Available online at www.sciencedirect.com
SCIENCE DIRECT®

Polymer 45 (2004) 687–697

polymer

www.elsevier.com/locate/polymer

Three-dimensional threading approach to protein structure recognition

Haibo Cao[a], Yungok Ihm[a], Cai-Zhuang Wang[b], James R. Morris[b], Mehmet Su[a], Drena Dobbs[c], Kai-Ming Ho[a,*]

[a]*Department of Physics and Astronomy, Iowa State University, Ames, IA 50011, USA*
[b]*Ames Laboratory-U.S. DOE., Iowa State University, Ames, IA 50011, USA*
[c]*Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA*

Received 9 July 2003; received in revised form 16 October 2003; accepted 16 October 2003

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 are 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.

Keywords: Protein structure; Threading; Alignment

1. Introduction

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 [2–4]. In the ASTRAL database [4], for example, more than 27,000 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 [6–10]. 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

[*] Corresponding author. Tel.: +1-515-294-1960; fax: +1-515-294-0689.
E-mail address: kmh@ameslab.gov (K.M. Ho).

0032-3861/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.091

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 [15–22,29,30,46,47]. 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 programming method [25–27] 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 [29,36,42,47]. Apart from this profiling approach, Jones et al. [16] used a double dynamic programming 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 neighbors 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.

2. Method

2.1. Energy functions

The interaction energy used in this paper follows the Li, Tang, Wingreen [38] parameterization 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 ith residue and jth residue are in contact, otherwise, the element is set to 0. We choose 6.5 Å 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

$$E = c_1(q_i + q_j) + c_2 q_i q_j + \text{constant} \tag{1}$$

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

$$E = c_2(q_i + a)(q_j + a) + K \tag{2}$$

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 Eq. (2) we can redefine modified q values as $q_i + a$, then Eq. (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 $\mathbf{s}$ 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

$$E = \sum_{i,j=1}^{n} q_i' C_{i,j} q_j' \tag{3}$$

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

2.2. Alignment

The problem of finding the best alignment of a query sequence $\mathbf{s}$ for a structure with contact matrix C is to find a transformation from $\mathbf{s}$ 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 occurring 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 $\mathbf{v}_0$

$$\frac{(\mathbf{s}' \cdot \mathbf{v}_0)^2}{(\mathbf{s}' \cdot \mathbf{s}')(\mathbf{v}_0 \cdot \mathbf{v}_0)} \tag{4}$$

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.

2.3. 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.

2.4. Gap penalty and size effects

For any method involving gapped alignment, the outcome is affected by the penalty for insertion/deletion.

Table 1
174 protein sequences used in self-recognition test

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						

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 Å 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.

2.5. 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: $f = N_+ - N_-/N_s$ 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: $E^{\text{modified}} = \alpha f E^{\text{threading}}$ where α is a parameter which can be optimized for accuracy of fold-recognition.

2.6. 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 $(E^{\text{raw}} - E^{\text{ave}})/\sigma$, 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 $E^{\text{rel}} = E^{\text{raw}} - E^{\text{ave}}$ 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).

3. 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 Å and a single peptide chain to avoid any possible inter-chain 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 Fig. 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.

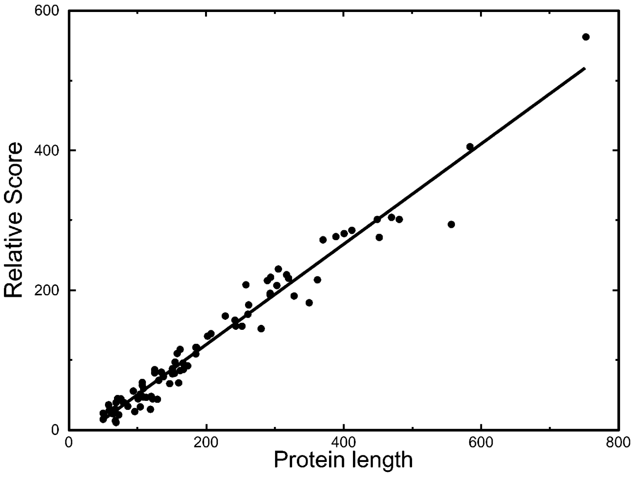

Fig. 1. Relationship between relative score E^{rel} and protein length. 174 randomly chosen proteins were self-threaded (see text). The relative score for each protein is plotted against the number of residues of that protein. A linear correlation between self-threading score and number of residues of the protein can be observed.

