Protein structure recognition: From eigenvector analysis to structural threading method

by

Haibo Cao

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Condensed Matter Physics

Program of Study Committee:
Kai-Ming Ho, Major Professor
Drena Dobbs
Amy Andreotti
Alan Goldman
Bruce Harmon
Joerg Schmalian

Iowa State University

Ames, Iowa

2003

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Haibo Cao

has met the dissertation requirements of Iowa State University

Major Professor

For the Major Program

TABLE OF CONTENTS

LIST OF FIGURES

Acknowledgments

I would like to take this opportunity to thank those people without whose help this thesis would not has been possible. My gratitude goes to my academic advisor, Professor KaiMing Ho, whose insigh, encouragement, guidance and supporting are invaluable to me. Thanks Dr. Caizhuang Wang for his inspiring discussions and encouragment during the time of my Ph. D. study in Iowa state university. My gratitude and appreciation also go to Dr. Drena Dobbs, whose biology knowledge is always a treasure to me. Her clear explaination and inspiring discussions are eccential to the acomplishent in this project.

I would like to thank my committee members, Professors Bruce Harmon, Alan Goldman, Joerg Schmalian, Amy Andreotti for their valuable comments and discussions. I would like to thank Ihm Youngok, who is my colaborator and group member, for her support and useful discussions. I would like also thank Tzs-Liang Chan, who is my officemate and group member, for his support, discussions and help in writing this thesis.

My deep appreciation must goes to our group adminestrator Rebecca Shivvers, who always works extremely hard to guide me through the maze of administrative obstacles and is always on hand to sort things out when I get it wrong. Thank Larry Stoltenberg for helping me preparing oral defence.

Thank all my many other friends who have made my life here ar Ames over the past several years enjoyable.

Thank Department of Physics and Astronomy, Institute for Physical Research and Technology, the Plant Science Institute, the Biotechnology Council, the Lawrence H. Baker Center for Bioinformatics and Biological Statistics at Iowa State University, and Ames laboratory for financial support, and the Scalable Computer Laboratory for computational support in this

project.

Abstract and Organization of Dissertation

In this work, we try to understand the protein folding problem using pair-wise hydrophobic interaction as the dominant interaction for the protein folding process. We found a strong correlation between amino acid sequence and the corresponding native structure of the protein. Some applications of this correlation were discussed in this disseration inlcude the domain partition and a new structural threading method as well as the performance of this method in the CASP5 competition.

In the first part, we give a brief introduction to the protein folding problem. Some essential knowledge and progress from other research groups was discussed. This part include discussions of interactions among amino acids residues, lattice HP model, and the designablity principle.

In the second part, we try to establish the correlation between amino acid sequence and the corresponding native structure of the protein. This correlation was observed in our eigenvector study of protein contact matrix. We belive the correlation is universal, thus it can be used in automatic partition of protein structures into folding domains.

In the third part, we discuss a threading method based on the correaltion between amino acid sequence and ominant eigenvector of the structure contact-matrix. A mathematically straightforward iteration scheme provides a self-consistent optimum global sequence-structure alignment. The computational efficiency of this method makes it possible to search whole protein structure databases for structural homology without relying on sequence similarity. The sensitivity and specificity of this method is discussed, along with a case of blind test prediction.

In the appendix, we list the overall perfomance of this threading method in CASP5 blind test in comparation with other existing approaches.

CHAPTER 1. Protein Structure And Building Blocks

In this part, I will briefly discuss some known properties of protein structures as well as basic knowledge important in discussing the protein folding problem.

Proteins

Proteins are chain-like polymers of small subunits (amino acids) [1, 2, 3, 4, 5, 6, 7, 8]. There are twenty different amino acids which occur in nature. The structures of these compounds are shown in Figure 1.1 [1]. Each amino acid has a central carbon atom (C_α) connected to an amino group(NH3+), a carboxyl group(COO-), a hydrogen atom(H), and a side chain. Different amino acids are distinguished by their side chains. The sequential arrangement of the amino acids in the protein gives each protein its unique character.

The amino acids are joined together in proteins via peptide bonds. This gives rise to the name polypeptide for a chain of amino acids. A protein can have one or more polypeptide chains. The atoms involved in forming covalent peptide (N-C_α) bonds and the C' atoms of all the amino acids in a protein are called the backbone of a protein. Because the peptide bond is planar, the backbone configuration of a protein is completely determined by the ϕ (between N and C_α), ψ (between C_α and C') angles as shown in Figure 1.2 [1].

A polypeptide chain has polarity with a free amino group left at one end called the amino terminus or N-terminus, and a free carboxyl group at the other end, called carboxyl terminus or C-terminus.

The linear order of amino acids consititutes a protein's primary structure. The way these amino acids interact locally with their neighbours give a protein its secondary structure [9, 10, 11, 12, 13, 14, 15, 16, 17]. The alpha-helix is a common form of secondary structure. It results

from hydrogen bonding between backbone atoms of near neighbour amino acids, as shown in Figure 1.3 [1]. Another common secondary structure found in proteins is the beta-pleated sheet which is shown in Figure 1.4 [1]. This involves extended protein chains, packed side by side, that interact by hydrogen bonding between backbone atoms.

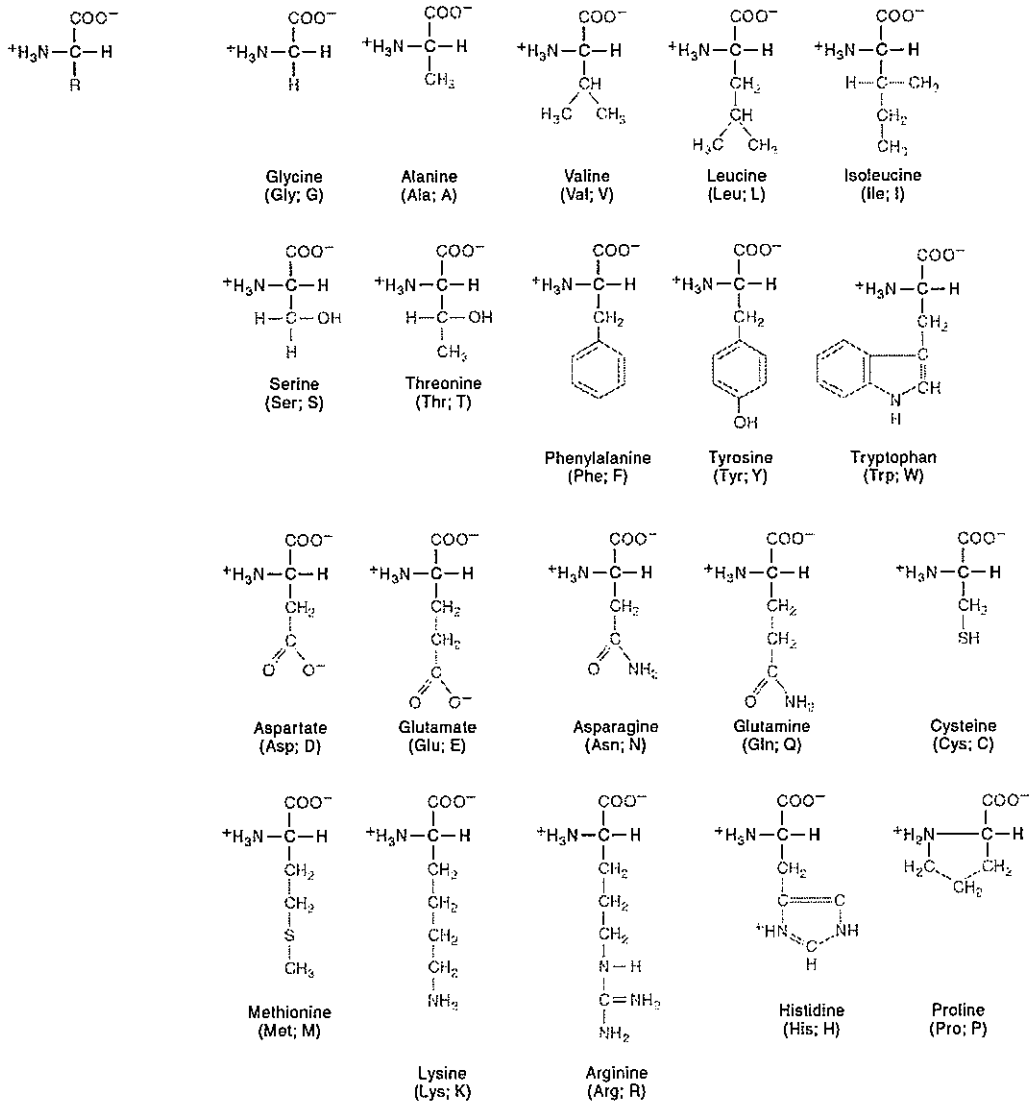

Figure 1.1 A list of natural occuring amino acids [1].

The total three-dimensional shape or "fold" of a polypeptide is its tertiary structure. Figure 1.5 [1] illustrates how the protein myoglobin folds up into its tertiary structure. The native structures for some proteins can contain more than one compact structural regions. Each of

these regions is called a domain. Quaternary structure [18, 19, 20, 21, 22] is the way two or more individual polypeptides fit together in a multi-unit protein.

Amino Acids

Amino acids are the building units of proteins which sometimes are also called residues. The 20 naturally occuring amino acids interact with each other via non-covelant interactions to form the global three dimensional structure of a protein. Figure 1.1 lists these amino acids with their names and abbreviations using three- letter or one-letter code. These amino acids can be divided into three classes according to the chemical properties of their side chains. ALA(A), VAL(V), LEU(L), ILE(I), PHE(F), MET(M) and PRO(P) are usually considered hydrophobic residues. SER(S), THR(T), CYS(C), ASN(N), GLN(Q), HIS(H), TYR(Y), TRP(W), and GLY(G) are polar, although CYS, TYR, TRP also have considerable hydrophobic character as well and HIS is sometimes charged. ASP(D), GLU(E), LYS(K) and ARG(R) are charged residues. These multi-atomic residue side chains may have different geometry in forming a protein's three dimensional structure. We analyzed the existing protein structure database to obtain a measurement of the sizes of the different amino acids. For every amino acid, we use the geometrical center of the side chain to represent its side chain position. The distance between the amino acid's C-alpha atom and its geometrical center is defined as the "size" of the residue. The average sizes of the 20 amino acids are listed in Table 1.1.

C	2.8	M	3.5	F	3.8	I	2.5	L	3.0
V	2.2	W	4.2	Y	4.2	A	1.5	G	1.5
T	2.2	S	2.4	N	2.9	Q	3.4	D	2.9
E	3.4	H	3.6	R	4.4	K	3.75	P	2.3

Table 1.1 Amino acid size obtained from PDB

The amino acid cysteine is special compared with other residues in constructing protein structures. Two cysteines can form a disulfide bridge via the following chemical reaction:

$$CH_2SH + CH_2SH + 1/2O_2 \longrightarrow CH_2 - S - S - CH_2 + H_2O \tag{1.1}$$

In some proteins, these covalent disulfide bonds are important to hold the polypeptide together in forming a functional structure (e.g. disulfide bonds are important in forming the three dimensional conformation of tumor necrosis factor(TNF)-receptors in human).

Secondary Structure

In 1958, the first X-ray crystallographic structure for a globular protein (myoglobin) was determined by John Kendrew [6]. Instead of the simple double-stranded helical structure found for DNA molecules, globular proteins adopt very complicated conformations. Proteins take on irregular and dramatically diverse structures in general. However, there are some common features: Globular proteins generally contain compact hydrophobic cores [20]. Most of the time, residues are so densely packed in the core that there is no space available for water molecules. In the few cases where there is a "hole" in the hydrophobic core, the water molecules inside are generally hydrogen-bonded with polar residues and thus can be viewed as part of the protein structure.

Another common feature in protein structures are local secondary structures. The most common secondary structures are α helix and β sheet.

α Helix

The α helical configuration of protein structure was first proposed by Linus Pauling [16, 17]. In forming α helix, a stretch of consecutive amino acids all have the ϕ, ψ angle pair approximately $-60°$ and $-50°$. The most common α helix has 3.6 residues per turn. For every residue n of a helix, its $C' = O$ is hydrogen bonded with the NH of the n+4 residue. All the residues in the helix are hydrogen-bonded together except the two boundary residues.

There are some variations of helical secondary structure in which the chain can be a little loose or tightly coiled. The hydrogen bonds in this case can be n to $n+5$, or n to $n+3$. They are called π helix and 3_{10} helix respectively. The 3_{10} helix has 3 residues per turn. Both π helix and 3_{10} helix occur rarely in protein structures.

The length of α helices in globular proteins can be varied. The average length of an α

helix is around 10 residues. Almost all α helices observed in proteins are right-handed except for, occasionally, a few short left-handed α helical regions (3-5 residues). The side chains of the amino acids are not directly involved in forming α helix. However, different amino acids have different propensities in forming α helical conformation. ALA(A), GLU(E), LEU(L) and MET(M) are believed to be good α helix formers, while GLY(G), TYR(Y) and SER(S) are rarely found in α helical formation. The amino acid PRO(P) is a α helix "breaker", since the main-chain atom N in proline forms a ring structure with the side chain, thus depriving it of the ability to form hydrogen bonds in a α helix[1].

β sheet

Another common secondary structure in proteins is the β sheet. Unlike α helices which are formed by a continuous stretch of amino acids, the β sheet is built from a combinations of several regions of a polypetide chain. These regions are often called β strands, and usually consist of 5 to 10 residues. Like α helices, hydrogen bonds play an important role in holding the strands together. Generally, in forming β sheet, the $C^{'} = O$ of one amino acid in a strand forms a hydrogen bond with the NH group of an amino acid in the adjacent strand.

There are two ways that β strands interact with each other to form a β sheet. If the amino acids in the aligned β strands all run in the same direction, it is called parallel β sheet. The directions of neighbouring β strands can also run in opposite directions. In this case, it is called antiparallel β sheet. β strands can also be combined into mixed β sheets with some of the β strand pairs parallel and some antiparallel. For known proteins, almost all β sheets have "twisted" strands, and the "twisted" strands are always twisted in a right-handed way.

Certain combinations of secondary structure elements with specific geometric arrangements has been found frequently in protein structures. They are often called structural "motifs". Some of the motifs are associated with particular biological functions such as DNA binding. Some examples of more frequent motifs are: the hairpin β motif, the Greek key motif, and the zinc finger motif.

Combinations of motifs and secondary structure regions can build up a structural "domain".

A structural domain is defined as part of a protein that can fold independently and is considered to be the building block of protein tertiary structure. A protein may have a single domain or multiple domains. Generally, a structural domain is also a unit of function. It is very often that different domains of a protein are associated with different functions[1].

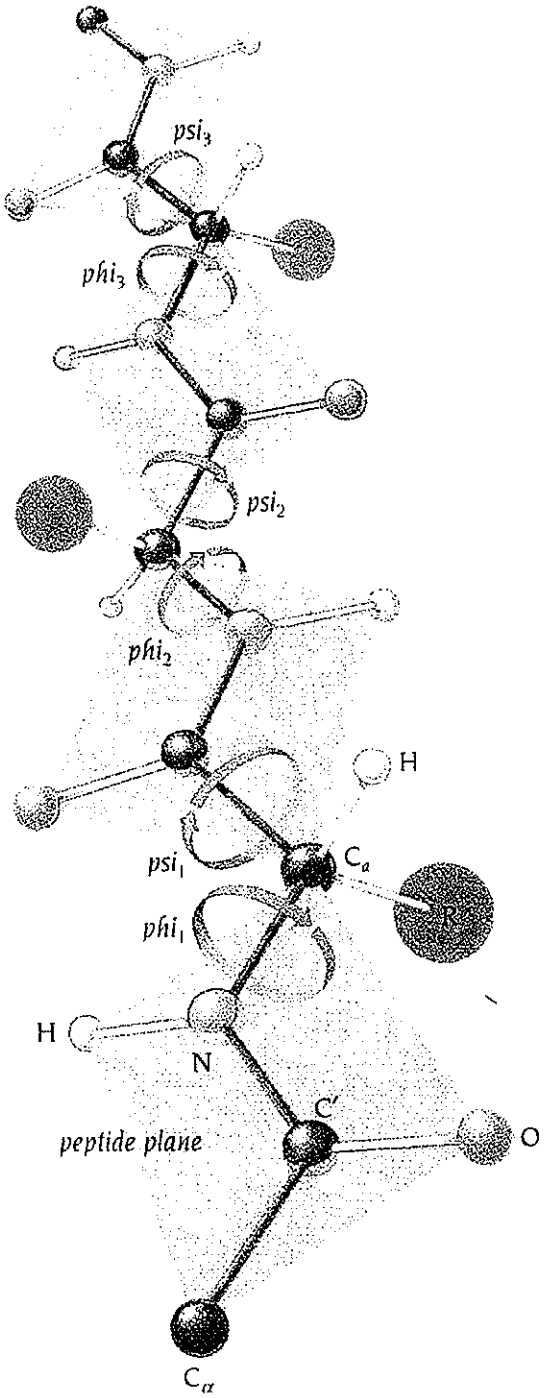

Figure 1.2 Backbone of protein [1]

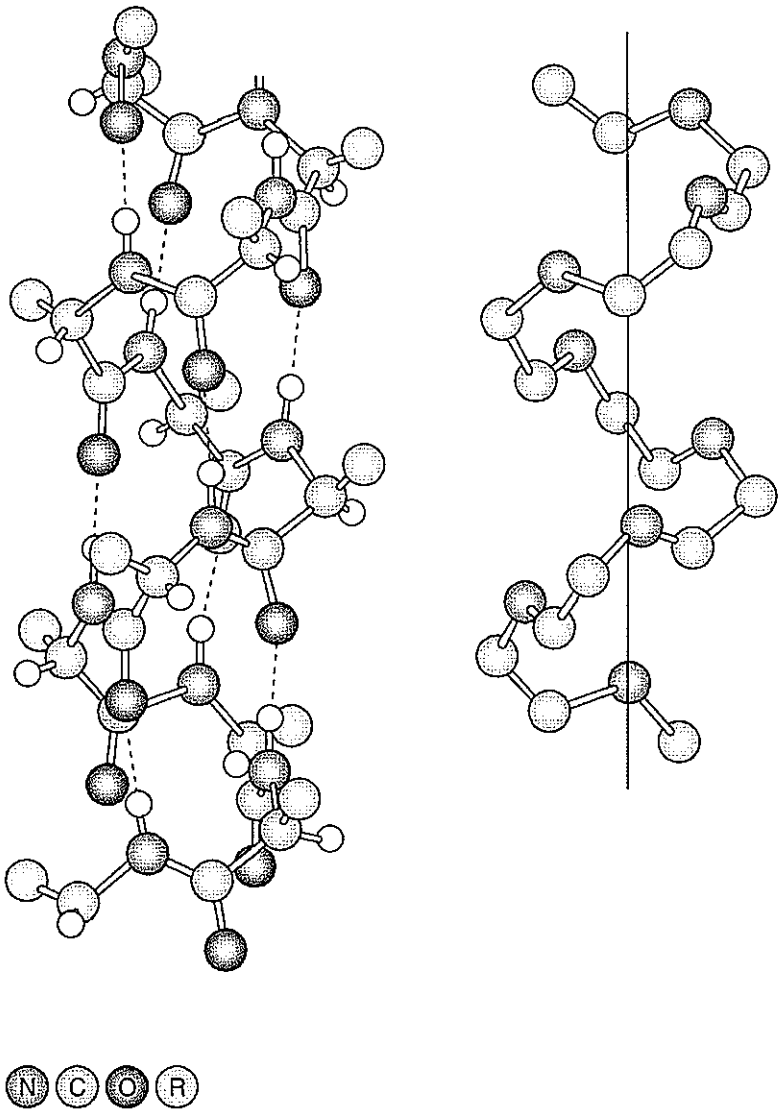

Figure 1.3 α helix formation of protein local structure[1].

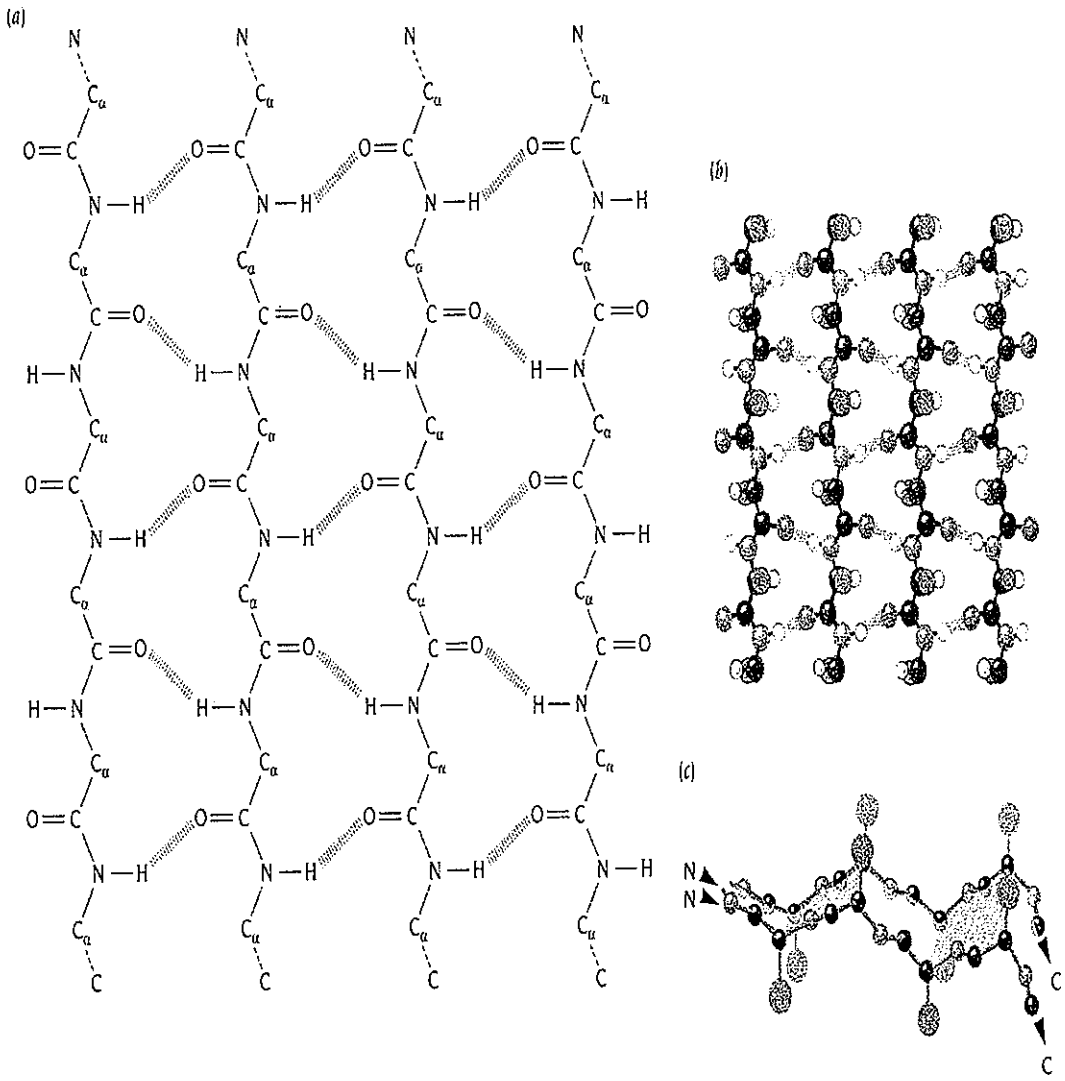

Figure 1.4 β formation of protein local structure[1].

