

Comprehensive statistical method for protein fold recognition

Jadwiga R. Bienkowska Lihua Yu Sophia Zarakovich Robert G. Rogers Jr.
Temple F. Smith

BioMolecular Engineering Research Center, College of Engineering,
Boston University, 36 Cummington Street, Boston, MA 02215, USA.

E-mail: jadwiga@darwin.bu.edu

Abstract

We present a protein fold recognition method that uses a comprehensive statistical interpretation of structural Hidden Markov Models (HMMs). The structure/fold recognition is done by summing the probabilities of all sequence-to-structure alignments. Conventionally, Boltzmann statistics dictate that the optimal alignment can give an estimate of the lowest free energy of the sequence conformation imposed by the structural model. The alignment is optimized for a scoring function that is interpreted as a free energy of an amino acid in a structural environment. Near-optimal alignments are ignored, regardless of how likely they might be compared to the optimal alignment. Here we investigate an alternative view. A structure model can be seen as a statistical representation of an ensemble of similar structures. The optimal alignment is always the most probable, but sub-optimal alignments may have comparable probabilities. These sub-optimal alignments can be interpreted as optimal alignments to the “other” structures from the ensemble or optimal alignments under minor fluctuations in the scoring function. Summing probabilities for all alignments gives an estimate of sequence-model compatibility. We have built a set of structural HMMs for 188 protein structures, and have compared two methods for identifying the structure compatible with a sequence: by the optimal alignment probability and by the total probability. Fold recognition by total probability was 40% more accurate than fold recognition by the optimal alignment probability.

1 Introduction

Protein fold recognition methods quickly evolve into viable tools that help to deduce the protein structure and function [13]. The ultimate goal of a fold recognition method is to predict the protein structure by identifying the correct fold (structural template) among already-solved protein structures or models and aligning the protein sequence correctly onto the structural model.

Most fold recognition methods use Boltzmann statistics to interpret probabilistic scoring functions [16, 3, 4, 18, 5, 11, 22, 19, 23, 21]. A sequence-to-structure alignment is evalu-

ated by a scoring function, and the score of the alignment is interpreted as the “free energy” of the sequence in the conformation imposed by the alignment. This interpretation dictates that the most probable sequence-to-structure alignment is the one with the lowest “free energy”. Thus determination of the compatibility of the structure model with a given sequence is based on the value of the probability of observing the sequence given the structure model and the optimal alignment $P(seq|Model, optimal - alignment)$. Here, we investigate a fold recognition method that evaluates the sequence-model compatibility using the sum of probabilities of all sequence-to-structure alignments. The mathematical justification of such method was discussed previously for HMMs [26] and in general terms of a threading approach to protein structure prediction [8, 7].

The usual approach to the fold recognition problem ignores the well-known fact that the optimal sequence-to-structure alignment is rarely the correct one [16, 10, 17]. This is true for predicting the alignment of the sequence to its native structure model as well as for predicting sequence-to-structure alignment for homologous and structurally similar proteins. Usually the correct/native alignment is suboptimal but nevertheless has a probability comparable to the probability of the optimal alignment. This situation reflects two facts about the mathematical models of protein structure and structure prediction by threading. First, scoring functions that are used to evaluate sequence-to-structure alignments are statistical approximations of the “true” scoring function or “free energy”. In consequence, the set of sub-optimal alignments should be seen as a set of optimal alignments under expected minor fluctuations in the scoring function. Second, a structure model should be seen as a statistical representation of an ensemble of similar structures or expected variations about a unique fold topology. A set of sub-optimal alignments can be interpreted as optimal alignments to structural variants of the same fold. According to the theory of Hidden Markov Models (HMMs), summing probabilities for all sequence-to-model alignments gives the rigorous probability of observing the sequence given the structure model $P(seq|Model)$ [26].

In our approach, the structure is modeled as a Discrete State Model (DSM) [26, 25], that is mathematically represented as an HMM. This is a linear representation of the 3D protein structure that is essentially equivalent to a set of structural profiles [2]. The distinction between a DSM and an HMM is that the state transitions are restricted to a minimal set and transition probabilities are designed to have a minimal bias. Other Hidden Markov Models for protein structure prediction or fold recognition were proposed

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee
RECOMB 2000 Tokyo Japan USA

Copyright ACM 2000 1-58113-186-0/00/04 \$5.00

recently [14]. Those HMMs are constructed from a generic HMM module. Subsequently, the generic HMM is trained, using a set of protein sequences that adopt similar 3D structures, to represent a unique structural fold. The HMMs that we propose here are not trained but are built directly from the 3D protein structures deposited in the PDB [1]. Models are built using an automatic analysis of protein structures. For each representative of a unique structural superfamily, as defined in the SCOP database [12], we build a unique Hidden Markov Model. Hidden states of the HMM represent states of structural positions. These states encode the secondary structure and the level of solvent exposure of a structural position. Each hidden (structural) state in the model is characterized by the amino acid preferences for this state. Hence, each HMM can be interpreted as a set of structural profiles.

The advantage of the HMM representation over the structural profile representation is the simple encoding of the structural variations observed among structures with the same fold. This variations are usually the variable length of the secondary structure elements and alternative loop types (tight turn, turn or coil) or loops with variable lengths connecting the secondary structure elements.

Since structure models that are derived from determined protein structures are not independent, a large fold model library requires a method that systematically addresses the problem of hierarchical classification of protein structures (structure models). Thus when calculating the posterior probabilities, $P(\text{Model}|\text{seq})$, which involves the normalization over all models from the library, one needs to account for the similarities among models at each level of the hierarchy. Here we employ the SCOP structural hierarchy. For example, for a fold represented by two superfamilies each populated by four structural families, the prior probability for each family model would be $P(\text{Model}) = \frac{1}{2} \times \frac{1}{4}$. The posterior probability of observing a particular structure model given the sequence is defined according to Bayes rule $P(\text{Model}|\text{seq}) = P(\text{Model}) \times P(\text{seq}|\text{Model})/P(\text{seq})$. In fold recognition methods, the posterior normalization of the structure model probabilities avoids overestimating the probability of observing a structural fold that is represented by many structure models when compared to the probability of the fold that is represented by only one structure model.

