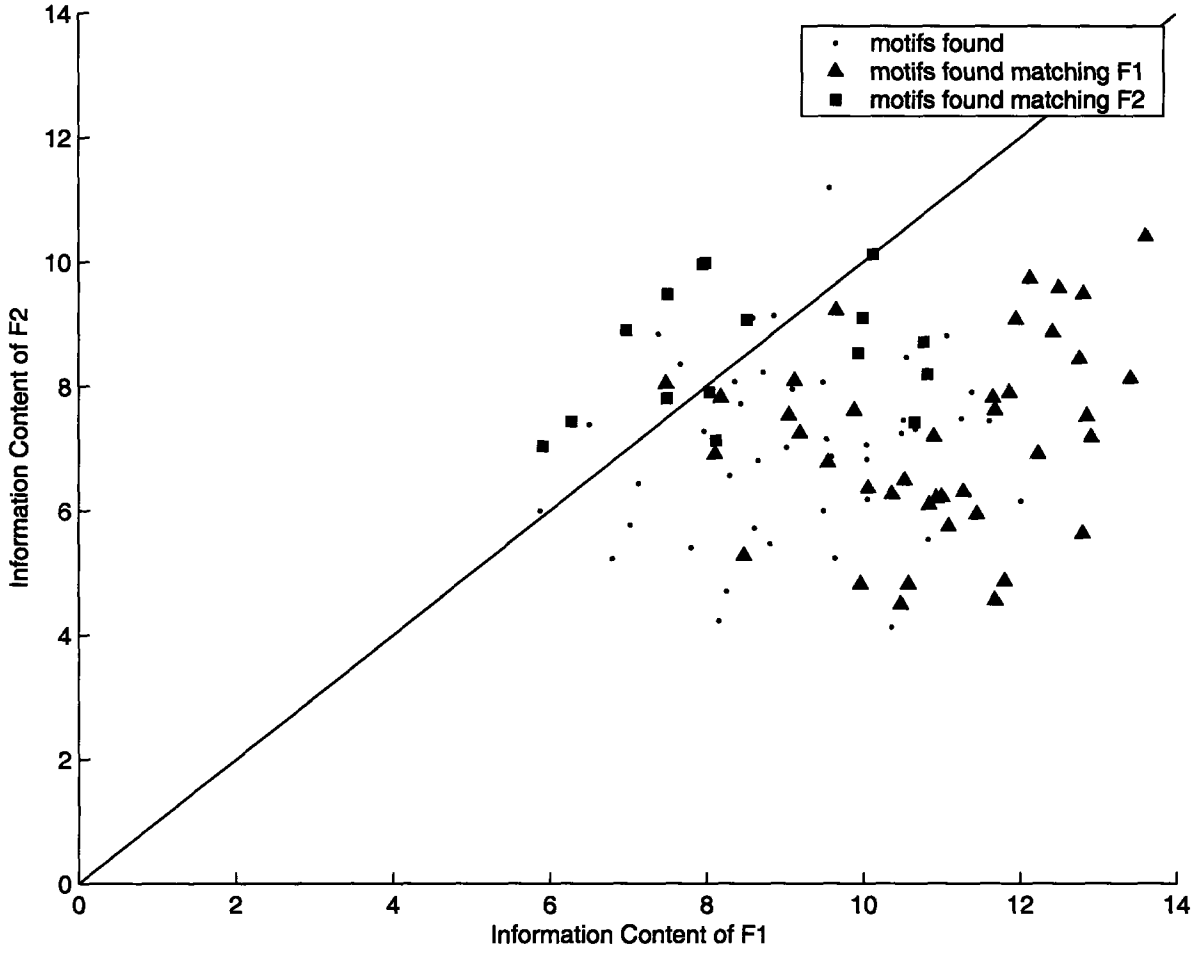


Figure 5-17: Results for the classical Bernoulli Gibbs sampler with only one motif occurrence per sequence. In the classical version, the second sequence set is ignored. The motif F1 is present in two sequence sets, but the motif F2 is only present in the first sequence set. The points marked by red squares and blue triangles are the output motifs of the program that matched the original planted F2 and F1 motif respectively. The green dots are the ones that the output motif did not match any of the planted motifs.