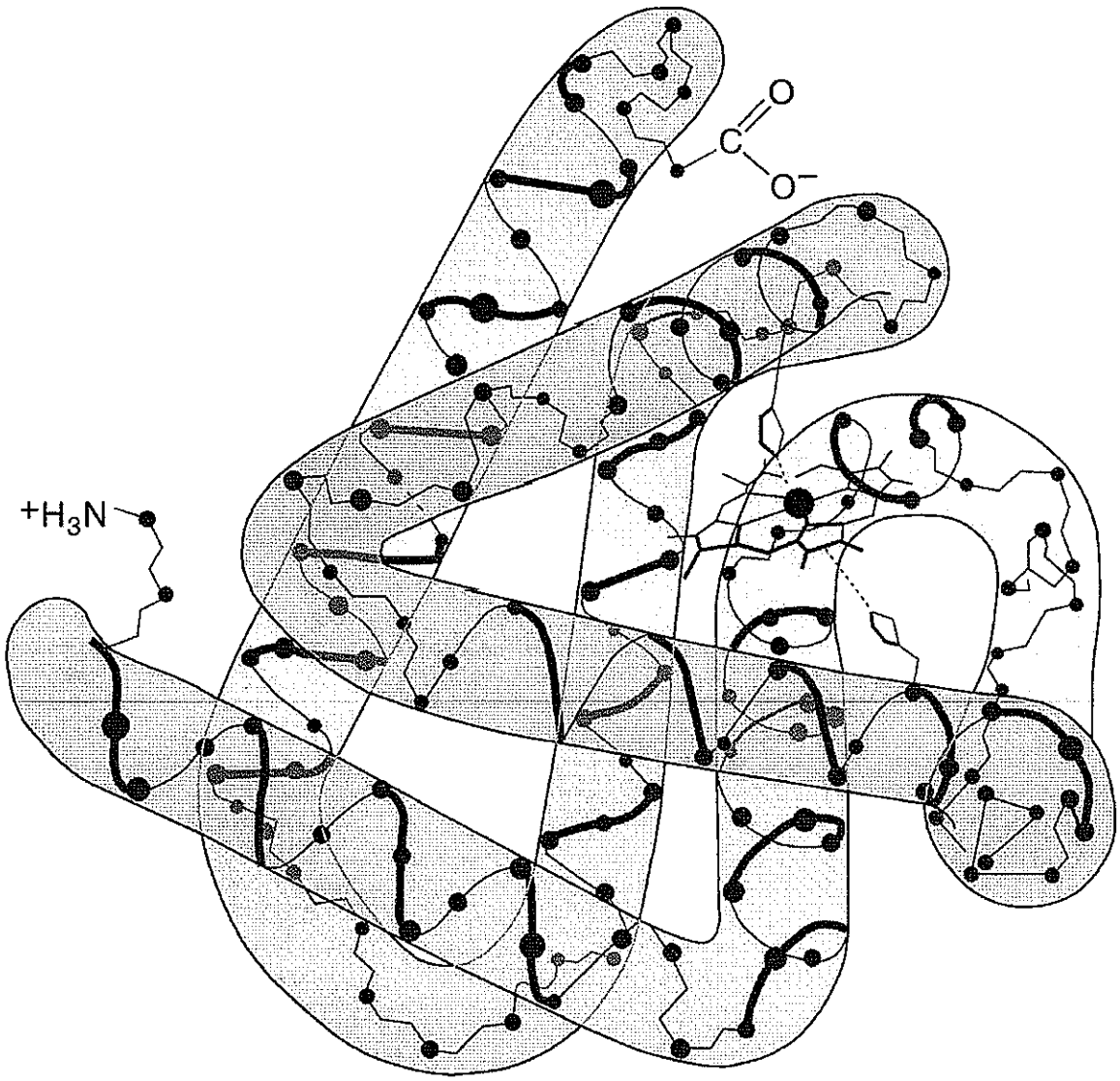

Figure 1.5 Tertiary structure of a protein[1].

Bibliography

[1] Branden, C. Tooze J. Introduction to protein structure Garland Publishing Inc. 1998

[2] Schulz GE. Structural rules for globular proteins Angew. Chem., int. 1977

[3] Rossmann, MG. Argos, P. Protein folding Annu. Rev. Biochem. **50** 497-532 1981

[4] Chothia, C. Principles that determine the structure of proteins. Annu. Rev. Biochem **53** 537-572 1984

[5] Doolittle, RF. Proteins. Sci. Am. **253** 88-99 1985

[6] Kendrew, JC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature **181** 662-666 1958

[7] Kendrew, JC. Structure of myoglobin. Nature **185** 422-427 1960

[8] Kendrew, JC. The three dimensional structure of a protein molecule. Sci. Am. **205** 96-110 1961

[9] Lesk, AM Themes and contrasts in protein structures. Trands Biochem. Sci. **9** June V, 1984

[10] Levitt, M. Chothia, C. Structural patterns in globular proteins. Nature **261** 552-558 1976

[11] Schulz, GE. Protein differentiation: emergence of novel proteins during evolution. Angew. Chem. Int 1981

[12] Chothia, C. Levitt, M. Richardson, D. Structure of proteins: packing of α helices and pleated sheets. Proce. Natl. Acad. Sci. USA **74** 4130-4134 1977

[13] Efimov, AV. Stereochemistry of α helices and β sheet packing in compact globule. J. Mol. Biol. **134** 23-40 1979

[14] Hol, WGJ. Van Duijnen, PT. Berendsen, HJC. The α helix dipole and the properties of proteins. Nature **273** 443-446 1978

[15] Levitt, M. Conformational preferences of amino acids in globular proteins. Biochemistry **17** 4277-4285 1978

[16] Pauling, L. Corey, RB. Configurations of polypeptide chains with favored orientation around single bonds: two pleated sheets. Proc. Natl. Acad. Sci. USA **37** 729-740 1951

[17] Pauling, L. Corey, RB. Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA **37** 205-211 1951

[18] Rao, ST. Rossmann, MG. Comparison of supersecondary structures in proteins. J. Mol. Biol. **76** 241-256 1973

[19] Rose, GD. Automatic recognition of domains in globular proteins. Methods Enzymol. **115** 430-440 1985

[20] Rose, GD. Prodiction of chain turns in globular proteins on a hydrophobic basis. Nature **272** 586-590 1978

[21] Sibanda, BL. Thornton, JM. β hairpin families in globular proteins. Nature **304** 654-657 1983

[22] Matthews, BW. Rossmann, MG. Comparison of protein structures. Method Enzymol. **115** 397-420 1985

CHAPTER 2. Protein Folding Problem

Proteins may be classified into three categories: fibrous,membrane and globular. Globular proteins form compact three dimensional configurations under natural conditions.

In 1960's Christan B. Anfinsen and co-workers show that globular proteins can fold reversibly [1]. It is now established that globular proteins in cells have unique three dimensional structures, which are called their "native structures". A protein's native configuration is important for its function [2]. For an example, the protein hemoglobin carries oxygen from lungs to remote areas of the body. A normal cell hemoglobin(HbA) and a sickle-cell hemoglobin(HbS) differ only by one amino acid which results in a small conformational difference between the two proteins. Normal hemoglobin remains soluble under ordinary physiological conditions, but the sickle-cell hemoglobin precipitates when the blood oxygen level falls, forming long, fibrous aggregates that distort the blood cell into the sickle shape. This may be fatal without medical attention. The native strcuture of a protein is uniquely determined by the protein's amino acids sequence as discussed in the previous part of this thesis.

Levinthal's Paradox

Thermodynamically, the native state of a protein is believed to be the lowest free energy state among all the conformations it can take. The conformational space of a polypeptide chain is huge. Considering an ordinary protein with length of 150 residues, and assume that for each residue, the ϕ, ψ angles bewteen the backbone atoms have around 10 possible distinct values, the total number of possible conformations will be around 10^{150} for the backbone alone. However, natural proteins fold in time from milliseconds to seconds. How can natural protein find their native structures in such a short time? This is the famous "Levinthal's paradox"

raised by Cyrus Levinthal in 1960s [3, 4]. Levinthal's argument suggests that there are "folding pathways" in the protein folding process. Soon after Levinthal's argument, experimentalist started to search for folding-intermediate states. In 1971, Ikai, Tanford [5] and Tsong, Baldwin and Elson [6] published their pioneering experimental paper [7, 8]. Those experiments used disulfide bonds to trap folding intermediates (e.g. in BPT1 [9]), or use proline isomerization [10] to study the process of folding. However, folding intermediates and complex kinetics can still be observed in the absence of disulfide bonds and incorrect proline isomers[11, 12]. In the classical view, it is assumed that there are intermediate states I_i for protein folding from a denatured state U to the native state N:

$$U \rightarrow I_1 \rightarrow I_2, \ldots, \rightarrow N \tag{2.1}$$

A "new view" proposed by Peter G. Wolynes [13, 14, 15, 16] and co-workers replaced the pathway concept of a sequence of events with a funnel concept of parallel events. They argue that in order for the protein to fold in such a short time, the "energy landscape" of a protein must be "funnel" like. The results from lattice HP models [15, 16] support this view. In the "new view", it is not neccessary for a denatured protein to go through given intermediates in a "pathway" to reach the final native structure, but it can "fall down the hill" quickly as long as the energy landscape is funnel-like. The funnel-like energy landscape can help us understand some of the experiments. For example, some proteins in cells must be assisted in their folding by "chaperone" proteins. When a protein falls into a misfolded local minimum, the "chaperone" protein causes the misfolded protein to unravel and overcome the energy barrier to start "falling down the hill" again. These "chaperone" proteins do not have to be specific to a particular protein in order to interact correctly, but rather give help by providing an "environment" to let misfolded proteins unfold. Thus, one kind of "chaperone" protein can help many different proteins fold.

Interactions

In order to determine the energy landscape of a protein, one must understand the interactions among the residues. There are various kind of interactions involved in the folding process. These interactions include hydrogen bonds, hydrophobic interactions, electrostatic interactions, van der Waals interactions and disulfide bonds. Even a small protein contains tens of thousands of atoms as well as surrounding water molecules. Because the vibrational motions of these bonded atoms have times of about 10^{-14} to 10^{-13} seconds, it is impossible for existing computers to simulate the dynamic process of folding which occurs on a time scale of around 10^{-1} to 10^{3} seconds. Thus, except for studies where the protein is restricted to vary around a given three dimentional structure(e.g. simulations in the final refinement of experimental NMR structures), the interactions used in theoretical protein folding study are generally described at the residue level. These effective interactions generally use one bead (C_α position) or two beads (C_α and center of side chain) in representing amino acids. Even at the residue level, there are still disagreements about what is the dominant force for protein folding. Historically, Linus Pauling [17] was the first to propose that hydrogen bond interactions drive the folding process. He based this conclusion on his experimental studies of membrane proteins. However, in 1950s, Walter Kauzmann [18] argued that hydrogen bonding would not strongly favor the folded state compared with the unfolded state, since water molecules can also form hydrogen bond with amino acids in the unfolded structures. From the observation that almost all globular proteins form a compact hydrophobic core, he proposed that the hydrophobic interaction is a stronger force for folding proteins. The mixing of nonpolar (oil-like or hydrophobic) molecules with water has a large positive free energy, and is disfavored by entropy near room temperature, leading to a large increase in heat capacity. Hydrophobic residues like to segregate to form a hydrophobic core, while the polar residues are more abundant on the surface. At present, it is believed that the driving force for protein folding is the hydrophobic interaction for the majority of globular proteins. However, hydrogen bonding interactions and other non-covalent interactions(e.g. electrostatic) are important in stabilizing the resulting native structure. In some special cases, covalent interactions like disulfide bonds can be crucial in determining the

native conformation when the protein is cysteine rich.

MJ matrix

In order to study hydrophobic interactions in detail at the residue level, some assumptions have to be made. First of all, since only the side chain of an amino acid is involved in this interaction, it is reasonable to represent the amino acid by a "bead" at the geometrical center of the side chain. This takes the form of a "contact" interaction. When the centers of two "bead" are within a cutoff distance (in our study, we choose $6.5\mathring{A}$), the two amino acids are "in contact". If the amino acids in contact are both hydrophobic, part of their surface areas are covered by each other and inaccessible to solvent molecules or other polar amino acids. A contact energy is assigned to this pair of residues according to the amino acids involved. The total conformational hydrophobic interaction energy of a protein is the sum of all the pair-wise contact energies for a given three dimensional structure. Because different amino acids differ in their hydrophobicity, the contact energies between different amino acid pairs are different. There are various interaction energy schemes to account for this difference in pair-wise contact energy. The simplest one is called the HP model. In the HP model, the 20 natural amino acids are classified into two groups: Hydrophobic(H) and polar(P). The contact energy for a polar residue with another polar residue (PP contact) is generally assigned to be 0, and HH contact is assigned to be 1. Different groups may choose different HP contact energy schemes. For example, Chan and Dill use 0 for the HP contact in their study, while Li, Tang and Wingreen used 1/2.3 . Because the HP model is generally used in qualitative studies (most of the groups using HP model also restrict the structure of protein to be on a lattice; lattice HP models will be described in more detail in chaper VI), this difference in assigning HP contact energies makes little difference as long as the HP interaction scheme satisfies the following condition:

$$E_{HH} > E_{HP} > E_{PP} \tag{2.2}$$

the results will be similar.

Obviously, the two alphabet HP model is over-simplified for the purpose of calculations of

realistic proteins, A more detailed pair-wise residue-residue interaction scheme was proposed by Sanzo Miyazawa and Robert Jernigan (the MJ matrix). The MJ matrix is a 20×20 matrix, with element $\epsilon_{i,j}$ as the contact energy between i type residue and j type residue. Miyazawa and Jernigan [19, 20, 21] studied the pair-frequency of real proteins in the Protein Data Bank(PDB). If we look at a specific protein structure, the frequency of two amino acids in contact is not only influenced by the interaction between these two amino acids, but also affected by the chain connectivity constraint. However, by including many different protein structures, the effect of chain connectivity might be averaged out. The basic assumption for getting the MJ matrix is that the frequency of two amino acids in contact is correlated with the strength of their interaction. In order to take into account the hard-core repulsions among residues and solvent molecules, they assume the residues occupy lattice sites in a linear chain fashion. The vacant sites are assumed to be occupied by solvent molecules, and for simplicity, they assume interactions occur only between nearest-neighbor pairs of amino acids and solvent molecules. Because different amino acids have different sizes, the number of nearest neighbors around a residue (the total number of contacts between residues or between residue and solvent molecule which is also called te coordination number of the residue) depends only on the size of the amino acid. The contacts between sequential neighbours are excluded, because these result from chain connectivity. If q_i is the coordination number of residue type i, and the number of residues of the ith type in all protein strcutures studied is n_i, then the following relationship should be satisfied:

$$q_i n_i / 2 = \sum_{i=0}^{20} n_{ij} \tag{2.3}$$

where $n_{ij} = n_{ji}$ is the total number of contact between ith type of residue and jth type residue. $i = 0$ represent the solvent molecule. If we define E_{ij} as the interaction energy between residue type i and residue type j, and relative energy e_{ij} as:

$$e_{ij} = E_{ij} + E_{00} - E_{i0} + E_{j0} \tag{2.4}$$

which is the energy "gain" for forming the i-j contact compared with exposing these two residues to solvent. Then, the total energy for th system is:

$$TotalEnergy = \sum_{i=0}\sum_{j=0} E_{ij}n_{ij} = \sum_{i=0}(2E_{i0} - E_{00})q_i n_i/2 + \sum_{i=1}\sum_{j=1} e_{ij}n_{ij} \tag{2.5}$$

The first summation in this equation is not relevant to the structures of protein (n_{ij}), and can be neglected in the following discussion. The partition function of the system in the Bethe approximation can be estimated as:

$$Z = constant \sum_{n_{ij}} \frac{n_{r0}!n_{0r}!n_{rr}!}{\Pi_{i=1}n_{i0}!\Pi_{j=1}n_{j0}!\Pi_{i=1}\Pi_{j=1}n_{ij}!} exp\left(-\sum_{i=1}\sum_{j=1} e_{ij}n_{ij}\right) \tag{2.6}$$

where n_{r0} is the total number of residue to solvent contacts and n_{rr} is the total number of contacts in the system. we use RT as the energy unit in the above equation. For this system, the n_{rr} and n_{r0} must satisfy the following constraints:

$$\sum_{i=1}\sum_{j=1} n_{ij} = n_{rr} \tag{2.7}$$

$$\sum_{i=1} n_{i0} = n_{r0} \tag{2.8}$$

$$\sum_{j=1} n_{j0} = n_{0r} \tag{2.9}$$

Maximizing the partition function under these contraints, the statistical average $< n_{ij} >$ of n_{ij} correlates with the interaction energy:

$$exp(-\Delta e_{ij}) = \frac{< n_{ij} > n_{r0}n_{0r}}{n_{rr} < n_{i0}n_{0i} >} \tag{2.10}$$

where $\Delta e_{ij} = e_{ij} - e_{rr}$, and the constant e_{rr} that is called the collapse energy is defined as:

$$exp(-e_{rr}) = \left(\frac{< n_{ij} > exp(e_{ij})}{n_{rr}}\right)^{-1} \tag{2.11}$$

The pair-wise contact freuencies obtained from the PDB are used to determine the values of e_{ij}. This MJ matrix is shown in Figure 2.1 (directly copied from Miyazawa and Jernigen's paper[19]). e_{ij} are the upper half triangle of this figure.

Sippl MJ. [22, 23] used the mean force approximation to obtain a similar empirical residue-residue interaction potential.

LTW parameterization

In 1997, Hao Li, Chao Tang and Ned S. Wingreen [24] completed an eigenvector analysis of the MJ matrix. Mathematically, a given $N \times N$ real symmetrical matrix M can be represented using the eigenvectors of M as:

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha V_{\alpha,i} V_{\alpha,j} \tag{2.12}$$

where λ_α is the αth eigenvalue of M, and V_α is the corresponding eigenvector. Tang and Wingreen found that the eigenvalue spectrum of MJ matrix is very abnormal. There are two dominant eigenvalues which are much larger in magnitude than the rest. $\lambda_1 = -22.49$ and $\lambda_2 = 18.62$. while the rest of the eigenvalues have absolute values between 2.17 and 0.013, which suggests that the MJ matrix can be accurately reconstructed using only two eigenvectors:

$$M_{ij} \simeq < M_{ij} > + \lambda_1 V_{1,i} V_{1,j} + \lambda_2 V_{2,i} V_{2,j} \tag{2.13}$$

They also found that the two eigenvectors V_1 and V_2 are correlated. Approximately, V_2 can be obtained from V_1 by a shift and rescaling:

$$V_{2,i} = \beta + \gamma V_{1,i} \tag{2.14}$$

Where $\beta = -0.30$ and $\gamma = -0.90$. The correlation coefficient is 0.986. This indicates that the MJ matrix can be effectively described using only 20 parameters,

$$M_{ij} \simeq C_0 + C_1(q_i + q_j) + C_2 q_i q_j \tag{2.15}$$

with the constants $C_0 = < M_{ij} > = -1.492$, $C_1 = 5.030$ and $C_2 = -7.400$. This reconstructed matrix reproduces the original MJ matrix with very high accuracy. The Figure 2.2 copied from Li, Tang and Wingreen's paper[24] shows the correlation between these two matrices. The regression line is $y = 0.999x + 0.008$ and the correlation coefficient is 0.989. The inset of the figure shows the distribution of the original MJ matrix elements.

Li, Tang and Wingreen found that there is an obvious "gap" in the q values of the 20 amino acids[24] (see Figure 2.3[24]).

The 20 residues can be roughly categorized into two groups of 8 and 12, respectively. The two groups are differentiated by their hydrophobicity. Rewriting the LTW parametrization:

$$M_{ij} = h_i + h_j - C_2(q_i - q_j)^2 \tag{2.16}$$

where h_i are defined as:

$$h_i = C_0/2 + C_1 q_i + (C_2/2)q_i^2 \tag{2.17}$$

In a "mixing" process, four residues i, i, j, j break initial contacts :i-i and j-j. to form two new contacts: i-j and i-j. , The energy gain for this "mixing" process is $\chi_{ij} = -C_2(q_i - q_j)^2$. This form is very similar to the mixing energy of two simple liquids as given by Hildebrand's solubility theory[25]. The h_i's defined above correspond to the hydrophobicities of the corresponding amino acids. Figure 2.4[24] (copied from Li, Tang and Wingreen's paper) clearly show this correlation. Thus, we believe the hydrophobic interaction is the dominant effect in the MJ matrix.

Contact Energies in RT Units Estimated by Method-D: $\Delta e_{ij} (\equiv \epsilon_{ij} - \epsilon_{rr})$ for Upper Triangular Half and Diagonal

	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO	
	-1.19	-0.61	-0.67	-0.64	-0.65	-0.59	-0.66	-0.39	-0.33	-0.31	-0.15	-0.13	-0.07	-0.01	0.20	0.12	-0.36	0.08	0.33	-0.18	CYS
CYS	-0.54	-0.70	-0.83	-0.66	-0.70	-0.51	-0.73	-0.56	-0.27	-0.17	-0.11	0.05	-0.06	0.04	0.12	0.30	-0.29	0.03	0.29	-0.13	MET
MET	-0.03	-0.19	-0.88	-0.73	-0.80	-0.67	-0.68	-0.58	-0.36	-0.19	-0.15	-0.12	-0.11	-0.01	0.14	0.18	-0.34	-0.05	0.19	-0.19	PHE
PHE	-0.02	-0.24	-0.22	-0.74	-0.81	-0.67	-0.60	-0.49	-0.37	-0.13	-0.15	0.03	-0.01	0.14	0.17	0.22	-0.13	0.00	0.24	-0.05	ILE
ILE	-0.03	-0.13	-0.11	-0.18	-0.84	-0.74	-0.62	-0.55	-0.38	-0.16	-0.15	-0.02	-0.04	0.04	0.17	0.27	-0.18	-0.04	0.22	-0.12	LEU
LEU	-0.00	-0.12	-0.14	-0.20	-0.19	-0.65	-0.51	-0.38	-0.32	-0.15	-0.07	0.04	0.08	0.12	0.26	0.36	-0.06	0.08	0.20	-0.05	VAL
VAL	-0.03	-0.03	-0.11	-0.16	-0.19	-0.20	-0.64	-0.49	-0.27	-0.25	-0.02	-0.01	-0.02	-0.10	-0.00	0.07	-0.37	-0.21	0.09	-0.37	TRP
TRP	-0.06	-0.21	-0.08	-0.05	-0.03	-0.02	-0.11	-0.45	-0.20	-0.22	-0.09	-0.08	-0.14	-0.11	-0.08	-0.07	-0.30	-0.25	-0.05	-0.25	TYR
TYR	0.16	-0.08	-0.02	0.02	0.00	0.08	0.01	0.00	-0.12	-0.08	0.04	0.10	0.22	0.15	0.38	0.27	0.07	0.24	0.41	0.15	ALA
ALA	0.01	0.00	-0.02	-0.07	-0.04	-0.07	0.01	0.05	-0.09	-0.29	-0.04	-0.01	0.13	-0.01	0.32	0.11	0.00	0.09	0.29	0.02	GLY
GLY	0.03	0.10	0.16	0.17	0.18	0.09	0.03	0.03	-0.05	-0.26	0.03	0.04	0.12	0.04	0.16	0.11	-0.03	0.11	0.33	0.13	THR
THR	0.12	0.09	0.13	0.08	0.13	0.10	0.19	0.08	0.00	-0.08	-0.08	0.05	0.22	0.09	0.18	0.10	0.04	0.16	0.36	0.20	SER
SER	0.08	0.19	0.10	0.20	0.20	0.16	0.15	0.04	0.00	-0.10	-0.13	-0.17	0.20	0.06	0.27	0.24	0.15	0.09	0.28	0.17	GLN
GLN	0.10	0.04	0.07	0.12	0.14	0.15	0.09	-0.06	0.08	-0.01	-0.09	-0.04	-0.10	-0.06	0.12	0.02	0.09	0.10	0.22	0.18	ASN
ASN	0.21	0.19	0.22	0.32	0.26	0.25	0.06	0.02	0.07	-0.10	-0.12	-0.12	-0.20	-0.27	0.46	0.44	0.00	-0.22	-0.06	0.37	GLU
GLU	0.31	0.16	0.26	0.24	0.28	0.27	0.05	-0.07	0.18	0.13	-0.11	-0.14	-0.10	-0.19	0.03	0.29	-0.10	-0.24	-0.01	0.33	ASP
ASP	0.26	0.37	0.33	0.32	0.41	0.41	0.15	-0.03	0.10	-0.06	-0.13	-0.20	-0.10	-0.27	0.04	-0.08	-0.40	0.05	0.38	0.01	HIS
HIS	-0.01	-0.02	0.00	0.18	0.17	0.19	-0.08	-0.05	0.10	0.04	-0.06	-0.06	0.02	-0.08	-0.19	-0.26	-0.36	0.19	0.66	0.17	ARG
ARG	0.33	0.21	0.20	0.21	0.21	0.22	-0.03	-0.10	0.18	0.02	-0.03	-0.04	-0.14	-0.08	-0.52	-0.51	-0.02	0.03	0.76	0.47	LYS
LYS	0.35	0.24	0.22	0.23	0.24	0.22	0.05	-0.12	0.13	0.01	-0.02	-0.05	-0.17	-0.18	-0.58	-0.49	0.10	0.28	0.16	0.11	PRO
PRO	0.03	0.01	0.03	0.12	0.09	0.06	-0.22	-0.13	0.05	-0.08	-0.03	-0.02	-0.10	-0.04	0.04	0.03	-0.08	-0.03	0.05	-0.11	
$e_{ij} - e_{rr}$	-0.32	-0.25	-0.33	-0.28	-0.32	-0.23	-0.27	-0.23	-0.02	-0.02	0.05	0.11	0.15	0.10	0.21	0.19	-0.02	0.08	0.30	0.11	

Figure 2.1 MJ matrix[19]

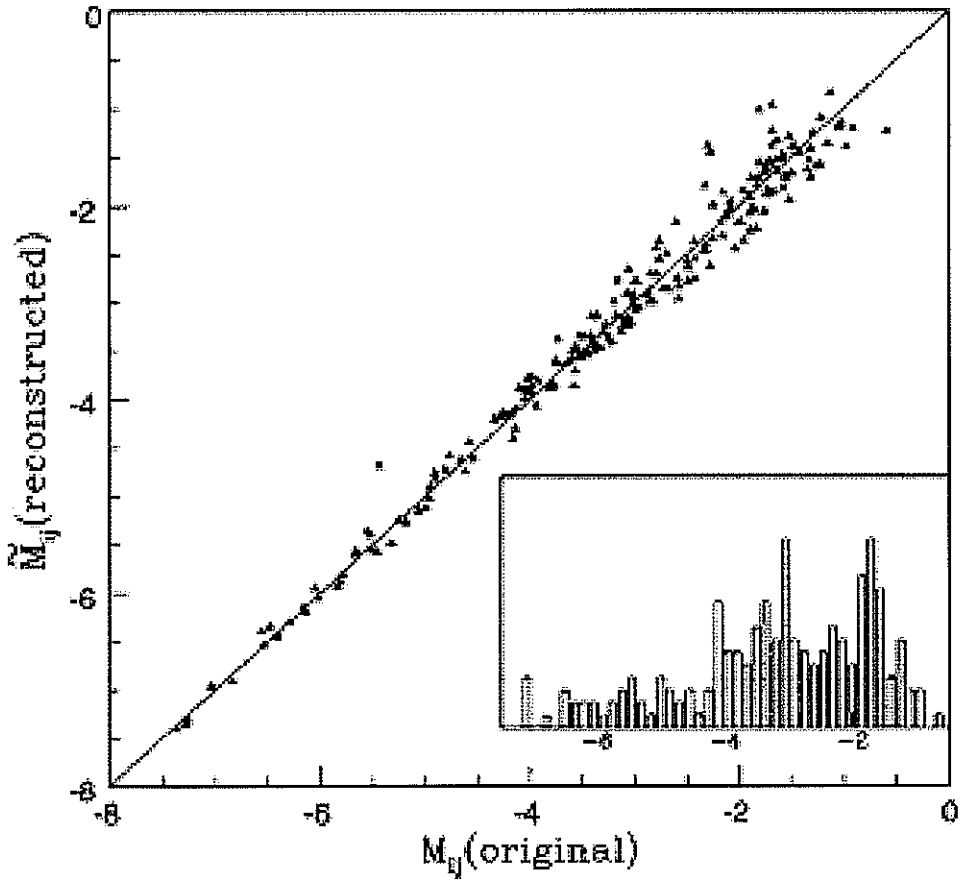

Figure 2.2 Correlation between the LTW parameterized MJ matrix and the original MJ matrix[24].