In a set of experiments, we compared the performance of two fold recognition methods. The first method identifies the best structure model for a sequence using the probability of the optimal sequence-to-model alignment. The second method identifies the best structure model for a sequence using the total sequence-to-model alignment probability. Our results demonstrate that the total probability method predicts the structure model compatible with a sequence 40% more accurately than the optimal alignment probability method. For both methods we used the hierarchical posterior normalization of structure model probabilities.

2 DSM Structure Models

We constructed our DSM library by selecting 188 protein structures from the SCOP database [12]. Those proteins have less than 40% sequence identity between any pair. From the original set of proteins classified in SCOP (release pdb40d.1.38), we eliminated structures identified as irregular, engineered or membrane proteins. We additionally restricted proteins to one representative per SCOP structural superfamily and required that each protein be a single struc-

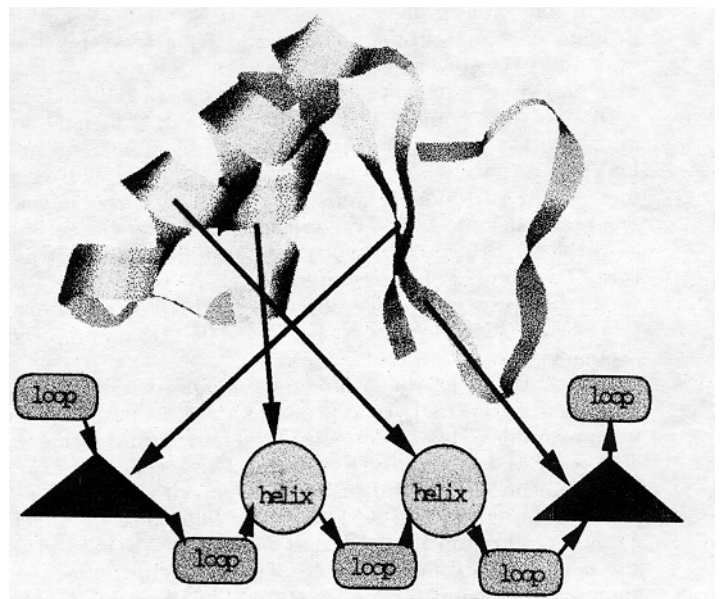


Figure 1: Encoding of a structure from the PDB into DSM modules. Each structural element is represented by a building module: strand, helix or loop. The internal parameters of each module, like the number and type of hidden states and transition matrix probabilities, are derived from the structural information as described in text.

tural domain. The PDB identifiers for all the structures are given in Table 2.

Each DSM is represented by three matrices: Φ , \mathbf{H} and \mathbf{x}_1 . The transition matrix Φ holds the conditional probabilities $\phi(s|s')$ of passing from any structural state s' to any other state s . The matrix \mathbf{H} holds the conditional probabilities $h(a|s)$ of emitting an amino acid a by a structural state s . The initial state-distribution matrix \mathbf{x}_1 is a vector holding the probabilities $x_1(s)$ that the Markov chain starts in any state s at the beginning of the sequence.

Our structure models comprise positions that have the secondary structure (SS) assigned by DSSP [6] as helix or strand. One-residue kinks in helices are smoothed over and assigned helix secondary structure. The distance between the end positions of consecutive secondary structure elements is recorded and used to determine if a tight turn or beta turn loops are geometrically possible connections between consecutive elements. Structural positions are constructed from the backbone atoms and the beta carbon (C_β) or modeled C_β for the positions occupied by glycine in the native structure. Each structural position environment is described by its secondary structure and Eisenberg-like solvent exposure of the position [2]. Solvent exposure is calculated for the poly-alanine chain and is independent of amino acids present in the native structure. Using the solvent exposure value we define three solvent exposure states: buried, partially buried and exposed. Thus we have six types of structural states for positions in helix or strand. The possible loop states are: tight turn (two residue loops), beta turn (four residue loops) and coil, loops longer than four residues.

$$\sum_{Model \in Library} P(Model|seq) = 1 \quad (2)$$

The prior probabilities $P(Model)$ for each structure model in the library are assigned following a structural classification hierarchy. In our library we have models belonging to one of four structural classes: α , β , α/β or $\alpha + \beta$. Each class is assigned a prior probability of 0.25. Each class is represented by a number of SCOP structural folds = $\#folds/class$. Each fold is represented by a number of SCOP structural superfamilies = $\#superfamilies/fold$. Each superfamily is represented by a number of SCOP structural families = $\#families/superfamilies$. Each family is represented by a number of Discrete Space Models belonging to the family = $\#Models/family$. Our library contains multiple models that were constructed from the same PDB entry with differences between DSMs resulting from alternative solvent exposure estimates, as was described in section 2. Thus even if there is only one PDB entry representing a SCOP superfamily, a structural superfamily may be represented by more than one DSM. The prior probability of a DSM classified by its class, fold, superfamily and family is:

$$P(Model) = 0.25 \times \frac{1}{\#folds/class} \times \frac{1}{\#superfamilies/fold} \times \frac{1}{\#families/superfamily} \times \frac{1}{\#Models/family} \quad (3)$$

Using the hierarchically assigned model priors we can rigorously answer the following question: What is the probability of observing a unique structural superfamily given a sequence? The posterior probability of observing a unique structural superfamily given a sequence is:

$$P(superfamily|seq) = \sum_{Model \in family \in superfamily} P(Model|seq) \quad (4)$$

Analogous equations apply for the posterior probabilities calculated for any level of the structural hierarchy: class, fold, superfamily or family.

We use a simple binary decision approach: either one model is preferred in comparison to all others or not. Thus as fold predictions we accept only the top folds/models with a posterior probability greater than 0.5.

4 Fold Recognition Experiments

We compared two alternative methods of fold recognition. In the first, the Viterbi method, we used the value of the optimal sequence-to-model alignment probability as calculated by the Viterbi algorithm. In the second, the Filtering method, used the value of the total probability as calculated by the Filtering algorithm. For both methods we used the hierarchical prior model probabilities to calculate the normalized posterior probability values and to make a fold prediction.

