Pacific Symposium on Biocomputing 5:104-115 (2000)

ELIMINATING SUPERFLUOUS NEIGHBOR PAIRS WHILE THREADING FOLD MODELS

J.R. BIEŃKOWSKA, R.G. ROGERS JR., and T.F. SMITH
BioMolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street, Boston, MA 02215.

In this paper we address the problem of identifying which of various possible spatial residue-residue neighbor pairs are plausible physical contacts without reference to the native structure side chain geometry. We propose an algorithm that eliminates most of the implausible physical contacts from the fold models. This algorithm exploits the correlations between the amino acid side chain rotamers and the direction of the physical contacts between the amino acid side chains. We use this algorithm to "filter" the score of the sequence-to-structure alignment. Filtering is dynamic, in the sense that the set of neighbor pairs contributing to the alignment score varies during threading. Whether or not a neighbor pair contributes to the score depends on the threaded amino acids. This score filtering improves the accuracy of the predicted sequence-to-structure alignment.

1 Introduction

Results of the recent CASP3 structure prediction contest [a] indicate that sequence similarity between the query and the sequence of the fold model is still the key aspect of successful structure prediction. Threading methods are still unable to predict the structure of a protein with the same fold as a protein of known structure but that is unrelated in sequence. The threading method described here does not rely on sequence similarity and attempts to improve the structure prediction for those proteins. The core of this method is the incorporation in the sequence-to-model-structure alignment score of only those contributions from residue-residue neighbor pairs that are plausible physical contacts. This method does not use any information about the physical contacts that are present in the native model structure.

The basic idea of threading is extremely simple. It relies on the observation that many nonsimilar protein sequences adopt the same basic three-dimensional structure [1]. This is seen in the limited number of distinct folds present in the PDB [2,3,4,5]. The principal components of the threading approach are: (1) a fold library; (2) a description of the three-dimensional structure environments for each fold and an associated scoring function; and (3) a sequence-to-structure alignment or threading procedure.

Many threading methods [6,7,8,9,10,11,12,13] use a scoring function that depends on the residue-residue neighbor preferences in a given structural environment. These terms are intended to model interactions between residues

[a] http://PredictionCenter.llnl.gov/casp3/papers/murzin

that are far apart in sequence. The structural environments describing bare (no side chain atoms) backbone structure necessarily lack the details of the actual packing of the amino acid side chains that affect the long-range physical contacts between the residues. On the other hand, the scoring functions that explicitly use some measure of the positions of the side chain atoms (e.g. center-of-mass of the side chain [14]) are not appropriate, since the fold model should not contain any information about the position of the native side chain in order to avoid bias against dissimilar sequences.

Two basic approaches to including the spatial neighbor preferences in the scoring scheme have been used in threading. In the first approach (see for example [8,9]) the amino acid structural positions are identified as neighbors solely by their spatial proximity. The neighbor pair contributes to the alignment score whether or not the amino acids threaded onto those positions could make a physical contact in the native structure. In the method proposed by Taylor [10], the contribution of the neighbor pair thus identified is mitigated by a shielding factor. The shielding factor is an additional structural environment parameter that describes the packing around the amino acid positions. A neighbor pair always contributes to the alignment score. In the second approach, the so-called frozen approximation (see for example [13]), only the amino acid positions that were in physical contact in the native structure are consider as neighbors. Clearly, in such a scheme, many likely contacts among query side chains will never be taken into account.

Whether or not two spatially neighboring positions can make a physical contact depends on the backbone atoms positioned between them, the distance between their beta carbon atoms and the space available to accommodate their side chains, but also depends on the amino acids occupying those positions and the orientation of their side chains. We address this problem by distinguishing three classes of pairs of structural positions. The term "neighbor pair" denotes any pair of spatially neighboring positions as defined in section 2. The term "physical contact" refers to a neighbor pair occupied by amino acids whose side chain atoms are in physical contact. The term "noncontacting" neighbor refers to a neighbor pair occupied by amino acids whose side chain atoms are not in physical contact.