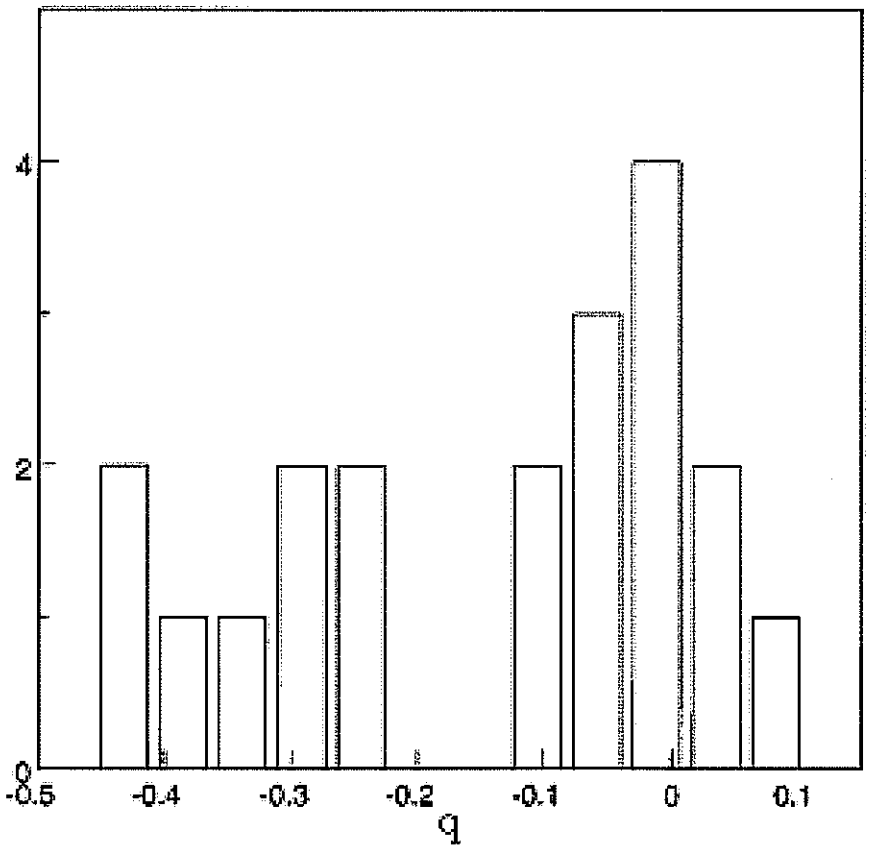

Figure 2.3 Distribution of LTW q_i values[24].

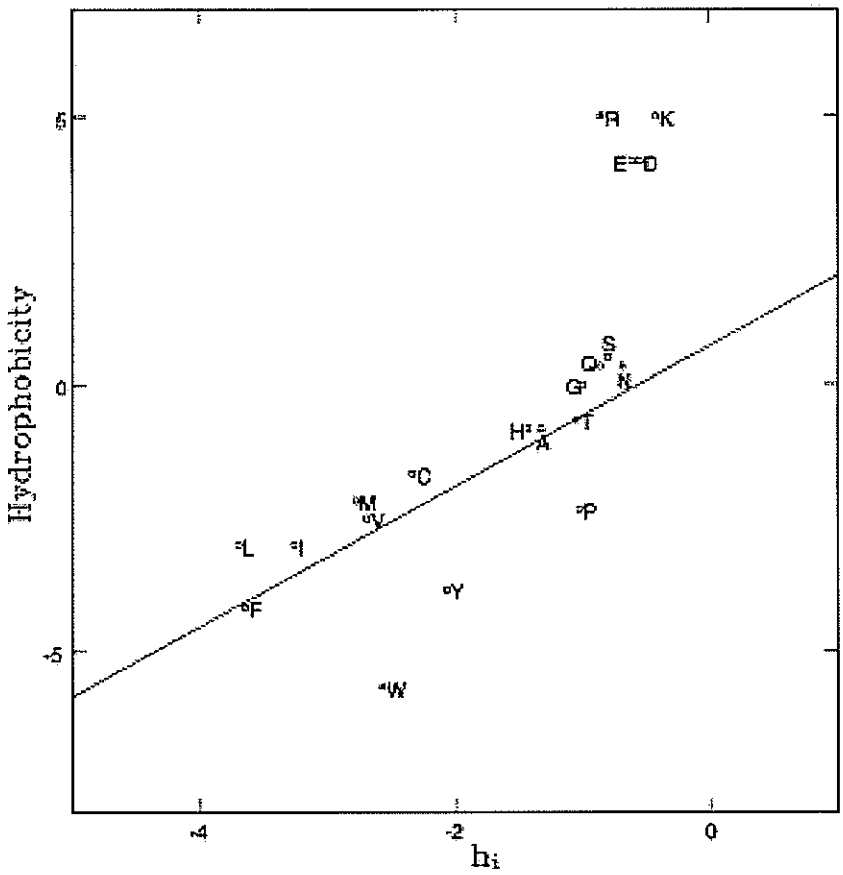

Figure 2.4 Coorelation between the value h_i calculated and the hydrophobicity obtained from experiments. Figure taken from Li, H. Tang, C. and Wingreen, NS. paper[24]

Bibliography

[1] Anfinsen, CB. Haber, E. Sela, M. White, FH. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA **181** 223-230 1961

[2] Weaver, R. Molecular biology The McGraw-Hill Companies, Inc. 1999

[3] Levinthal, C. Are there pathway for protein folding? J. Chem. Phys. **65** 44-45 1968

[4] Levinthal, C. Mossbauer Spectroscopy in biological system. Preceedings of a meeting held at Allerton house, Monticello, Illinois. 1969

[5] Ikai, A. Tanford, C. Kinetic eveidence for incorrectly folded intermediate states in the refolding of denatured proteins. Nature **230** 100-102 1971

[6] Tsong, TY. Baldwin, RL. Elson, EL. The sequential unfolding of ribonucleases A: Detection of a fast initial phase in the kinetic of unfolding. Proc. Natl. Acad. Sci. USA **68** 2712-2715 1971

[7] Creighton, TE. Conformational restrictions on the pathway of folding and unfolding of the pancreatic trypsin inhibitor. J. Mol. Biol. **113** 275-293 1977

[8] Creighton, TE. Experimental studies of protein folding and unfolding. Prog. Biophys. Mol. Biol. **33** 231-297 1978

[9] Weissman, JS. Kim, PS. Reexamination of the folding of BPTI: Predominance of native intermediates. Science **253** 1386-1393 1991

[10] Brandts, JF. Halvorson, HR. Brennan, M. Consideration of the possibility that the slow step in protein denaturation reaction is due to cis-tans isomerism of proline residue. Biochemistry **14** 4953-4963 1975

[11] Kim, PS. Baldwin, RL. Intermediates in the folding reactions of small proteins and mechanism of protein folding. Annu. Rev. Biochem. **59** 631-660 1990

[12] Mann, CJ. Shao, X. Matthews, CR. Characterization of the slow folding reactions of trp aporepresssor from Escherichia coli by mutational analysis of prolines and catalysis by a peptidyl-prolyl isomerase. Biochemistry **34** 14573-14580 1995

[13] Wolynes, PG. Onuchic, JN. Thirumalai, D. Navigating the folding route. Science **267** 1619-1620 1995

[14] Onuchic, JN. Wolynes, PG. Luthey-Schulten, Z. Socci, ND. Towards an outline of the topography of a realistic folding funnel. Proc. Natl. Acad. Sci. USA **92** 3626-3630 1995

[15] Chan, HS. Dill, KA. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. **100** 9238-9257 1994

[16] Dill, KA. Principle of protein folding: A perspective from simple exact models. Protein Sci. **4** 561-602 1995

[17] Mirsky, AE. Pauling, L Proc. Natl. Acad. Sci. USA **22** 439 1936

[18] Kauzmann, W. Adv. Protein Chem. **14** 1 1959

[19] Miyazawa, S. Jernigan, RL. Identifying sequence-structure pairs undetected by sequence alignments. Protein Engineering **13** 459-475 2000

[20] Miyazawa, S. Jernigan, RL. Estimation of effective interresidue contact energy from protein crystal structures: quasi-chemical approximation. Macromolecules **18** 534-552 1985

[21] Miyazawa, S. Jernigan, RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins **34** 49-68 1999

[22] Sippl, MJ. Calculation of conformational ensembles from potentials of mean force. J. Mol. Biol. **213** 859-883 1990

[23] Sippl, MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the determineation of protein structures. J. Comput. Aids Mol. Des. **7** 473-501 1993

[24] Li, H. Tang, C. Wingreen, NS. Nature of driving force for protein folding: A result from analyzing the statistical potential. Phys. Rev. Lett. **79** 765-768 1997

[25] Hildebrand, JH. Scott, RL. The Solubility of Non-electrolytes. Reinhold Publishing Corporation, New York 1950.

CHAPTER 3. Simple Exact Model

HP model

As described earlier, the three dimensional structures of proteins are extremely variable. To simplify the geometry in modeling, a simple lattice HP model was first proposed by HueSun Chan and Ken A. Dill [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13] in 1989. In this model, each amino acid is represented by a single "bead". A protein is modeled as a chain of "beads" on a lattice. A lattice site can be either empty or filled by one bead. In this way, the bond angles have only a few discrete values, depending on the lattice being used. Most studies used a two dimensional rectangular lattice or three dimensional cubic lattice.

In the HP model, there are two types of amino acids: H for hydrophobic residues, and P for polar residues. Interactions are pair-wise contacts only. There are only three parameters in this interaction scheme: HH interaction energy, PP interaction energy and HP interaction energy. Generally, the contact energy of P-P contact is set to be 0. The most popular scheme is the (1,0,0) scheme in which HH contact energy is set to be 1 and HP contact energy is 0. According to Chan and Dill, the HP model is supported by the following experimental evidence:

1. The water-to-oil transfer free energy is large and negative for nonpolar amino acids [14], which is in consistent with the fact that they are buried in the protein core to avoid contact with solvent molecules. The average transfer free energy of a nonpolar amino acid is around -2 kcal/mol.

2. Large positive changes of heat capacity are observed for most proteins in the unfolding process [15, 16, 17]. Though unfolding the backbone also involves heat capacity changes, it cannot be the basis for the folding code, because it is insensitive to the amino acid sequence.

3. The free energy for helix formation is small [18, 19, 20]. Most of the amino acids are unfavorable for forming helical conformation, except for alanine(A) which is a strong helix-former. Helix stability increases with length, but most helices in globular proteins are not long enough to be stabilized by themselves [21]. The observation that most helices in globular proteins are amphipathic suggests that hydrophobic interaction may also be important in stabilizing helices in globular proteins.

4. β sheets in general have few local interactions. Hydrogen bond interactions cannot fully explain the folding of sheet proteins. Because β sheets generally involve strands that are far apart in primary sequence, nonlocal interactions should dominate the folding of sheet proteins.

5. Electrostatic interactions in proteins generally contribute little to stability, as determined by the general insensitivity of native structures to pH and salt[22, 23].

6. Polypeptides can be designed to fold to helical bundles by designing only the hydrophobic and polar residues without considering their helical propensities, side chain packing and charge[24, 25]. There is evidence that the tendencies to form helical configurations depend more on solvent than on residue sequence[26].

In the lattice HP model, when two amino acids are geometrical neighbors on the lattice they are in contact. Because residues adjacent in sequence are always connected on lattice model, they are not counted as contacts. The total "free energy" of a configuration for a lattice protein is the summation of all contacts. For example, in (1,0,0) scheme, it is the total number of HH contacts in the conformation. A globular protein in nature has a unique three dimensional structure. The total number of gene sequences discovered is of the order of 10^7, while for a polypeptide with around 200 residues, the total number of possible polypeptide sequences is of the order of 20^{200}. Yet few human ab inito designed polypetide seqeuneces has folded into unique structures. This is because a random polypetide chain does not usually have a unique ground state structure. A sequence is viewed as "protein-like" when its lowest energy structure is unique.

The lattice model has the advantage of simplicity. For short chains (lengths $\leq$ 34 for

three dimension cubic lattice[27]) the conformations of a given lattice protein can be exactly enumerated.

Even for such a simple model, the number of possible sequences and conformations increases exponentially with the number of residues. Often researchers restrict the conformational space of proteins by restricting the protein chain to a $n \times m$ two dimensional lattice or $n \times n \times n$ cubic lattice(most of the time it is $3 \times 3 \times 3$). In this way, people can study lattice proteins with relatively long chain lengths. A Kloczkowski and RL. Jernigan [28, 29] enumerated the possible protein-like configurations on some $n \times m$ two dimensional lattices (see Figure 3.1).

Figure 3.2 gives an example of protein-like sequences for lattice models in 2D(A), 3D cubic(B) and perturbed homopolymer on 3D(C) lattices. Black beads represent H residues and white beads represent P residues[30] (Figure 3.2(C) is obtained by using (3 3 1) scheme).

Despite the simplification in the lattice HP model, many general qualitative results about the protein folding problem have been obtained.

In the HP model, it can be shown that the requirement of "unique ground state" is a strong constraint for choosing seqeunce. The fraction of protein-like HP sequences in the 2D lattice is about $2.1 - 2.4\%$, depending slightly on the chain length. Figure 3.3 is the relation between this percentage and the chain length obtained by Chan and Dill[30].

Protein structures are highly compact, but not perfect spheres. This feature has been reproduced in lattice model, which is an important result for the protein "design" problem. Designing an amino acid sequence to fold to a desired target conformation has two aspects: (1) positive design which is to ensure that the sequence will fold to the target structure(energetically favorable). (2) negative design which is to ensure that the sequence does not fold to an alternative conformation(which means that in the whole conformation space of the sequence, there does not exist another structure that is energetically more favorable that the desired structure). The results from lattice HP model showed the importance of negative design. For an example, in a Harvard/UCSF collaboration, the Harvard group chose 10 different three dimensional lattice target conformations of 48 residues. They designed the sequence to fold on these structures by a Monte Carlo method[31, 32] without explicit negative design. The UCSF group used different

search strategies(CHCC algorithm and hydrophobic zippers) to search the conformation space of the designed sequence provided by the Harvard group. For 9 out of 10 sequences provided by the Harvard group, the UCSF group found a conformation with lower contact energy. Though the desired conformations are maximally compact, the designed sequences invariably folded to more stable conformations that were not maximally compact. Figure 3.4 show one of the cases.

One surprising result from lattice model relates to secondary structure formation. Traditionally, secondary structures are believed to be stabilized by hydrogen bonds which are local interactions. Because the lattice HP model does not include the backbone hydrogen bond interaction in its energy scheme, it is expected that lattice proteins will not be able to produce local structures like α helices and β sheets which are special features for proteins. But it has been observed that there is an unexpected abundance of "protein-like secondary structures" in the conformations of lattice proteins. They generally have more ordered structures compared with random configurations on lattice, although because of the simple geometry of lattice, these "α helices" and "β sheets" are distorted. It is now believed that the "hydrophobic collapse" resulting in compact conformations, also helps in building local structures like those in α helices. Different interactions might help the proteins form the native structure cooperatively, even though how this is achieved by nature is still unclear. Figure 3.5 show the distributions of secondary structures in real proteins and proteins on lattice.

Lattice models are also useful in studying folding kinetics. For short lattice proteins, all intermediate states can be enumerated, thus qualitative results about protein folding process can be obtained and compared with experimental results for real proteins. Figure 3.6 copied from Chan and Dill's paper[30] shows the multipathway feature of funnel like energy landscape for a 13 residues protein folding to its native state.

Designability principle

Another important result from studies of lattice HP model is the "designability principle" proposed by Li, Tang and Wingreen[8, 9]. For a given protein structure, the designability (N_s)

is defined as the total number of different sequences that have this structure as their lowest energy structure. Li, Tang and Wingreen found that there is a small class of lattice structures which are extrordinary because they have very high designabilities.

Figure 3.7 obtained by Li, Tang and Wingreen shows histograms of N_s for a $3 \times 3 \times 3$ cubic lattice(a) and a 6×5 two dimension rectangular lattice. The energy scheme used is (2.3, 1, 0). In both of these two figures, there appeared a long tail at the high designability end. For example, on the $3 \times 3 \times 3$ cubic lattice, there are 3794 different sequences that fold on to the same native configuration($N_s = 3794$). If we assume that the native structures on this cubic lattice are randomly distributed for all protein sequences, the expectation value for a structure with designability $N_s > 120$ is 1.76×10^{-6}. There are 51704 different structures unrelated by rotation, reflection or reverse labeling on this cubic lattice. Considering the exponential decay tendency of the expectation value with respect to designability, the high N_s tail is very interesting.

Li, Tang and Wingreen also observed that highly designable structures have secondary structures that are absent in random compact structures. For highly designable structures, there are many more parallel running lines folded in a regular way than in an average random structure. Figure 3.8 [9] gives an example of a highly designable structure in 2D and in 3D. The number of straight lines (three amino acids in a row) found in the high N_s structures is 8 or 9, while the average structure has only 5.4 straight lines.

The average energy gap δ_S for a structure S is defined as the minimum energy required to change the native state structure to a different compact structure averaged over all N_s sequences that design on it, At the high N_s region, there is a higher energy gap between the native state and the lowest misfolded state (see Figure 3.9). Thus highly designable structures are thermodynamically more stable than random compact structures. In order for a natural protein to be able to survive mutations and evolution, it has to be relatively robust against sequence variation. Li, Tang and Wingreen proposed that only highly designable structures on lattice can be considered as "proteins".

n	$m=2$	$m=3$	$m=4$	$m=5$
2	4	8	14	22
3	8	20	62	132
4	14	62	276	1 006
5	22	132	1 006	4 324
6	32	336	3 610	26 996
7	44	688	12 010	109 722
8	58	1 578	38 984	602 804
9	74	3 190	122 188	2 434 670
10	92	6 902	375 122	12 287 118
11	112	13 878	1 128 446	49 852 352
12	134	29 038	3 342 794	237 425 498
13	158	58 238	9 767 588	969 300 694
14	184	119 518	28 217 820	4 434 629 912
15	212	239 390	80 709 424	18 203 944 458
16	242	485 822	228 864 620	80 978 858 522
17	274	972 414	664 060 262	333 840 165 288
18	308	1 960 830	1 800 346 140	1 456 084 764 388
19	344	3 923 326	5 002 457 832	6 021 921 661 718
20	382	7 882 494	13 825 549 136	25 904 211 802 080

n	$m=6$	$m=7$	$m=8$
2	32	44	58
3	336	688	1 578
4	3 610	12 010	38 984
5	26 996	109 722	602 804
6	229 348	1 620 034	12 071 462
7	1 620 034	13 535 280	175 905 310
8	12 071 462	175 905 310	3 023 313 284
9	82 550 864	1 449 655 468	43 551 685 370
10	572 479 244	17 198 428 572	682 958 971 778
11	3 808 019 582	142 545 533 336	9 735 477 214 522
12	25 304 433 030	1 580 868 297 042	144 397 808 917 246
13	164 452 629 818	13 246 916 541 978	2 033 155 413 979 838
14	1 062 773 834 046	139 620 415 865 920	29 105 375 742 858 518
15	6 777 328 517 896	1 183 338 916 049 852	404 654 754 079 984 324
16	42 944 798 886 570	11 997 107 474 280 224	5 656 098 437 704 094 140
17	269 706 791 277 978	102 719 325 162 193 010	77 710 312 229 803 403 554
18	1 683 956 271 732 804	1 010 824 101 911 587 178	1 067 886 114 091 399 967 842
19	10 445 800 698 724 066	8 728 784 450 632 453 306	14 517 649 840 508 475 301 004
20	64 470 330 298 173 718	83 947 749 266 911 632 982	196 974 144 293 101 997 656 968

Figure 3.1 Enmumeration of all possible configurations of chains in $m \times n$ rectangular lattice. (A Klockowski and RL. Jernigan[29]

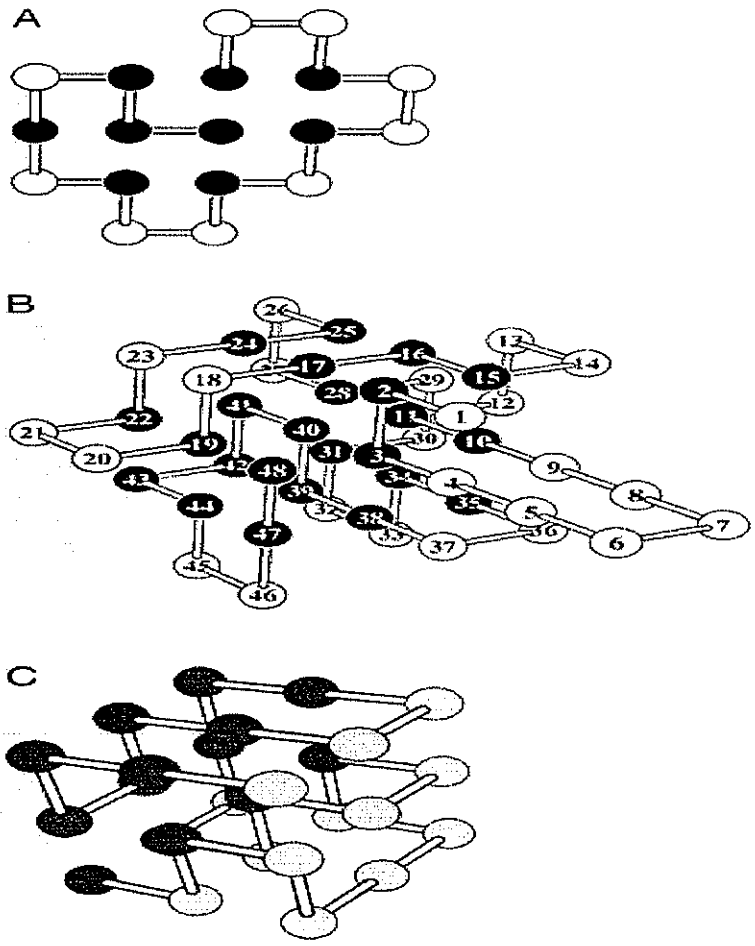

Figure 3.2 Examples of native structures on lattice.
A: a HP model in 2D rectangular lattice(Chan and Dill 1994).
B: a HP model in 3D cubic lattice(Yue et al. 1995).
C: a perturbed homopolymer model(Shakhnovich and Gutin 1993).
Energy scheme for A and B is (1,0,0), C is (3,3,1)

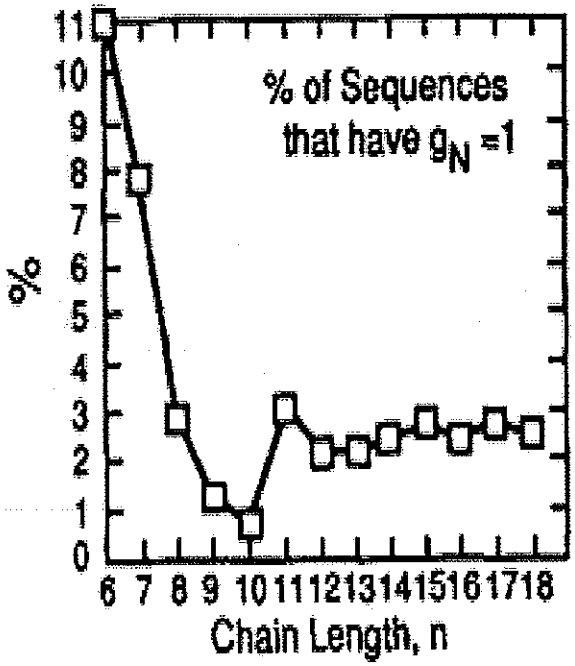

Figure 3.3 Percentage of HP sequences that have unique native structures on 2D lattices as a function of chain length (Chan and Dill 1991)

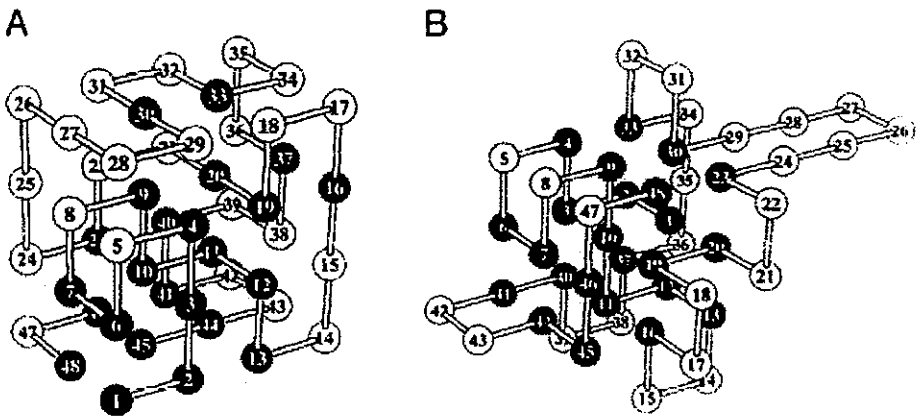

Figure 3.4 The importance of "negative design":
A: lattice model-designed protein and its HP sequences. Protein is designed by the Monte Carlo method of Harvard group(Shakhnovich and Gutin 1993).
B: a lower energy configuration found by UCSF group use CHCC conformational search method (Yue et al. 199)[27].
H residues are represented black beads, P residues white beads.