In the first set of experiments we calculated $P(family|seq)$, the posterior probability for each structural family, for 188 native sequences on 350 DSMs from our library. The results of the fold prediction results for the Viterbi and for the Filtering method are presented in Table 2. The posterior probabilities of observing a structural family given the sequence were normalized according to equations 1, 2 and 3. With the Filtering fold recognition method, 152 out of 188 sequences

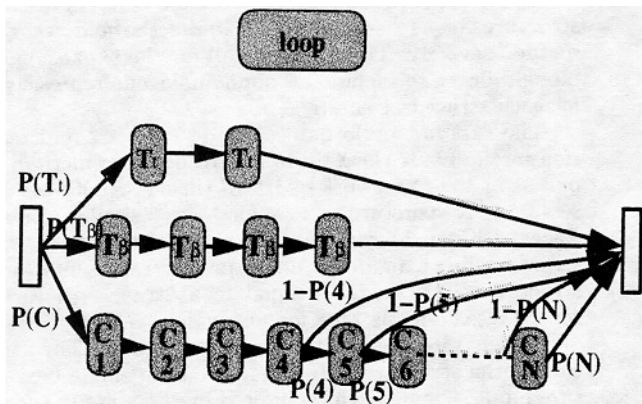


Figure 3: A generic DSM loop module. Three types of loops connecting the secondary structure elements are possible: tight turn (T_t), beta turn (T_β) and coil or irregular loop (C). Arrows connecting states represent the nonzero transition matrix elements and numbers assigned to each arrow represent the transition probabilities. The arrows with no numbers associated with them have a transition probability equal to one. The transition probabilities $P(T_t)$ - the probability of a tight turn, $P(T_\beta)$ - probability of a beta turn or $P(C)$ - probability of coil are determined from the geometry of consecutive SS elements. If all three loops are possible $P(T_t) = P(T_\beta) = P(C) = 1/3$. If tight turn is not allowed $P(T_t) = 0$, $P(T_\beta) = P(C) = 1/2$ and $P(T_t) = 0$. The transition probabilities $P(N)$, $1-P(N)$ for a coil loop states are set to represent a uniform loop length distribution between minimal loop length L_{min} and maximal loop length L_{max} . The loops can be extended beyond the maximum length L_{max} with the exponentially decreasing probability. When only a coil loop is allowed, the minimal loop length is determined by the distance between the ends of consecutive SS elements, otherwise it is set to four when additionally a beta turn is allowed, and it is set to two when a tight turn is also allowed. When the native structure loop is shorter than ten residues, L_{max} is set to ten. Otherwise L_{max} is equal to the length of the native structure loop. The N- and C-terminal loop modules have the $P(T_t) = P(T_\beta) = 0$. The N- and C-terminal loop module can have zero length, begin as a coil or begin as an amphipathic alpha helix at least 5 residues long, each with probability 1/3. The loop module does not contain any information about the solvent exposure of loop positions.

ranked the native structural family with the highest probability. With the Viterbi fold recognition method, 110 out of 188 sequences ranked the native structural family with the highest probability. The “acceptable” predictions with the top structural family having the probability of at least 0.5 had results as follows. For the Filtering fold recognition method, there were 161 predictions and 145 were correct - a 90% success rate. For the Viterbi fold recognition method, there were 168 predictions and 107 were correct - a 64% success rate. These results demonstrate that the Filtering fold recognition method is 40% more accurate than the Viterbi fold recognition method.

It may seem surprising that the native model recognition by Filtering has an accuracy of 90% only. For most recent fold recognition methods self-recognition is an easy task because all of the currently evaluated fold recognition methods [13] use scoring functions that include a measure of sequence similarity between the query and the native sequence of model. Clearly, for those fold recognition methods the self-recognition is not a challenge and the ultimate test is a structural homolog recognition. Analogously, it was demonstrated for DSMs [27], that even minimal information about the native sequence of the model greatly improves distant homolog recognition. However, the self-recognition is not 100% accurate for fold recognition methods that do not include any native sequence information in the model. This is true even when additional structural information about the amino acid dependent solvent exposure, the native positions of side chains and contacts between side chains is included [9].

In the second set of experiments, we tested performance of the fold recognition methods in recognizing the structure of proteins that do not necessarily share sequence similarity with the proteins used to generate our set of 350 DSMs, but nevertheless, have similar structural folds. For testing we used a set of proteins classified into 10 SCOP structural folds that were recently used for testing the Recursive Dynamic Programming (RDP) threading fold recognition method [23]. We had to remove the cysteine-knot cytokines fold from the original RDP set of 11 structural folds because it is an irregular fold with very few secondary structure elements. Our DSMs are based primarily on the SS and solvent exposure preferences and the irregular structures do not produce specific DSMs when only such preferences are used. In these experiments we calculated $P(\text{fold}|\text{seq})$, the posterior probability for each structural fold, for 71 sequences listed in Table 3 on our library of 350 DSMs. The results of the fold recognition results for the Viterbi and for the Filtering method are presented in Table 3. With the Filtering fold recognition method, 33 out of 71 sequences ranked the correct structural fold with the highest probability. With the Viterbi fold recognition method, 31 out of 71 sequences ranked the correct structural fold with the highest probability. The “acceptable” predictions (with the top structural fold having the probability of at least 0.5) had results as follows. For the Filtering fold recognition method, there were 53 fold predictions and 32 were correct - a 60% success rate. For the Viterbi fold recognition method, there were 66 predictions and 29 were correct - a 44% success rate. These results confirm that the Filtering fold recognition method is 36% more accurate than the Viterbi fold recognition method in recognizing the correct structural fold. The filtering fold recognition method has a fold recognition rate of 60% that is slightly better but comparable to the fold recognition rate of 57% reported for the RDP threading [23].

The worst fold recognition rate by the Filtering method

was obtained for the α/β hydrolases (a 0% fold recognition rate) represented in our library by only one structure (3lip) and for the viral coat and capsid protein fold (a 17% fold recognition rate) also represented in our library by only one structure (2stv). Both of these folds are adopted by proteins with highly variable sequence lengths as reported by Tiele *et al.* [23]. The length of α/β hydrolases varies from 265 to 534 amino acids and the length of viral coat proteins varies from 175 to 548 amino acids. Our DSMs allow for only small variations in the secondary structure segment lengths. Thus it is not surprising that having only one representative structure for such diverse folds limits the fold recognition method severely. These results call for the expansion of our model library to include all nonhomologous representatives for each structural family.