It has been observed recently [15] that scoring functions that reflect frequencies of spatial residue-residue neighbors are essentially random and that inclusion of those neighbor pair preferences does not improve threading structure predictions [16]. The problem of misrepresenting residue-residue neighbor preferences is most notable for the same-charge polar amino acid pairs [7], but is clearly present for other amino acid pairs. For example, two alanines will never make physical contact if their beta carbon (C_β) atoms are further than

5Å apart, even if they are frequently observed in positions separated by much greater distances. In comparison the alanine and phenylalanine side chains can actually make contact when their C_β atoms are 7.5Å away.

This simple observation suggests that besides having a good description of the structural environment one needs also to devise a method of distinguishing which neighbor pairs are plausible physical contacts when those positions are occupied by particular amino acids. One could attempt to construct the fold model from many similar structures and include in the scoring function contributions from all physical contacts or from a set of conserved physical contacts that are observed in those structures. However, for most folds only one or two representative structures are available [3]. Even for the folds currently most populated in the structural database there are no means of checking whether the fold representation is complete.

Here, we present a threading method that attempts to overcome these problems. We distinguish those amino acid structural neighbors that can make physical contacts from those that cannot and we exploit this information in a score filtering scheme. We have identified a partition of a multidimensional space of structural environment parameters that separates physical contacts from other neighbor pairs. Using this partition, we have developed a dynamic score filtering method for protein threading. During the search for the optimal sequence-to-structure alignment (threading), this filtering allows us to assign a score only to those amino acid pairs that, when placed in structural environments, are likely to make physical contacts.

In the sections below we describe: the local structure description, the scoring function and the structural environment states that determine the scoring function. We introduce the partition of the structural environment space of the neighbor pair into physical contacts and noncontacting neighbors and we describe the neighbor pair score filtering method that uses this partition. The last section compares the results of threading experiments performed using the "standard" neighbor pair scoring method, the new neighbor pair score filtering threading method, and threading with randomly filtered pair scores.

2 Local structure description

The three backbone atoms C, C_α and N, and the beta carbon (C_β) of any amino acid uniquely define a local reference frame centered at C_β (modeled C_β for GLY positions).

Local coordinates may be seen as corresponding to an idealized side chain rotamer tetrahedron [17,18], centered at C_β. Its orientation is fixed as shown in Figure 1. The observed side chain rotamers can be assigned a discrete value (2, 3 or 4) by the face through which the beta to gamma carbon (or oxygen in

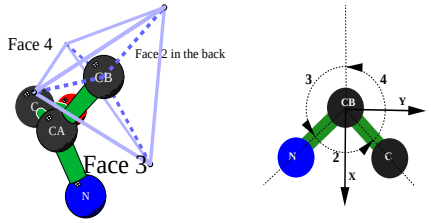

Figure 1: An example of the idealized tetrahedron constructed around a beta carbon with faces numbered 1, 2, 3 and 4, the $C_\alpha - C_\beta$ bond pierces face 1. Projection onto a local X, Y plane with respective space regions indexed by tetrahedral faces. Vectors $\overline{C_\alpha - C_\beta}$, $\overline{C_\alpha - N}$, $\overline{C_\alpha - C}$ define a local orthonormal frame $\hat{\mathbf{z}}_\mathbf{0} = \overline{C_\alpha - C_\beta}/\|\overline{C_\alpha - C_\beta}\|$, $\hat{\mathbf{y}}_\mathbf{0} = \hat{\mathbf{z}}_\mathbf{0} \times (\frac{\overline{C_\alpha - N}}{\|\overline{C_\alpha - N}\|} + \frac{\overline{C_\alpha - C}}{\|\overline{C_\alpha - C}\|})/\|\frac{\overline{C_\alpha - N}}{\|\overline{C_\alpha - N}\|} + \frac{\overline{C_\alpha - C}}{\|\overline{C_\alpha - C}\|}\|$ and $\hat{\mathbf{x}}_\mathbf{0} = \hat{\mathbf{y}}_\mathbf{0} \times \hat{\mathbf{z}}_\mathbf{0}$. Local spherical coordinates (r, θ, φ) are defined by: $x_0 = r\cos\theta\cos\varphi$, $y_0 = r\cos\theta\sin\varphi$ and $z_0 = r\sin\theta$.

the case of SER and THR) vector protrudes. These rotamer states correspond to the g^+, g^-, t rotamer states that are classified by the side chain dihedral angle χ_1 (definitions as in Dunbrack and Karplus [19]).