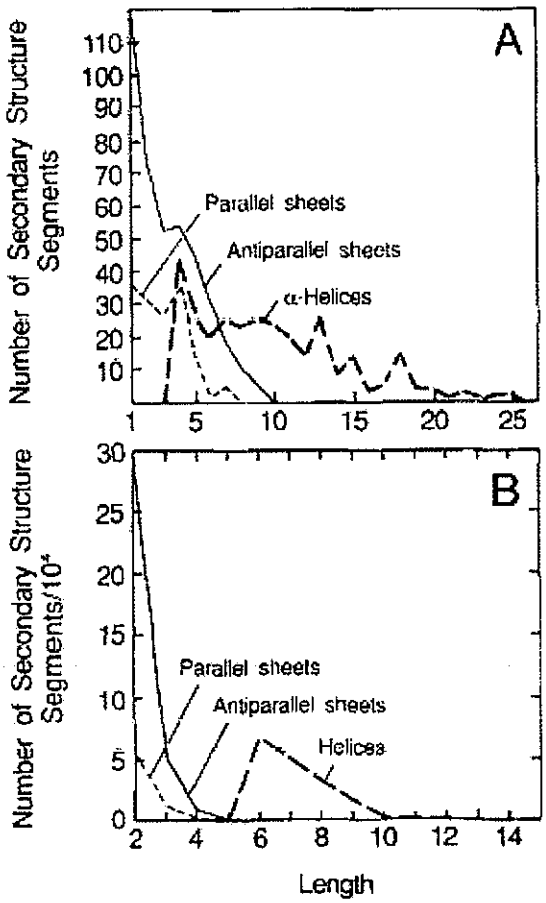

Figure 3.5 Length distribution of secondary structures.
A: Database observations of Kabsch and Sander(1983)
B: exhaustive simulations of maximally compact chains of 26 residues on 2D rectangular lattice(Chan and Dill 1990)[30].

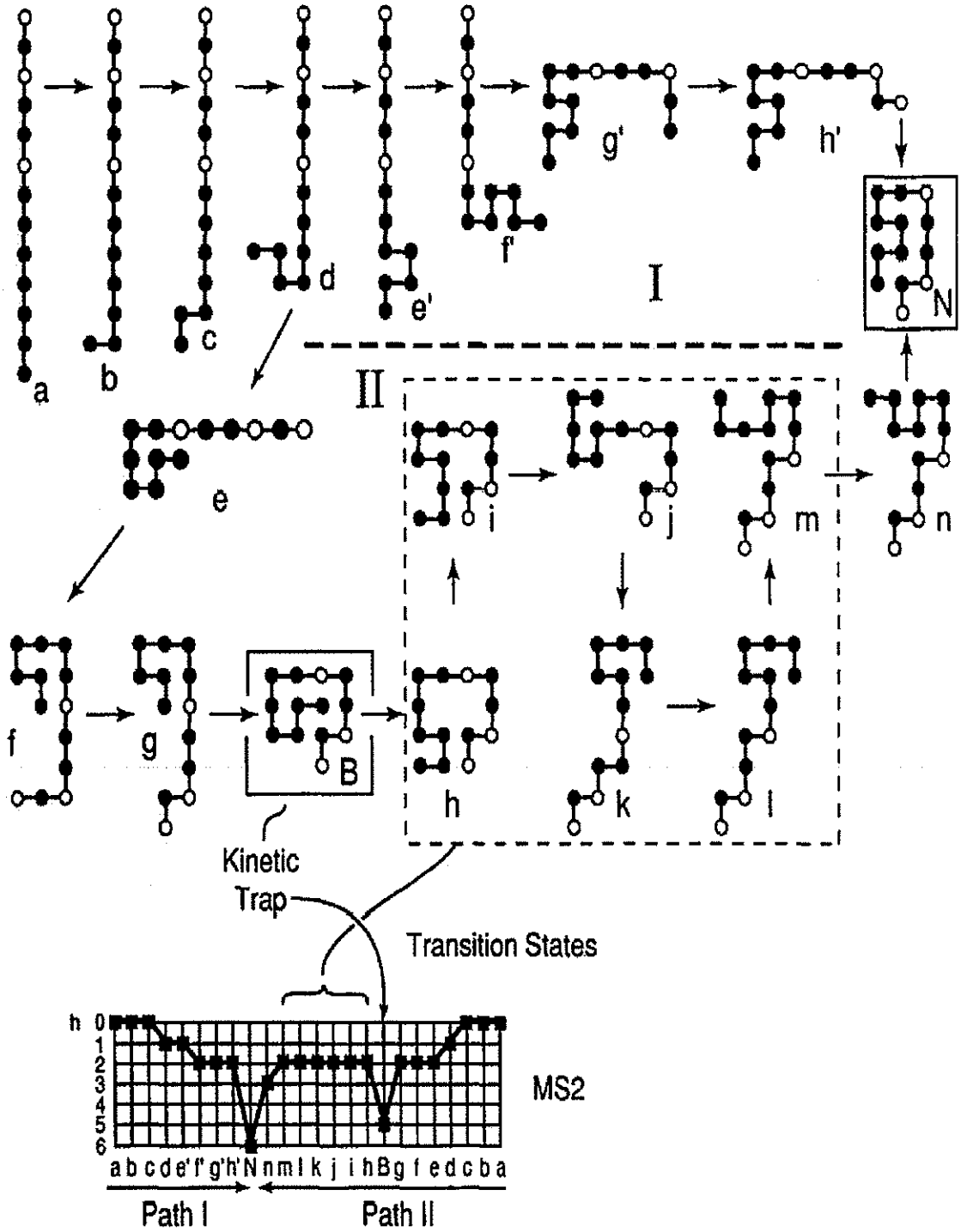

Figure 3.6 Folding pathways and energy landscape of a 13mers lattice protein. (Chan and Dill 1994)

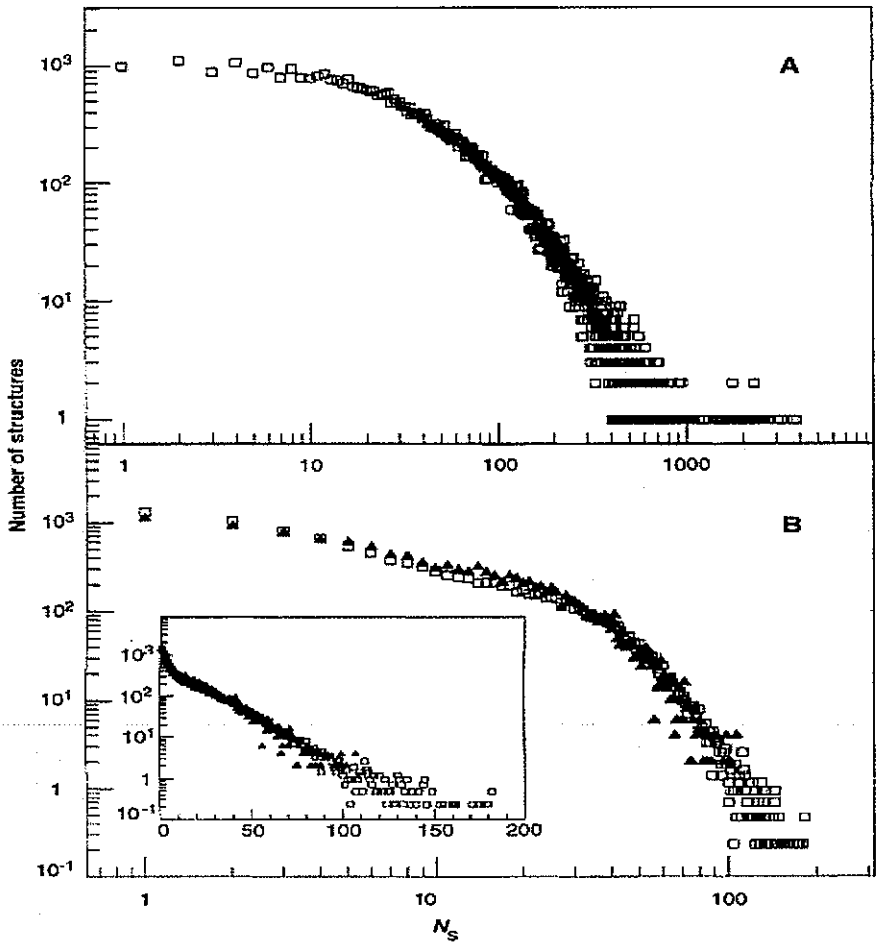

Figure 3.7 A: Histogram of numbers of structures with a given number of associated sequences N_s for $3 \times 3 \times 3$ cubic lattice.
B: Histogram of numbers of structures with a given number of associated sequences N_s for 2D 6×5 maximally compact ractangular lattice.
Inset: similar result for 2D 6×5 ractangular lattice.

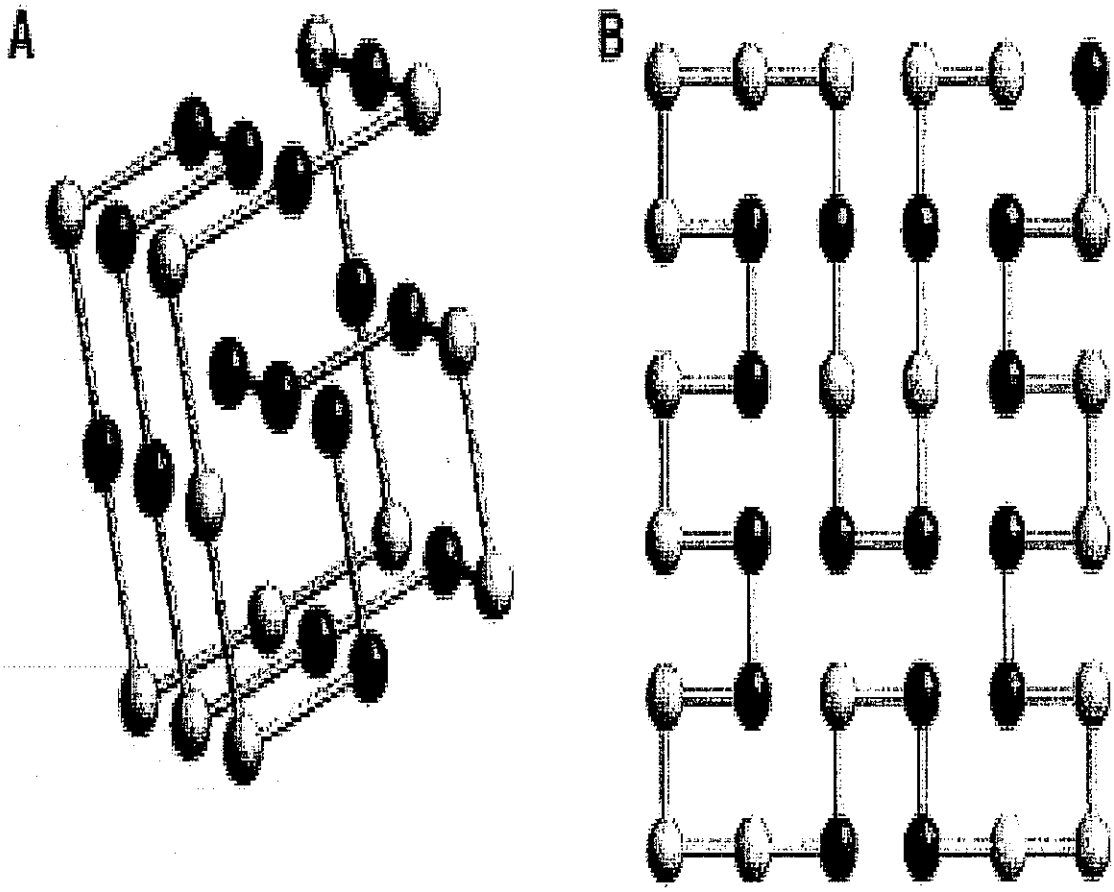

Figure 3.8 Structures with the largest number of N_s for 3D $3 \times 3 \times 3$ (top) cubic and 2d 6×6 rectangular lattice. Black beads represent H residues and white beads represent P residues. energy scheme is (2.3, 1, 0)[9]

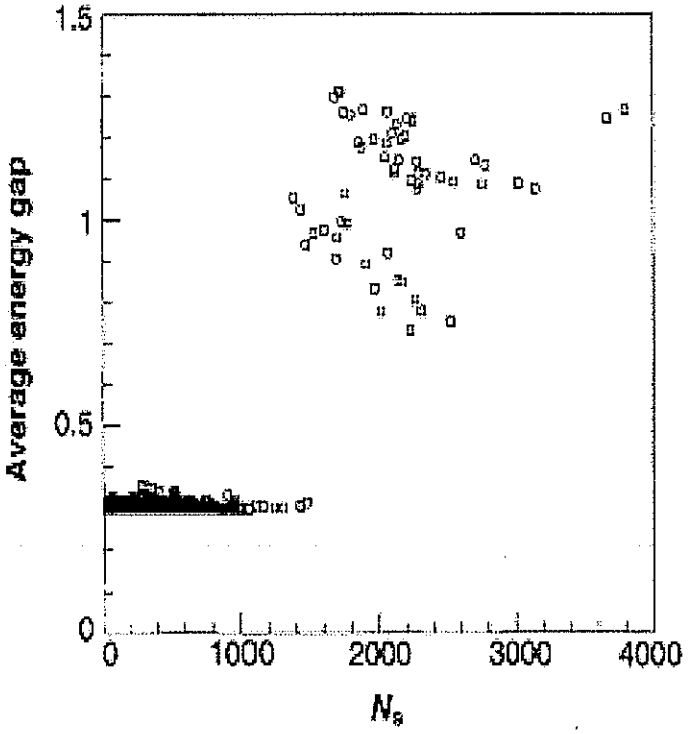

Figure 3.9 Averge energy gap of $3 \times 3 \times 3$ cubic lattice structures vs. N_s of the structure.

Bibliography

[1] Chan, HS. Dill, KA. Inrachain loops in polymers: Effects of excluded volume. J Chem Phys **90** 492-509 1989

[2] Chan, HS. Dill, KA. Compact Polymers. Macromolecules **22** 4559-4573 1989

[3] Chan, HS. Dill, KA. The effects of internal contstrains on the configuration of chain molecules. J Chem Phys **92** 3118-3135 1990

[4] Chan, HS. Dill, KA. Origins of structure in globular proteins. Proc Natl Acad Sci USA **87** 6388-6392 1990

[5] Chan, HS. Dill, KA. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem **20** 447-490 1991

[6] Chan, HS. Dill, KA. "Sequence space soup" of proteins and copolymers. J Chem Phys **95** 3775-3787 1991

[7] Chan, HS. Dill, KA. The protein folding problem. Phys Today**46(2)** 24-32 1993

[8] Chan, HS. Dill, KA. Energy landscape and the collapse dynamics of homopolymers. J Chem Phys **99** 2116-2127 1993

[9] Chan, HS. Dill, KA. Transitions states and folding dynamics of proteins and heteropolymers J Chem Phys **100** 9238-9257 1994

[10] Dill, KA. Dominant forces in protein folding Biochemistry **29** 7133-7155 1990

[11] Dill, KA. Folding proteins: Finding a needle in a haystack Curr Opin Struct Biol **3** 99-103 1993

[12] Dill, KA. Alonso, DOV. Hutchinson, K. Thermal stabilities of globular proteins Biochemistry **28** 5439-5449 1989

[13] Dill, KA. Fiebig, KM. Chan, HS. Cooperativity in protein folding kinetics Proc Natl Acad Sci USA **90** 1942-1946 1993

[14] Nozaki, Y. Tanford, C. The solubility of amino acids and two glyscine polypeptides in aqucous ethanol and dioxane solutions J Biol Chem **246** 2211-2217 1971

[15] Privalov, PL. Stablilty of proteins: small globular proteins. Adv Protein Chem **33** 167-241 1979

[16] Privalov, PL. Gill, SJ. Stability of protein structure and hydrophobic interaction. Adv Protein Chem **39** 191-234 1988

[17] Privalov, PL. Makhatadze, GI. Contribution of hydration to protein folding thermodynamics. J Mol Biol **232** 660-679 1993

[18] Suki, M. Lee, S. Power, SP. Denton, JB. Konishi, Y. Scheraga, HA. Helic-coil stability constants for naturally occuring amino acids in water. Macromolecules **17** 148-155 1984

[19] Chakrabartty, A. Schellman, JA. Baldwin, RL. Large differences in the helix propensities of alanine and glycine. Nature **351** 586-588 1991

[20] Chakrabartty, A. Kortemme, T. Baldwin, RL. Helix propensities of amino acids measured in alanine based peptides without helix-stabilizing side-chain interactions. Protein Sci. **3** 843-852 1994

[21] Kabsch, W. Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22** 2577-2637 1983

[22] Goto, Y. Nishikiori, S. Role of electrostatic repulsion in the acidic molten globule of cytochrome c. J. Mol. Biol. **222** 679-686 1991

[23] Hagihara, Y. Tan, Y. Goto, Y. Comparison of the conformational stability of the molten globule and native states of horse cytochrome c. Effects of acetylation, heat,urea and guanidine-hydrochloride. J. Mol. Biol. **237** 336-348 1994

[24] Kamtekar, S. Schiffer, JM. Xiong, H. Babik, JM. Hecht, MH. Protein design by binary patterning of polar and nonpolar amino acids. Science **262** 1680-1685 1993

[25] Munson, M. OBrien, R. Sturtevant, JM. Regan, L. Redesigning the hydrophobic core of a four-helix-boundle protein. Protein Sci. **3** 2015-2022 1994

[26] Waterhous, DV. Johnson, WC. Importance of environment in determining secondary structure in proteins. Biochemistry **33** 2121-2128 1994

[27] Yue, K. Fiebig, KM. Thomas, PD. Chan, HS. Shakhnovich, EL. Dill, KA. A test of lattice protein folding algorithms. Proc. Natl. Acad. Sci. USA **92** 325-329 1995

[28] Kloczkowski, A. Jernigan, RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. I. suqare lattice. J. Chem. Phys. **109** 5134-5146 1998

[29] Kloczkowski, A. Jernigan, RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. J. Chem. Phys. **109** 5151-5159 1998

[30] Dill, KA. Bromberg, S. Yue, K. Fiebig, K. Yee, DP. Thomas, PD. Chan, HS. Principles of protein folding-A perspective from simple exact models. Protein Sci. **4** 561-602 1995

[31] Shakhnovich, EI. Gutin, AM. A new approach to design of stable protein. Protein Engineering **6** 793-800 1993

[32] Shakhnovich, EI. Gutin, AM. Engineering of stable and fast-folding sequence of model proteins. Proc. Natl. Acad. Sci. USA **90** 7195-7199 1993

[33] Li, H. Tang, C. Wingreen, N. Emergence of prefered structures in a simple model of protein folding. Science **273** 666-669 1996

[34] Li, H. Tang, C. Wingreen, N. Designability of protein structures: A lattice-model study using Miyazawa-Jernigan matrix. Proteins **49** 403-412 2002

CHAPTER 4. Protein Structure Prodiction

At present, methods in protein structure prediction can be classified into 3 categories: sequence-based approaches, structure-based approaches and approaches starting from first principles(ab initio).

Sequence Based Approach

Proteins with similar sequences generally have similar native structures. These proteins often come from the same ancestor. Such evolutionarily-related proteins are said to be homologous to each other. In order to find a homlogous protein from the data base for a query protein, it is necessary to align the query protein sequence to template protein sequences from the database. There are various kinds of sequence alignment methods that can search the whole sequence database rapidly. These methods, in general, use dynamic programming algorithms to generate optimum alignments. Dynamic programming is mathematically proven to provide the global optimum sequence alignment for a given similarity matrix(or substitution matrix, which is a matrix for substituting an amino acid for a different related amino acid). The most popular software for sequence-based identification of homologous structures are BLAST and its variations (e.g. PSI-BLAST).

Dynamic Programing

Suppose that we want to align two sequences which are represented by two sequence vectors $\mathbf{A} = \mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_n}$ and $\mathbf{B} = \mathbf{b_1}, \mathbf{b_2}, ...\mathbf{b_m}$. We define a function $S(\mathbf{a}, \mathbf{b})$ as the best(largest) score from aligning $\mathbf{A}$ and $\mathbf{B}$. We define:

$$H_{i,j} = max\{0, S(a_x, a_{x+1}, ...a_i \rightarrow b_y, b_{y+1}, ...b_j)\} \tag{4.1}$$

where $1 \leq x \leq i, 1 \leq y \leq j$.

$H_{i,j}$ is the best score of any alignment ending at a_i and b_j or 0, whichever is larger. By doing this, we mean that the alignment score for two similar sequences should be positive. An alignment score of zero means there is no sequence similarity, and the alignment equal to random(which is assigned as 0). By starting with $H_{i,0} = H_{0,j} = 0$ for $1 \leq i \leq n, 1 \leq j \leq m$ we have

$$H_{i,j} = max\{0, H_{i-1,j-1} + s(a_i, b_j), E_{i,j}, F_{i,j}\} \tag{4.2}$$

where $s(a_i, b_j)$ is the contribution of aligning residue a_i on residue b_j which is an element of the substitution matrix, and $E_{i,j} = max\{H_{i,j-k} - w(k)\}$ and $F_{i,j} = max\{H_{i-k,j} - w(k)\}$ are the maximum scores for having an insertion/deletion at place (i, j). $w(k)$ is the gap penalty, usually defined as linear with the gap length $w(k) = u + vk$, where $k + 1$ is the total length of continuous gaps. u is the gap initiation penalty, v is the the gap elongation penalty. For the upper corner $H_{0,0}$, the best alignment and score $H_{i,j}$ can be generated systematically.

However, sequence-based methods for identifying structural homologs are only reliable when the query protein has a sequence above 30% identical with the template sequence. When sequence identity falls below 20%, the results from this method are unstable. This sequence similarity region is called the "twilight zone"$(20 - 30\%)$. When sequence similarity is lower than 15%, results are highly unreliable.

