Triage protein fold prediction

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · SEPTEMBER 2002

Impact Factor: 2.63 · DOI: 10.1002/prot.10194 · Source: PubMed

3 AUTHORS:

Hongxian He
Dupont
5 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Gregory Mcallister
Novartis
11 PUBLICATIONS **169** CITATIONS
SEE PROFILE

Temple F. Smith
Boston University
210 PUBLICATIONS **19,423** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 48:654-663 (2002)

Triage Protein Fold Prediction

Hongxian He, Gregory McAllister, Temple F. Smith*
BioMolecular Engineering Research Center, Biomedical Engineering Department, Boston University, Boston, Massachusetts

ABSTRACT **We have constructed, in a completely automated fashion, a new structure template library for threading that represents 358 distinct SCOP folds where each model is mathematically represented as a Hidden Markov model (HMM). Because the large number of models in the library can potentially dilute the prediction measure, a new triage method for fold prediction is employed. In the first step of the triage method, the most probable structural class is predicted using a set of manually constructed, high-level, generalized structural HMMs that represent seven general protein structural classes: all-α, all-β, α/β, α+β, irregular small metal-binding, transmembrane β-barrel, and transmembrane α-helical. In the second step, only those fold models belonging to the determined structural class are selected for the final fold prediction. This triage method gave more predictions as well as more correct predictions compared with a simple prediction method that lacks the initial classification step. Two different schemes of assigning Bayesian model priors are presented and discussed. Proteins 2002;48: 654-663.** © 2002 Wiley-Liss, Inc.

Key words: protein fold prediction; threading; HMM; DSM; triage

INTRODUCTION

Threading methods have proven to be powerful tools for protein fold recognition, especially when there is no detectable sequence similarity to a protein of known structure. Generally these methods search a set of structural templates, which are abstract descriptions of the three-dimensional coordinates of known structures, for an optimal sequence-structure alignment.[1–6] This alignment is evaluated by means of an empirical energy function or probabilistic scoring scheme where it is presumed that the maximum score identifies the optimal alignment to a given structural model. It is generally assumed that the model having the highest maximum alignment score predicts the correct fold. Since the correct prediction is possible only if an appropriate structural model is present in the library, such methods require the construction of a comprehensive library of structural templates.

A threading method previously developed in our center uses discrete state-space models (DSMs) to represent tertiary structural classes.[7,8] DSMs are mathematically represented as hidden Markov models (HMMs) but differ in that the model parameters are designed with our knowledge of protein structures, as opposed to being trained on a particular training set, as is customary in HMM construction.[9,10] A predefined set of DSMs was constructed manually following the classification of structural domains by Jane Richardson[11] and represents twenty-four structural classes, called macro-classes, such as 4-helix bundle proteins and TIM-barrel proteins. Macro-classes were grouped under four major folding super-classes of globular single-domain proteins: α, β, α/β, and irregular. The tertiary structure prediction is made by finding the macro-class (and super-class) with the highest score, which is expressed as the posterior probability of a given macro-class for the given sequence. The macro-class probabilities are obtained by summing the probabilities of all DSMs belonging to the given macro-class, and the super-classes probabilities are obtained by summing up the probabilities of all underlying macro-classes. The probability of each individual DSM for a given sequence is computed using a Bayesian formula.[7]

The large number of currently known structures present in the PDB database[12] renders the manual construction of models impractical. We have, therefore, built a new DSM library according to the SCOP database[13] via an automated procedure similar to that of Bienkowska et al.[14] Each DSM in the library, representing a distinct SCOP fold, was automatically constructed from the determined atomic coordinates of a representative structure from the PDB database, along with a set of generalized rules. The reliance of the structural prediction on a model library imposes several problems. First, the large number of models increases the likelihood that the correct model will exhibit a small posterior probability since the calculation involves the normalization over all the models in the library. Second, the calculation of posterior probabilities always results in one model being the most probable, even if the appropriate structural model is not present in the library. In order to reduce these problems, we have devised a triage method that first classifies the protein sequence into a general structural class prior to searching the full DSM library. This approach relies on a set of manually constructed, high-level generic DSMs that model major structural classes, such as all-α, all-β, α/β, and membrane

Grant sponsor: Department of Energy; Grant number: DE-FG02-98ER62558.

*Correspondence to: Temple F. Smith, BioMolecular Engineering Research Center, Biomedical Engineering Department, Boston University, 36 Cummington St., Boston, MA 02215.
E-mail: tsmith@darwin.bu.edu

Received 10 September 2001; Accepted 2 May 2002

proteins. This two-step method eliminated many of the incompatible models from the model library before the fold posterior probability distribution is calculated, therefore increasing the chance that the correct structural template will be identified. It also helps to reduce the number of potential false positives by filtering out the models that do not fall into the same structural class through the initial classification.

METHODS

Posterior Model Probability

Given a set of N structural models, $S = M_i$ ($i = 1, \ldots, N$), each of which defines a structural hypothesis, the goal of structure prediction is to find the best hypothesis to explain the given sequence. This is quantified by the posterior probability $P(M_i|seq)$ computed according to Bayes' rule:

$$P(M_i|seq) = \frac{P(seq|M_i)P(M_i)}{\sum_{k=1}^{N} P(seq|M_k)P(M_k)}, \quad (1)$$

where N is the total number of models in the library.

$P(seq|M_i)$, the total probability that the observed sequence would have been generated by model M_i, is calculated using an optimal filtering algorithm or the forward algorithm.[15,16] This algorithm differs from the more commonly used Viterbi algorithm in that it sums over all possible paths through the model as opposed to only the optimal path. Here, $P(M_i)$ is the prior probability of model M_i. The method of assigning priors, $P(M_i)$, will be discussed later.

We apply a conservative binary decision rule for prediction: if a model has a posterior probability greater than 0.5, a prediction is made.