The secondary structure SS (as defined by DSSP [20]), the nonoccluded volume, the number and distances to other C_β and backbone atoms seen within each third of the solid angle provides a detailed local description of the native fold around each residue position i. This description is independent of the detailed placement of the side chains.

A given pair of residue positions i and j is also characterized by:

• A distance $D(i,j)$ - distance between the beta carbons.

• Line-of-sight neighbors - positions i, j are called line-of-sight neighbors when the vector $\overline{C_\beta(i) - C_\beta(j)}$ does not pass through any other atom (see Figure 2). Each line-of-sight neighbor pair is indexed by the tetrahedral faces $F(i,j)$ and $F(j,i)$. Local spherical coordinate $\varphi(i)$, $\varphi(j)$ of the vector $\overline{C_\beta(i) - C_\beta(j)}$ uniquely corresponds to a region of space around $C_\beta(i)$, $C_\beta(j)$ indexed by $F(i,j)$, $F(j,i)$.

• Neighbor pairs - Within the same strand only the positions $(i, i+2)$ are called neighbors. Within the same helix only the positions $(i, i+1)$, $(i, i+3)$ and $(i, i+4)$ are called neighbors. For the intersegment pairs the line-of-sight neighbors are called neighbors. We have chosen the cutoff distance $D(i,j)$ = 11.2Å, the threshold where the number of physical contacts starts to drop down significantly.

• Physical contact - a pair of residue positions occupied by amino acids whose side chains are in physical contact.

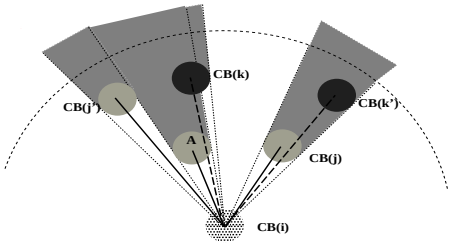

Figure 2: Two-dimensional representation of the line-of-sight and visible volume concept. The space in gray is considered "not visible" from C_β(i) position. Atoms A, C_β(j), and C_β(j') (light gray) are line-of-sight neighbors of C_β(i). Atoms C_β(k), C_β(k') (dark grey) are not line-of-sight neighbors of C_β(i). The limiting sphere within which the visible volume and atoms are calculated is illustrated as the outer dashed arc. We use atom radii: R(N)=1.01Å, R(C_α)=1.24Å, R(C)=1.13Å, R(O)=1.01Å, R(C_β)=1.24Å(scaled down by 25% from standard values). The visible volume is calculated as percentage of the maximum possible visible volume - the visible volume in the space occupied only by the backbone atoms of that residue.

Representative structures

We selected a set of 368 nonsimilar protein structures from the PDB [2]. These sequences were checked for similarity using BLAST-p [21] with the upper bound probability of 10^{-10} and for shared functional definitions. Some of these proteins are multidomain, thus this set represents 417 unique single-domain SCOP [3] structural superfamilies. The set was reduced by eliminating small folds with fewer than four secondary structure elements, membrane proteins, and designed proteins. We call these 368 proteins the scoring-function-training set. The complete list of the PDB four-letter locus name and a chain identifier is available by e-mail request to: jadwiga@darwin.bu.edu.

3 Scoring Function

In our approach to threading methodology we adopt the description of the sequence-to-structure alignment given by the Markov Random Field (MRF) [22]. A structural environment state is assigned to each position and to each neighbor pair. For each structural environment state, a probability distribution characterizes the amino acid or amino acids pair preferences for that state. The probability of observing a given sequence-to-structure alignment (assignment of an amino acid to each position) is equal to:

$$\begin{aligned} P(\mathbf{a}) = \quad & \tfrac{1}{N} \prod_{l \in \mathcal{L}} P_l(a_l | Env(l)) \prod_{i \in \mathcal{V}} P_i(a_i | Env(i)) \times \\ \times \quad & \prod_{\{i,j\} \in \mathcal{E}} \frac{P_{i,j}(a_i, a_j | Env(\{i,j\}))}{P^{(i)}_{i,j}(a_i | Env(\{i,j\})) \cdot P^{(j)}_{i,j}(a_j | Env(\{i,j\}))} \end{aligned} \tag{1}$$