Ab Initio Approach

Approaches starting from first principles are still not mature. Due to the large conformation space of protein structures and complicated interactions involved in the protein folding process (in atomic detail), it is very hard for an ab initio method to search for a global optimum even using simple interaction schemes (like a simple bulk potential). At present, this approach generally has to be combined with part of the results from either sequence-based or structural-based approaches. One of the most sucessful approaches has been developed by

Baker's group, called fragment assembly, which performed well in the CASP5 competition. In the Baker method, a query protein sequence is chopped into continuous segments of nine residues. Possible configurations for each segment are obtained from aligning the segment to existing protein structures. Monte Carlo simulations are used to assemble the segments to generate candidate global conformations. These candidates are then clustered according to their structural similarity, resulting in a final conformation which has the largest global conformation cluster (the conformation that has the highest number of similar candidate structures). In general, the ab initio approach is not as good as the other two approaches (in terms of both speed and accuracy) when there are existing sequence homologs or structure homologs in the database for a query sequence. The method developed by Baker's group(a server called ROSETTA) starts by using a sequence-based approach to search database for homologous proteins for a query sequence. If no sequence homologs are found, the ab-initio method will be used to generate predictions. Despite the inaccuracy of the ab initio approach, it is the only existing approach with which is possible to generate a "new" fold when both structural and sequencial similarity are absent. Recently many popular "meta servers" have emerged, which use clustering methods to generate a "popular" structure which is compatible with the majority of the results from different methods for a query protein sequence. We believe the success of these method may be due to some inclusion of entropy effects in the protein folding process by structural clustering, even though it may not be explicitly considered when the methods are developed.

Structural Threading

Structural threading is an approach which uses both sequence and structural information of existing proteins to predict the native structure of a query protein. The query protein's amino acid sequence is assumed to take the 3-d structure of a template protein structure. This is the sequence to structure "alignment" step. After an alignment is obtained, the conformational energy is then calculated (the scoring step). By ranking the threading energy of the query sequence threaded on different template structues, the one with the lowest conformational

energy is predicted to be the structure of the query protein. The basic assumption of this method is that the native state of a protein has the lowest free energy compared with all the other physical conformations the query protein polypetide chain can take. It is also assumed that structures similar to the native structure of a query protein will have low conformational energy (i.e. the native state has a wide basin)

This approach is slower for database searches compared with sequence-based approach because the sequence to structure alignment step is a 1-d to 3-d alignment process rather than a simple 1-d to 1-d alignment used in sequence to sequence alignment. However, structural threading does not depend on similarity in sequence, which makes it useful for sequences in or below the "twilight zone". The method we developed belongs to this category. I will discuss our method and some of its applications in the remainder of this thesis.

CHAPTER 5. Eigenvector Analysis of Protein Structures: A Manuscript To Be Submitted To Phys. Rev. Lett.

by Haibo Cao, Yungok Ihm, Cai-Zhuang Wang,Drena Dobbs and Kai-Ming Ho

Abstract

We study the sequence-structure relation of protein using pairwise residue-residue hydrophobic interaction. A strong correlation between amino acid sequence and the corresponding native structure of the protein is observed through the analysis of the eigen-spectrum of the contact matrix of the native structure. We show that, in the first approximation, the dominant eigenvector of the contact matrix provides a better representation of the sequence profile of structurally similar protein than the amino acids sequence. Contributions from higher rank eigenvectors (rank > 4) are found to be sequence blind. A method to determine protein domain boundary based on maximizing this correlation yields results in good agreement with biological results.

Eigenvector analysis of protein structures

Globular proteins fold into unique three dimensional structures under natural conditions. These native structures are primarily determined by the protein's amino acid sequences [1]. However, to predict the native structure of a protein from its amino acid sequence remains one of the most challenging problem in biophysics. Nature is extremely selective in choosing the polypeptide sequence and structure of proteins. Among $\approx 20^{300}$ theoretically possible polypeptide sequences, only about 10^7 occurr in nature. These natural proteins further condensed to

around 2000 distinct structural families [4]. Very few polypeptides designed from first principles have ever successfully folded under natural conditions. It is still not clear what makes protein sequences and structures so special and different from ordinary polypeptides. Understanding the origin of this specificity can provide valuable information towards the ultimate solution of the protein folding problem.

Chan and Dill have tried to understand the specificity of protein sequences using a lattice H-P model [11, 12]. In their model, there are only two type of residues: H(hydrophobic) and P(polar). The structures of these HP-residue sequences are restricted to be on lattice. Their studies show that on a two-dimensional lattice, among all possible combinations of two letter alphabet sequences of length 6-18, only a small fraction have nondegenerate lowest energy structures. Li, Tang, and Wingreen have used a similar HP lattice model to study the specificity of protein structures [8, 9]. They showed that a very small fraction of the lattice structures distinguish themselves from other possible "native structures" in their ability to be native structures of many different HP sequences. This ability is named "designability" of a given structure. Structures with high designability generally are more stable thermodynamically and often process "secondary structures" similar to those of real proteins. They suggest that only highly "designable" structures can be native structures of proteins in nature.

The native state of a protein is believed, thermodynamically, to be the lowest free energy state among all physical conformations the protein can take. Thus, the results from the above HP lattice models are consistent with the funnel-like energy landscape of protein conformation proposed by Wolynes and co-workers [13] who among many others have tried to explain why protein sequence can "search" such a big conformational space and achieve a native fold in a very short time(milliseconds to seconds). Lattice model studies suggest that both naturally selected amino acid sequences and protein native structures are atypical, so that the funnel like energy landscape of protein folding can be achieved [6]. However, the interplay between these two "specificity"(or atypical) is still not well understood.

In this paper we study the correlation between real protein sequences and structures under the assumption that the hydrophobic interactions dominate the protein folding process. By de-

composing the contact matrix corresponding to a protein native structure into its eigenvectors, we observed a strong overlap between protein sequence and the first dominant eigenvector. We believe the above mentioned specificity comes from the fact that nature is restricted by this rule in choosing the sequence and structure of a protein. As a result, the dominant eigenvector provides a better representation of the protein structure than the protein sequence which is generally used in structural threading. Our study also shows that this correlation also applies to individual protein domains, thus domain boundaries can be predicted using this correlation if a protein structure is known.

A protein is a complex system where thousands of atoms interact with each other and with water molecules. For our purpose of studying the sequence-structure correlation, we restrict ourselves to a coarser residue-level representation for describing interactions. In our work, the pairwise hydrophobic interaction takes the form of "contact" interaction. When the centers of two residues are geometrically within a certain cutoff distance, the two residues are in contact and a contact energy is assigned, according to the residue type. The total hydrophobic interaction energy for a given protein structure is the summation of all pairwise contact energies of the conformation. Under this interaction scheme, the three dimensional structure of a protein can be reduced to a contact matrix H which is a $n \times n$ matrix if the protein has n residues. The element $H_{i,j}$ of H is assigned a value of 1 if the ith residue and jth residue are in contact, otherwise, it is 0.

There are various ways to weight contact energies for different residue pairs from the 20 naturally occurring amino acids. The simplest is the HP model in which the amino acids are classified as H type (Hydrophobic) and P type (Polar). Pairwise contact energy is 1 if both the residues involved are H type, otherwise, it is 0. The statistical potential obtained by Miyazawa and Jernigan is a 20×20 matrix which can be written in the following form after Li, Tang and Wingreen parameterization[38]:

$$E = c_2(q_i + a)(q_j + a) + constant \tag{5.1}$$

where q_i is the q value of residue type i. By replacing the value q_i by $q_i^{'} = q_i + a$ where a

is a constant, Eq. (1) can be rewritten as

$$E_{i,j} = c_2 q_i' q_j' + constant \tag{5.2}$$

We will refer to the modified q' as q value in the rest of this paper. Since the constant in Eq. (2) is irrelevant to the residue identity and hence has no effect in our study of sequence-structure relationship, we will neglect it in the rest of this paper. The constant c_2 in Eq. (2) can be viewed as a unit in calculating energy, thus, it is set to be 1.

Under these interaction scheme(HP or LTW parameterized MJ matrix), the sequence of a protein can be represented by a sequence vector **S** whose elements are either 1 or 0 for HP model, or q values of the residues if using the LTW parameterized MJ matrix. For a protein with H as the native contact matrix, the conformational contact energy can be written as

$$E = \langle S \mid H \mid S \rangle \tag{5.3}$$

The energy form of Eq. (3) is similar to a standard quantum system with H as its Hamiltonian. The difference is in the the vector space. For a quantum system, elements in vector **S** can be any complex number, while for the protein system, the elements in vector **S** are limited to the 20 q values (LTW) or (1,0) (HP model). Note that H can be decomposed into its eigenstates $\mid V_i \rangle$, i.e.,

$$H = \sum_i \lambda_i \mid V_i \rangle \langle V_i \mid \tag{5.4}$$

where $H \mid V_i \rangle = \lambda_i \mid V_i \rangle$. Thus the total conformational contact energy can be expressed as the summation of the individual contribution of the eigenvectors of H :

$$E = \sum_i \lambda_i W_i \tag{5.5}$$

where $W_i = \mid \langle S \mid V \rangle \mid^2$.

For a quantum system, the ground state is $\mid V_0 \rangle$, with the W_i spectrum : $W_0 = 1$, $W_i = 0$ if $i \neq 0$. For the protein, however, the vector space $\mid S \rangle$ is restricted by the 20 naturally occurring amino acids, Thus the $\mid V_0 \rangle$ is generally unreachable for the protein sequence vector, and W_i might be different from that of the quantum system even the protein folding process is indeed optimized the hydrophobic interaction.

Figure 5.1 and Figure 5.2 show the W_i spectrum of a protein (1a0b from Protein Data Bank (PDB)) using HP sequence(Figure 5.1) and LTW q values(Figure 5.2). Clearly, the dominant contribution to the energy is from the first eigenvector. The contributions from other eigenvectors can be viewed as result of the restriction on $\mid S\rangle$ described above, or may partially be a result of the inaccuracy of the energy scheme used.

The spectrum of W_i we get from 1a0b is not a special case. In order to characterize the feature of W_i spectrum of proteins, we randomly chose 174 proteins from the PDB. The only requirement for these proteins is they should have good experimental structure resolution($< 1.5\mathring{A}$) and should consist of a single chain. W_i is calculated for each of the proteins, and the average of W_i over these 174 proteins is plotted in Figure 5.3. For the purpose of comparison , we randomly shuffled the native sequence(HP) of each protein. W_i of the shuffled sequences were averaged and plotted with the average W_i spectrum obtained from the native sequences. The difference between the native sequence and the shuffled sequence mainly comes from the first eigenvector. Because even the simple two letter alphabet HP sequences exhibit this character, we believe this reflects the fact that the dominant factor in protein folding is optimization of hydrophobic interactions. Any interaction scheme which grasp the hydrophobic interaction as the driving force for protein collapse should be able to reproduce this feature. The difference between W_i of the native and shuffled sequence drops quickly as the rank of the eiegenvector increases. When $i > 4$, this W_i difference is negligible, which implies that the majority of the eigenvectors are residue-ordering "blind". This means that in the process of assessing the sequence-structure fitness for a given amino acid sequence and a given protein structure, one only needs to exam the effect of the several eigenvectors rather than those of entire contact matrix, and that the dominate eigenvector $\mid V_0\rangle$ is the first order of approximation.

Structural "threading" is a widely used methods for protein structure recognition and prediction. A common approach in structural threading is to use a structural "profile" to represent a template of known protein structure. Generally, the native sequence of the template protein plays an important role in generating the profile. Due to the strong correlation between

the sequence and the dominant eigenvector of the contact matrix of the protein structure as discussed above, we believe that the dominant eigenvector is a better representation than the native sequence as a profile for structural threading studies. It is well known that homologous proteins generally have similar structure, but the structural similarity does not necessarily require similarity in sequence. Due to evolution, some protein's homologous sequence can diverge so much that the sequence similarity is undetectable, while they still share similar three dimensional structures (e.g. TNF family). Also, very different sequences can give rise to similar structures through convergent evolution. Correlation between the sequence and the dominant-eigenvector is a consequence of the interaction among residues, and probably, is not directly effected by evolution. Thus, for a given structure, its dominant may eiegenvector provide more fundamental structural information than its the sequence, especially when one want to establish the linkage between proteins with very little sequence similarity.

In order to compare the efficiency of native sequence vs eigenvector as the structural profile, we have done a threading test using the threading scheme described in another paper. Protein sequences were chosen from the ASTRAL data base of protein structural domain(the detail of this database will be discussed in the later part of this thesis). Figure 5.4 show the frequency at which the optimum threading energy were obtained using different profiles. The dominant eigenvector has 17% more chance in finding the energy minimum compared with the native seqeunce.

The sequence and eigenvector correlation we get above hold when we use the entire protein sequence. The strong correlation generally does not exist if we instead use only part of the sequence of a protein. This is in agreement with experimental observations that partial sequences of proteins often unfold (or are unstable) under normal conditions. However, there are abundant examples in which protein fragments can fold independently. These partial sequences generally correspond to "domains" of proteins. We believe these individual structural domains have the feature we described above that distinguishing them from random amino acid sequence, and hence can fold correctly. To test this assumption, we use the overlap of a multidomain protein sequence and 1st eigenvector as an index to test whether this value

Protein name	domain boundaries(experiment)	domain boundaries(predicted)
1cfb	101	108
1cid	106	105
1dru	128	133
1fdr	100	100
1grj	79	74
1hjp	65	66
1kaz	189	218

Table 5.1 Comparation of predicted domain boundaries with biological determined domain boundaries

can identify the positions of domain boundaries. Seven two-domain proteins were randomly selected from PDB. For each of the protein sequence, we collect all continuous (sequence-wise) segments with length > 30 residues and $< 2/3$ of the total sequence. The sub-matrix of the protein contact matrix corresponding to segment is diagonalized, and the overlap between the subsequence and the dominant eigenvector of the sub-matrix is calculated. The function Ph(i) is defined as the largest overlap for all segments starting at residue i. Figure 5.6 shows the Ph(i) value for a single domain protein, Figure 5.5 show the Ph(i) value for a two-domain protein. The obvious difference between the single domain Ph(i) and the two-domain Ph(i) is that the two-domain protein present a strong peak at certain positions. This suggested that there is a segment starting at that position which exhibits a strong sequence and dominant eigenvector correlation. We believe this implies a sub-domain in the protein which has a domain boundary around the special position. Table 5.1 is the list of the domain boundaries proposed by the eigenvector study (using the starting position of largest segment-eigenvector overlap) compared with the actual domain boundaries determined biologically(ASTRAL). For six two domain proteins, the predicted and experimental boundaries are in close agreement.

In summary, we found a strong correlation between protein sequence and the dominate eigenvector of its structure contact matrix. High ranking eigenvectors are sequence blind. This correlation hold for protein sub-domains. The dominate eigenvector provides a good representation of a protein's three dimensional structure, thus a gapped structural threading method can be built using this principle. We believe this correlation is a result of the hydrophobic interaction dominating a protein's folding process. In order to fulfill this principle,

protein sequences and structures in nature are restricted, which result in the specificity in natural protein sequences and structures. Although it is impossible to enumerate all possible contact matrice of natural proteins at present, to enumerate the contact matrice of compact lattice proteins is within our reach if the sequences are relatively short. Since the underlying interaction in lattice HP model is only hydrophobic interaction, we believe the principle we suggest for real proteins will remain true for HP model lattice protein. We are testing this in a designability study using the dominate eigenvector overlap as a measurement of distance between protein structures on lattice. Highly designable lattice protein structures should have a low overlap with other structures. We believe this study on lattice is generalizable to real three dimensional structures because contact matrix does not differentiate between lattice and real space off-lattice proteins. Such studies are underway.

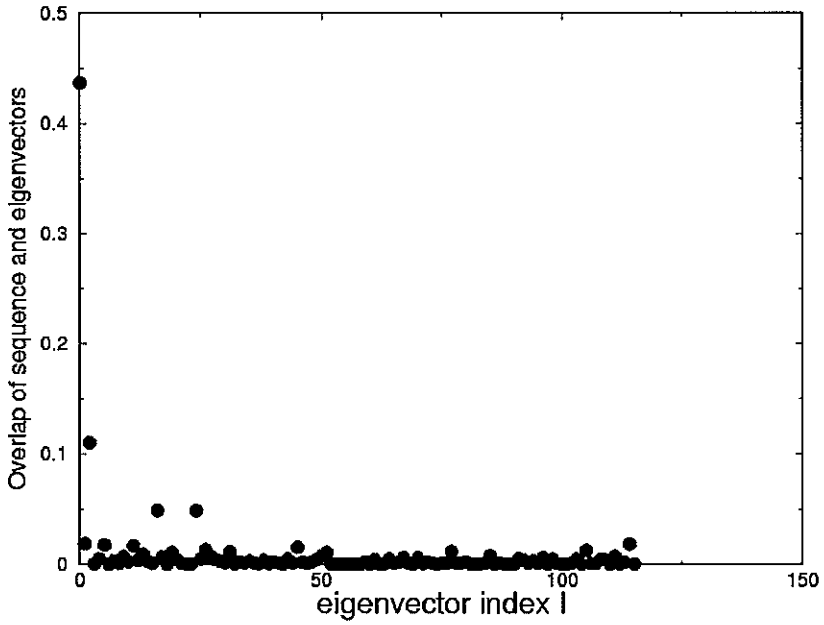

Figure 5.1 Overlap between protein (1a0b) sequence (using HP model) and eigenvectors of the protein's native structure contact matrix.

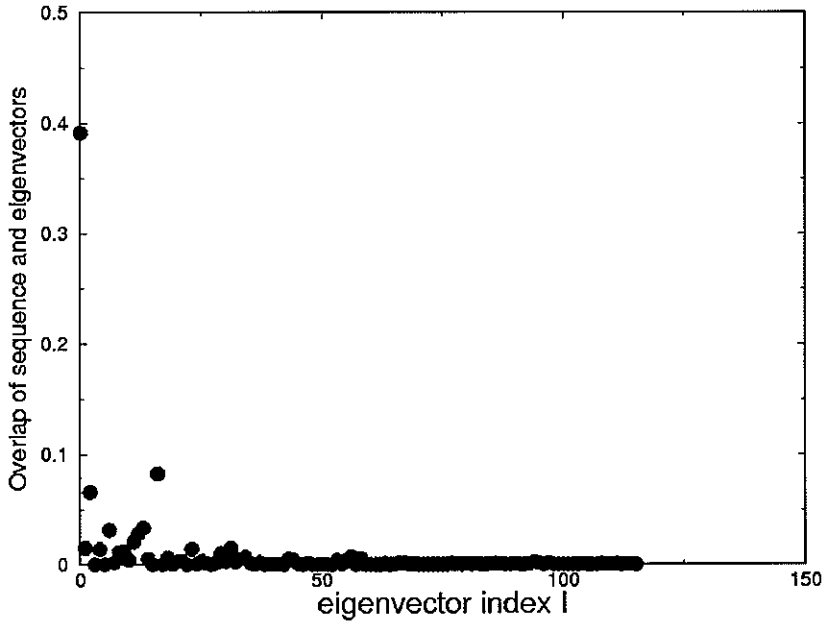

Figure 5.2 Overlap between protein (1a0b) sequence (using LTW parameterized MJ matrix) and eigenvectors of the protein's native structure contact matrix.

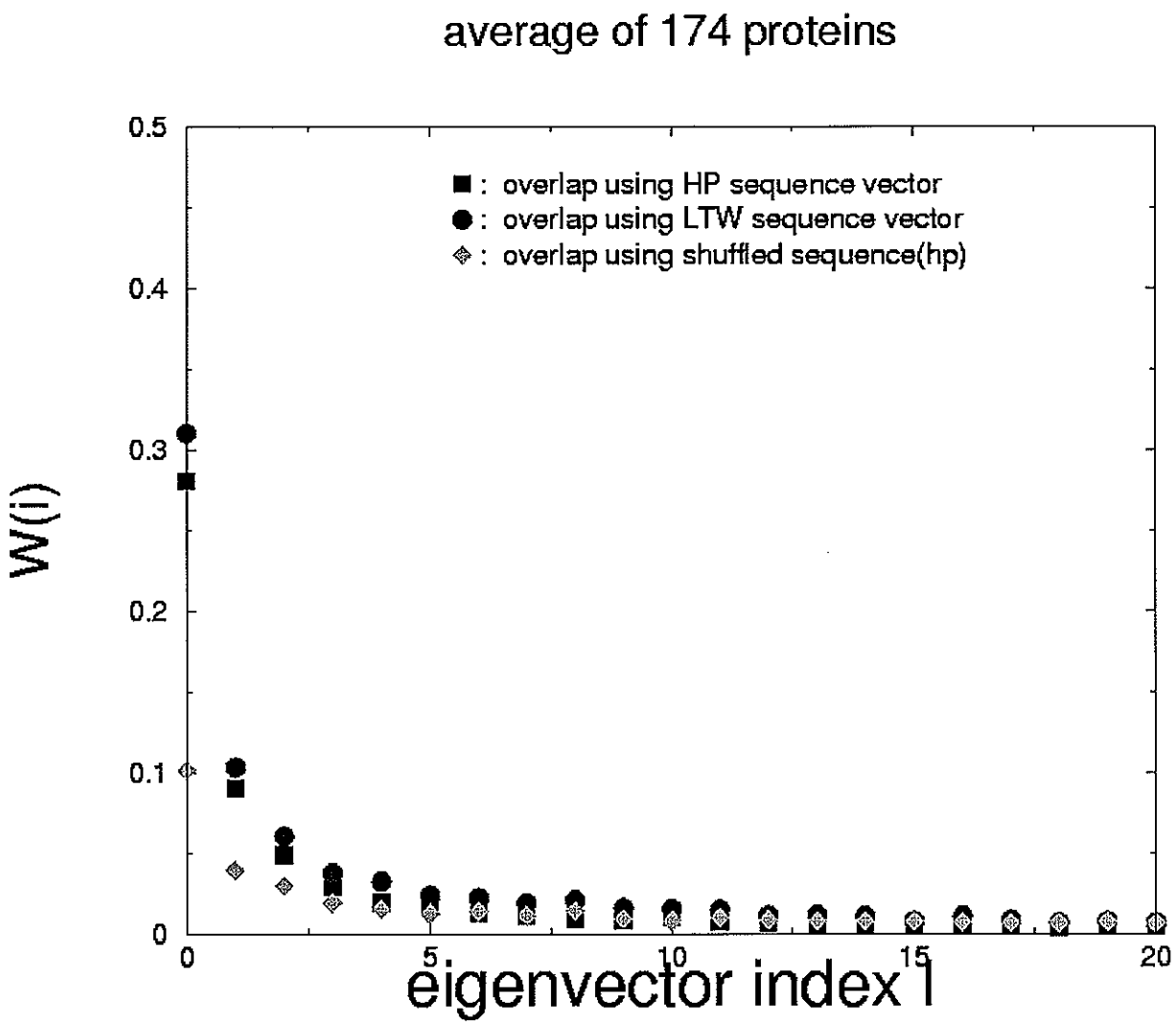

Figure 5.3 Average overlap between a protein sequence and eigenvectors of its native structure. The Average is over 174 randomly selected proteins from PDB. Only resluts of the top 30 eigenvectors are plotted. Average overlap using shuffled HP sequences is also plotted for comparison.

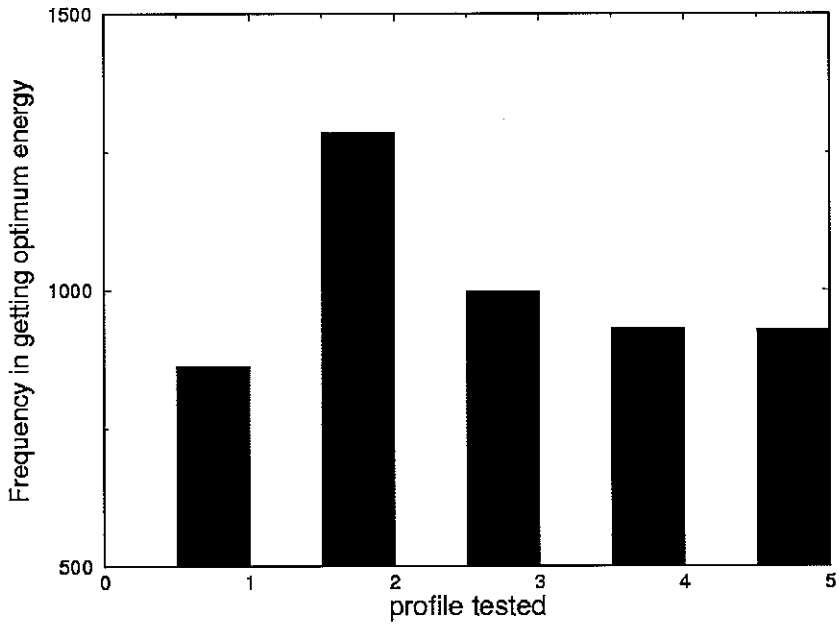

Figure 5.4 Comparison of efficiency of threading using different profiles. Profiles used were:
1: native sequence 2: $|V_0\rangle$. 3: $|V_1\rangle$. 4: $|V_2\rangle$. 5: $|V_3\rangle$.
34 protein sequences chosen from 3 different families in ASTRAL were threaded on protein structures belong to the same superfamily to produce these statistics.