Automated Construction of a PDB Domain DSM Library

A discrete state-space model (DSM) is a linear representation of structural states that correspond to the structural positions of a protein. A structural state is defined by the type of secondary structure (helix, strand, loop, or turn) and degree of solvent exposure (buried, partially-buried, or exposed). Each state is associated with a characteristic emission or occupancy probability distribution over the 20 amino acids derived from the statistical analysis of a large set of representative protein structures (unpublished data). Structural states are connected to form a Markov chain where the state transition probabilities are designed to model the important constraints on secondary structure imposed by the specific tertiary structure being modeled.

In our automated model generation,[14] each DSM was constructed from a solved protein structure in a hierarchical fashion. First, secondary structural elements (SSE) were determined using the DSSP program[17] with the degree of solvent exposure for each residue position calculated according to Eisenberg et al.[18] Second, each SSE was mapped into a DSM module, which is a sequence of structural states defined by the determined secondary structure and calculated degree of solvent exposure. Additional structural states and state-to-state transitions were added at both ends of each module. This allows for length variations (extension/deletion by one or two residues) while maintaining the minimum length of a secondary structure element: two residues for a strand and five residues for a helix. Finally, these secondary structural modules were connected by loop modules according to their topological order in the native protein structure. The type of loop module was determined by the geometrical distance between the ends of two consecutive SSEs.

The set of representative structures was selected according to the SCOP classification (Release 1.48). The level of classification in which we are interested is the "fold," which is defined as a group of structural domains having the same major secondary structural elements with the same topology. All structures having repeated or intertwined domains, in addition to all membrane proteins, were excluded, resulting in 358 unique SCOP folds. For every chosen SCOP fold, a structure was selected from each of its constituent "families" for a total of 881 representatives. For structural folds whose PDB records include multiple subunits, domains, or bound cofactors, multiple DSMs were constructed accounting for differing solvent exposure profiles corresponding to the presence or absence of additional elements. There are, therefore, multiple models representing each SCOP fold. The final library consists of 1,282 models, denoted as PDB domain DSMs.

Generic DSMs

As a feature of our triage method, a set of generic DSMs was designed to represent all known gross structural classes. These models represent the general structural features of a particular class while allowing a large range of anticipated secondary structure variations. Seven generic DSMs have been manually built for the following structural classes, where class 1–5 refer to soluble proteins only: (1) all-α proteins having only α-helical secondary structure, (2) all-β proteins having mainly β-strand secondary structure, (3) α/β proteins having alternating α-helical and β-strand secondary structures, (4) α+β proteins having local clusters of α-helices and β-strands segregated along the primary sequence, (5) irregular proteins having few secondary structure elements and a high content of disulfide bonds or metal ligands, (6) transmembrane β-barrel proteins, and (7) transmembrane α-helical proteins. These classes are considered to be mutually exclusive and exhaustive.

Generic All-α DSM

Figure 1 illustrates the model structure of the generic all-α DSM. Basic secondary structure modules have been combined to form larger secondary structure complexes, called *plexes* for short. Each oval represents either a secondary structure module or a *plex*. The overall model consists of two identical α-plexes, each constructed from a α-helix module (amphipathic or buried) followed by a turn or loop module. A secondary structure module contains a series of structural states that represent structural posi-

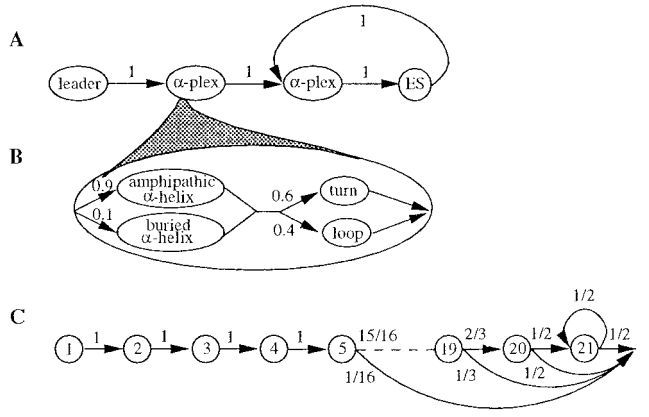

Fig. 1. Schematic of the generic model for all-α proteins. All paths are labeled with transition probabilities from one structural module to another. **A:** The DSM organization at the α-plex level of detail. All the α-plexes are identical. The circle labeled with ES is the *end state*. **B:** Each α-plex consists of an α-helix (amphipathic or buried) followed by the connecting loop or turn. **C:** The structure of the loop module. Each circle represents a loop state and all the states are identical. The numbers in the circle are the indices of the states.

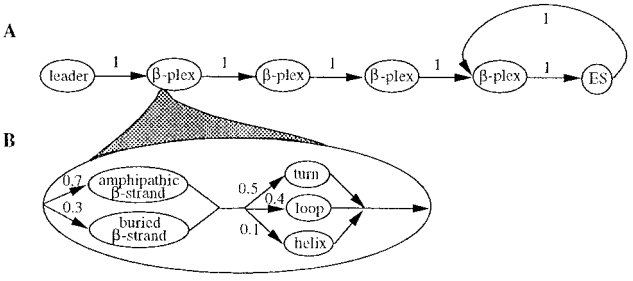

Fig. 2. Schematic of the generic model for all-β proteins. **A:** The DSM organization at the β-plex level. All the β-plexes are identical. **B:** Each β-plex consists of a β-strand (amphipathic or buried) and connecting region, which can be a loop, turn, or short helix.

tions in a protein. The arrows indicate the transitions from one structural state or module to another with the transition probabilities assigned based on our knowledge of the relative frequencies of these secondary structure elements in all-α proteins. For example, there are no all-α proteins with fewer than two helices, amphipathic helices are more common than completely buried helices in globular proteins, and turns occur more often between helices than loops. We explicitly model the length distribution for these modules according to the typical lengths observed in such proteins. The α-helix module has a uniform length distribution between 5 and 30 residues and exponential distribution over a length greater than 30 residues. The loop has a uniform length distribution between 5 and 20 residues and an exponential distribution for more than 20 residues. The turn can be 2-, 3-, or 4-residues in length with equal probability.