Here $\mathbf{a} = (a_1 ... a_n)$ is the sequence of amino acids, where a_i denotes an amino acid at position i. $\mathcal{V}$ is the set of residue positions in helix or strand and $Env(i)$ is the i-th position structural environment. $\mathcal{E}$ is the set of neighbor pairs and $Env(\{i,j\})$ is the $\{i,j\}$th neighbor pair structural environment. a_l denotes an amino acid in a loop position l. $\mathcal{L}$ is a set of loop positions and $Env(l)$ is the loop environment. In equation 1, $P^{(i)}_{i,j}\big(a_i \big| Env(\{i,j\})\big)$ denotes the sum $\sum_{a_j} P_{i,j}\big(a_i, a_j \big| Env(\{i,j\})\big)$. $\mathcal{N}$ is the overall normalization constant of the MRF probability distribution. The $\log(P(\mathbf{a}))$ represents the scoring function.

The scoring function is represented as two score tables $S(a_i, Env(l))$ and $S(a_i, a_j, Env(k))$. For $l = 1 \ldots L$ singleton environments and $k = 1 \ldots K$ pairwise environments the score tables are:

$$S(a_i, Env(l)) = -\log P_i(a_i | Env(l)) \tag{2}$$

$$S(a_i, a_j, Env(k)) = -\log \frac{P_{i,j}(a_i, a_j | Env(k))}{P^{(i)}_{i,j}(a_i | Env(k)) \cdot P^{(j)}_{i,j}(a_j | Env(k))} \tag{3}$$

For the scoring-function-training set of proteins we record the amino acid and amino acid pair occurrences in each structural environment state. This gives us a set of conditional probability distributions that are then translated into the score tables.

4 Structural Environment States

The scoring function is defined in terms of the structural environment states. We use a new method to select structural environment states that maximizes the information content of the amino acid probability distribution that determines the scoring function[23]. This method rigorously selects various parameter thresholds that define structural environment states.

The structural environment state of a single position i is characterized by its secondary structure $SS(i)$ (alpha-helix or beta-strand) and its visible volume $VVE(i)$ within a sphere of 14Å radius. The visible volume is partitioned into ten discrete states, requiring nine visible volume threshold parameters $\mathbf{vv_{tr}}$=(38, 43, 47, 52, 56, 60, 67, 71, 79). The structural environment of the loop is characterized by the loop length and is represented by two states: loops shorter than six residues and loops at least six residues in length.

Each pair of neighboring residue positions i, j is characterized by:

• pair secondary structures $SS(i)$ and $SS(j)$. Based on distinct secondary structure geometries we identify six pairwise states of the residues in: same-strand, same-helix, same-sheet-different-strand, different-sheets, different-helices, sheet-helix.

• pair solvent exposure assignments $EXP(i)$, $EXP(j)$. Two exposure states of the fold position i are defined the visible volume $VVE(i)$ and the exposed-buried threshold value $\mathrm{vv}_{\mathrm{bur-exp}} = 60.0$. Thus there are four exposure states for a neighbor pair.

• the distance between C_βs $D(i,j)$. The simplest classification of neighbor pairs based on the C_β distance divides them into near, $D(i,j)$ less than d_{tr} and far, $D(i,j)$ greater than d_{tr}. Simple geometry suggests that small amino acids will preferentially interact with each other when their C_β atoms are closer, while the big amino acids will more likely make physical contact when their C_β's are further apart. We define two distance-dependent structural environment states: $DIST(i,j) = 1$(**far**) if $D(i,j) \geq d_{tr}$ and $DIST(i,j) = 2$(**near**) if $D(i,j) \leq d_{tr}$. Where $d_{tr} = 6.0$Å for same-strand, $d_{tr} = 5.6$Å for same-helix, $d_{tr} = 5.7$Å for same-sheet-different-strand, $d_{tr} = 4.5$Å for different-sheets, $d_{tr} = 5.6$Å for different-helices, $d_{tr} = 6.6$Å for sheet-helix.