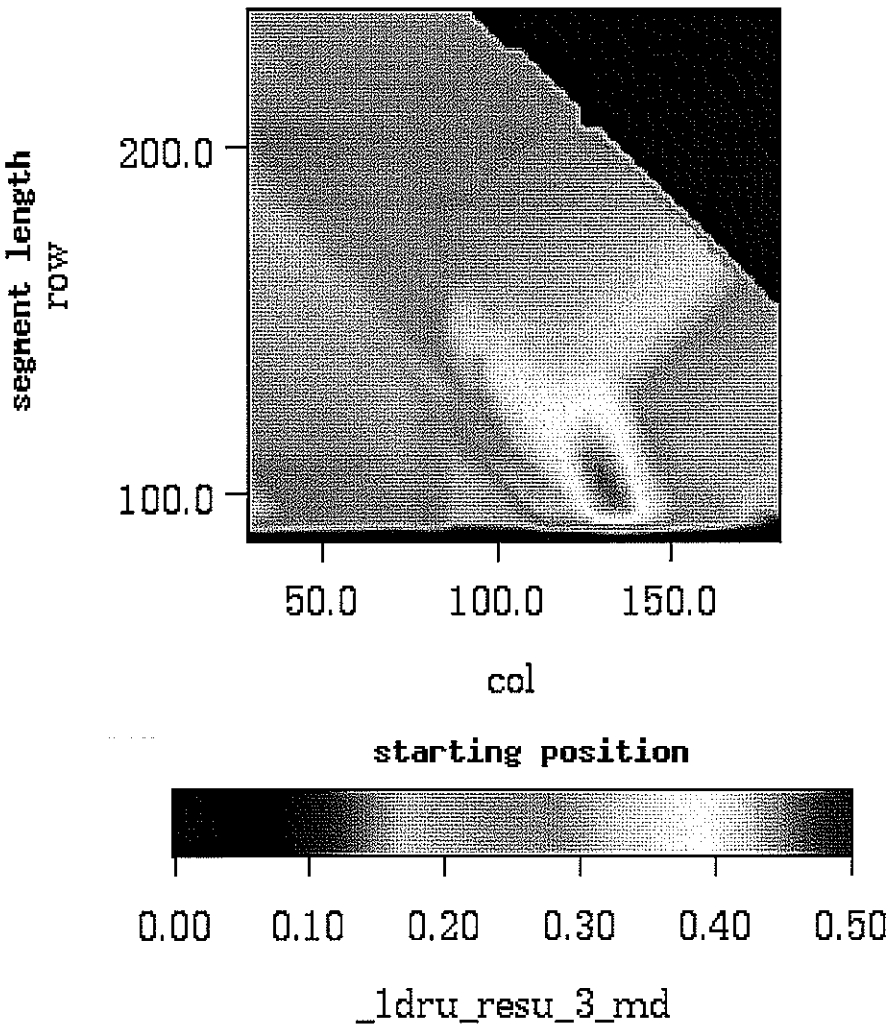

Figure 5.5 Overlap of partial sequences with dominant eigenvectors from their corresponding native structure contact matrice: two-domain protein. Protein id: 1dru(two domains, domain boundary at residue 127 according to ASTRAL)

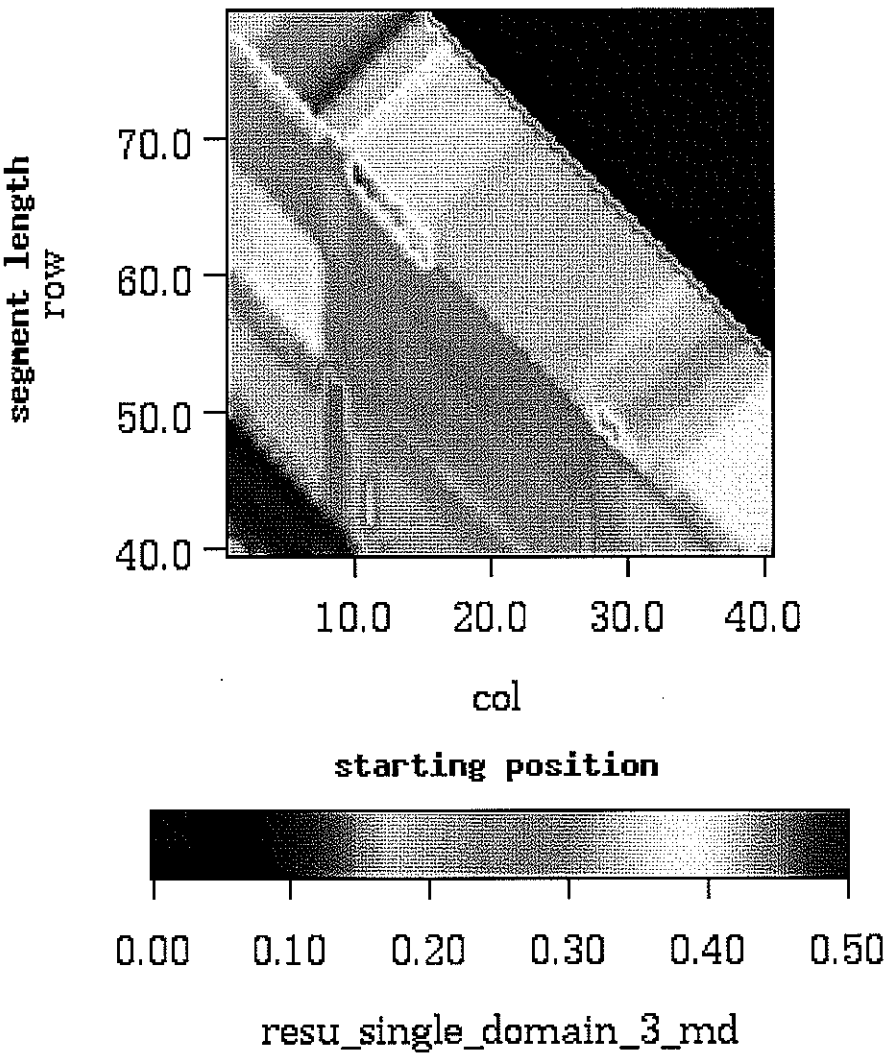

Figure 5.6 Overlap of partial sequence with dominant eigenvector from their corresponding native structure contact matrix: single domain protein. Protein id: 1a0b.

Bibliography

[1] Anfinsen, CB. Haber, E. Sela, M. White, FH. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA **181** 223-230 1961

[2] Chandonia J.M., Walker N.S. Lo Conte L., Koehl P., Levitt M., Brenner S.E. Nucleic Acids Research **30**:260-263 (2002). Brenner S.E., Koehl P., Levitt M. Nucleic Acids Research **28**:254-256 (2000).

[3] HueSun Chan, K.A. Dill J. Chem. Phys. **92** (5) :3118 (1990).

[4] K.F. Lau, K.A. Dill Macromolecules **22** 3986 (1989).

[5] J.D Bryngelson, J.N Onuchic, N.D Socci, P.G Wolynes Proteins **21**:167-195 (1995).

[6] K.A. Ken, S. Bromberg, K. Yue, K.M. Feibig, D.P. Yee, P.D. Thomas, H.S. Chan Principle of protein folding-A perspective from simple exact model. Protein Sci. 4: 561-602 (1995).

[7] Hao Li, Chao Tang, N.S. Wingreen Phy. Rev. Lett. **79**:765-768 (1997).

[8] Li, H. Tang, C. Wingreen, N. Science **273** 666-669 1996

[9] Li, H. Tang, C. Wingreen, N. Proteins **49** 403-412 2002

CHAPTER 6. Three-Dimensional Threading Approach To Protein Structure Recognition: A Paper Been Accepted By Polymer.

by Haibo Cao, Yungok Ihm, Cai-Zhuang Wang, Drena Dobbs and Kai-Ming Ho

abstract

We describe a gapped structural threading method starting from aligning the query protein sequence to the dominant eigenvector of the structure contact-matrix. A mathematically straightforward iteration scheme provides a self-consistent optimum global sequence-structure alignment. The computational efficiency of this method makes it possible to search whole protein structure databases for structural homology without relying on sequence similarity. The sensitivity and specificity of this method is discussed, along with a case of blind test prediction. This method will provide a versatile tool for protein structure prediction and protein domain recognition complementary to existing tools that rely on sequence homology.

Review

Globular proteins form unique three dimensional structures under natural conditions. With few exceptions, the native structure of a protein is determined only by its amino acid sequence [1]. Nevertheless, to predict the unique native structure of a protein given its amino acid sequence (i.e., protein folding problem) remains an outstanding challenge.

Although naturally occurring proteins can have dramatically different structures, related groups of proteins often share a global folding topology. A number of databases exploit this to classify known proteins according to their structural similarities [2, 3, 4] . In the ASTRAL database [4], for example, more than 27000 known proteins are classified in a hierarchical way.

The five structural levels assigned by this database are protein subfamilies, families, superfamilies, folds, and classes in the order of decreasing similarity among members. When two proteins belong to the same family, they generally share similar biological functions and exhibit significant sequence similarity which can be detected by sequence comparison tools like PSIBLAST[5]. The average root mean square deviation (RMSD) between different protein structures from the same family is usually under 1 Å. At the superfamily level, proteins have much higher RMSD (around 5 Å) and generally low sequence similarity even though they share a similar global folding topology. When a sequence alignment method is used among these proteins, the sequence identity generally falls into the "twilight zone" (below 20% amino acid identity) where the linkages among these remotely homologous structures cannot be established. The structural threading method we introduce in this paper aims to identify these remotely homologous structures from other unrelated known structures.

When a protein is in its natural environment, it is generally believed that the native state corresponds to the global minimum of the free-energy of the protein molecule. Studies of the protein folding process suggest a global collapse followed by fine tuning of the structure around the native global free-energy minimum [6, 7, 8, 9, 10]. From studies of lattice models, Chan and Dill [11, 12] proposed that proteins correspond to highly atypical polymer sequences with a well-defined unique free-energy minimum configuration separated from other configurations by a relatively large energy gap. A funnel-like energy landscape for protein folding was also proposed by Wolynes and co-workers [13]. Therefore, it is reasonable to assume that, when a protein folds into a three-dimensional structure similar to its native structure, it should have lower free energy compared with misfolded structures. Thus, the native structure for a given protein sequence can be inferred by threading the sequence on known protein structures and calculating the energy for each threading. If a target protein's native structure is similar to a known structure in the database, then the threading energy should be lower than those of other structures in the database. Thus the global fold of the protein can be recognized.

Hendlich et al [14] introduced, in 1990, a threading method to test sequence-structure compatibility. A number of schemes for structural threading have been proposed over the past 13

years [15, 16, 17, 18, 19, 20, 21, 22, 29, 30, 46, 47]. The basic idea of threading is to assume that a query protein sequence takes on the three-dimensional conformation of a template structure. This is a one-dimensional to three-dimensional (1D-3D) alignment since the ordering of the original sequence is required to remain unchanged in the threading process. The difficulty of this problem depends on whether "gaps" are allowed in the alignment process or not. Early work generally involved gapless threading [17, 20] in which insertions and deletions were not considered. For gapless threading, it is possible to enumerate all possible alignments, however, generally this approach cannot provide competitive decoys [23, 24]. While it can pick out the native fold from a collection of structures, it is not good at identifying closely-related proteins even when the structural similarity is high. When gaps are introduced in the alignment process, a simple dynamic programing method [25, 26, 27] cannot be used without significant modifications due to the long-range (in terms of sequence separation) interactions of the residues in the threaded structure. Godzik and Skolnick [18] proposed the "frozen approximation," in which the residue's environment is evaluated using the native sequence of the threaded structure instead of the query sequence. Then, a conventional dynamic programming method can be used for the sequence-structure alignment. This approach can be viewed as a way to make a 1D structural profile on which the sequence can be aligned. By modifying the structural profile according to the alignment obtained in the previous step, the threading result can be improved in an iterative manner [28]. A number of threading schemes have been proposed using various ways to obtain structural profiles [29, 36, 47, 42]. Apart from this profiling approach, Jones et al [16] used a double dynamic programing method to find the optimum sequence-structure alignment. A search algorithm for getting global optimum threading method was also devised by Lathrop and Smith [30] using a branch-and-bound approach.

When the optimum sequence-structure alignment is achieved, the accuracy of the threading method depends on the interaction scheme used for calculating the free energy of the system. Many kinds of interactions are involved in the protein folding process, including hydrophobic interactions, hydrogen bond interactions, electrostatic interactions and covalent bond interactions. An interaction scheme which involves atomic details is not suitable for the purpose of

structural threading because amino acids on the template structure are replaced by different types of amino acids from the sequence of query protein. Also, because threading studies may examine many (20,000 or more) sequence-structure pairs, an effective residue-residue interaction that captures the dominant interaction of the protein folding process is important for this purpose.

The driving force for protein folding has been the topic of many discussions. Mirsky and Pauling proposed in 1936 that hydrogen bonds determine the structure of proteins[31]. In 1950s, Walter Kauzmann proposed that the dominant driving force for protein collapse is the hydrophobic interaction [32]. This point of view is adopted in lattice-protein-models studies by Chan and Dill [11, 12, 39, 40]. In the simple H-P model, the interaction energy is a two letter alphabet (H for hydrophobic residues and P for polar residues) pairwise contact energy. When two residues are within a specified cutoff distance (in lattice models, contact is defined as when the two residues are neighbours to each other), a contact energy is assigned according to the characters of the residue pair (e.g., hydrophobic-hydrophobic (H-H) contacts have energy -1, polar-polar (P-P) and hydrophobic-polar (H-P) contacts have energy 0). The total energy is the sum of all pairwise contact energies of the conformation. A more detailed 20 alphabet residue-residue interaction was proposed by Miyazawa and Jernigan[33, 34]. They applied a quasi-chemical approximation to the relative abundance of different types of residue-residue contacts in existing structures in the protein data bank (PDB) to produce a table of residue-residue contact energies among the 20 amino acids : the MJ matrix [33, 34]. Various other empirical interaction energy forms have also been proposed and tested by different groups[20]. Li, Tang, and Wingreen showed that the Miyazawa-Jernigan (MJ) matrix can be factorized and interpret the resulting form of the interaction to show that hydrophobic interaction is the dominant factor in the MJ interaction matrix [38]. Local interactions to stabilize secondary structures in the native state of the protein are also important in determining the three-dimensional structures of proteins. Miyazawa and Jernigan [35] showed that it is possible to distinguish native structures from other decoy structures using a gapless threading method when the secondary structure energy is included [35]. Here we propose a two-step structural

threading method. In the first step, the query sequence is aligned onto the target structure by optimizing the overlap of the sequence vector and the dominant eigenvector of the target structure contact matrix. In the second step, the threading energy is calculated based on the alignment obtained in the first step.

Method

Energy Functions

The interaction energy used in this paper follows the Li, Tang, Wingreen [38] parametrization of the MJ matrix. In the HP and the MJ models, the interactions are "contact" interactions. In calculations of the free energy, a three-dimensional protein structure can be represented by a contact map. For a protein containing N residues, the contact map is a $N \times N$ matrix with element (i,j) whose value is 1 if the i^{th} residue and j^{th} residue are in contact, otherwise, the element is set to 0. We choose $6.5\mathring{A}$ as the contact cutoff distance in accordance with the MJ matrix.

Through eigenvector analysis of the MJ matrix, Li, Tang, and Wingreen showed that the interaction energy can be written in the form

$$E = c_1(q_i + q_j) + c_2 q_i q_j + constant \tag{6.1}$$

Thus, the 210 different residue-residue interactions in the MJ matrix are not entirely independent but can be described approximately by 20 parameters. This can be written in a factorized form:

$$E = c_2(q_i + a)(q_j + a) + K \tag{6.2}$$

where K and a are constants independent of residue type. The additive constant K has no effect on the output of the structural threading and will be eliminated hereafter in this paper. From equation (2) we can redefine modified q values as $q_i + a$, then equation (2) can be written as : $E = c_2 q_i q_j + K$. We will refer to this modified q value as q_i in the rest of this

paper. If we represent a protein sequence vector s by the q values of its amino acids q_i, for a given alignment after threading a sequence on a template structure, the conformation energy can be written as:

$$E = \sum_{i,j=1}^{n} q_i' C_{i,j} q_j' \tag{6.3}$$

where $C_{i,j}$ is the contact matrix of the structure and q_i' is the aligned sequence vector $\mathbf{s}'$.

Alignment

The problem of finding the best alignment of a query sequence s for a structure with contact matrix C is to find a transformation from s to $\mathbf{s}'$ that optimizes the free-energy function (3). The transformation has to be performed under the following restrictions:

1: $|\mathbf{s}'| \leq |\mathbf{s}|$ i.e. , no added residues can be introduced.

2: the ordering of the sequence must be kept.

Mathematically, if the residue types are not restricted to the 20 naturally occuring amino acids and the two threading restrictions are ignored, the sequence vector can span the whole N dimensional real space. This modified problem is readily solved: The optimum $\mathbf{s}'$ is the dominant eigenvector $\mathbf{v_0}$ of the contact matrix C (see Appendix). Under the threading restrictions, the phase space of $\mathbf{s}'$ consists of discrete points in the N dimensional space. If the native structure of the query protein is similar to that of the template structure being considered, we may expect the resulting transformed vector $\mathbf{s}'$ to be located close to $\mathbf{v_0}$. We will discuss in detail the evidence for the correlation between a protein sequence and the dominant eigenvector of its native structure's contact matrix in another publication. Here we propose that the transformation we are seeking can be obtained by maximizing the correlation between $\mathbf{s}'$ and $\mathbf{v_0}$:

$$\frac{(\mathbf{s}' \cdot \mathbf{v_0})^2}{(\mathbf{s}' \cdot \mathbf{s}')(\mathbf{v_0} \cdot \mathbf{v_0})} \tag{6.4}$$

This is an alignment problem, and the dynamic programming method in sequence alignment can be readily adopted to solve this problem. The process can also be viewed as using $\mathbf{v_0}$ as

a profile.

Iteration

The step of aligning with $\mathbf{v_0}$ will produce a transformed vector $\mathbf{s'}$ which is close to $\mathbf{v_0}$. The ultimate solution $\mathbf{s^{max}}$ also sits close to $\mathbf{v_0}$. This makes us believe that the transformation we get is close to the optimum solution. Further improvements can be achieved by an iteration scheme described below. The contact matrix energy function (3) can be rewritten as: $E = \mathbf{s'} \cdot \mathbf{A}$, where $\mathbf{A} = C \cdot \mathbf{s'}$. If the vector $\mathbf{A}$ is known, the transformation from $\mathbf{s}$ to $\mathbf{s'}$ is an alignment problem. On the other hand, $\mathbf{A}$ can be found by using the contact matrix C to transform the vector $\mathbf{s'}$. This makes it possible to use an iterative method to optimize the $\mathbf{s} \Rightarrow \mathbf{s'}$ transformation we need. Starting with $\mathbf{v_0}$ as the initial guess for $\mathbf{A_0}$, alignment with sequence vector $\mathbf{s}$ gives $\mathbf{s'_1}$. From $\mathbf{s'_1}$ transformed by C, we can get $\mathbf{A_1} = C \cdot \mathbf{s'_1}$, and repeat the process of alignment. This iterative procedure can be repeated until $\mathbf{A_n}$ and $\mathbf{A_{n+1}}$ converge. This iteration process is similar to commonly-used iterative methods for finding the eigenvectors of a symmetrical matrix [41]. Because of the involvement of the alignment process and the restrictions on the choice of $\mathbf{s'}$, the convergence of the iterative process is not mathematically guaranteed. In order to get a final converged alignment, the initial guess is important. In our work, we used for initial guesses not only the eigenvector with the largest eigenvalue but also repeat the calculation with each of top four eigenvectors of the contact matrix as well as the vector corresponding to the frozen approximation. This improves the chance of getting a converged result.

Gap penalty and size effects

For any method involving gapped alignment, the outcome is affected by the penalty for insertion/deletion. In the work of Lathrop and Smith [30], the structure is divided into two regions: regions with well-defined secondary structures and loop regions. Insertions and deletions are forbidden in the secondary structure regions and no gap penalties are assessed in the loops. We follow a similar approach. In our work, the threading is divided into two steps. In

the first step, the sequence is aligned to the vector **A**, and then in the second step, the score is calculated using the resultant alignment. After some tests, we found that the performance of the scheme is optimized when we include gap penalties only in the alignment step and not in the energy calculation step. In the alignment step, insertion/deletion in the coil region have small penalties, while gaps in the secondary structure region are strongly penalized. Using this gap penalty system, we allow the possibility of making big "jumps" in the threaded structure without serious disruption of the secondary structure. Using our threading method, a substantial portion of the threaded structure can be removed without severe penalty as long as the contact score stays high.

We adopt a similar treatment of size effects. Size penalties are included only in the alignment step and not in the final score calculation. We obtained an average size for each amino acid from the PDB. If a residue in the template structure is replaced by a residue in query sequence whose size differs by $0.5\mathring{A}$ or more in radius, the alignment contribution for that alignment pair is reduced if that residue has three or more contacts in the threaded structure. The alignment score penalty is bigger as the discrepancy in size increases.

The process of including gap and size penalties only in the alignment step has the advantage of removing threading alignments with unphysical gaps and packing from consideration without putting too many parameters into the energy calculations.

Secondary structure energy

Hydrogen-bonds in the secondary structure region play an important role in helping to stabilize the native structure[16]. Miyazawa and Jernigan pointed out in their paper [34] that inclusion of secondary structure energy helps to distinguish native structures from other decoy structures. In this work, we use a "global fitness" factor to take this interaction into account. To calculate this factor, we first obtain a secondary structure prediction for the query sequence using secondary structure predictors such as PSIPRED, PROF, JPRED, and SAM. The "global fitness" is then defined as: $f = \frac{N_+ - N_-}{N_s}$ where N_+ is the total number of matches between the predicted secondary structure and the threaded structure. N_- is the total number

of mismatches, and N_s is the total number of residues in the threaded structure selected in the alignment. We define a modified energy of the form : $E^{modified} = \alpha f E^{threading}$ where α is a parameter which can be optimized for accuracy of fold-recognition.

Raw score and relative score

The negative of the modified energy obtained above is taken to be the raw score for the threading. Thus, a high score denotes a structure with favorable energy. The raw score can contain systematic biases that lead to inaccuracies in identifying sequence-structure relationships. In comparing different structures, structures with more contacts tend to have higher scores than structures with fewer contacts. In comparing different sequences, sequences that have a higher percentage of hydrophobic residues tend to have higher scores. Thus, a high raw score does not automatically mean a high compatibility between the sequence and the threaded structure.

Work by Bryant and Altschul [37] and Meller and Elber [42] showed that the accuracy of threading method can be improved by using the Z-score instead of the raw score for the selection of candidates. In this approach, after a sequence-structure threading is obtained, the query sequence is randomly shuffled and threaded again on the same structure. The Z-score is obtained by $(E^{raw} - E^{ave})/\sigma$, where E^{raw} is the result of the sequence-structure threading, and E^{ave} and σ are the average and standard deviation respectively of the results from the randomly shuffled sequences. In order to eliminate some of the biases inherent in raw scores, we take an approach similar to the Z-score scheme by computing a relative score which we use for our selection criterion. The "relative score" is defined by $E^{rel} = E^{raw} - E^{ave}$ where E^{ave} is the average score obtained by randomly shuffling the protein sequence and threading again on the target structure. We find that relative scores give better discrimination among structures. The use of the relative score may be rationalized from the thermodynamics of protein folding. When a protein folds, it is not the raw final energy which makes the structure different from its denatured states, but the energy difference between the native energy and that of the molten-globule states. For a randomly-shuffled sequence, we would expect the native structure to have

a free energy similar to the molten-globule configurations. Thus, we can model the average energy of the molten globule by the average of the threaded energies of the randomly shuffled sequences on the native structure. E^{rel} can be viewed as the "energy gap" between the native structure and its molten-globule competitors. E^{rel} is obviously closely related to the Z-score used in other threading studies. However, operationally, relative scores converge much more rapidly with the number of shuffled sequences than the Z-score because E^{rel} does not involve the standard deviation (which converges much more slowly than the average score).

Results and Discussion

We have performed a series of tests to benchmark the above method and scoring scheme. In the first test, we randomly selected 174 proteins from PDB. These proteins are listed in Table 1. We restricted ourselves to those proteins which have experimental resolution better than $1.5\mathring{A}$ and a single peptide chain to avoid any possible interchain interactions. For each protein sequence in this set, we perform threading calculations on all of the 174 template structures , a process we call "cross threading". The self-threading score is compared with the best decoy threading score. We found that the native structures always give better scores (higher E^{rel} values) than any decoys in this selected protein set. The self-threading score exhibits a well-defined linear relationship as a function of the sequence length as shown in Figure 1. The reason for the linear correlation is that the number of contacts of a native protein structure is roughly proportional to its sequence length. By taking this into account, we can compare threading results of proteins with different lengths.