The generic DSM shown in Figure 1 differs from a typical DSM, which is called a left-to-right model or a Bakis model.[19] When a left-to-right model is used to generate a state sequence, the state index increases or stays the same as time (or sequence position in the DSM analysis) increases. A DSM representing a protein structure proceeds from its N-terminal to its C-terminal, i.e., the structural states are numbered from the amino end to the carboxy end of the protein structure. A sequence is always filtered through the model from the first structural state, corresponding to the N-terminal residue position, to the last state, corresponding to the C-terminal residue position. Such a model imposes a strict probabilistic length distribution including a minimum length that can be threaded onto or generated by the model. However, our generic model requires, by definition, the accommodation of sequences of near arbitrary length. We achieved this goal by introducing a special state, the *end state*, and the *end symbol*. The latter is an artificial symbol placed at the end of the query sequence as the last amino acid. The end state is positioned after the last plex in the model and has a transition path back to the first state in the last plex with a probability of 1 (Fig. 1A). This state plays a dual role in the filtering process. During iteration of the forward filtering process, this end state is equivalent to a loop state, emitting the 20 amino acids according to the same amino acid emission probability distribution. When the end symbol is filtered through the end state, it acts as if its emission probabilities were zero for all the 20 amino acids and one for the end symbol. Since all of the other states in the DSM have zero emission probabilities for the end symbol, this forces the end of the sequence to be aligned to the end of the model.

Generic All-β DSM

The generic all-β model is similar in architecture to the all-α model, but with the α-plexes replaced by β-plexes (Fig. 2). Additionally, a minimum of four β-plexes is required, there is a slightly higher probability for a buried strand, and the connecting region between adjacent strands can be a short helix, in addition to a loop or turn. The β-strand module has a uniform length distribution between 3 to 14 residues. The loop module is the same as used in the generic all-α model.

Generic α/β DSM.

Figure 3 depicts the generic α/β model, which is designed to allow an alternating helix/strand architecture while permitting two consecutive repeats of each with small probabilities. It is more probable to (1) find a strand at the N-terminal of such proteins, and (2) have two consecutive strands than two consecutive helices. Module lengths in this model have a more constrained distribution: the lengths of the strands are uniformly distributed between 3 to 10 residues and the lengths of the helices are uniformly distributed between 4 to 23 residues. Note that the model has two end states but the computation of model likelihood remains the same.

Generic α+β DSM and β+α DSM

The α+β structural class has segregated helical and strand regions within its structure and often within the primary sequence as well. We have designed two DSMs to represent the two principle types of α+β structures: the

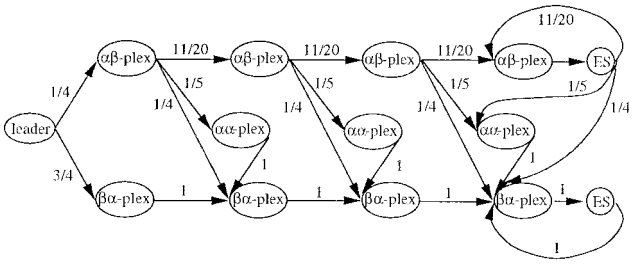

Fig. 3. Schematic of the generic model for α/β proteins. Each αβ-plex consists of a α-helix followed by a β-strand. Each βα-plex consists of a β-strand followed by a α-helix. An αα-plex contains two consecutive helices. In a αβ-plex or βα-plex, there is a parallel path with the helix (not shown), allowing a loop to replace the helix between two adjacent strands with a small probability. All the α-helix modules are identical, as are the β-strand modules.

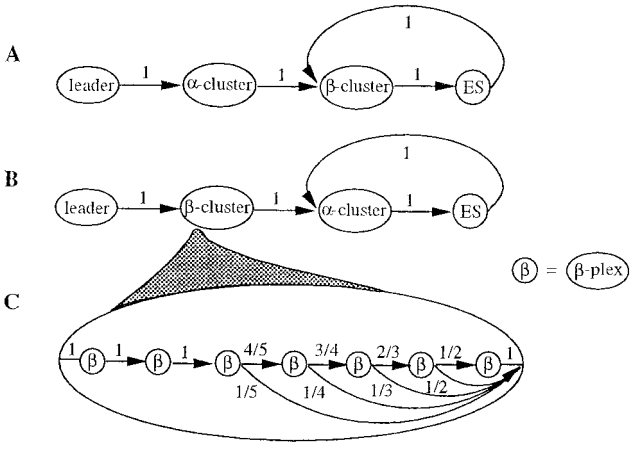

Fig. 4. Schematic of the generic model for α+β proteins. A: Generic α+β model contains an α-cluster followed by a β-cluster. B: Generic β+α model contains a β-cluster followed by an α-cluster. C: The structure of a β-cluster, which generates a subsequence of 3 to 7 strands with equal probabilities. Similarly, an α-cluster generates a subsequence of 4 to 7 helices with equal probabilities.

generic α+β model allows a α-cluster (minimum of 4 helices) followed by a β-cluster, while the generic β+α allows a β-cluster followed by an α-cluster. Both models represent two separate structural classes since they are independent of each other in the model space (Fig. 4).

There are a few protein folds in this category that we did not attempt to model. One example is a protein fold with a α-cluster - β-cluster - α-cluster topology. A model for this fold will have many helical elements and will be expected to overlap with the all-α model to a large extent. In fact, most of these proteins can usually be further divided into separate domains after visual inspection.