We also compare the differential motif finder to the classical motif finder. The classical version only takes one sequence set as an input. As a result, it is not sensitive to the differentially enriched motifs. The results are shown in figure 5-17. As can be seen, most of the motifs in the lower right section of the figure match F1, and most of the motifs in the upper right correspond to F2. There are several instances where motifs in the lower right corner match F2 and ones in the upper left match F1. The figure does not look completely symmetric because for most of the data sets, F1 has more information content than F2.

The differential motif-finding method can further be improved by using a better background model to capture the dependencies in neighboring sites. This could improve performance on biological examples. One of the common problems in Monte Carlo methods is failure to converge to a fixed state. This deficiency is present in the differential motif finder as well. One way to solve this problem is to use simulated annealing methods; the temperature is lowered as the Gibbs sampler approaches its global minima. One way to introduce temperature into the Gibbs sampling method is as follows:

$$p(\xi_i = 1) = \frac{qe^{\frac{s_i}{T_{\text{eff}}}}}{qe^{\frac{s_i}{T_{\text{eff}}}} + (1-q)} \tag{5.40}$$

Decreasing T_{eff} could increase the chance of sampling for sites with high scores, resulting in convergence as the temperature is lowered.

5.4 Conclusions

In this chapter, an analytical method was developed to find differentially enriched motifs in DNA sequences. The classical Gibbs sampling method was expanded to take advantage of the information in a second set of sequences, or "negative set." By penalizing motifs found in the negative set, the differential Gibbs sampler avoids false positives, functionally-irrelevant motifs found in both the positive and negative set. The method uses the concept of a partition function to detect the presence of a motif in the negative set. Using several approximations, the change in the partition function in the negative sequence was estimated by sampling a new site in the positive sequence set in a fast and efficient way. This allowed the definition of an "effective" weight matrix that includes the motif occurrences from the first set and a contribution coming from the second set. This effective weight matrix can be used to search for new motif occurrences in the first data set, as in the classical Gibbs sampler. The method was offered in two versions. In one, based on the classical Bernoulli Gibbs sampler, multiple motif occurrences in each sequence are possible. In the other, each sequence has only one motif occurrence. The assumptions in the method were examined by testing on constructed data sets that included differential enriched motifs. It was shown that this method finds differentially enriched motifs even when they are not as strong as the motifs that are present in both sequence sets.

Bibliography

[1] C. B. Anfinsen. Principles that govern folding of protein chains. *Science*, 181(4096):223–230, 1973.

[2] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, 21(1-2):51–80, 1995.

[3] F. C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. Protein data bank - computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112(3):535–542, 1977.

[4] K. M. Biswas, D. R. Devido, and J. G. Dorsey. Evaluation of methods for measuring amino acid hydrophobicities and interactions. *J. Chromatogr. A*, 1000:637–655, 2003.

[5] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, 5(2):279–305, 1998.

[6] S. E. Brenner, C. Chothia, and T. J. P. Hubbard. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.*, 7(3):369, 1997.

[7] N. E. G. Buchler and R.A. Goldstein. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus. *J. Chem. Phys.*, 112(5):2533–2547, 2000.

[8] NEG. Buchler and RA. Goldstein. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins-Structure Function and Genetics*, 41(1):113, 1999.

[9] C. J. Camacho and D. Thirumalai. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.*, 71(15):2505–2508, 1993.

[10] L. R. Cardon and G.D. Stormo. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned dna fragments. *J. Mol. Biol.*, 223(1):159–170, 1992.

[11] H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, N. Wingreen, and C. Tang. Fast tree search for enumeration of a lattice model of protein folding. *J Chem Phys*, 116(1):352–359, 2002.

[12] C. Chothia. Proteins - 1000 families for the molecular biologist. *Nature*, 357(6379):543–544, 1992.