The superior performance of the Filtering fold recognition method over the Viterbi fold recognition method is apparent in the framework of HMM theory. A DSM is an ensemble of N structure models for N very similar structures. Each path can be viewed as one model for one particular structure but a model is not a perfect representation of a structure. The model is equal to a “true” structure plus some noise. This is true for all models, including the best one which Viterbi algorithm might recognize. The attempt to find the structure for a sequence using single best structure model (single path) is limited by the noise in the structure model. Using the sum of all structure models should improve the signal to noise ratio by a factor of \sqrt{N} .

5 Conclusion

In our fold recognition method we have incorporated a hierarchical structure classification scheme that allows a rigorous assignment of posterior fold/model probabilities. Each unique structural class represented by different DSMs has an assigned posterior probability for that particular class. The Bayesian assignment of posterior probabilities systematically addresses the problem of interpreting fold recognition results that use a library of diverse and persistently interdependent models. We have used a classification scheme where the model priors are assigned following an independent structural hierarchy. Alternatively, the priors could be assigned according to the overlap among the DSMs in the space of structural states.

We have presented here relatively simple structural Hidden Markov Models, the Discrete State Models. These models are built automatically from the protein structures deposited in the PDB. The DSMs represent amino acid preferences for a small set of structural states. Our DSMs encode only six SS/solvent-exposure structural states and seven loop states. As such these models can be seen as an alternative and very simple representation of a structural profile. The HMM representation has two advantages over the structural profile representation used previously by many fold recognition methods [16, 23]. The first advantage of the HMM representation is the incorporation of structural variations such as: variable secondary structure element length and variable loop states that connect the secondary structure elements. The second advantage comes directly from the HMM theory. The compatibility of the query sequence with a model can be rigorously calculated as the total (summed over all sequence-to-model alignments) probability of the model. We have demonstrated that the fold recognition method that uses the total probability is 40% more accurate than the “standard” fold recognition method that uses the probability of the optimal sequence-to-model

alignment.

Acknowledgment

We thank Jim White for many discussions about the HMMs and statistics, and continuing support for the DSMs project. We also thank Scott Mohr for carefully reading the manuscript and making many helpful suggestions.

References

- [1] BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T., AND TASUMI, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112 (1977), 535-542. Brookhaven Protein Data Bank release 80.
- [2] BOWIE, J. U., LUTHY, R., AND EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253 (1991), 164-170.
- [3] BRYANT, S. H., AND LAWRENCE, C. E. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function and Genetics* 16 (1993), 92-112.
- [4] GODZIK, A., SKOLNICK, J., AND KOLINSKI, A. A topology fingerprint approach to the inverse folding problem. *J. Mol. Biol.* 227 (1992), 227-238.
- [5] JERNIGAN, R. L., AND BAHAR, I. Structure-derived potentials and protein simulations. *Current Opinion in Structural Biology* 6 (1996), 195-209.
- [6] KABSCH, W., AND SANDER, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (1983), 2577-2637.
- [7] LATHROP, R. H., ROGERS JR., R. G., SMITH, T. F., AND WHITE, J. V. A bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology.* 60 (1998), 1-33.
- [8] LATHROP, R. H., ROGERS JR., R. G., BIENKOWSKA, J. R., BRYANT, B. K. M., BUTUROVIĆ, L. J., GAITATZES, C. NAMUDRIPAD, R., WHITE, J. V., AND SMITH, T. F. *Analysis and Algorithms for Protein Sequence-Structure Alignment*. S. Salzberg, D. Searls and S. Kasif, Elsevier Press, Amsterdam, Netherlands, 1998, pp. 227-283.
- [9] LEMER, C., ROOMAN, M. J., AND WODAK, S. Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* 23 (1995), 337-355.
- [10] LEVITT, M. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins: Structure, Function and Genetics Suppl.* 1 (1997), 92-104.
- [11] MIYAZAWA, S., AND JERNIGAN, R. L. Residue-residue potentials with a favorable contact pair term and unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256 (1996), 623-644.
- [12] MURZIN, A., BRENNER, S. E., HUBBARD, T., AND CHOTHIA, C. SCOP: a structural classification of proteins database for the investigation of the sequences and structures. *J. Mol. Biol.* 247 (1995), 536-540.
- [13] MURZIN, A. G. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Structure, Function and Genetics Suppl.* 3 (1999), 88-103.
- [14] PARK, J., KARPLUS, K., BARRETT, C., HUGHEY, R., HAUSSLER, D., HUBBARD, T., AND CHOTHIA, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284 (December 1998), 1201-1210.
- [15] RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings IEEE* 77 (1989), 257-286.
- [16] ROST, B., SCHNEIDER, R., AND SANDER, C. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270 (1997), 471-480.
- [17] RUSSELL, R. B., COPLEY, R. R., AND BARTON, G. J. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* 259 (1996), 349-365.
- [18] SIPPL, M. J. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology* 5 (1995), 229-235.
- [19] SKOLNICK, J., JAROSZEWSKI, L., KOLINSKI, A., AND GODZIK, A. Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct? *Protein Science* 6 (1997), 676-688.
- [20] SMITH, T. F., AND WATERMAN, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* 147 (1980), 195-197.
- [21] T., J. D. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 4 (April 1999), 797-815.
- [22] TAYLOR, W. R. Multiple sequence threading: An analysis of alignment quality and stability. *J. Mol. Biol.* 269 (1997), 902-943.
- [23] THIELE, R., ZIMMER, R., AND LENGAUER, T. Protein threading by recursive dynamic programming. *J. Mol. Biol.* 290 (July 1999), 757-779.
- [24] VITERBI, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Information Theory IT-13* (April 1967), 260-269.
- [25] WHITE, J. V. *Bayesian analysis of time series and dynamic models*. Marcel Dekker, New York, NY USA, 1988, pp. 255-283.
- [26] WHITE, J. V., STULTZ, C. M., AND SMITH, T. F. Protein classification by stochastic modeling and optimal filtering of amino acid sequences. *Bulletin of Mathematical Biosciences* 119 (1994), 35-75.
- [27] YU, L., WHITE, J. V., AND SMITH, T. F. A homology identification method that combines sequence and structure information. *Protein Science* 7 (1998), 2499-2510.