Generic DSM for irregular proteins

The generic model for small, irregular/metal-binding proteins consists mainly of loops and turns. The amino-acid emission probabilities for the loop state in this model differ from those for loop states in other generic models in that they have much higher probabilities for the amino acids histidine and cysteine (Fig. 5).

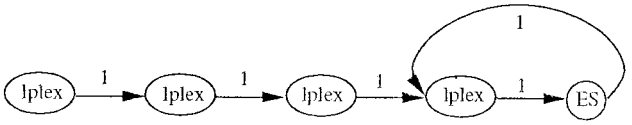

Fig. 5. Schematic of the generic model for irregular/metal-binding proteins. Each lplex is composed of a loop followed by a turn module.

Generic DSM for transmembrane β-barrel proteins

Transmembrane β-barrel proteins, which are found in the outer membranes of bacteria, mitochondria, and chloroplasts, fall into three main categories of solved structure: outer membrane proteins, outer membrane phospholipase A, and porins.[20] The size of these proteins ranges from small eight-stranded to large twenty-four-stranded β-barrels. There are three common features for this type of protein: (1) the number of β-strands is even, (2) strand connections on the periplasmic side of the membrane are generally short turns, while on the external side they are generally long loops or very short helices, and (3) transmembrane β-strands are amphipathic, composed of alternative polar (inside barrel) and non-polar (outside barrel) residues.

The generic model for this class has an architecture very similar to the generic all-β model with the principle differences accounting for the features as described above. The minimum number of strands is eight and an even number of strands is required. The residue positions at the barrel inside and outside are modeled by two special structural states whose emission probabilities were pre-calculated for this class of proteins.

Generic DSM for transmembrane α-helical proteins

A generic DSM representing α-helical membrane-spanning proteins was designed based on known structural features of these proteins. It contains transmembrane helical regions connected by intervening regions allowing loops, turns, and helices. The structural module for the transmembrane helical region consists of the hydrophobic core region flanked by membrane interface regions at both ends (Fig. 6). Two structural states were defined to model the membrane core and interface. Their emission probability distributions were calculated from a statistical analysis of a set of 56 known transmembrane proteins from ABC transporter complexes extracted from the BMERC all-genome database (http://bmerc-www.bu.edu/information/all_genomes.shtml). The transmembrane helix has a uniform length distribution between 20 and 38 residues, corresponding to the approximate length of a helix needed to span a typical bilayer.[21]

Models were also constructed at the fold level for α-helical transmembrane proteins. Due to the paucity of membrane protein structures in the PDB, the automated method cannot be relied on for a complete fold coverage of this class. Therefore, 56 individual models were manually designed for proteins containing between 5 and 12 transmembrane helical regions with the intervening regions allowing loops, turns, or helices.[22] These models do not represent the SCOP folds and are classified based on the

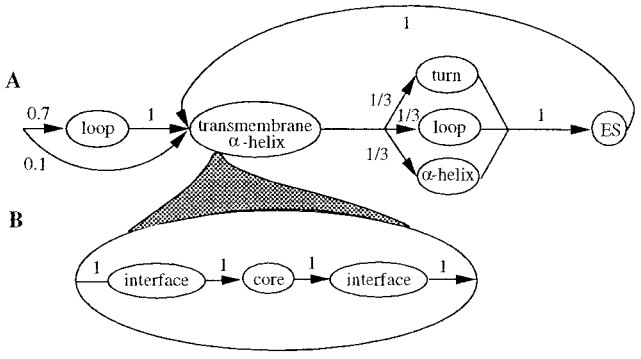

Fig. 6. Schematic of the generic model for the transmembrane helical proteins. **A:** The model contains a transmembrane helix followed by either a loop, turn, or non-membrane-spanning helix. **B:** The structural module for the transmembrane helix, where the membrane core region is flanked by the interface regions.

number of transmembrane helices being modeled. These specific models can be used to predict individual transmembrane regions for a given protein sequence using the smoothing algorithm, as explained.[16,8,22]

Triage Method for Fold Prediction

Step 1. Protein structure classification

The first step of the triage method is to attempt to identify the most probable structural class of a query sequence using only the set of generic DSMs. The model likelihood is computed for each generic model using the optimal filtering algorithm. The posterior probability of each model is then calculated according to Eq. 1 with equal model prior probabilities. This prior assignment models the minimum bias for structural class. Our aim is to reduce the number of PDB domain DSMs to be considered for subsequent fold prediction.

Step 2. Protein fold prediction

All of the PDB domain DSMs are considered for each structural class for which the generic model has a posterior probability greater than 0.3 (an empirical value). The method of categorizing the PDB domain DSMs into different structural classes will be discussed later. If the structural class for a query sequence is uniquely determined to be a transmembrane β-barrel protein, no further analysis is made. Otherwise the fold prediction is made based on the set of selected PDB domain models.

Consistent with previous DSM analysis,[7] we consider each PDB domain model as a structural hypothesis and compute the probability that the primary sequence came from each of the candidate hypotheses, $P(M|seq)$, using the model priors and likelihoods (Eq. 1). Fold prediction is made according to the probabilities of the SCOP folds, which are obtained by summing the probabilities of all PDB domain models that belong to the given fold (union of hypotheses):

Formula 1 (hierarchical scheme):

$$P(fold_i|seq) = \sum_{M\epsilon\, fold_i} P(M|seq), 1 \le i \le N_f, M\epsilon\{M_j\}$$

$$for: \quad 1 \le j \le N_m, \quad (2)$$

where N_f is the number of unique folds being considered and N_m is the total number of PDB domain DSMs being considered.