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 nine families from the ASTRAL database from which we selected 86 proteins

Table 2
86 proteins from nine families in ASTRAL database used in homolog-recognition test

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 3
Protein structures used in remote-homolog-recognition test

Domain family	Sequence in the family				Superfamily	Structures in the Superfamily			
a.1.1.2	1flp	1kfrA	2hbg	1a6m	a.1.1	1phnA	1cpcA	1i7yA	
	1eco	2gdm	1d8uA	1irdA		1gh0A	1allA	1b33A	
	1babB	1ch4A	1it2A	2lhb		1liaA	1b8dA	1qgwC	
	1ash	1ithA	1hlb	1vhbA		1kr7A			
	1ew6A								
b.1.1.1	1ah1	1ah1	1akjD	1eajA	b.1.1	1frtA1	1bmg	1a6zB1	1ld9A1
	1fo0A	1gya	1i85A	1neu		1b3jA1	1c16B1	1exuA1	1exuB1
	1qfoA					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	1e3eA2	c.2.1	1kvq	1fjhA	1bdb	1a4uA
	1e3jA2	1gpjA2	1kevA2	1qorA2		1gcoA	1h5qC	1i01A	1ae1A
	1ykfC2					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	

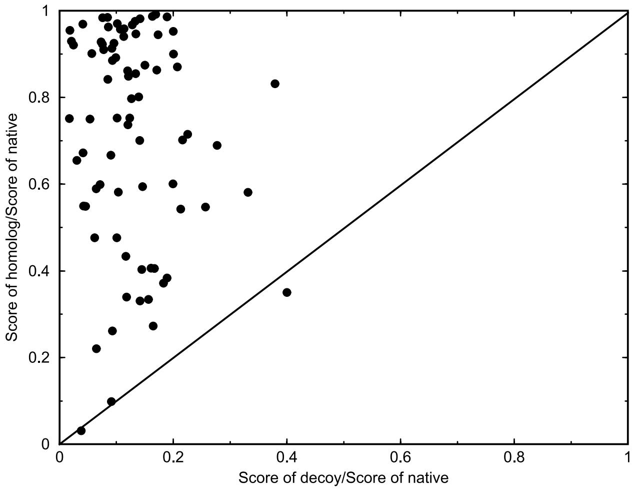

Fig. 2. Cross threading test of homolog recognition. 86 protein sequences are chosen from nine different families in the ASTRAL database. Each sequence is threaded on the structure of the other 85 proteins. The highest threading score obtained when a sequence is thread on protein structures in its own 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.

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. PSI-BLAST) 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 $E_i^{\mathrm{hom}}/E_i^{\mathrm{nat}}$ against $E_i^{\mathrm{dec}}/E_i^{\mathrm{nat}}$ is shown in Fig. 2. For 83 out of 86 cases, E^{hom} is clearly much higher than E^{dec}. For the remaining three 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 cannot 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 three superfamilies (a.1.1, b.1.1, c.2.1) from ASTRAL database. They belong to three 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 Fig. 3. Comparing Fig. 3(a) and (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 Fig. 3(b) with Fig. 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.