Conditional probabilities of Amino Acids

AA	Hb	Hpb	He	Sb	Spb	Se	T _{t1}	T _{t2}	T _{β1}	T _{β2}	T _{β3}	T _{β4}	C
A	0.147	0.083	0.100	0.081	0.044	0.040	0.030	0.023	0.074	0.066	0.046	0.074	0.074
C	0.017	0.004	0.001	0.029	0.018	0.006	0.003	0.003	0.021	0.011	0.013	0.021	0.021
D	0.022	0.050	0.082	0.017	0.043	0.057	0.165	0.144	0.069	0.069	0.098	0.069	0.069
E	0.033	0.099	0.148	0.013	0.068	0.096	0.079	0.032	0.047	0.065	0.046	0.047	0.047
F	0.071	0.035	0.009	0.081	0.049	0.027	0.008	0.006	0.034	0.019	0.024	0.034	0.034
G	0.042	0.021	0.031	0.058	0.035	0.043	0.300	0.422	0.109	0.090	0.280	0.109	0.109
H	0.014	0.029	0.017	0.018	0.030	0.020	0.017	0.006	0.024	0.012	0.022	0.024	0.024
I	0.098	0.049	0.023	0.127	0.070	0.051	0.008	0.006	0.036	0.022	0.006	0.036	0.036
K	0.018	0.108	0.125	0.017	0.084	0.095	0.048	0.030	0.056	0.066	0.045	0.056	0.056
L	0.176	0.095	0.043	0.122	0.067	0.050	0.009	0.009	0.061	0.035	0.021	0.061	0.061
M	0.034	0.023	0.013	0.024	0.019	0.011	0.008	0.003	0.015	0.008	0.015	0.015	0.015
N	0.018	0.040	0.060	0.017	0.040	0.049	0.147	0.089	0.062	0.050	0.115	0.062	0.062
P	0.010	0.015	0.029	0.013	0.014	0.031	0.036	0.012	0.076	0.199	0.014	0.076	0.076
Q	0.026	0.075	0.086	0.018	0.039	0.054	0.024	0.026	0.035	0.022	0.039	0.035	0.035
R	0.029	0.088	0.072	0.022	0.066	0.052	0.026	0.024	0.038	0.034	0.040	0.038	0.038
S	0.036	0.039	0.066	0.046	0.058	0.078	0.064	0.070	0.077	0.102	0.093	0.077	0.077
T	0.041	0.051	0.045	0.043	0.089	0.116	0.006	0.071	0.071	0.052	0.037	0.071	0.071
V	0.098	0.051	0.034	0.167	0.104	0.081	0.009	0.005	0.052	0.038	0.008	0.052	0.052
W	0.019	0.015	0.005	0.020	0.016	0.012	0.002	0.006	0.012	0.012	0.013	0.012	0.012
Y	0.048	0.033	0.011	0.067	0.047	0.029	0.014	0.014	0.032	0.028	0.024	0.032	0.032

Table 1: First column (AA) denotes the amino acids. The remaining columns represent conditional probabilities of observing amino acids given a structural environment state. Hb, Hpb and He are buried, partially buried and solvent exposed states in a helix. Sb, Spb and Se are buried, partially buried and solvent exposed states in a strand. T_{t1}, T_{t2} are positions 1 and 2 in a tight turn loop. T_{β1}, T_{β2}, T_{β3} and T_{β4} are positions 1, 2, 3 and 4 in a beta turn loop. C is any position in a coil loop. The data were collected from a set 474 solvable proteins with less than 25% sequence identity between any pair. The secondary structure was assigned by DSSP [6]. The list of PDB codes is available upon request.

Results of the native fold recognition experiments

Sequence PDB code	Fold recognition method						Sequence PDB code	Fold recognition method					
	Filtering			Viterbi				Filtering			Viterbi		
	top family probability	top family PDB code	native family rank	top family probability	top family PDB code	native family rank		top family probability	top family PDB code	native family rank	top family probability	top family PDB code	native family rank
153l	1.00	153l	1	1.00	153l	1	1a17	1.00	1a17	1	1.00	1a17	1
1a1x	1.00	1cmcA	18	1.00	1nsgB	26	1a32	1.00	1a32	1	1.00	1a32	1
1a62	1.00	1a62	1	1.00	2mhr	4	1a68	1.00	1a68	1	1.00	1nsgB	7
1a6jA	1.00	1a6jA	1	1.00	1ffp	2	1a9t	1.00	1a9t	1	1.00	1a9t	1
1aa7B	1.00	1aa7B	1	1.00	1aa7B	1	1aazA	1.00	1aazA	1	1.00	1cmcA	4
1ab8B	1.00	1ab8B	1	1.00	1nfn	2	1acx	1.00	1acx	1	1.00	1lkkA	29
1add	1.00	1add	1	1.00	1add	1	1ae9B	1.00	1ae9B	1	1.00	1nfn	2
1aep	1.00	1aep	1	1.00	1aep	1	1aerB	1.00	1aerB	1	1.00	1aerB	1
1af5	1.00	1af5	1	1.00	256bA	10	1ahq	1.00	1ravA	7	1.00	1ffp	8
1air	1.00	1air	1	1.00	1cem	4	1aj2	1.00	1aj2	1	1.00	1ribA	2
1ako	1.00	1ako	1	1.00	1ako	1	1alkA	1.00	1alkA	1	1.00	1ft1A	2
1am2	1.00	1am2	1	1.00	1am2	1	1amk	1.00	1gky	7	1.00	2lbd	17
1amp	1.00	1amp	1	1.00	1amp	1	1amx	1.00	1amx	1	1.00	1amx	1
1an7A	1.00	1an7A	1	1.00	2asr	3	1aol	1.00	1aol	1	1.00	1aol	1
1apa	1.00	1apa	1	1.00	1apa	1	1av6A	1.00	1av6A	1	1.00	1av6A	1
1awd	1.00	1awd	1	1.00	1awd	1	1axn	1.00	1axn	1	1.00	1axn	1
1ay9B	1.00	1ay9B	1	1.00	1cewI	4	1ayi	1.00	1ayi	1	1.00	1ayi	1
1ba7A	1.00	1cghA	4	1.00	1nfn	65	1bam	1.00	1bam	1	1.00	1nfn	2
1bgc	1.00	1bgc	1	1.00	1bgc	1	1bkrA	1.00	1bkrA	1	1.00	1bkrA	1
1ble	1.00	1cyw	2	1.00	1aep	7	1bme	1.00	1bme	1	1.00	1cby	2
1btn	1.00	1btn	1	1.00	256bA	8	1bv1	1.00	1bv1	1	1.00	1ffp	2
1c52	1.00	1c52	1	1.00	1c52	1	1cby	1.00	1cby	1	1.00	1cby	1
1cem	1.00	1cem	1	1.00	1cem	1	1cewI	1.00	1cewI	1	1.00	1cewI	1