The model priors, $P(M)$, are assigned in a hierarchical manner as in White et al.[7] The underlying assumption is that all models at the same hierarchical level are exclusive, all-inclusive, and equally possible. The SCOP hierarchy is used here. This approach has also been used by Bienkowska et al.[14] for a small set of automatically generated DSMs. We denote this approach as the hierarchical scheme.

Here we propose an additional approach to fold prediction. Because all the models under a specific SCOP fold represent the same fold, i.e., they are overlapping models, it is not proper to treat them as individual hypotheses in the Bayesian formula. Instead, we view the hypothesis space as composed of distinct SCOP folds, that is, a given sequence is considered as coming from one, and only one, of the SCOP folds. Since there is more than one PDB domain DSM that represents a given fold, we only select the constituent DSM with the highest model likelihood to represent the fold and compute the probability of the SCOP fold using Bayes' rule:

Formula 2 (equal-fold-probability scheme):

$$P(fold_i|seq) = \frac{P(seq|fold_i)P(fold_i)}{\sum_{k=1}^{N_f} P(seq|fold_k)P(fold_k)}. \quad (3)$$

Prior probabilities are uniformly assigned to each fold selected, $P(fold_k) = 1/N_f$, to minimize the bias for different folds. $P(seq|fold_i)$ is the model likelihood of the representative DSM for $fold_i$:

$$P(seq|fold_i) = max\{P(seq|M), M\epsilon\, fold_i\}, 1 \le i \le N_f. \quad (4)$$

We name this approach as the equal-fold-probability scheme.

Finally, fold prediction is made if the most probable SCOP fold has a posterior probability, $P(fold|seq)$, greater than 0.5 according to either Formula 1 (Eq. 2) or Formula 2 (Eq. 3).

Categorization of PDB Domain DSMs

The success of the triage method will largely depend on the success of the initial classification. The grouping of the PDB domain DSMs into different generic DSM classes should be consistent with the classification of a given sequence, thus allowing the correct domain model to be selected in the subsequent fold prediction.

Initially, the library of PDB domain models was organized according to the SCOP classification, so it seems to be straightforward to group them into different structural classes according to their SCOP *class* assignment. However, in the top-level of triage, the structural classes defined by our generic DSMs do not universally correspond to the SCOP classification. The generic DSM design emphasizes the composition of secondary structure elements and, very importantly, their topological order, while the SCOP

TABLE I. Categorization of PDB Domain DSMs Into Structural Classes

	Mainly-α	Mainly-β	α/β	Irregular	TM-β	TM-α
N_{DSMs}	652	536	318	7	0	6

TABLE II. Overlap Between Generic DSMs†

Sequence generator	α	β	α/β	α+β	β+α	Irregular	TM-β	TM-α
α	**195**	0	3	2	0	0	0	0
β	1	**163**	3	6	20	0	0	0
α/β	10	1	**174**	3	9	0	0	0
α+β	9	18	7	**145**	10	0	0	0
β+α	13	4	0	0	**176**	0	0	0
Irregular	6	0	0	0	0	**186**	0	0
TM-β	0	1	0	1	0	0	**198**	0
TM-α	2	0	1	0	0	0	0	**195**

†Each row shows the total number of predictions made for each structural class given the 200 simulated sequences generated from the generic model in the left most column.
The numbers in bold indicate the number of correct class predictions for each class.

classification is mainly based on the spatial arrangement of secondary structure elements as judged by human experts. It is, thus, not surprising that the structural classification using our generic DSMs does not always agree with the SCOP classification (unpublished data). We have, therefore, categorized each PDB domain DSM into the predicted structural class(es) (for which the generic model has a posterior probability greater than 0.3) based on its template PDB sequence.

Using the above procedure for the 1,282 PDB domain DSMs in the library, 1,047 (82%) were mapped into a single class, 233 (18%) into two classes, and two into three structural classes. The classification of the model library can be found at http://bmerc-www.bu.edu/hxian/triage/. Table I shows the number of PDB domain models in each structural class. Note that the structural classes listed in Table I are slightly different from those introduced in the above section. This difference will be explained in Results in addition to the three misclassifications that resulted in classifying six domain models into the transmembrane α-helical class.

RESULTS

Probabilistic Overlap Between Generic DSMs

Each DSM can be viewed as a sequence generator that can generate a string of amino acids with a certain probability. The generation procedure draws amino acids randomly based on the state transition probability matrix and emission probability matrix of a particular model. We have used this random sequence generation method to measure the overlap or degree of independence between generic DSMs, as described below.

To validate the use of generic DSMs for the initial classification of triage, we measured the overlap between these models through the following simulation. Two hundred artificial amino acid sequences were generated from each generic model. Each of the 200 sequences was then threaded through the full set of generic models and the model posterior probabilities were computed. If a model has a posterior probability above 0.5 for a sequence, a prediction was recorded. The results are shown in Table II. Generic models for irregular proteins and transmembrane proteins show little overlap with other generic models. The all-α model and all-β model are also clearly separated from each other, although a small overlap is found between the all-α and α/β, and between the all-β and α/β model. A large overlap is found between the all-α and β+α model, and between all-β model and α+β model. When we combined each of the above pairs of overlapping models into a single class, mainly-α and mainly-β, better results were obtained. To compute the posterior probability for mainly-α or mainly-β class, the likelihood for the class takes the higher likelihood value of the two generic models from that class. The overlap between different structural classes after the above adjustment is shown in Table III.

Protein Fold Prediction

We first tested the triage method in the self-threading using the set of 882 template proteins. The results of fold prediction are shown in Table IV. To illustrate the triage method, we compare it with a simple prediction method, where the query sequence is filtered through every model in the library and the posterior probabilities are calculated using all the models. Furthermore, for each of the above two methods, we test the two formulas of computing the fold probabilities according to Eq. 2 and Eq. 3, respectively.