• the visible volume $VV(F(i,j))(i)$ and $VV(F(j,i))(j)$ - amount of space visible through face $F(i,j)$ from position i (through face $F(j,i)$ from position j) within a sphere of 7.5Å radius. The visible volume vector contains considerable information about the rotamer state of an aromatic amino acid and, to a lesser extent, of other amino acids with big side chains [17]. The information contained in the visible-volume-dependent probability distribution of rotamer states suggests that the visible volume within the sphere of 7.5Å radius best predicts the rotamer state, consequently, the plausible physical contacts. We define two additional states associated with a pair of residue positions: state $VVP(i,j) = 1$(**large**) if $VV(F(i,j))(i) \geq VV_{tr}$ and $VV(F(j,i))(j) \geq VV_{tr}$, and otherwise state $VVP(i,j) = 2$(**small**). We found that the threshold values of VV_{tr}(strand) $= 69.7$ and VV_{tr}(helix)=74.3 provide best predictions of the discretized rotamer states (data not shown).

5 Identification of plausible physical contacts

In native structures, physical contacts can be identified using the minimal distance between side chain atoms or an energy-like variable describing the strength of interaction. However, the description of the fold model is purely geometrical and, as mentioned before, any *a posteriori* identification of the plausible physical contacts should rely only on the side-chain-independent variables. To determine the geometrical characteristics that distinguish the physical contacts from noncontacting neighbors we analyzed all neighbor pairs from the scoring-function-training set of proteins (our structural database).

For each pair of neighbors (i,j) characterized by secondary structure, $D(i,j)$ and $VV(F(i,j))(i)$, $VV(F(j,i))(j)$ variables we considered six additional variables:

• $VVE(i)$, $VVE(j)$ - the visible volume within a sphere of 14Å radius from positions i and j.
•$VCB(F(i,j))(i)$, $VCB(F(j,i))(j)$- number of C_β atoms visible through face $F(i,j)$ from position i (through face $F(j,i)$ from position j) within a sphere of 7.5Å radius.
• $VBB(F(i,j))(i)$, $VBB(F(j,i))(j)$ - number of backbone atoms visible through face $F(i,j)$ from position i (through face $F(j,i)$ from position j) within a sphere of 7.5Å radius.

Since our structural database is relatively small and the physical contacts between the amino acids depend on the side chain size, charge, etc., we pooled amino acids into seven classes according to the side chain size:1={ALA, PRO, and SER}; 2={PHE, HIS, TRP, and TYR}; 3={ASP, GLU, ASN, and GLN}; 4={LYS and ARG}; 5={CYS}; 6={ILE, LEU, and MET}; 7={THR and VAL}. The size of the side chain is an important factor in determining whether a physical contact between neighboring positions is possible. Thus 28 amino acid pair classes are identified. This amino acid pooling is used only for the partition of the neighbor environment space. We describe below the algorithm that partitions the neighbor environment space for each amino acid pair class.

A neighbor environment is a 9-dimensional vector $\vec{X}(i,j)$ of real or discrete variables. Let $\mathbf{\Omega}$ be the space of all possible neighbor environments. We map the set of all observed pairs of neighboring positions in the structural database onto $\mathbf{\Omega}$. All observed pairs are characterized as physical contacts or noncontacting neighbors.

The following procedure finds the analytical formula for the hyper-plane $\mathbf{\Sigma}$ that best partitions the neighbor environment space $\mathbf{\Omega}$ into: physical contacts, $\mathbf{\Omega_0}$, and noncontacting neighbors, $\mathbf{\Omega} - \mathbf{\Omega_0}$. Let us define the variable: $C(i,j) = 1$ for physical contact and $C(i,j) = 0$ otherwise. Let N be the number of pairs of neighboring positions in the database. Let $\mathbf{A}$ represent an N×9 matrix of all pairs of neighboring positions $\vec{X}_n(i,j), \quad n = 1 \ldots \mathrm{N}$ over our structural database. Let the N-vector $\mathbf{b}$ represent the values of the $C(i,j)$ for neighboring positions. In the space $\mathbf{\Omega}$ the vector $\mathbf{v}$ perpendicular to the hyper-plane $\mathbf{\Sigma}$ is given by the minimization of the following χ^2 function: $\chi^2 = |\mathbf{A} \cdot \mathbf{v} - \mathbf{b}|^2$. The value of $\mathbf{v}$ is obtained using any general linear least squares method [24]. Sliding the $\mathbf{\Sigma}$ along the $\mathbf{v}$, we select the position of $\mathbf{\Sigma}$ such that at least 90% (more if the number of physical contacts is small) of the structurally determined physical contacts are in the $\mathbf{\Omega_0}$ region.