continued on next page

<i>continued from previous page</i>													
1cex	1.00	1cex	1	1.00	1cex	1	1cghA	1.00	1cghA	1	1.00	1nfn	22
1cgmE	1.00	1cfd1	18	1.00	2asr	5	1chd	1.00	1chd	1	1.00	1chd	1
1cmcA	1.00	1cmcA	1	1.00	256bA	3	1cpt	1.00	1cpt	1	1.00	1cpt	1
1cyw	1.00	1cyw	1	1.00	1nfn	2	1deaB	1.00	1deaB	1	1.00	1deaB	1
1dhpA	1.00	1dhpA	1	1.00	1dhpA	1	1div	1.00	1div	1	1.00	1lis	2
1dosA	1.00	1dosA	1	1.00	1dosA	1	1ecmB	1.00	1ecmB	1	1.00	1ecmB	1
1ema	1.00	1ema	1	1.00	1ema	1	1exnB	1.00	1exnB	1	1.00	1exnB	1
1fiaB	1.00	1fiaB	1	1.00	1fiaB	1	1fkd	1.00	1fkd	1	1.00	2mhr	2
1ffp	1.00	256bA	2	1.00	1ffp	1	1fmb	1.00	1fmb	1	1.00	1tul	2
1fmcA	1.00	1fmcA	1	1.00	1fmcA	1	1fna	1.00	1fna	1	1.00	1fua	1
1fps	1.00	1fps	1	1.00	1fps	1	1frb	1.00	1frb	1	1.00	1frb	1
1ft1A	1.00	1ft1A	1	1.00	1ft1A	1	1fua	1.00	1fua	1	1.00	1fua	1
1garB	1.00	1garB	1	1.00	1garB	1	1gen	1.00	1gen	1	1.00	1gen	1
1gky	1.00	1gky	1	1.00	1nfn	2	1gox	1.00	1gox	1	1.00	1ribA	9
1gpr	1.00	1gpr	1	1.00	1gtqA	50	1gtqA	1.00	1gtqA	1	1.00	1gtqA	1
1hfc	1.00	1hfc	1	1.00	153l	5	1htp	1.00	1jpc	2	1.00	1htp	1
1hus	1.00	1hus	1	1.00	256bA	4	1ido	1.00	1ido	1	1.00	1ido	1
1lfc	1.00	1lfc	1	1.00	1lfc	1	1iibA	1.00	1ycqA	27	1.00	1nsgB	15
1lipsA	1.00	1lipsA	1	1.00	1ipsA	1	1jdw	1.00	1jdw	1	1.00	1jdw	1
1jpc	1.00	1jpc	1	1.00	1xsoB	5	1knb	1.00	1amx	3	1.00	1ema	6
1kpf	1.00	1jpc	2	1.00	256bA	9	1ksaA	1.00	1ksaA	1	1.00	1cem	3
1lba	1.00	1lba	1	1.00	1lba	1	1lfb	1.00	1fiaB	4	1.00	1lfb	1
1lis	1.00	1pdo	2	1.00	1lis	1	1lkkA	1.00	1lkkA	1	1.00	1lkkA	1
1lrv	1.00	1rgp	2	1.00	1rgp	3	1lxa	1.00	2prk	5	1.00	1lxa	1
1mkaB	1.00	1mkaB	1	1.00	1mkaB	1	1msk	1.00	1msk	1	1.00	1msk	1
1mspB	1.00	1ravA	2	1.00	1ravA	10	1mugA	1.00	1mugA	1	1.00	1aep	5
1nar	1.00	1nar	1	1.00	1nar	1	1nbcB	1.00	1nbcB	1	1.00	1nbcB	1
1nfn	1.00	1nfn	1	1.00	1nfn	1	1nfp	1.00	1nfp	1	1.00	1nfp	1
1nls	1.00	1nls	1	1.00	1ema	4	1npk	1.00	1c52	26	1.00	1cex	26
1nsgB	1.00	2mhr	2	1.00	1nsgB	1	1nsj	1.00	1nsj	1	1.00	1nsj	1
1nsyA	1.00	1nsyA	1	1.00	1nsyA	1	1opy	1.00	2msbA	2	1.00	1opy	1
1oroA	1.00	1oroA	1	1.00	1oroA	1	1osa	1.00	1osa	1	1.00	1osa	1
1pdo	1.00	1pdo	1	1.00	1pdo	1	1phr	1.00	1bv1	3	1.00	1cgmE	6
1pmi	1.00	1pmi	1	1.00	1pmi	1	1pne	1.00	2rhe	25	1.00	1bkrA	12
1poh	1.00	1poh	1	1.00	1poh	1	1pud	1.00	1pud	1	1.00	1pud	1
1ravA	1.00	1fmb	3	1.00	1iris	22	1regY	1.00	1a62	2	1.00	2spcB	16
1rgeA	1.00	1acx	9	1.00	1lfb	17	1rgp	1.00	1rgp	1	1.00	1rgp	1
1rhs	1.00	1rhs	1	1.00	1rhs	1	1ribA	1.00	1ribA	1	1.00	1ribA	1
1rie	1.00	1rie	1	1.00	2msbA	16	1iris	1.00	1iris	1	1.00	1iris	1
1rkd	1.00	1rkd	1	1.00	1rkd	1	1rpa	1.00	1rpa	1	1.00	1rpa	1
1rsy	1.00	1rsy	1	1.00	1rsy	1	1sfp	1.00	1alx	4	1.00	1alx	5
1smnB	1.00	1smnB	1	1.00	2cyp	2	1snc	1.00	1snc	1	1.00	2asr	12
1tig	1.00	1tig	1	1.00	1a32	2	1tlcB	1.00	1tlcB	1	1.00	1tlcB	1
1tml	1.00	1tml	1	1.00	1tml	1	1tmy	1.00	1tmy	1	1.00	1tmy	1
1toh	1.00	1toh	1	1.00	1fps	2	1ttaB	1.00	1ttaB	1	1.00	1lis	5
1tul	1.00	1tul	1	1.00	1ecmB	9	1uch	1.00	1uch	1	1.00	1uch	1
1udiI	1.00	1ycqA	4	1.00	1a32	12	1vhh	1.00	1pdo	9	1.00	1ffp	8
1wab	1.00	1wab	1	1.00	1wab	1	1wgjB	1.00	1wgjB	1	1.00	1wgjB	1
1whi	1.00	1af5	2	1.00	1hus	11	1who	1.00	1fna	3	1.00	2end	3
1wpoB	1.00	1wpoB	1	1.00	1lxa	5	1xaa	1.00	1xaa	1	1.00	1cem	2
1xsoB	1.00	1xsoB	1	1.00	1xsoB	1	1ycqA	1.00	1aazA	3	1.00	1ecmB	10
1ycsA	1.00	1ycsA	1	1.00	1amx	17	1yer	1.00	1yer	1	1.00	1yer	1
1ygs	1.00	1ygs	1	1.00	1ygs	1	1ytw	1.00	1ytw	1	1.00	1ytw	1
256bA	1.00	256bA	1	1.00	256bA	1	2a0b	1.00	2a0b	1	1.00	2a0b	1
2aacA	1.00	2aacA	1	1.00	2aacA	1	2aak	1.00	2aak	1	1.00	2aak	1
2acy	1.00	2acy	1	1.00	1bkrA	5	2asr	1.00	2asr	1	1.00	2asr	1
2bopA	1.00	2bopA	1	1.00	1faB	2	2cba	1.00	2cba	1	1.00	1lrv	2
2cpl	1.00	2cpl	1	1.00	2cpl	1	2cyp	1.00	2cyp	1	1.00	2cyp	1
2dkb	1.00	2dkb	1	1.00	2dkb	1	2dri	1.00	1nsj	8	1.00	2dri	1
2end	1.00	2end	1	1.00	2end	1	2hts	1.00	2hts	1	1.00	2end	4
2lbd	1.00	2lbd	1	1.00	2lbd	1	2mhr	1.00	2hts	2	1.00	2mhr	1
2msbA	1.00	2msbA	1	1.00	1lfb	3	2phy	1.00	1opy	3	1.00	1opy	4
2plc	1.00	2plc	1	1.00	1rgp	3	2prk	1.00	2prk	1	1.00	1cem	2
2pth	1.00	1npk	2	1.00	2pth	1	2rhe	1.00	2rhe	1	1.00	1cmcA	10