Among the 882 template proteins, 703 were uniquely classified, 154 classified into two classes, and one into three classes. On average, there were 672 PDB domain models being considered in the final fold prediction for each sequence. This number is only half of the size of the PDB domain model library. The triage method, as compared to the simple method, gave many more predictions (hits) using either the hierarchical scheme (Formula 1) or the equal-fold-probability scheme (Formula 2), when using 0.5 as the cutoff for the fold probability. This cutoff reflects our confidence level for prediction. The triage method also gave more correct predictions (true hits), resulting in a

TABLE III. Overlap Between Structural Classes With Reduced Number of Classes†

Sequence generator	Mainly-α	Mainly-β	α/β	Irregular	TM-β	TM-α
Mainly-α	**197**	0	3	0	0	0
Mainly-β	22	**172**	4	0	0	0
α/β	19	4	**175**	0	0	0
Irregular	6	0	0	**186**	0	0
TM-β	0	2	0	0	**198**	0
TM-α	2	0	1	0	0	**196**

†Each row shows the total number of predictions made for each structural class given the 200 simulated sequences generated from the generic model in the left most column.
The numbers in bold indicate the number of correct class predictions for each class.

TABLE IV. Results of Fold Prediction†

Self-threading				Distant-homolog-threading			
Triage method		Simple method		Triage method		Simple method	
Formula 1	Formula 2	Formula 1	Formula 2	Formula 1	Formula 2	Formula 1	Formula 2
502	**539**	443	483	**40**	43	31	39
423/79	**458/81**	373/70	414/69	**22/18**	21/22	17/14	19/20

†In the first row is the number of total hits using $P_{fold} \geq 0.5$ as the cutoff. In the second row is the number of true positives versus false positives, using $P_{fold} \geq 0.5$ as the cutoff.
The bold values indicate the best result from all four methods used.

similar true/false hits ratio as the simple prediction method. Thus, the triage method is more sensitive by predicting more true hits without sacrificing specificity. Additionally, the use of the equal-fold-probability scheme in computing the fold probabilities resuted in more true hits as compared to the hierarchical scheme. Overall, the triage method combined with the equal-fold-probability scheme gave the best performance.

Second, we applied the triage method to a set of 110 proteins identified as distant homologs to the model template proteins. Each member of this set (target protein) has a structural similarity to one of the template proteins as classified by SCOP, while having a low sequence similarity (cannot be identified by normal BLAST search with e-value cutoff of $1e^{-3}$). The results of fold prediction are shown in Table IV. Similarly, the triage method gave the most correct predictions. Unlike the self-threading test, the hierarchcial scheme performed slightly better than the equal-fold-probability scheme. This most likely results from the divergence of the target protein to the template protein. By assigning hierarchical priors to every constituent model under a particular fold and summing the posterior probabilities over all the constituent models as the fold probability, more true hits and fewer false hits resulted.

Most of the successful methods of identifying distant homologs utilize the conserved sequence information in multiple sequence alignments, while our method works without reference to the sequences of the template proteins. There is, therefore, no benefit in comparing them directly. However, our method provides an independent resource for protein structure prediction by modeling only the structural preferences but not the native sequence information in a structural model. To show this, we searched the PDB database for the corresponding template protein of the target proteins using a PSI-BLAST search.[23] The target sequence was used as a query to search against the nonredundant sequence database at National Center for Biotechnology Information (NCBI) for up to five iterations. The position-specific matrix generated from the last round was then saved and used to search against the PDB database, with the e-value cutoff set as $1e^{-5}$. This PSI-BLAST search yielded 61 correct predictions. Our method gave only 21 correct predictions (using Formula 2), five of which were not detected by the PSI-BLAST search. This proves that the structural preferences encoded in the DSM can drive the correct fold identification when no sequence similarity can be detected.

As mentioned above, our fold prediction method works in the complete absence of any template sequence information, which may explain the low number of fold predictions. It has been shown that including the sequence information in the threading approach can greatly improve the fold prediction accuracy. Yu et al. have shown that the sensitivity of the DSMs can be improved through the embedding of conserved sequence pattern within a structural model.[24] At the time of writing this article, we have implemented a protocol to automatically embed a minimum pattern in a PDB domain DSM.[25] Out of the total, 1,066 domain models had a counterpart with an embedded sequence pattern, resulting in the final model library consisting of 2,348 PDB domain DSMs with or without embedded patterns. Since a PDB domain DSM obviously overlaps with its pattern-embedded counterpart, the fold probability is computed according to Formular 2, the equal-fold-probability scheme. We ran the distant-homolog-threading test with this enlarged library, and the results of fold prediction were compared with the PSI-BLAST search (Table V). The two methods yielded a similar number of true positives, where each method

TABLE V. Results of Fold Prediction Using Pattern-Embedded DSMs†

Pattern-embedded DSMs		
Triage method	Simple method	PSI-BLAST
55/12	54/10	61/16

The numbers shown here are the number of true hits vs. false hits using $P_{fold} \geq 0.5$ as the cutoff.

TABLE VI. Results of Structural Classification

	Structural classification		
	Soluble protein	TM-α protein	TM-β protein
Soluble proteins (882)	881	1	0
TM-β (12)	0	0	12
TM-α (16)	1	15	0

correctly predicted some unique proteins. Of the 55 true positives identified by pattern-embedded DSMs, twelve were not found by PSI-BLAST.

Classification Between Transmembrane and Soluble Proteins

We took all the non-identical transmembrane proteins from the PDB to test the ability of the set of generic DSMs to differentiate soluble from transmembrane proteins. This set includes 16 transmembrane α-helical proteins and 12 transmembrane β-barrel proteins. The 882 template soluble proteins were used as negative controls. Table VI summarizes the results of structural classification between the soluble proteins and transmembrane proteins.