[13] K. A. Dill. Theory for the folding and stability of globular-proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[14] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[15] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The helical hydrophobic moment - a measure of the amphiphilicity of a helix. *Nature*, 299(5881):371–374, 1982.

[16] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Nat. Acad. Sci. US*, 81(1):140–144, 1984.

[17] D. Eisenberg, W. Wilcox, and A. D. Mclachlan. Hydrophobicity and amphiphilicity in protein-structure. *J. Cell Biochem.*, 31(1):11–17, 1986.

[18] F. Eisenhaber, F. Imperiale, P. Argos, and C. Frommel. Prediction of secondary structural content of proteins from their amino acid composition alone .1. new analytic vector decomposition methods. *Protein-Struct. Funct. Genet.*, 25(2):157–168, 1996.

[19] MR. Ejtehadi, N. Hamadani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad. Highly designable protein structures and inter-monomer interactions. *Journal of Physics A*, 31(29):6141–6155, 1998.

[20] MR. Ejtehadi, N. Hamadani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad. Stability of preferable structures for a hydrophobic-polar model of protein folding. *PRE*, 57(3):3298–3301, 1998.

[21] W. G. Fairbrother, R.F. Yeh, P.A. Sharp, and C.B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002.

[22] J. Fauchere and V. Pliska. Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of n-acetyl amino-acid amides. *Eur. J. Med. Chem.*, 18:369–375, 1983.

[23] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin. Why do protein architectures have boltzmann-like statistics. *Protein-Struct. Funct. Genet.*, 23(2):142–150, 1995.

[24] A. V. Finkelstein, A.M. Gutun, and A.Y. Badretdinov. Why are the same protein folds used to perform different functions. *FEBS Lett.*, 325(1-2):23–28, 1993.

[25] A. V. Finkelstein and O.B. Ptitsyn. Why do globular-proteins fit the limited set of folding patterns. *Prog. Biophys. Mol. Biol.*, 50(3):171–190, 1987.

[26] B. Futcher. Microarrays and cell cycle transcription in yeast. *Curr. Opin. Cell Biol.*, 12(6):710–715, 2000.

[27] X. Gallet, B. Charloteaux, A. Thomas, and R. Brasseur. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, 302(4):917–926, 2000.

[28] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restortion of images. *IEEE transactions on pattern analysis and machine intelligence*, 6(6):721–741, 1984.

[29] G. Getz, M. Vendruscolo, D. Sachs, and E. Domany. Automated assignment of scop and cath protein structure classifications from fssp scores. *Protein-Struct. Funct. Genet.*, 46(4):405–415, 2002.

[30] S. Govindarajan and R.A. Goldstein. Searching for foldable protein structures using optimized energy functions. *Biopolymers*, 36(1):43–51, 1995.

[31] S. Govindarajan and R.A. Goldstein. Why are some protein structures so common? *Proc. Natl. Acad. Sci. U. S. A.*, 93(8):3341–3345, 1996.

[32] S. Govindarajan and R.A. Goldstein. The foldability landscape of model proteins. *Biopolymers*, 42(4):427–438, 1997.

[33] S. Govindarajan, R. Recabarren, and R. K. Goldstein. Estimating the total number of protein folds. *Protein Struct. Funct. Genet.*, 35(4):408, 1999.

[34] E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Y. Grosberg. Primary sequences of proteinlike copolymers: levy-flight-type long-range correlations. *Phys. Rev. E*, 64(4):R40903, 2001.

[35] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[36] R. Helling, H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. The designability of protein structures. *J. Mol. Graph.*, 19(1):157–167, 2001.

[37] L. Holm and C. Sander. Touring protein fold space with dali/fssp. *Nucl. Acid Res.*, 26(1):316–319, 1998.

[38] B. Honig and A.S. Yang. Free-energy balance in protein-folding. *Adv.Protein Chem.*, 46:27–58, 1995.

[39] S. J. Hubbard, S. F. Campbell, and J. M. Thornton. Molecular recognition - conformational-analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, 220(2):507–530, 1991.