A more challenging test for the threading method is homolog-recognition. The above test of self-recognition depends more on the scoring function than alignment process because a gapless threading method would be able to provide similar results. We choose 9 families from the ASTRAL database from which we selected 86 proteins listed in Table 2. Proteins belonging to the same family are homologous and generally have greater than 20% sequence identity, thus a sequence alignment method (e.g. PSIBLAST) can detect the similarity among them. We performed a cross threading test using this 86 protein set. For each query sequence, E_i^{hom} is

defined as the highest threading score among the homologous structures, E_i^{dec} is defined as the best threading score among all the rest of the decoy structures. We rescale E_i^{hom} and E_i^{dec} according to their threading score on native structures E_i^{nat}. A plot of E_i^{hom}/E_i^{nat} against E_i^{dec}/E_i^{nat} is shown in Figure 2. For 83 out of 86 cases, E^{hom} is clearly much higher than E^{dec}. For the remaining 3 cases, the native structure cannot be distinguished from the best decoy structure. This might be a result of inaccuracy of the scoring function we used.

The above tests give us confidence that when a given template structure has a native sequence which is similar to the query protein sequence, our method can distinguish it from random decoy structures without using the sequence information. In the next test, we want to investigate the fold recognition capability of our method for proteins with low sequence similarity. It is well known that structural similarity does not necessarily require sequence similarity. Proteins in the ASTRAL database which belong to the same superfamily but different families generally share similar global structure, but have low sequence identity not detectable by sequence comparison methods. In some cases, even proteins in the same family have such divergent sequences that the structural homology can not be detected by sequence-based recognition methods. For example, the TNF-like family includes both tumor necrosis factor (TNF) ligand domains as well as complement 1q (c1q) proteins. The structural relationship between these two families of proteins was not recognized by sequence-based methods such as PSIBLAST and hidden-Markov-model methods such as PFAM. Because we designed our method to use only structural information, we believe that it can distinguish such similar structures from random decoy structures. To test this, we chose 3 superfamilies (a.1.1, b.1.1, c.2.1) from ASTRAL database. They belong to 3 different folding classes: all alpha (a), all beta (b) and mixture of alpha/beta (c, which is mainly beta sheets). One family is chosen from each of these superfamilies: a.1.1.2, b.1.1.1, and c.2.1.1 respectively. A test set of 34 sequences listed in Table 3 were chosen from the three selected families. Structures belonging to the same superfamily but different families are selected as structural homologs (see Table 3). Each sequence from the chosen sequence set is threaded on all the chosen structures. For each sequence in the test set, we define E^{hom} as the threading score obtained when the sequence

is threaded on structures within the same family. In order to assess the noise background, we used the 86 protein structures in the homolog-recognition test to provide decoy structures. E^{dec} is the highest threading score among all decoy structures (i.e. structures not in the same superfamily as the test sequence). The remote homologous threading score E^{remote} is the highest threading score obtained on structures within the same superfamily but not in the same family. Histograms of E^{hom}, E^{remote}, E^{dec} normalized by the self-threading score are plotted in Figure 3. Comparing Figure 3 (a) and Figure 3 (c), we can see that the distribution of E^{hom} is well separated from the E^{dec} distribution. This result is very similar to that obtained in the homolog-recognition test described above. The wide distribution of the E^{hom} could be the result of the inaccuracy in either the alignment step or the scoring scheme.

The result of the remote homolog recognition can be seen by comparing Figure 3 (b) with Figure 3 (c). The center of distribution of E^{remote} is well separated from that of E^{dec}, although the high score tail of E^{dec} overlaps with the low score tail of E^{remote}. Thus, at least half of the remote structural homologs can be recognized using this structural-threading method.

Because the above tests are done using an existing database of proteins with known structures, we cannot ignore the fact that the results may be to some extent biased by the existence of the final structure in the known database. The CASP5 [44] competition provided us with a chance to do a "blind test" of our threading method. In CASP5, sequence of target proteins whose structures have not yet been published are given to participants for prediction. We will discuss one of our successful predictions. The target T174 is one of the difficult targets according to the CASP5 assessment. There are two domains in this protein structure: T174_1 and T174_2. Of all the predictions submitted to CASP5 by various groups, domain T174_1 has the lowest average score and correct alignment percentage, and the T174_2 domain ranks in the lowest 11% of average scores among the 83 domains predicted in CASP5.

Structurally, the T174_2 domain belongs to the d.14.1.5 ASTRAL family, but has very low sequence identity(10%) with its structural homologs. In our blind test prediction of T174, we prepared a representative structure database for threading by selecting structures from the ASTRAL database . When a family in ASTRAL database has more than 20 protein

structures, we randomly choose 20 among them to reduce the redundancy but retain enough representatives to collect sufficient statistics to overcome the noise from decoy structures. Around 15,000 structures were included in our template structure dataset.

In the CASP5 blind test, the entire T174 sequence is provided without any knowledge of the domain boundary. We selected all continuous 120 amino acid segments of T174 sequence shifted by intervals of 5 residues. The choice of 120 is based on examination of the number of ASTRAL domains as a function of domain size. There is a peak in the distribution around 120. Thus we have a good chance of including a large portion of a single domain of the T174 sequence in some of our cuts. Every segment is threaded against all of the template structures to produce a segment-structure alignment score. For each structure, the threading energies of all segments on that structure are compared. The highest E^{max} score is used to represent the threading score of the structure. A histogram showing the distribution of E^{max} scores is plotted in Figure 4. The histogram takes a shape similar to a normal distribution. The best score was obtained by threading one of the partial sequences on a domain structure which belongs to the ASTRAL family d.14.1.5 . The high score end of the histogram is plotted in the inset of Figure 4. The abundance of the d.14.1.5 family structures (indicated in black) in the high end of the distribution indicates that the high threading score for d.14.1.5 is not due to statistical noise. The aligned part of the segment is then extended to the whole sequence and submitted to the CASP5 as our prediction for the T174 structure. Figure 5 compares the experimentally determined structure (a) of the T174_2 domain with our prediction (b) . There are clear global similarities, with close arrangements of α helix and β sheet. The Dali Z-score for structural similarity between the two structures is 8.9 (The higher the Dali Z-score the more similar the structures. A Dali Z-scores of 2.0 or higher indicates structure similarity between the two structures being compared). The alignment is not completely right, about 34% of the residues are aligned in the correct positions.

In order to analyze the sensitivity and specificity of this method, we used the 34 proteins from the remote homolog recognition test as our query sequences, and the representative structures used in CASP5 as a structure database. Structures do not belong to the same

superfamily as a query protein's native structure, it is treated as a decoy structure for the query protein. We excluded decoy structures with significent structural similarity to native structures (i.e. Dali Z-score greater than 2.8) of the query proteins (if the target structure is not in the same superfamily as the query sequence). This resulted in a set of more than 10,000 structures with much more competitive decoy structures than the dataset used in the remote homolog recognition test. We rescaled the score for each query sequence threaded on a template structure according to its threading score on its native structure. For a given cutoff score, A "true positive" is obtained when a query sequence threaded on a remote homolog structure (within the same superfamily as the query sequence in ASTRAL database, but in a different structural family) results in a score higher than the cutoff. Otherwise, it is treated as a false negative. Similarly, when a query sequence threaded on a decoy (i.e. not similar) structure results in a score higher than the cutoff, it is treated as a false positive. Otherwise, it is treated as a true negative. We define sensitivity=$\frac{TP}{TP+FN}$ and specificity=$\frac{TP}{TN+FP}$, where TP, TN, FP, FN stand for true positive, true negative, false positive, and false negative respectively[43]. We plot the sensitivity and specificity vs rescaled score for each of the three superfamilies separately in Figure 6. According to Figure 6, if a query protein sequence has no sequence-homolog in the ASTRAL database but a structural-homolog is present, our method has roughly 35% chance to detect it under optimum conditions.

Conclusion

In this paper, we propose a structural threading method which can be used to perform whole database or genome-wide searches. The method is designed to focus predominantly on structural information, making it particularly useful for establishing linkages between structurally similar proteins that have very low sequence similarity. This tool can provide valuable information complementary to existing sequence-based methods. Also, other groups interested in testing their energy schemes can use this method to generate competitive decoy sets as long as the dominant factor of their energy form can be factorized. With some modifications, the method we propose can also be used in the study of protein-protein interfaces.

Acknowledgment

We would like to thank Kent VanderVelden, Robert Jernigan, and Amy Andreotti for helpful discussions. We would like to thank the Institute for Physical Research and Technology, the Plant Science Institute, the Biotechnology Council, and the Lawrence H. Baker Center for Bioinformatics and Biological Statistics at Iowa State University for financial support and the Scalable Computer Laboratory for computational support in this project.

Appendix

Eigenvectors and Eigenvalues of Contact Matrix

Given a $n \times n$ symmetrical matrix C, its eigenvectors $\mathbf{v_i}$ and eigenvalues λ_i satisfy the following relation:

$$C\mathbf{v_i} = \lambda_i \mathbf{v_i} \tag{6.5}$$

Where index i goes from 1 to n.

For simplicity, we only consider matrix with nondegenerate eigenvalues, by which we mean $\lambda_i \neq \lambda_j$ if $i \neq j$. In this case, the eigenvectors are orthogonal to each other. Because a constant times an eigenvector remains an eigenvector of the same matrix, eigenvectors can be normalized, therefor :

$$\mathbf{v_i} \cdot \mathbf{v_j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{6.6}$$

If C is a real and symetrical matrix (i.e. $C_{i,j} = C_{j,i}$), any n dimension real vector s can be decomposed using $\mathbf{v_i}$:

$$\mathbf{s} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \omega_\mathbf{i} \mathbf{v_i} \tag{6.7}$$

where $\omega_i = \mathbf{s} \cdot \mathbf{v_i}$ is the overlap between vector $\mathbf{v_i}$ and s

The matrix C can also be decomposed into the contribution of its eigenvectors:

$$C = \sum \lambda_i \mathbf{v_i} \times \mathbf{v_i^T} \tag{6.8}$$

In structural threading, the score has the form:

$$E = \mathbf{s} \cdot \mathbf{C} \cdot \mathbf{s} \tag{6.9}$$

We are interested in which unit vector $\mathbf{s}$ ($\mathbf{s} \cdot \mathbf{s} = 1$) will maximize E. we can rearrange vector indices i so that the eigenvalues are in decreasing order: $\lambda_1 < \lambda_2 < ... < \lambda_n$. Because vector $\mathbf{s}$ is unitary, the overlaps satisfy the following equation:

$$1 = \mathbf{s} \cdot \mathbf{s} = \sum_{\mathbf{i,j}} \omega_\mathbf{i} \omega_\mathbf{j} \mathbf{v_i} \cdot \mathbf{v_j} \tag{6.10}$$

using equation (6), we get

$$\sum_i \omega_i^2 = 1 \tag{6.11}$$

Using equations (7) and (8), we can decompose E (equation 9) into contributions of different eigenvectors:

$$E = \sum_i \lambda_i \omega_i^2 \tag{6.12}$$

Because eigenvalues are in decreasing order: $\lambda_1 < \lambda_2 < ... < \lambda_n$, we have $E = \sum \lambda_i \omega_i^2 \leq \sum \lambda_1 \omega_i^2 = \lambda_1$. We used the unitary condition in the last step. Note that the equal sign can be achieved only when ω_i satisfies the following conditions: $\omega_1 = 1$, and $\omega_i = 0$ if $i \neq 1$. By putting this ω_i into equation (7), we get:

$$\mathbf{s} = 1 \cdot \mathbf{v_1} + 0 \cdot \mathbf{v_1} + ... + 0 \cdot \mathbf{v_n} = \mathbf{v_1} \tag{6.13}$$

which means that the dominant eigenvector maximizes the score.

Bibliography

[1] Anfinsen C., Science **181** 223 (1973).

[2] Murzin A.G., Brenner S.E., Hubbard T., Chothia C. J. Mol. Biol. **247**, 536-540.

[3] Dengler U., Siddiqui A.S., Barton G.J (2001) Proteins **42**, 332-344 (2001). Siddiqui A.S., Dengler U., Barton G.J. (2001). Bioinformatics **17**, 200-201 (2001).

[4] Chandonia J.M., Walker N.S. Lo Conte L., Koehl P., Levitt M., Brenner S.E. Nucleic Acids Research **30**:260-263 (2002). Brenner S.E., Koehl P., Levitt M. Nucleic Acids Research **28**:254-256 (2000).

[5] Altschul S.F., Thomas L.M., Alejandro A.S, Zhang J., Zhang Z., Webb M., David J.L. Nucleic Acids Res. **25**:3389-3402 (1997).

[6] Flory P.J. Principle of polymer chemistry. Cornell U.P, Ithaca NY (1953).

[7] Ptitsyn O.B., Kron A.K., Yu.Y. Eizner J Polymer Sci. C **16**,3509 (1968).

[8] de Gennes P.G. J Phys. Lett(Pairs) **36**,L55 (1975).

[9] Post C.B., Zimm B.H. Biopolymers **18** 1487 (1979).

[10] Sanchez I.C. Macromolecules **12** 980 (1979).

[11] Chan H., Dill K.A. J. Chem. Phys. **92** (5) :3118 (1990).

[12] Lau K.F., Dill K.A. Macromolecules **22** 3986 (1989).

[13] Bryngelson J.D., Onuchic J.N., Socci N.D., Wolynes P.G. Proteins **21**:167-195 (1995).

[14] Hendlich M., Lackner P.,Weitckus S. ,Floechner H., Froschauer R.,Gottsbachner k. Casari G.,Sippl M.J. J. Mol. Biol. **216** 167-180 (1990).

[15] Bowie J.U., Luthy R., Eisenberg D. Science **253**:164-170 (1991).

[16] Jones D.T., Taylor W.R., Thornton J.M. Nature bf 358:86-89 (1992).

[17] Sippl M.J., Weitckus S. Proteins **13**:258-271 (1992).

[18] Godzik A., Kolinski A., Skolnick J. J. Mol. Biol. **227**:227-238 (1992).

[19] Ouzounis C., Sander C., Scharf M., Schneider R. J. Mol. Biol. **232**:805-825 (1993).

[20] Bryant S.H., Lawrence C.E. Proteins **16**:92-112 (1993).

[21] Matsuo Y., Nishikawa K. Protein Sci. **3**:2055-2063 (1994).

[22] Mirny L.A., Shakhovich E.I. J. Mol. Biol. **283**:507-526 (1998).

[23] Park B.H., Levitt M. J. Mol. Biol. **258**:367-392 (1996).

[24] Park B.H., Huang E.S., Levitt M. J. Mol. Biol. **266**:831-846 (1997).

[25] Needleman S.B., Wunsch C.D. J. Mol. Biol. **48**:443-453 (1970).

[26] Smith T.F., Waterman M.S. J. Mol. Biol. 1981;**147**:195-197 (1981).

[27] M.S. Waterman, M. Eggert J. Mol. Biol. **197**:723-728 (1987).

[28] Miyazawa S., Jernigan R.L. Protein Engineering **13**:459-475 (2000).

[29] Elofsson A., Fischer D., Rice D.W., Le Grand S., Eisenberg D. Folding Design 1:451-461 (1998).

[30] Lathrop R.H., Smith T.F. J. Mol. Biol. **255**:641-665 (1996).

[31] Mirsky A.E., Pauling L. Proc. Natl. Acad. Sci. USA **22** 439 (1936).

[32] Kauzmann W. Adv. Protein Chem. **14** 1 (1959).

[33] Miyazawa S., Jernigan R.L. Macromolecules **18**:534-552 (1985).

[34] Miyazawa S., Jernigan R.L. J. Mol. Biol. **256**:623-644 (1996).

[35] Miyazawa S. Jernigan R.L. Protein **36**:347-356 (1999).

[36] Jones D.T. J. Mol. Biol. **287**:797-815 (1999).

[37] Bryant S.H., Altschul S.F. Current Opinion in structural biology **5**:236-244 (1995).

[38] Li H., Tang C., Wingreen N.S. Phy. Rev. Lett. **79**:765-768 (1997).

[39] Lau K.F., Dill K.A. Proc Natl Acad Sci USA **87** 638 (1990).

[40] Shortle D., Chan H., K.A. Dill K.A. Protein Sci **1**,201 (1992).

[41] Press W.H., Teukolsy S.A., Vetterling W.T., Flannery B.P. Numerical Recipes in Fortran 77 Cambridge U.P. NY (1999).

[42] Meller J., Elber R. Proteins **45**:241-261 (2001).

[43] Pierre B., Brunak S., Yves C., Claus A.F., Andersen, Herik N. Bioinformatics **16**:no 5,412-424 (2000).

[44] The detail of CASP5 experiment is shown in the official CASP5 website: http://predictioncenter.llnl.gov/casp5/Casp5.html

[45] Fisher D., Elofsson A., Rice D., Eisenberg D. Pacific Symposium on Biocomputing, Hawaii,1996;300-318

[46] CAFASP-1: Critical Assessment of Fully Automated Structure Prediction Methods Fischer, D., Barret C., Bryson K., Elofsson A., Godzik A., Jones, D., Karplus K.J., Kelley L.A., Maccallum R.M., Pawowski K., Rost B., Rychlewski L. and Sternberg M.J. Proteins: Structure, Function and Genetics, Suppl 3:209-217 (1999).

[47] Rost B., Schneider N., Sander C. J. Mol. Biol. **270** 471-480 (1997).

153l	1a0b	1a0i	1aa0	1aa2	1aa3	1aac	1aba	1ad2	1ads
1af7	1ag2	1ag4	1ah7	1ahk	1aho	1ajj	1ak1	1ako	1akz
1al3	1aly	1amp	1anu	1anv	1aol	1aop	1arv	1aua	1awd
1awj	1axn	1bdo	1beo	1bgp	1bkf	1bor	1bp1	1btn	1bv1
1cem	1cfb	1chd	1cid	1csh	1ctj	1cyx	1dad	1ddf	1dhs
1div	1dru	1eca	1ehs	1erv	1eur	1fbr	1fdr	1fkx	1fna
1gai	1gin	1goh	1gpc	1grj	1gvp	1hcd	1hfc	1hjp	1hoe
1htp	1hxn	1idk	1ido	1igd	1irk	1irl	1iso	1itg	1ixh
1jdw	1jli	1kaz	1kid	1knb	1kte	1kuh	1kvu	1lba	1lbu
1lcl	1lit	1lki	1ll1	1lml	1lxa	1mml	1mrj	1mrp	1msk
1mxa	1mzm	1nif	1nls	1nom	1nox	1npk	1nre	1ois	1opd
1opr	1pax	1pbn	1pex	1phc	1php	1pkn	1plc	1plr	1pmi
1poc	1pot	1ppn	1ppt	1prr	1pta	1ptq	1quf	1ra9	1rcf
1res	1rgs	1rie	1rlw	1rmd	1rmg	1rnl	1rss	1ryt	1sig
1sly	1sra	1svb	1tca	1tfe	1tfr	1tib	1tif	1tml	1tsg
1tul	1ubi	1uby	1uch	1utg	1uxy	1vcc	1vhh	1vif	1vin
1vls	1vsd	1vvc	1wer	1whi	1xnb	1ysc	1ytw	1yub	1zid
						1zin	1zxq	2i1b	5p21

Table 6.1 174 protein sequences used in self-recognition test

[48] Luz J.G., Hassig C.A., Pickle C. , Godzik A., Meyer B.J., Wilson I.A. Genes Dev. **17** 977 (2003)

Domain Family	Protein sequence chosen				
a.1.1.2	1a6m	1ash	1babB	1ch4A	1d8uA
	1eca	1eco	1ew6A	1flp	1hlb
	1irdA	1it2A	1ithA	1kfrA	1vhbA
			2gdm	2hbg	2lhb
a.3.1.1	1c6oA	1cie	1crg	1ctj	1f1cA
			1hh7A	1irv	1yeb
b.1.1.1	1ah1	1akjD	1eajA	1fo0A	1gya
			1i85A	1neu	1qfoA
b.3.1.1	1a47-2	1ac0	1b90A1	1cdg-2	1cqyA
	1cyg-2	1d7fA2	1qhoA2	5bcaB1	8cgtA2
c.2.1.1	1a71A2	1agnA2	1cdoB2	1e3eA2	1e3jA2
		1gpjA2	1kevA2	1qorA2	1ykfC2
c.3.1.1		1cjcA1	1djnA2	1h7wA3	1h7xA3
d.1.1.1	1aqzA	1ay7A	1bu4	1bujA	1fus
			1rds	1rtu	1yvs
d.3.1.1	1aec	1aim	1atk	1bp4	1cjl
	1cpjA	1cqdD	1cv8	1dkiB	1fh0A
	1gecE	1meg	1pbh	1qdqA	1yal
e.1.1.1	1a7cA	1atu	1imvA	1jtiA	1qmnA
					1sek

Table 6.2 86 proteins from 9 families in ASTRAL database used in homolog-recognition test

Family	Sequence in the family			Superfamily	Structures in Superfamily		
a.1.1.2	1flp	1kfrA	2hbg	a.1.1	1phnA	1cpcA	1i7yA
	1a6m	1eco	2gdm		1gh0A	1allA	1b33A
	1d8uA	1irdA	1babB		1liaA	1b8dA	1qgwC
	1ch4A	1it2A	2lhb				1kr7A
	1ash	1ithA	1hlb				
	1vhbA	1ew6A					
b.1.1.1	1ah1	1ah1	1akjD	b.1.1	1frtA1	1bmg	1a6zB1
	1eajA	1fo0A	1gya		1ld9A1	1b3jA1	1c16B1
	1i85A	1neu	1qfoA		1exuA1	1exuB1	1igtA2
					1ij9A1	2ncm	1tlk
					1tnn	1wiu	1tiu
					1gl4B	1qtyY	1wwaX
					1hcfX	1fcgA1	1f2qA1
					1efxD1	1g0xA1	1f45A1
					1jbjA2	1eh9A1	1evuA1
					1cc0E	1gdf	1ksgB
					1ayrA1	1a02N1	1bftB
					1h6uA1	1ehxA	1im3P
							1jjuA3
c.2.1.1	1a71A2	1agnA2	1cdoB2	c.2.1	1kvq	1fjhA	1bdb
	1e3eA2	1gpjA2	1kevA2		1a4uA	1gcoA	1h5qC
	1qorA2	1ykfC2			1i01A	1ae1A	1dohA
					1hdoA	1hu4A	1gpdG1
					1brmA1	1dapA1	1arzA1
					2nacA1	1qp8A1	1gdhA1
					1psdA1	1sayA1	1f8gA1
					1b3rA1	1mldA1	1hyhB1
					1hyeA1	1qmgA2	1f0yA2
					1dljA2	1evyA2	1ks9A2
					1jaxA	1bgvA1	1lehA1
					1bw9A1	1a4iB1	1ee9A1
					1do8A1	1id1A	1cqiA1

Table 6.3 Protein structures used in structural-homolog-recognition test

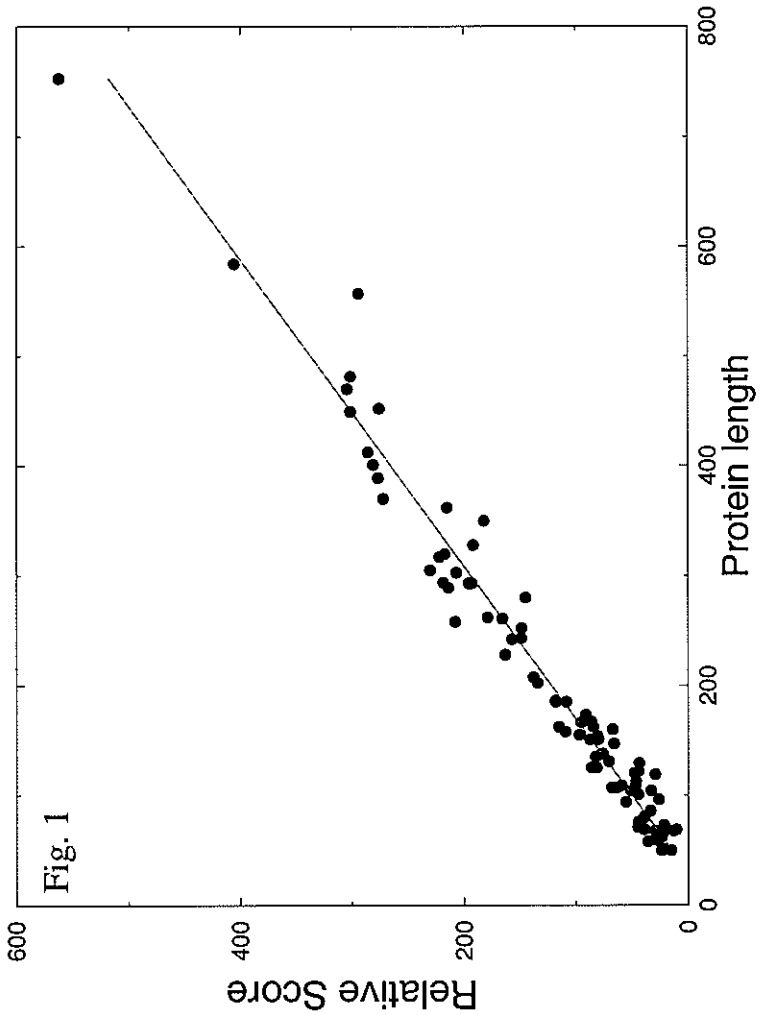

Figure 6.1 Relationship between relative score E^{rel} and protein length. 174 randomly chosen proteins were self-threaded (see text). The relative score for each protein is plotted against the number of residues of that protein. A linear correlation between self-threading score and number of residues of the protein can be observed.