Among the 882 template proteins, only one (SCOP ID: 1uag_1) was uniquely predicted to be a transmembrane helical protein. This sequence is the N-terminal domain (residue 1-93) of a D-glutamate ligase and has a stretch of 20 hydrophobic residues in a helix, which possibly explains the misclassification. Only one transmembrane helical protein (SCOP ID: 1eulA) was misclassified as a soluble protein. The structure reveals that it has two large inserted globular domains (143 and 444 residues) between adjacent transmembrane helices; thus, it is not surprising that a generic model for soluble proteins had a higher probability (Note: we are not currently modeling the globular domains that often appear within the adjacent membrane-spanning helices in the transmembrane helical protein model). There are also two cases where the template protein was classified into both the transmembrane helical protein structural class and a soluble structural class with low probabilities. Since their corresponding domain models are not uniquely classified into the structural class of transmembrane helical proteins, these misclassifications were ignored.

Secondary Structure Prediction

Once the most probable PDB domain model (the one that represents the given fold) is determined for a given sequence, it can be used to compute the probability of each residue being in a particular secondary structure element. These probabilities are calculated using the smoothing or forward-backward algorithm.[7,16] The smoothing algorithm computes the probability of each residue being in a structural state given the entire sequence. The probability of a residue being in a secondary structure element is then obtained by summing the probabilities of the residue being in all the structural states belonging to the secondary structure element. The result can be visualized in a contour map, as shown in Figure 7. Since each PDB domain model represents a real PDB structure with the structural states corresponding to the structural positions in the structure, the smoothing results can be used to guide the alignment of the query sequence to the PDB structure. This sequence-structure alignment can then provide a starting point for a detailed modeling for the query sequence.

DISCUSSION

We present an extension of the DSM approach[7,8] that provides a greatly expanded structural template library covering most of the known structural folds, as classified by the SCOP database. This involves two major innovations: first, an automated procedure for model generation and, second, a triage method for protein fold prediction that is applicable to any such large model library.

The automated method of model generation constructs a structural DSM directly from a protein structure deposited in the PDB. It inherits many advantages of the DSM design methodology. It requires only a single structure to build the model and thus is able to construct a structural model for a protein with no or few homologs. In contrast, the conventional profile HMM construction usually requires about 20 homologous sequences to train the model. [9,10] The DSM design also explicitly incorporates expert knowledge into the model design.[14] The automation of this method gives us the ability to expand the library as each new structure is identified.

The triage method relies on a set of manually designed high-level or generic models. These generic models are used to classify a given sequence into its probable structural class before the fold prediction is made, improving the chance that the correct fold can be identified from within the large model library. This approach is particularly useful when using a Bayesian or similar likelihood evaluation approach.

Since our fold prediction method uses a Bayesian estimation approach, which will always predict one model as the most probable, this requires the model library to be as complete as possible. However, this very completeness can result in a small posterior probability for the correct model when all PDB domain models in the library are used for calculating the posterior probabilities. The triage method limits this effect by reducing the final number of PDB domain models used in computing the posterior probabilities. This reduction in the model set also reduces the computational cost. To ensure that the set of generic DSMs is suitable for the initial structural classification, we have shown that these generic models are exclusive from each

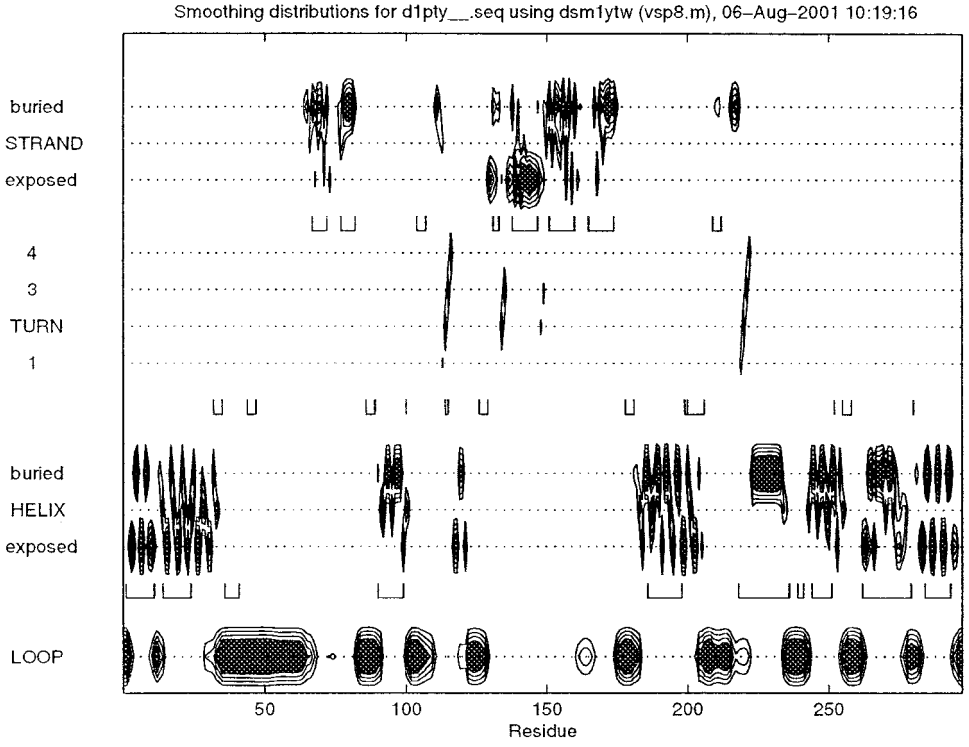

Fig. 7. Contour map of secondary structure probabilities calculated by the smoothing algorithm for the sequence of tyrosine phosphatase (PDB ID: 1pty). The model used was the best fold model predicted for this sequence using the triage method described in the text. The template protein has less than 30% sequence identity to the query sequence. The contours were drawn at 0.2, 0.3, 0.4, and 0.5, and the solid areas correspond to the region with probability above 0.5. The *y*-axis denotes the secondary structure element types and the *x*-axis denotes the residue position along the sequence. The helices and strands in the known structure, identified by DSSP, are denoted with solid lines.

other and cover most of the known structures. We expect that other methods can also be used to make the initial structural classification in the proposed triage method. For example, one can predict the structural class of a protein from its amino acid composition[26] or any secondary structure prediction method.[27–29] We propose that better results can be achieved by applying the same prediction method in a consistent manner both for classifying the PDB domain DSMs and for the initial structural classification.