[40] A. Irback, C. Peterson, and F. Potthast. Evidence for nonrandom hydrophobicity structures in protein chains. *Proc. Nat. Acad. Sci. USA*, 93(18):9533–9538, 1996.

[41] A. Irback, C. Peterson, and F. Potthast. Identification of amino acid sequences with good folding properties in an off-lattice model. *Phys. Rev. E.*, 55:860–867, 1997.

[42] A. Irback and E. Sandelin. On hydrophobicity correlations in protein chains. *Biophysical Journal*, 79(5):2252–2258, 2000.

[43] A. Kabakcioglu, I. Kanter, M. Vendruscolo, and E. Domany. Statistical properties of contact vectors. *Phys. Rev. E*, 65(4), 2002.

[44] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.*, 16:1–62, 1959.

[45] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. W. Wyckoff, and D. C.Phihips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–668, 1958.

[46] A. R. Khokhlov and P. G. Khalatur. Protein-like copolymers: computer simulation. *Phyisca A*, 249(1-4):253, 1998.

[47] A. R. Khokhlov and P. G. Khalatur. Conformation-dependent sequence design (engineering) of ab copolymers. *Phys. Rev. Lett.*, 82(17):3456–3459, 1999.

[48] E. L. Kussell and E.I. Shakhnovich. Analytical approach to the protein design problem. *Phys. Rev. Lett.*, 83(21):4437–4440, 1999.

[49] K. F. Lau and K.A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[50] K. F. Lau and K.A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 87(2):638–642, 1990.

[51] C. E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, F. Neuwald, and J.C. Wootton. Detecting subtle sequence signals - a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[52] B. K. Lee and F. M. Richards. The interpretation of protein structures. estimation of static accessibility. *J. Mol. Biol*, 55:379–400, 1971.

[53] M. Levitt and C. Chothia. Structural patterns in globular proteins. *Nature*, 261:552–558, 1976.

[54] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of proteins. *Scinece*, 273(5275):666–669, 1996.

[55] H. Li, C. Tang, and N. S. Wingreen. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Lett.*, 79(4):765, 1997.

[56] H. Li, C. Tang, and N.S. Wingreen. Are protein folds atypical? *Proc. Natl. Acad. Sci. U. S. A.*, 95(9):4987–4990, 1998.

[57] H. Li, C. Tang, and N.S. Wingreen. Designability of protein structures: a lattice-model study using the miyazawa-jernigan matrix. *Proteins*, 49(3):403–412, 2002.

[58] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene-regulation problem. *J. Am. Stat. Assoc.*, 89(427):958–966, 1994.

[59] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*, chapter The Gibbs Sampler, pages 129–151. Springer, 2001.

[60] Lodish, Berk, Zipursky, Matsudaira, Baltimore, and Darnell. *Molecular Cell Biology*, chapter Regulation of Transcription Initiation, pages 341–403. W. H. Freeman and Company, 2001.

[61] R. Melin, H. Li, N.S. Wingreen, and C. Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. Chem. Phys.*, 110(2):1252–1262, 1999.

[62] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1091, 1953.

[63] J. Miller, C. Zeng, N.S. Wingreen, and C. Tang. Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins*, 47(4):506–512, 2002.

[64] S. Miller, J. Janin, A. M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196(3):641–656, 1987.

[65] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3):623, 1996.

[66] S. Moelbert, E. Emberly, and C. Tang. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci.*, 13(3):752–762, 2004.

[67] A. G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, 1995.

[68] H. Naderi-manesh, M. Sadeghi, S. Arab, and A. AM. Movahedi. Prediction of protein surface accessibility with information theory. *Protein-Struct. Funct. Genet.*, 42(4):452–459, 2001.

[69] A. F. Neuwald, J.S. Liu, and C.E. Lawrence. Gibbs motif sampling - detection of bacterial outer-membrane protein repeats. *Protein Sci.*, 4(8):1618–1632, 1995.