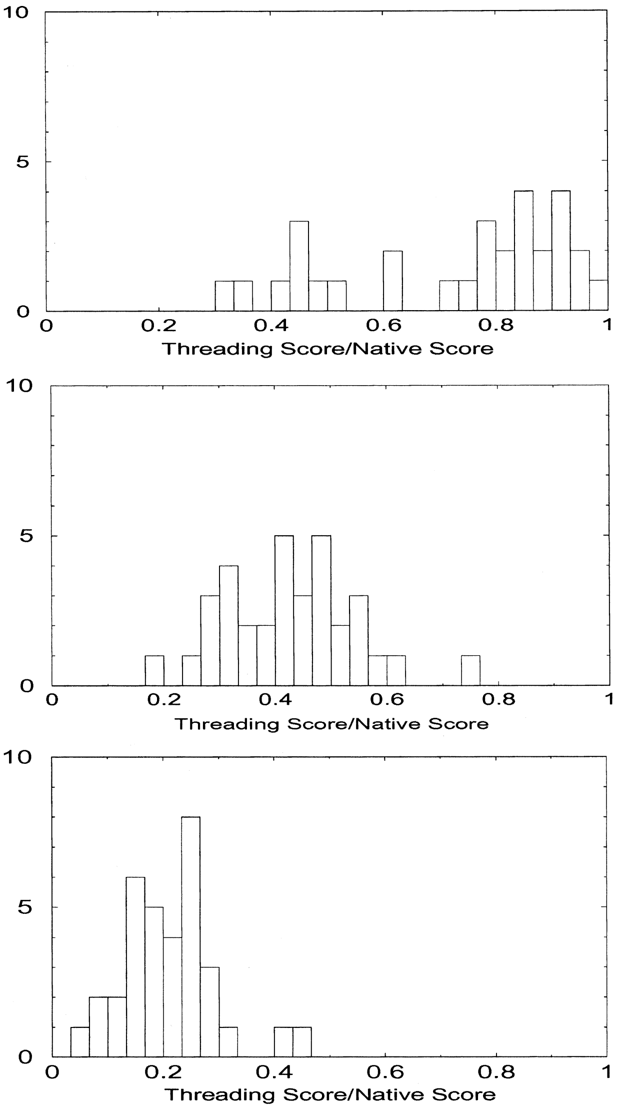

Fig. 3. Cross threading test of remote homolog recognition. 34 protein sequences belonging to three different families are chosen from ASTRAL database. Histograms of E^{hom} (a) E^{remote} (b) and E^{dec} (c) normalized by self-threading score are shown. (a) E^{hom} : each of the 34 protein sequences 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.

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

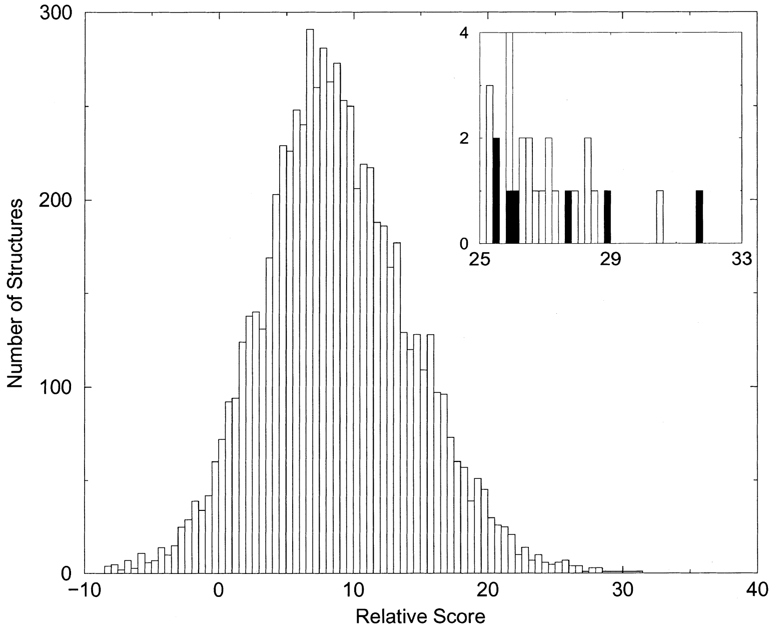

Fig. 4. Distribution of E^{rel} scores for CASP5 targer T174_2. Segments of T174 sequence with length 120 (continuous) were threaded on representative ASTRAL database structures (see text). For each structures, the highest segment-structure alignment E^{rel} score is used to represent the threading score 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.