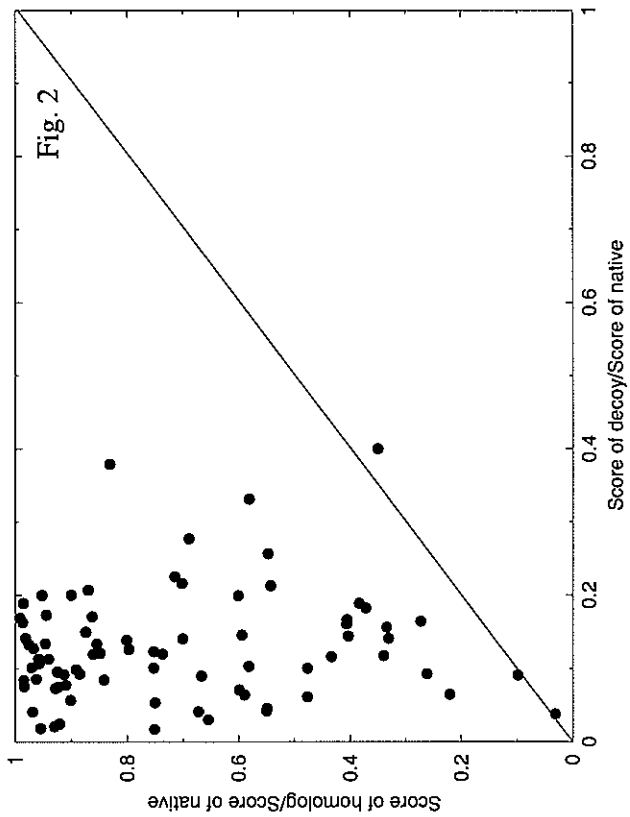

Figure 6.2 Cross threading test of homolog recognition. 86 protein sequences are chosen from 9 different families in the ASTRAL database. Each sequence is threaded on the structure of the other 85 proteins. The highest threading score obtained when a sequence is thread on protein structures in its own family is used to represent the homologous threading score E^{hom}. The decoy threading score E^{dec}, is the highest threading score obtained when the sequence is threaded on decoy structures (not in the same family). Homologous threading score E^{hom} is plotted against decoy threading score E^{dec} using the self-threading score E^{nat} as unit for each sequence. Points above the diagonal represent cases in which structural homologos are distinguished from decoys.

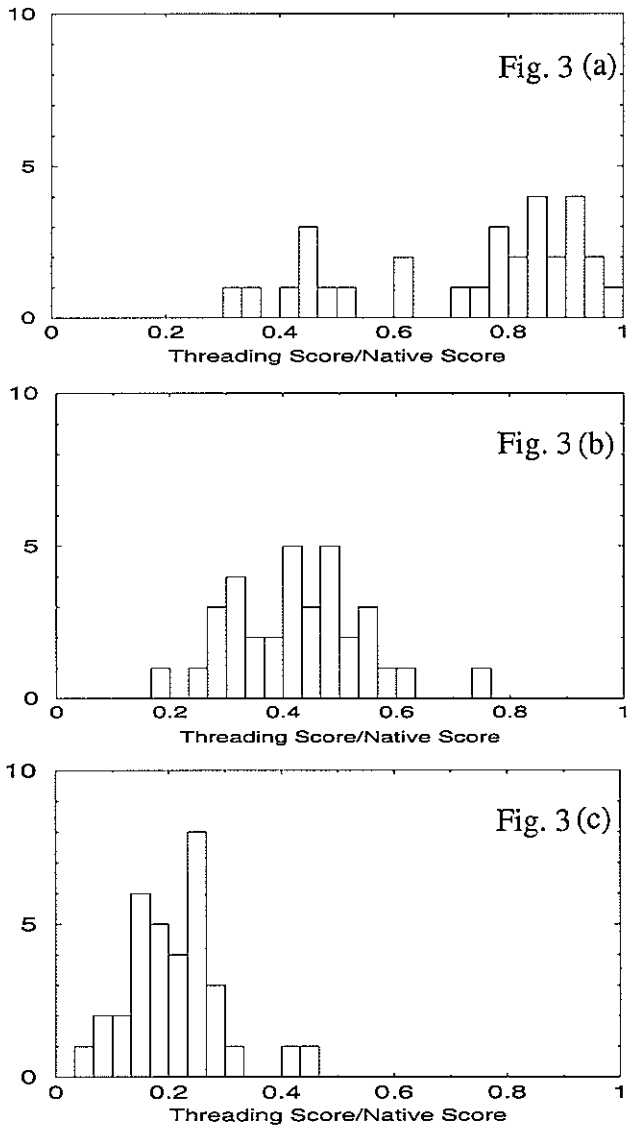

Figure 6.3 Cross threading test of remote homolog recognition. 34 protein sequences belonging to 3 different families are chosen from ASTRAL database. Histograms of E^{hom} (a) E^{remote} (b) and E^{dec} (c) normalized by self-threading score are shown. (a) E^{hom}: Each of the 34 protein sequences is threaded on protein structures in its own family. The highest threading score of each sequence is plotted in this histogram. (b) E^{remote}: Each of the 34 protein sequences is threaded on protein structures belonging to the same superfamily but different family (i.e. remote homologs). (c) E^{dec}: Each of the 34 protein sequences are threaded on structures randomly chosen from other superfamilies (decoys). In all histograms, the highest threading score for each sequence is plotted.

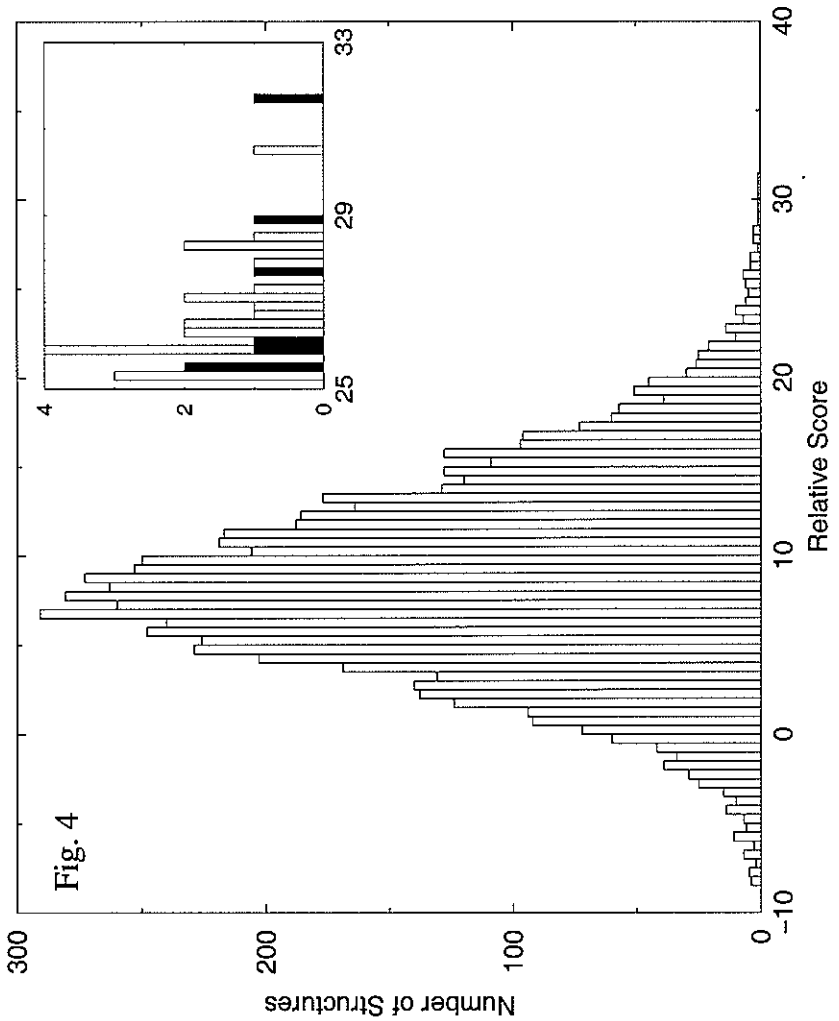

Figure 6.4 Distribution of E^{rel} scores for CASP5 targer T174_2. Segments of T174 sequence with length 120 (continuous) were threaded on representative ASTRAL database structures (see text). For each structures, the highest segment-structure alignment E^{rel} score is used to represent the threading score of that structure. Histogram of the threading energies of all the representative database structures is plotted. The high relative score tail of this histogram is enlarged in the inset. The dark bins in the inset belong to the structures from ASTRAL family d.14.1.5.

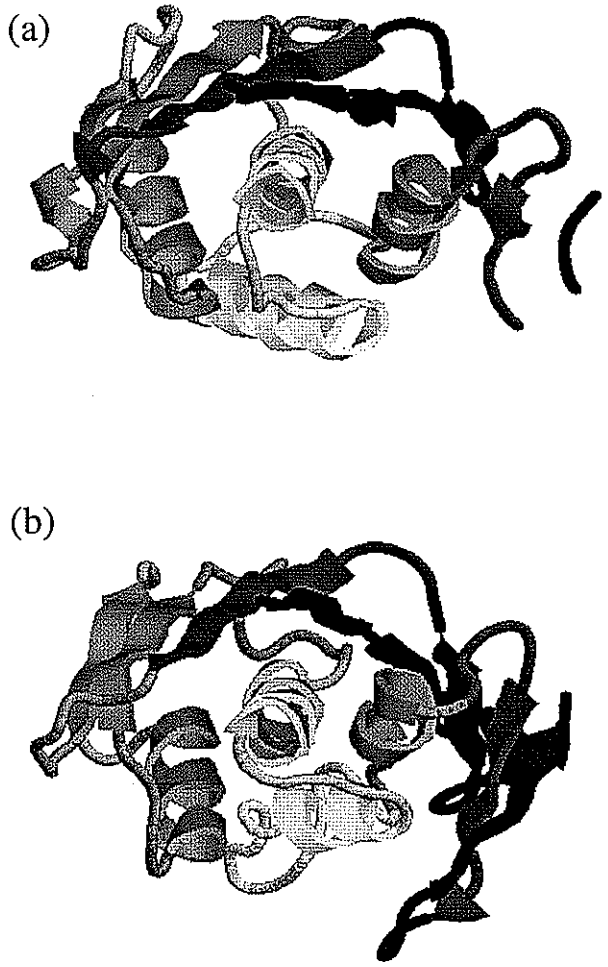

Fig. 5

Figure 6.5 Comparison of experimental and predicted structure of CASP5 target T174_2 domain. (a) T174_2 domain structure experimentally determined by J. G. Luz et al [48]. (b) T174_2 domain structure submitted to CASP5 by our group.

Fig. 6 (a)

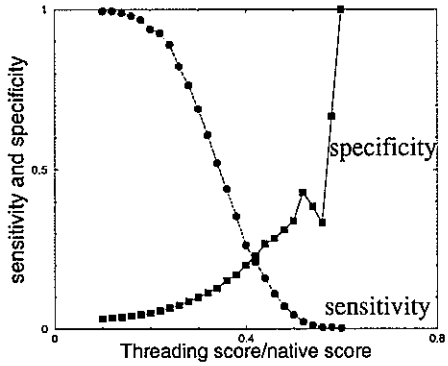

Fig. 6 (b)

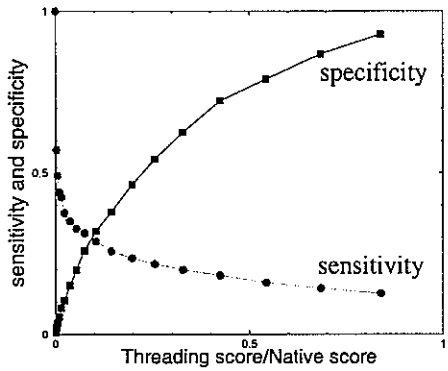

Fig. 6 (c)

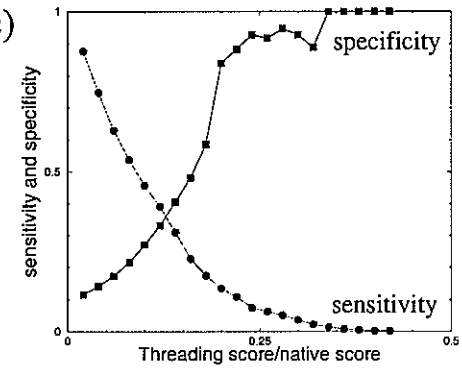

Figure 6.6 Sensitivity and specificity of threading method. performance of threading method was evaluated using 34 protein sequences belonging to 3 different supperfamilies thread on a representative ASTRAL database of structures (see text). (a) Sensitivity and specificity as a function of E^{rel} for superfamily a.1.1 (b) Sensitivity and specificity as a function of E^{rel} for superfamily b.1.1 (c) Sensitivity and specificity as a function of E^{rel} for superfamily c.2.1

CHAPTER 7. Conclusion

1. From the discussion above, we believed that the hydrophobic interaction dominate the protein folding process. With the help of local secondary structure information, the native structures of proteins show significant energy gaps compared with decoy conformations. Given the sequence information of a protein, it is possible to use this energy scheme to distinguish homologous structures of the protein from decoy structures.

2. The dominant eigenvector of a protein structure contact matrix shows strong correlation with the protein seqeunce vector. This correlation limits the choice of protein sequences in nature. The strength of this correlation can be used as an index for the fitness of a sequence to a structure. A strong seqeunce-eigenvector correlation was also found in the protein domains. Thus, it is possible to automatically partition a given multi-domain protein structure into separated domains useing this correlation.

3. High rank eigenvectors of the contact matrix are sequence-order blind. Their contribution should be removed from the calculation of fitness of a seqeunce to a structure. This explains why the commonly used Z-score can improve the prediction in many threading methods. We propose to use the relative energy as a replacement of the Z-score becasue it is more meaningful physically and faster in calculation.

4. A threading algorithm has been developed useing the correlation between seqeunce and dominant eigenvector. The results from this algorithms can be improved by an iterative method which is mathmetically straightforward. Our method can distinguish protein structure homologs from decoy structures without the requirement of high sequence similarity. This new method performed reasonably well in the CASP5 blind test in the fold-recognition and new-fold categories. Our method can be used cooperatively with most of exsiting approaches. When

working along, this method can provide information to biologist about the global configuration of a query protein even when there are no sequence similarity between the query protein and existing protein structures in the database. It can be combined with existing ab intio approaches as a good screening tool before clustering. It can also be used as input to meta servers to combine with results from other approaches.

CHAPTER 8. Appendix: A CASP5 Automatic Assessment By Michael Levitt

Abstract

In this appendix, I directly copied a CASP5 Automatic Assessment done by Michael Levitt. Because our method was designed for recognizing proteins without high sequence similarity to existing protein sequences, only the Fold recognition and New fold part are in listed. Our group name in CASP5 is Ho-Kai-Ming and group id number G437.

An Automatic Assessment of all CASP5 categories follows. More complete data including the perl script and data that produced these tables is available at http://csb.stanford.edu/levitt/

Automatic Assessment of CASP5 (by Michael Levitt 18 January 2003)

(1) All the assessment is based on the official GDT_TS scores calculated and released at the Asilomar meeting.

(2) The targets in each of the three categories are those selected by the three human assessors.

(3) Ranks are based on a total Z-Score. This total is the sum of Z-scores above 1.0 for all models of a particular group. Groups ID's are from the official score file. The names were added manually and those servers that were in CAFASP are indicated in bold.

(4) For Comparative Modeling and Fold Recognition the first model is the only one used. For New Folds, all targets are assessed. When more than a single target is assessed, the weights of the models of a particular group and particular target are weighted to have a total weight of 1.0 (i.e. the same as that when just one target is assessed).

(5) The Z-Score of a particular model for a particular target is calculated as: Z-Score = (GDT_TS - mean) / SD.

(6) The mean and SD of the model are calculated without assuming a score distribution.

The mean value is the value of GDT_TS with 52.5% of the data having a lower value. For a normal distribution, this is equivalent to eliminating the data worse than 2 standard deviations (in lowest 2.5%) and then taking the mean of the remaining data.

The SD value is calculated as the GDT_TS value that has 15.105% of the scores above it. For a normal distribution, this would just be the standard deviation.

Some General Observations

Comparative Modeling

The Polish group (Leszek Rychlewski) dominates this section with human entries in rank 1 (Bujnicki-Janus), 2 (Ginalski) and 3 (GeneSilico) and their automatic server in rank 4 (BIOINFO.PL). Honig in rank 4 is the best non-server related entry in the top nine ranks. The best real server (non-meta) is ORNL-PROSPECT (rank 6); no other non-meta server makes the top-40 list.

Fold Recognition

The Polish group also dominates this section with human entries in rank 1 (Ginalski) and their automatic server BIOINFO.PL in rank 4. The Skolnick and Baker entries are at ranks 2 and 3 and the Baker server is at rank 5 (BAKER-ROBETTA). Shortle did well at rank 5 and Brooks at rank 7.

The best real server (non-meta) is ARBY-SCAI in rank 16.

New Folds

This is the only CASP5 category not dominated by the BIOINFO.PL server and associated groups.

Surprise entries on the list are I-sites/Bystrof (rank 3) and Chimera (rank 8).

The meta-servers, BAKER-ROBETTA and PMODEL3 are tied in ranks 10 and 11.

The best non-meta server is SAMUDRALA-NF (which is essentially the same as PROTINFO-AB).

Ranking

Fold Recognition (targets manually assessed by Nick Grishin)

targets: 134 138 156 157 174 193_1 135 147 148 162_1 162_2 187_2 191_1

Use the first model for all the targets.

Total Z-Scores Above 1.0 for CASP5 All Targets Listed Above: Table 8.1

New Folds (targets manually assessed by Rob Russell)

Targets: 129 149_2 161 162_3 181 146_1 146_2 146_3 172_2 173 186_3 187_1 170

Use all the models weighted to have a total weight of 1.0

Total Z-Scores Above 1.0 for CASP5 All Targets Listed Above: Table 8.2

User Contributed Links:

MichaelLevitt: url : http://csb.stanford.edu/levitt/

MichaelLevitt: url : http://csb.stanford.edu/levitt/CASP5_AutoAssessor/

FR	Rank Group	Z-Score	Ngood	Npred	NgNW	NpNW	Group-name
FR	1 G453	24.26	9.00	12.00	9	12	Ginalski
FR	2 G010	21.64	7.00	12.00	7	12	Skolnick-Kolinski
FR	3 G002	19.55	8.00	12.50	9	14	Baker
FR	4 G006	16.88	6.00	10.00	6	10	BIOINFO.PL
FR	5 G349	15.25	7.00	7.00	7	7	Shortle
FR	6 G029	14.56	6.50	11.50	7	13	BAKER-ROBETTA
FR	7 G373	13.49	4.00	11.00	4	11	Brooks
FR	8 G437	11.34	3.00	6.00	3	6	Ho-Kai-Ming
FR	9 G068	10.45	3.00	5.50	3	6	Jones-NewFold
FR	10 G001	9.61	5.00	8.00	5	8	Sam-T02-human
FR	11 G067	9.19	4.00	9.00	4	9	Jones
FR	12 G427	9.04	5.00	10.00	5	10	Fischer
FR	13 G045	9.00	4.00	9.00	4	9	PMODEL3
FR	14 G028	8.43	5.00	6.00	6	7	Celltech
FR	15 G224	8.09	5.00	10.00	5	10	3D-SHOTGUN-3DS5
FR	16 G183	7.86	4.50	7.50	5	8	ARBY-SCAI
FR	17 G223	7.29	4.00	10.00	4	10	3D-SHOTGUN-3DS3
FR	18 G096	7.29	3.00	10.50	3	11	Bates-Paul
FR	19 G110	7.11	3.00	8.00	3	9	Honig
FR	20 G020	6.79	5.00	7.00	5	7	Bujnicki-Janusz
FR	21 G046	6.73	4.00	9.00	4	9	PCOMB
FR	22 G222	6.72	4.00	10.00	4	10	3D-SHOTGUN-INBGU
FR	23 G214	6.50	3.00	7.00	3	7	Advanced-Onizuka
FR	24 G012	6.33	5.00	11.00	5	11	ORNL-PROSPECT
FR	25 G517	6.25	4.00	11.00	4	11	GeneSilico
FR	26 G368	5.86	3.00	7.00	3	7	Jose
FR	27 G464	5.60	2.00	6.00	2	6	Atome
FR	28 G476	5.50	3.00	7.00	3	7	123d-server
FR	29 G423	5.28	2.00	8.00	2	8	Taylor
FR	30 G450	4.82	2.50	8.00	3	10	Tome
FR	31 G078	4.47	2.00	7.00	2	7	Rost
FR	32 G041	4.39	2.00	9.00	2	9	Cbrc
FR	33 G537	4.34	1.00	5.00	1	5	Nec-asogawa
FR	34 G435	4.30	3.00	7.00	3	7	Fujita
FR	35 G417	3.72	3.00	6.00	3	6	Cbsu
FR	36 G288	3.68	3.00	9.00	3	9	Lomize-Andrei
FR	37 G242	3.59	3.00	8.00	3	8	Genesilico.pl-servers-on1
FR	38 G265	3.54	2.00	9.00	2	9	Sasson-Iris
FR	39 G447	3.54	2.00	7.00	2	7	Cam-Biochem
FR	40 G039	3.54	1.00	6.00	1	6	PCONS3

Table 8.1 NgNW is the number of good predictions without weighting for multiple models.
NpNW is the number of total predictions without weighting for multiple models.
All-uppercase names are of CAFASP registered servers.

NF	Rank Group	Z-Score	Ngood	Npred	NgNW	NpNW	Group-name
NF	1 G002	25.72	9.33	12.50	47	63	Baker
NF	2 G349	17.57	8.25	10.00	14	17	Shortle
NF	3 G132	13.46	5.00	10.00	5	10	I-sites/Bystroff
NF	4 G010	11.78	5.69	11.51	29	59	Skolnick-Kolinski
NF	5 G001	11.31	5.53	11.33	20	38	Sam-T02-human
NF	6 G016	10.50	6.70	9.30	26	36	Levitt
NF	7 G068	10.39	5.40	7.00	27	35	Jones-NewFold
NF	8 G153	9.07	4.00	6.00	4	8	Chimera
NF	9 G450	7.64	4.03	9.60	13	29	Tome
NF	10 G029	7.59	4.50	11.80	28	58	BAKER-ROBETTA
NF	11 G045	7.59	4.75	10.15	15	41	PMODEL3
NF	12 G051	7.29	3.80	9.27	19	49	SAMUDRALA-NF
NF	13 G517	7.08	2.75	8.50	6	14	GeneSilico
NF	14 G140	7.04	3.60	8.30	18	44	PROTINFO-AB
NF	15 G040	6.94	4.93	11.67	22	49	PMODEL
NF	16 G453	6.81	3.50	10.50	4	11	Ginalski
NF	17 G373	6.57	3.95	11.30	16	46	Brooks
NF	18 G020	6.40	3.50	5.00	4	6	Bujnicki-Janusz
NF	19 G112	6.05	4.77	9.67	13	30	Friesner
NF	20 G437	5.94	3.92	9.17	9	18	Ho-Kai-Ming
NF	21 G475	5.84	3.00	4.00	3	4	Bionomix
NF	22 G170	5.74	4.00	9.00	4	9	Chimerax
NF	23 G006	5.39	3.00	11.00	3	12	BIOINFO.PL
NF	24 G531	5.17	2.80	5.60	14	24	Kias
NF	25 G067	5.02	3.00	7.50	3	8	Jones
NF	26 G099	4.83	2.00	3.67	6	10	Camacho-Carlos
NF	27 G224	4.45	3.00	8.00	3	8	3D-SHOTGUN-3DS5
NF	28 G314	4.31	2.40	4.00	12	20	Scheraga-Harold
NF	29 G427	4.18	2.75	8.00	4	15	Fischer
NF	30 G105	4.10	3.00	9.00	3	9	Sternberg
NF	31 G084	3.88	2.83	11.50	5	18	Sbc
NF	32 G429	3.68	2.20	3.50	11	19	Keasar
NF	33 G096	3.50	3.00	9.00	3	9	Bates-Paul
NF	34 G423	3.37	2.00	4.67	2	6	Taylor
NF	35 G214	3.34	2.50	8.50	5	15	Advanced-Onizuka
NF	36 G288	3.31	2.00	4.00	2	4	Lomize-Andrei
NF	37 G516	3.22	2.50	4.50	4	6	Burnham
NF	38 G464	3.18	2.20	8.03	9	36	Atome
NF	39 G265	3.10	2.00	8.00	2	8	Sasson-Iris
NF	40 G039	3.07	2.28	8.07	7	35	PCONS3

Table 8.2 NgNW is the number of good predictions without weighting for multiple models.
NpNW is the number of total predictions without weighting for multiple models.
All-uppercase names are of CAFASP registered servers.