[70] U. Ohler and H. Niemann. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, 17(2):56–60, 2001.

[71] C. A. Orengo, D.T. Jones, and J.M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634, 1994.

[72] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[73] V. S. Pande, A. Y. Grosberg, C. Joerg, M. Kardar, and T. Tanaka. Freezing transition of compact polyampholytes. *PRL*, 77(17):3565–3568, 1996.

[74] V. S. Pande, A. Y. Grosberg, and T. Tanaka. Nonrandomness in protein sequences - evidence for a physically driven stage of evolution. *Proc. Nat. Acad. Sci. USA*, 91(26):12972–12975, 1994.

[75] V. S. Pande, A. Y. Grosberg, and T. Tanaka. Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Mod. Phys.*, 72(1):259–314, 2000.

[76] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 37:235–240, 1951.

[77] R. Pollock and R. Treisman. A sensitive method for the determination of protein-dna binding specificities. *Nucleic Acids Res.*, 18(21):6197–6204, 1990.

[78] S. Rackovsky. "hidden" sequence periodicities and protein architecture. *Proc. Nat. Acad. Sci. USA*, 95(15):8580–8584, 1998.

[79] F. P. Roth, J.D. Hughes, P.W. Estep, and G.M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat. Biotechnol.*, 16(10):939–945, 1998.

[80] E. Roulet, S. Busso, A.A. Camargo, A.J.G. Simpson, N. Mermod, and P. Bucher. High-throughput selex-sage method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, 20(8):831–835, 2002.

[81] R. Schwartz, S. Istrail, and J. King. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci.*, 10(5):1023–1031, 2001.

[82] V. Shahrezaei and M. R. Ejtehadi. Geometry selects highly designable structures. *J. Chem. Phys.*, 113(15):6437, 2000.

[83] C. T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, and H.C. Lee. Mean-field hp model, designability and alpha-helices in protein structures. *Phys. Rev. Lett.*, 84(2):386–389, 2000.

[84] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 30(24):5549–5560, 2002.

[85] P. T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.

[86] G. D. Stormo and G.W. Hartzell. Identifying protein-binding sites from unaligned dna fragments. *Proc. Natl. Acad. Sci. U. S. A.*, 86(4):1183–1187, 1989.

[87] G. D. Stormo, T.D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the perceptron algorithm to distinguish translational initiation sites in escherichia-coli. *Nucleic Acids Res.*, 10(9):2997–3011, 1982.

[88] B. J. Strait and T. G. Dewey. Multifractals and decoded walks - applications to protein-sequence correlations. *PRE*, 52(6):6588–6592, 1995.

[89] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genet.*, 22(3):281–285, 1999.

[90] J. van helden, B. Andre, and J. Collado-vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281(5):827–842, 1998.

[91] T. R. Wang, J. Miller, N.S. Wingreen, C. Tang, and K.A. Dill. Symmetry and designability for lattice protein models. *J. Chem. Phys.*, 113(18):8329–8336, 2000.

[92] Z. X. Wang. How many fold types of protein are there in nature? *Proteins*, 26(2):186–191, 1996.

[93] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287, 2004.

[94] T. R. Weikl and K. A. Dill. Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.*, 332(4):953–963, 2003.

[95] O. Weiss and H. Herzel. Correlations in protein sequences and property codes. *Journal of Theoretical Biology*, 190(4):341–353, 1998.

[96] J. Wilder and E. I. Shakhnovich. Proteins with selected sequences: a heteropolymeric study. *Phys. Rev. E*, 62(5):7100–7110, 2000.

[97] C. T. Workman and G. D. Stormo. Ann-spec: a method for discovering transcription factor binding sites with improved specificity. *Pacific Symposium on Biocomputing*, 5, 2000.

[98] M. Yahyanejad, M. Kardar, and C. Tang. Structure space of model proteins: a principle component analysis. *J. Chem. Phys.*, 118(9):4277–4284, 2003.

[99] K. Yue and K.A. Dill. Forces of tertiary structural organization in globular-proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 92(1):146–150, 1995.