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 five 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 Fig. 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 Fig. 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. Fig. 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 [45] 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 that do not belong to the same superfamily as a query protein's native structure, are treated as decoy structures for the query protein. We excluded decoy structures with significant 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

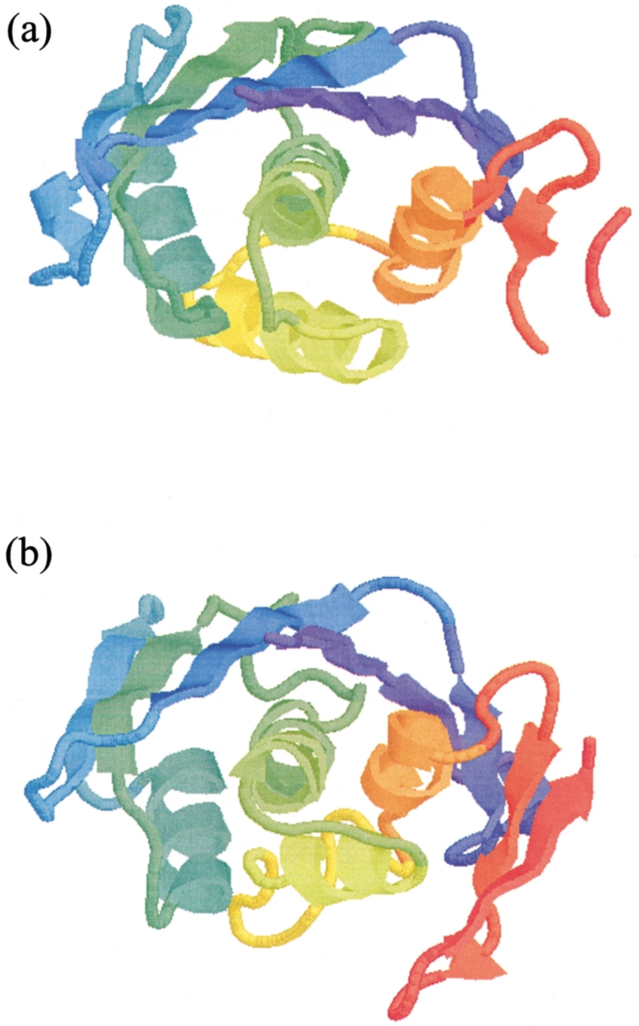

Fig. 5. Comparison of experimental and predicted structure of CASP5 target T174_2 domain. (a) T174_2 domain structure experimentally determined by J.G. Luz et al. [48]. (b) T174_2 domain structure submitted to CASP5 by our group.

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 = TP/TP + FN and specificity = 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 Fig. 6. According to Fig. 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.