If the pair of neighboring positions is mapped onto the $\mathbf{\Omega_0}$ region of the neighbor environment space, it is identified as a plausible physical contact, otherwise it is identified as an implausible physical contact. The selection

of the position of the hyper-plane is an essential part of the procedure. For example, the covariance matrix group classification method [24] typically leaves more than 30% of the physical contacts out of the plausible physical contact region Ω_0.

Scoring residue-residue neighbor pairs

Each pair of neighboring positions is characterized by a detailed side-chain-independent structural environment. For a given sequence-to-structure alignment, those positions are occupied by specific amino acids. Using the partition of the neighbor environment space, the threading algorithm checks whether the neighbor environment is classified as a plausible physical contact for those amino acids. If so, the neighbor pair is scored normally using the score for that amino acid pair in the corresponding structural environment state; if not, the neighbor pair does not contribute to the alignment score. The alignment score calculation does not depend on the identification method of the plausible physical contact region. We call this method of scoring neighbor pairs filtered neighbors threading (FNT).

6 Comparison of Performance of Threading Methods

We compared three threading scoring methods: the usual unfiltered neighbors threading (UNT), the FNT method and the randomly filtered neighbors threading (RFNT). In the RFNT, the amino acid pairs that contributed to the alignment score were selected randomly such that the number of contributing amino acid pairs that were selected for each neighbor pair was the same as were chosen via filtering by the FNT procedure.

We assessed the performance of the threading scoring method by comparing its sequence-to-structure alignments to the alignments reported by the Dali/FSSP database [5]. We evaluated the threading accuracy using measures defined for the CASP2 competition [25]: the *alignment sensitivity (ASns)* and the *alignment sensitivity* ± 4 *(ASn4)*. The *ASns* and *ASn4* were always calculated in the same manner, over the whole sequence-to-structure alignment. Each threading experiment was fully cross-validated by eliminating from the scoring-function-training set any member with similar sequence (BLAST-p [21] score not lower than 10^{-10}) and/or belonging to the same functional family as the threaded protein or the native protein of the fold model.

We tested the performance of the threading method using a set of 57 pairs of fold models and structurally homologous protein sequences. This set of fold models represents globular proteins selected previously for testing branch-and-bound algorithm and a variety of scoring functions [6]. Using the FSSP and SCOP databases, we selected structural homologues with lowest sequence similarity to the native sequence of the fold model. 52 out of 57 threadings

converged to the optimal alignment within the time limit that was set to eight hours per threading. The detailed results of these experiments are reported in Table 1. The UNT method had on average the alignment accuracy with *ASns*=15.8% and*ASn4*=42.6%. The FNT method had on average alignment accuracy with *ASns* = 27.6% and *ASn4* = 52.1%. The RFNT method had on average alignment accuracy with *ASns* = 22.9% and *ASn4* = 52.9%. RFNT gives worse results than FNT because RFNT eliminates the "wrong" amino acid pairs, but it does better than UNT because it is eliminating amino acid pairs between the same neighbors as FNT, resulting in a net reduction of noise. Comparison of both *ASns* and *ASn4* shows that on average the FNT method gives almost twice as accurate sequence-to-structure alignments as the usual unfiltered neighbors threading. Similar results are obtained using as the alignment accuracy measure *alignment specificity (ASpc)* [25] (data not shown). There is no correlation between the degree of sequence identity and the accuracy of the alignment. This result suggests that the FNT method may have captured the residue-residue interaction preferences relevant for structure recognition.

7 Conclusions

The construction of threading fold models and, consequently, threading potentials, requires careful assignments of neighbor pair preferences. In addition to the faithful geometric description of the 3D surroundings of the C_β atom position, one must include information about physical contacts made between the amino acids' side chains. The natural way is to include the rotamer preferences for the amino acid side chain and other geometrically described preferences that will allow the elimination of the implausible physical contacts between the residue positions.