continued on next page

continued from previous page

2rn2	1.00	2rn2	1	1.00	2rn2	1	2sak	1.00	2sak	1	1.00	2sak	1
2sic1	1.00	2sic1	1	1.00	2sic1	1	2sil	1.00	2sil	1	1.00	2sil	1
2snil	1.00	2snil	1	1.00	2snil	1	2spcB	1.00	2spcB	1	1.00	2spcB	1
2stv	1.00	2stv	1	1.00	2stv	1	3b5c	1.00	1fkd	6	1.00	1a32	22
3bct	1.00	3bct	1	1.00	3bct	1	3cla	1.00	3cla	1	1.00	1aep	4
3lip	1.00	1ako	9	1.00	1cem	5	4fgf	1.00	4fgf	1	1.00	1vhh	6
4xis	1.00	4xis	1	1.00	4xis	1	6fd1	1.00	6fd1	1	1.00	6fd1	1

Table 2: Results of the fold recognition experiments for 188 SCOP superfamily representatives. The structure prediction was done at the SCOP structural family level. The prediction is equivalent to structural superfamily prediction since only one family represents each superfamily. Top family indicates the most probable family according to the posterior probability value.

Results of the structural homolog fold recognition experiments

Sequence PDB code	Sequence length	Correct fold PDB code	Fold recognition method					
			Filtering			Viterbi		
			top fold probability	top fold PDB codes	correct fold rank	top fold probability	top fold PDB codes	correct fold rank
SCOP fold classification: OB fold								
1gpc	218	1a62	0.34	1aerB	14	0.97	1lxa	39
1snc	149	1a62	0.39	1gpr 1htp	2	0.94	4-helix bundle	10
1prtF	98	1a62	0.42	1lfb 2hts	40	0.78	4-helix bundle	32
1prtD	110	1a62	0.47	2sic1	38	0.39	2end	13
1a62	130	1a62	0.77	1a62 1snc 1wgjB	1	0.69	1cis	5
1pyp	285	1a62	0.99	1a62 1snc 1wgjB	1	0.84	1a62 1snc 1wgjB	1
1wgjA	286	1a62	0.99	1a62 1snc 1wgjB	1	0.94	1a62 1snc 1wgjB	1
2prd	174	1a62	0.99	1lxa	20	0.63	α/α superhelix	34
SCOP fold classification: four-helical cytokines								
1lki	180	1bgc	0.71	1bgc	1	1.00	1bgc	1
1ilk	151	1bgc	0.96	4-helix bundle	23	1.00	4-helix bundle	4
1huw	166	1bgc	0.97	1ahq	21	0.77	1ffp	3
1bgc	174	1bgc	0.99	1bgc	1	1.00	1bgc	1
SCOP fold classification: globin-like								
1eca	136	1ffp	0.35	1ahq	36	0.54	1ffp	1
1cpcA	162	1ffp	0.39	1hfc	41	0.70	4-helix bundle	2
1pbxA	143	1ffp	0.45	1cewI 1opy 1udil	10	0.94	1ffp	1
2gdm	153	1ffp	0.50	1ffp	1	0.85	4-helix bundle	2
2fal	146	1ffp	0.66	1ffp	1	0.91	1ffp	1
2hbg	147	1ffp	0.67	1ae9B	23	0.45	1ffp	1
1ffp	142	1ffp	0.75	1ffp	1	0.99	1ffp	1
1h1b	157	1ffp	0.85	1ffp	1	0.51	1ffp	1
1hrm	153	1ffp	0.96	1ffp	1	0.99	1ffp	1
1ash	150	1ffp	0.97	1ffp	1	0.99	1ffp	1
3sdhA	146	1ffp	0.99	1a62 1snc 1wgjB	57	0.92	1aep	2
1cpcB	172	1ffp	0.99	1ffp	1	1.00	1ffp	1
SCOP fold classification: lipocalins								
1mup	166	1lfc	0.30	1lfc	1	0.61	1aep	31
1bbpA	173	1lfc	0.38	1knb	2	0.22	1cex 1tmy 1wab	17
1epaA	164	1lfc	0.92	1lfc	1	0.80	1aep	17
1hbq	183	1lfc	0.96	1ble	3	0.85	1aep	39
1lfc	132	1lfc	0.99	1lfc	1	1.00	1lfc	1
1hmt	132	1lfc	0.99	1lfc	1	1.00	1lfc	1
SCOP fold classification: α/β TIM-barrel								
1ubsA	268	1nar	0.24	1rhs	8	0.95	2cyp	19
1fbaA	360	1nar	0.41	2dri	2	1.00	1fps	7
5tima	250	1nar	0.49	1chd	3	0.68	1cem	2