4. 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

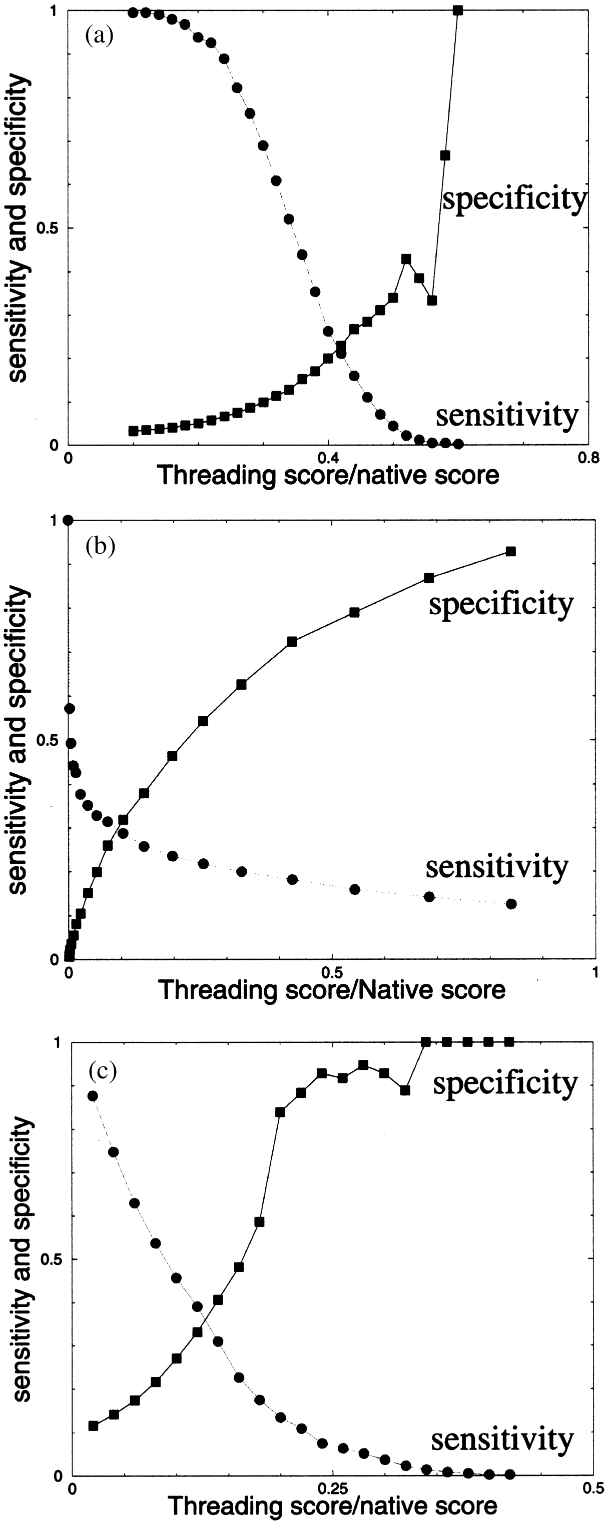

Fig. 6. Sensitivity and specificity of threading method. Performance of threading method was evaluated using 34 protein sequences belonging to three different superfamilies thread on a representative ASTRAL database of structures (see text). (a) Sensitivity and specificity as a function of E^{rel} for 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.

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.

Acknowledgements

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 A. 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

$$C\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{5}$$

Where index i goes from 1 to n.

For simplicity, we only consider matrix with non-degenerate 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, therefore

$$\mathbf{v}_i \cdot \mathbf{v}_j = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases} \tag{6}$$

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

$$\mathbf{s} = \sum_{i=1}^{n} \omega_i \mathbf{v}_i \tag{7}$$

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

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

$$C = \sum \lambda_i \mathbf{v}_i \times \mathbf{v}_i^{\mathrm{T}} \tag{8}$$

In structural threading, the score has the form

$$E = \mathbf{s} \cdot \mathbf{C} \cdot \mathbf{s} \tag{9}$$

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: $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. Because vector $\mathbf{s}$ is unitary, the overlaps satisfy the following equation

$$1 = \mathbf{s} \cdot \mathbf{s} = \sum_{i,j} \omega_i \omega_j \mathbf{v}_i \cdot \mathbf{v}_j \tag{10}$$

using Eq. (6), we get

$$\sum_i \omega_i^2 = 1 \tag{11}$$

Using Eqs. (7) and (8), we can decompose E (Eq. (9)) into contributions of different eigenvectors

$$E = \sum_i \lambda_i \omega_i^2 \tag{12}$$

Because eigenvalues are in decreasing order: $\lambda_1 < \lambda_2 < \ldots < \lambda_n$, 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 Eq. (7), we get

$$\mathbf{s} = 1 \cdot \mathbf{v}_1 + 0 \cdot \mathbf{v}_1 + \ldots + 0 \cdot \mathbf{v}_n = \mathbf{v}_1 \tag{13}$$

which means that the dominant eigenvector maximizes the score.

References

[1] Anfinsen C. Science 1973;181:223.

[2] Murzin AG, Brenner SE, Hubbard T, Chothia C. J Mol Biol, 1995; 247, 536–540.

[3] (a) Dengler U, Siddiqui AS, Barton G. Proteins 2001;42:332–44. (b) Siddiqui AS, Dengler U, Barton GJ. Bioinformatics 2001;17:200–1.

[4] (a) Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. Nucl Acids Res 2002;30:260–3. (b) Brenner SE, Koehl P, Levitt M. Nucl Acids Res 2000;28:254–6.

[5] Altschul SF, Thomas LM, Alejandro AS, Zhang J, Zhang Z, Webb M, David JL. Nucl Acids Res 1997;25:3389–402.