We present a new scheme, the equal-fold-probability scheme, for assigning model priors and calculating fold probabilities in addition to the previous hierarchical scheme. Our results show that both schemes work but have different strengths. The hierarchical scheme minimizes the bias within and between structural folds. This performs better when none of the constituent models closely resembles the structure of the query sequence. Under the equal-fold-probability scheme, only one constituent model is used to represent the fold for calculating fold probabilities, and since the model likelihood is the only evidence for a given model in light of the data (observed sequence), the model with the highest likelihood is selected as the fold representative. This scheme works better when there is one model, among all those belonging to a particular fold, that most closely represents the structure of the query sequence. However, only the equal-fold-probability scheme is used in the case of the pattern-embedded PDB domain models. Both schemes are available on our web server at http://bmerc-www.bu.edu/psa-new/.

As mentioned earlier, our fold prediction method works in the complete absence of any template sequence information. It has been shown that the sensitivity of the DSMs can be significantly improved by embedding conserved sequence patterns within a structural model.[24] After including the pattern-embedded PDB domain models in our model library,[25] we achieved much greater prediction accuracy. The performance is comparable to that of PSI-BLAST. However, embedding a conserved pattern in a structural model restricts the utility of such a model to only the functional domain family containing the specific sequence motif. Therefore, such models are less useful in predicting the folds of proteins with few homologs.

One limitation to our method is that these automatically constructed DSMs can only detect proteins with fairly similar topology. Although the models allow varying lengths of loops and secondary structural elements as well as additional structural states modeling the N- and C-termini, the number of secondary structural segments in a model is fixed. This is inadequate since many distant homologs can have large insertions/deletions that can result in variations of the number of secondary structural elements. In our current model library, we have generated

models representing each SCOP family under each representative SCOP fold in order to capture the structural variations within each fold. However, this is limited by the available structural families within a particular fold. Another potential solution might be to generate *consensus models* allowing all the structural variations observed and anticipated within each SCOP fold, including variations in both the number and length of secondary structural elements. A second limitation results from the fact that many models in our current library are quite similar in terms of the linear sequence of secondary structural elements, thus there is a potential for multiple models to overlap. As an example, three SCOP folds, the NAD(P)-binding Rossmann-fold domain, the FAD/NAD(P)-binding domain, and the nucleotide-binding domain all adopt a three-layer, α/β/α structure with a central parallel β-sheet of 5/6 strands. Models generated from these similar proteins will overlap. The competition between similar models will result in the reduction of their individual posterior probabilities. This overlap offers a possible explanation for the low number of fold predictions in both the self-threading and the distant-homolog threading. The construction of a consensus model is expected to ease this overlap problem.

ACKNOWLEDGMENTS

We thank Jadwiga Bienkowska for her helpful discussions about the model construction and Robert Rogers, Jr., for setting up and maintaining the web server. We also thank Nancy Sands for careful proofreading of the manuscript.

REFERENCES

1. Bowie JU, Clarke ND, Pabo CO, Sauer RT. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 1990;7:257–264.
2. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–253.
3. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
4. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 1992;13:258–271.
5. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 1992;227:227–238.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
7. White JV, Stultz CM, Smith TF. Protein classification by stochastic modeling and optimal filtering of amino acid sequences. Math Biosci 1994;119:35–75.
8. Stultz CM, White JV, Smith TF. Structural analysis based on state-space modeling. Protein Sci 1993;2:305–314.
9. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 1994;235:1501–1531.
10. Eddy SR. Hidden Markov models. Curr Opin Struc Biol 1996;6:361–365.
11. Richardson JS. The anatomy and taxonomy of protein structures. Adv Protein Chem 1981;34:167–339.
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Wesissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of protein database for the investigation of the sequences and structures. J Mol Biol 1995;247:536–540.
14. Bienkowska JR, Yu L, Zarakhovich S, Rogers Jr RG, Smith TF. Protein fold recognition by total alignment probability. Protein Sci 2000;40:451–462.
15. White JV. Modeling and filtering for discretely valued time series. In: Spall JC, editor. Bayesian analysis of time series and dynamic models. New York: Marcel Dekker. 1988. p 255–283.
16. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc Inst Electric Electron Eng 1989;77:257–286.
17. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
18. Eisenberg D, MacLachlan AD. Solvation energy in protein folding and binding. Nature Jan, 1986319:199–203.
19. Bakis R. Continuous speech word recognition via centisecond acoustic states. In: Proc ASA Meeting, Washington, DC. April, 1976.
20. Schulz GE. β-Barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.
21. Engelman D, Steitz T, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 1986;5:321.
22. McAllister GD. Modeling of transmembrane proteins using discrete state-space models. Master's thesis, Boston University, Boston, MA, June 2001.
23. Altschul SF, Madden T, Schaffer A, Zhang J,-Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402.
24. Yu L, White JV, Smith TF. A homology identification method that combines protein sequence and structure information. Protein Sci 1998;7:2499–2510.
25. Bienkowska JR, He H, Smith TF. Automatic pattern embedding in protein structure models. IEEE Intell Syst 2001;16:21–25.
26. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem Tokyo 1986;99:153–162.
27. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 1988;202:865–884.
28. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 1994;90:7558–7562.
29. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.