We have proposed a method of eliminating the superfluous residue-residue neighbor contributions from the scoring function by identifying the stereochemical restrictions imposed on the neighbor pairs that are physical contacts. This elimination procedure is implemented automatically in the threading algorithm and does not imprint the fold model with the native sequence or the native physical contacts. Our results show that the explicit elimination of noncontacting neighbor pairs, which introduce noise to the scoring function, substantially improves the sequence-to-structure alignment accuracy.

References

1. J. Richardson. *Advan. Protein Chem.*, 34:167–339, 1981.
2. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. *J.*

Mol. Biol., 112:535–542, 1977. Brookhaven Protein Data Bank release 80.

3. A. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. *J. Mol Biol.*, 247:536–540, 1995.
4. C. A. Orengo, A. D. Michie, S Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. *Structure*, 5:1093–1108, 1997.
5. L. Holm and C. Sander. *Nucleic Acids Research*, 26:316–319, 1998.
6. R. H. Lathrop and T. F. Smith. *J. Mol. Biol.*, 255:641–665, 1996.
7. R. L. Jernigan and I. Bahar. *Current Opinion in Structural Biology*, 6:195–209, 1996.
8. S. H. Bryant and C. E. Lawrence. *Proteins: Structure, Function and Genetics*, 16:92–112, 1993.
9. S. Miyazawa and R. L. Jernigan. *J. Mol. Biol.*, 256:623–644, 1996.
10. W. R. Taylor. *J. Mol. Biol.*, 269:902–943, 1997.
11. M. J. Sippl. *J. Computer-Aided Mol. Design*, 7:473–501, 1993.
12. D. Jones, W. Taylor, and J. Thornton. *Nature*, 358:86–89, 1992.
13. A. Godzik, J. Skolnick, and A. Kolinski. *J. Mol. Biol.*, 227:227–238, 1992.
14. I. Bahar and R. L. Jernigan. *J. Mol. Biol.*, 266:195–214, 1997.
15. L. Zhang and J. Skolnick. *Protein Science*, 7:112–122, 1998.
16. L. Jaroszewski, L. Rychlewski, B. Zhang, and A. Godzik. *Protein Science*, 6:676–688, 1997.
17. T. F. Smith, L. Lo Conte, J. R. Bienkowska, R. G. Rogers Jr, C. Gaitatzes, and R. H. Lathrop. In *RECOMB97 Proceedings of the First Annual International Conference on Computational Molecular Biology*, Santa Fe, New Mexico USA, January 1997. ACM Press.
18. L. Lo Conte and T. F. Smith. *J. Mol. Biol.*, 273(1):338–348, 1997.
19. R. L. Dunbrack Jr. and M. Karplus. *J. Mol. Biol.*, 230:543–574, 1993.
20. W. Kabsch and C. Sander. *Biopolymers*, 22:2577–2637, 1983.
21. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. *J. Mol. Biol.*, 215:403–410, 1990.
22. J. V. White, I Muchnik, and T. F. Smith. *Mathematical Biosciences*, 124:149–179, 1994.
23. J. R. Bieńkowska, R. G. Rogers Jr, and T. F. Smith. In *RECOMB99 Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 25–32, Lyon, France, April 1999. ACM Press.
24. A. C. Rencher. John Wiley & Sons, Inc., 1995.
25. A. Marchler-Bauer and S. H. Bryant. *Proteins: Structure, Function and Genetics*, Suppl. 1:74–82, 1997.