1pbgA	468	1nar	0.49	1jdw	4	0.65	1ribA	5
1xyzA	347	1nar	0.53	α/β TIM-barrel	1	1.00	1ribA	9
1oyc	400	1nar	0.64	1rpa	2	0.62	1fps	7
1byb	495	1nar	0.92	1alkA	6	1.00	α/α superhelix	5
1nar	290	1nar	0.99	α/β TIM-barrel	1	1.00	α/β TIM-barrel	1
1nfp	228	1nar	0.99	α/β TIM-barrel	1	1.00	α/β TIM-barrel	1

continued on next page

continued from previous page								
2ebn	289	1nar	0.99	α/β TIM-barrel	1	0.74	α/β TIM-barrel	1
2acq	315	1nar	1.00	α/β TIM-barrel	1	1.00	α/β TIM-barrel	1
SCOP fold classification: four-helical up-and-down bundle								
2hmzA	113	1nfn	0.53	4-helix bundle	1	1.00	4-helix bundle	1
2tmvP	158	1nfn	0.76	1gpr 1htp	9	0.58	4-helix bundle	1
1was	146	1nfn	0.99	4-helix bundle	1	1.00	4-helix bundle	1
2ccyA	128	1nfn	0.99	4-helix bundle	1	0.87	2spcB	2
1lpe	144	1nfn	0.99	4-helix bundle	1	1.00	4-helix bundle	1
1nfn	191	1nfn	0.99	4-helix bundle	1	1.00	4-helix bundle	1
256bA	106	1nfn	0.99	4-helix bundle	1	1.00	4-helix bundle	1
SCOP fold classification: flavodoxin-like								
3tmy	120	1tmy	0.94	1cex 1tmy 1wab	1	0.68	1cex 1tmy 1wab	1
3chy	128	1tmy	0.99	1cex 1tmy 1wab	1	0.92	1cex 1tmy 1wab	1
2fox	138	1tmy	0.99	1cex 1tmy 1wab	1	0.74	1cex 1tmy 1wab	1
1cus	200	1tmy	0.99	1cex 1tmy 1wab	1	1.00	1cex 1tmy 1wab	1
1rcf	169	1tmy	0.99	1mugA	4	0.80	1bgc	12
SCOP fold classification: viral coat and capsid proteins								
4rhv3	236	2stv	0.45	3cla	4	0.88	1ema	5
1bbt3	220	2stv	0.48	1cghA	29	0.75	153l	21
2bpa2	175	2stv	0.54	1ba7A 4fgf	21	0.78	1ema	17
1btt1	213	2stv	0.63	1gky	38	0.96	4-helix bundle	42
1btt2	218	2stv	0.84	1am2	5	1.00	1ema	5
2stv	195	2stv	0.97	2stv	1	0.54	2stv	1
SCOP fold classification: α/β hydrolases								
3tgl	269	3lip	0.28	1wpoB	48	0.68	ferredoxin-like	49
1ede	310	3lip	0.42	1tml	22	1.00	2lbd	24
1thtA	305	3lip	0.62	α/β TIM-barrel	8	0.45	1cem	16
1tahB	318	3lip	0.63	1av6A	5	0.85	α/α superhelix	4
3lip	320	3lip	0.66	1ako	6	1.00	1cem	5
1tca	317	3lip	0.82	2prk	18	0.91	1chy	18
SCOP fold classification: ferredoxin-like								
1regX	122	6fd1	0.29	1a62 1snc 1wgjB	7	0.93	2spcB	9
1aps	98	6fd1	0.50	β sandwich	3	0.69	1fiaB	4
2bopA	85	6fd1	0.61	ferredoxin-like	1	0.85	1fiaB	2
6fd1	106	6fd1	0.62	1kpf	2	0.45	ferredoxin-like	1
1nhkR	144	6fd1	0.74	1pdo	8	0.94	1aep	32
1pba	81	6fd1	0.79	1a68	13	0.72	1a32	3

Table 3: Results of the fold recognition experiments for 10 SCOP structural folds. The structure prediction was done at the structural fold level. Top fold indicates the most probable structural fold according to the posterior probability value. For folds represented by more than three PDB entries we used the SCOP names as follows. α/β TIM-barrel: 1add, 1aj2, 1amk, 1dhpA, 1dosA, 1frb, 1gox, 1nar, 1nfp, 1nsj, 1pud, 2plc and 4xis. 4-helix bundle: 1cgmE, 1nfn, 1nsgB, 256bA, 2a0b, 2asr and 2mhr. Ferredoxin-like: 1ab8B, 1npg, 1regY, 1ris, 2acy, 2bopA and 6fd1. α/α superhelix: 1a17, 1ft1A, 1hrv and 3bct.