[6] Flory PJ. Principle of polymer chemistry. Ithaca, NY: Cornell University Press; 1953.

[7] Ptitsyn OB, Kron AK, Yu Y, Eizner YY. J Polym Sci C 1968;16:3509.

[8] de Gennes PG. J Phys Lett (Pairs) 1975;36:L55.

[9] Post CB, Zimm BH. Biopolymers 1979;18:1487.

[10] Sanchez IC. Macromolecules 1979;12:980.

[11] Chan H, Dill KA. J Chem Phys 1990;92(5):3118.

[12] Lau KF, Dill KA. Macromolecules 1989;22:3986.

[13] Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Proteins 1995; 21:167–95.

[14] Hendlich M, Lackner P, Weitckus S, Floechner H, Froschauer R, Gottsbachner K, Casari G, Sippl MJ. J Mol Biol 1990;216:167–80.

[15] Bowie JU, Luthy R, Eisenberg D. Science 1991;253:164–70.

[16] Jones DT, Taylor WR, Thornton JM. Nature 1992;358:86–9.

[17] Sippl MJ, Weitckus S. Proteins 1992;13:258–71.

[18] Godzik A, Kolinski A, Skolnick J. J Mol Biol 1992;227:227–38.

[19] Ouzounis C, Sander C, Scharf M, Schneider R. J Mol Biol 1993;232: 805–25.

[20] Bryant SH, Lawrence CE. Proteins 1993;16:92–112.

[21] Matsuo Y, Nishikawa K. Protein Sci 1994;3:2055–63.

[22] Mirny LA, Shakhovich EI. J Mol Biol 1998;283:507–26.
[23] Park BH, Levitt M. J Mol Biol 1996;258:367–92.
[24] Park BH, Huang ES, Levitt M. J Mol Biol 1997;266:831–46.
[25] Needleman SB, Wunsch CD. J Mol Biol 1970;48:443–53.
[26] Smith TF, Waterman MS. J Mol Biol 1981;147:195–7.
[27] Waterman MS, Eggert M. J Mol Biol 1987;197:723–8.
[28] Miyazawa S, Jernigan RL. Protein Engng 2000;13:459–75.
[29] Elofsson A, Fischer D, Rice DW, Le Grand S, Eisenberg D. Fold Design 1998;1:451–61.
[30] Lathrop RH, Smith TF. J Mol Biol 1996;255:641–65.
[31] Mirsky AE, Pauling L. Proc Natl Acad Sci USA 1936;22:439.
[32] Kauzmann W. Adv Protein Chem 1959;14:1.
[33] Miyazawa S, Jernigan RL. Macromolecules 1985;18:534–52.
[34] Miyazawa S, Jernigan RL. J Mol Biol 1996;256:623–44.
[35] Miyazawa S, Jernigan RL. Protein 1999;36:347–56.
[36] Jones DT. J Mol Biol 1999;287:797–815.
[37] Bryant SH, Altschul SF. Curr Opin Struct Biol 1995;5:236–44.
[38] Li H, Tang C, Wingreen NS. Phys Rev Lett 1997;79:765–8.
[39] Lau KF, Dill KA. Proc Natl Acad Sci USA 1990;87:638.
[40] Shortle D, Chan H, Dill KA. Protein Sci 1992;1:201.
[41] Press WH, Teukolsy SA, Vetterling WT, Flannery BP. Numerical recipes in fortran 77. New York: Cambridge University Press; 1999.
[42] Meller J, Elber R. Proteins 2001;45:241–61.
[43] Pierre B, Brunak S, Yves C, Claus AF. Bioinformatics 2000;5(16): 412–24.
[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. p. 300–18.
[46] CAFASP-1: critical assessment of fully automated structure prediction methods Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, Maccallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg MJ. Proteins: Struct Funct Genet 1999;(Suppl. 3):209–17.
[47] Rost B, Schneider N, Sander C. J Mol Biol 1997;270:471–80.
[48] Luz JG, Hassig CA, Pickle C, Godzik A, Meyer BJ, Wilson IA. Genes Dev 2003;17:977.