FOLD			Threading method					
			unfiltered		randomly filtered		filtered	
structure pdb code	sequence pdb code	% sequence identity	*ASns*	*ASn4*	*ASns*	*ASn4*	*ASns*	*ASn4*
1alc	153l	11	0.0	0.0	0.0	0.0	0.0	0.0
2lzm	153l	12	0.0	19.6	0.0	19.6	0.0	19.6
1bgc	1alu	16	73.1	100.0	73.1	100.0	50.5	100.0
2baa	1am7A	14	0.0	37.7	0.0	37.7	22.6	22.6
9rnt	1aqzA	26	0.0	8.3	0.0	0.0	0.0	0.0
1rec	1auiB	26	14.9	45.9	0.0	41.9	33.8	43.2
2sns	1bcpD	2	0.0	24.1	13.8	41.4	13.8	37.9
2cyp	1bgp	21	0.0	0.0	35.7	49.1	22.3	35.7
1lis	1br0	6	0.0	0.0	0.0	0.0	0.0	0.0
2had	1brt	17	41.5	71.7	40.6	77.4	30.2	58.5
351c	1c52	18	0.0	0.0	17.6	32.4	17.6	17.6
5cpv	1cll	35	62.0	88.0	22.0	72.0	62.0	88.0
1rcb	1cnt3	12	74.3	100.0	33.8	33.8	51.4	74.3
1dhr	1cydA	19	19.5	50.4	19.5	71.7	25.7	64.6
5cytR	1cyj	19	71.4	71.4	71.4	71.4	71.4	71.4
1pkp†	1dar	15	TIME-OUT		0.0	94.7	31.6	94.7
4tgl	1din	14	0.0	32.8	19.7	68.9	23.0	42.6
1aba	1erv	14	0.0	68.4	7.9	50.0	7.9	50.0
5tmnE†	1ezm	31	0.0	3.9	3.9	5.5	3.9	33.1
1cde	1fmtA	16	14.3	38.8	11.2	59.2	20.4	38.8
3adk	1gky	16	33.8	62.0	43.7	85.9	52.1	70.4
1f3g	1hcz	18	TIME-OUT		0.0	0.0	0.0	0.0
1mbd	1ithA	15	13.5	30.3	13.5	30.3	13.5	30.3
2ca2	1kopA	34	0.0	36.8	50.0	76.5	64.7	76.5
1byh	1led	11	0.0	16.1	3.4	16.1	0.0	12.6
1ifc	1eal	21	17.8	89.0	61.6	100.0	56.2	100.0
1ubq	1lxdA	11	32.3	80.6	19.4	100.0	32.3	80.6
1cewI	1molA	20	0.0	25.6	0.0	17.9	33.3	33.3
1apa	1mrj	28	11.6	41.9	31.8	67.4	40.3	65.1
2end	1mtyG	4	0.0	23.1	0.0	23.1	0.0	23.1
2mhr	1nfn	8	0.0	0.0	0.0	0.0	0.0	0.0
7rsa	1onc	27	0.0	40.7	79.6	100.0	0.0	85.2
1atu	1ovaA	30	31.8	47.0	27.3	62.1	42.4	42.4
2hpr	1pfh	35	70.8	100.0	72.9	91.7	50.0	100.0
1bp2	1poc	27	0.0	0.0	0.0	0.0	0.0	0.0
5nll	1rcf	23	0.0	23.4	45.3	95.3	70.3	95.3
1yat	1rot	27	31.9	83.0	31.9	91.5	61.7	100.0
3chy	1srrA	26	13.7	100.0	69.9	100.0	47.9	100.0
3est	1svpA	12	0.0	31.1	0.0	57.8	0.0	31.1
2act	1theA	28	0.0	8.7	27.5	48.7	36.2	57.5
2mcm	1tvdB	9	9.7	22.6	0.0	41.9	0.0	41.9
8dfr	1vdrA	24	29.9	68.7	29.9	44.8	40.3	61.2
1hoe	1wkt	6	8.3	41.7	0.0	16.7	0.0	83.3
5fd1	1xer	31	0.0	0.0	0.0	0.0	0.0	0.0
1lec	2ayh	11	0.0	38.2	0.0	47.2	13.5	39.3
256bA	2ccyA	17	28.6	77.9	0.0	19.5	28.6	58.4
1tie	2i1b	11	0.0	10.0	0.0	24.0	0.0	28.0
4fgf	2i1b	14	0.0	21.7	0.0	45.7	0.0	8.7
2cpl	2nul	30	0.0	20.0	35.0	35.0	35.0	35.0
1s01	2pkc	38	TIME-OUT		16.2	86.5	51.4	91.0
2aak	2uce	33	0.0	43.7	0.0	43.7	0.0	20.8
1plc	7paz	25	0.0	26.7	0.0	0.0	0.0	0.0
Average			15.8	42.6	22.9	52.9	27.6	52.1

Table 1: Comparison of three threading methods: the UNT, the RFNT and the FNT method. "TIME-OUT" indicates threadings that did not converge within the time limit. The † indicates structures defined as multidomain by SCOP.