Structural analysis based on state-space modeling

COLLIN M. STULTZ,^{1,2} JAMES V. WHITE,³ AND TEMPLE F. SMITH²

¹Committee on Higher Degrees on Biophysics, Harvard University, Cambridge, Massachusetts 02138

²BMERC, Boston University, 36 Cummington Street, Boston, Massachusetts 02215

³TASC, 55 Walkers Brook Drive, Reading, Massachusetts 01867

(RECEIVED September 10, 1992; REVISED MANUSCRIPT RECEIVED November 16, 1992)

Abstract

A new method has been developed to compute the probability that each amino acid in a protein sequence is in a particular secondary structural element. Each of these probabilities is computed using the entire sequence and a set of predefined structural class models. This set of structural classes is patterned after Jane Richardson's taxonomy for the domains of globular proteins. For each structural class considered, a mathematical model is constructed to represent constraints on the pattern of secondary structural elements characteristic of that class. These are stochastic models having discrete state spaces (referred to as hidden Markov models by researchers in signal processing and automatic speech recognition). Each model is a mathematical generator of amino acid sequences; the sequence under consideration is modeled as having been generated by one model in the set of candidates. The probability that each model generated the given sequence is computed using a filtering algorithm. The protein is then classified as belonging to the structural class having the most probable model. The secondary structure of the sequence is then analyzed using a "smoothing" algorithm that is optimal for that structural class model. For each residue position in the sequence, the smoother computes the probability that the residue is contained within each of the defined secondary structural elements of the model. This method has two important advantages: (1) the probability of each residue being in each of the modeled secondary structural elements is computed using the totality of the amino acid sequence, and (2) these probabilities are consistent with prior knowledge of realizable domain folds as encoded in each model. As an example of the method's utility, we present its application to flavodoxin, a prototypical α/β protein having a central β -sheet, and to thioredoxin, which belongs to a similar structural class but shares no significant sequence similarity.

Keywords: flavodoxin; hidden Markov model; secondary structure; state space modeling; tertiary structure classification; thioredoxin

The number of known protein sequences greatly exceeds the number of directly determined structures. This is due in part to the relative ease of protein sequence determination by DNA sequencing as compared to the difficulty and expense of structure determination by X-ray crystallography and NMR spectroscopy. Although the number of known structures continues to grow, the number of protein sequences is expected to greatly exceed the number of known structures for the foreseeable future. Because knowledge of the structure of a protein is essential to understanding its function, the elucidation of any aspect of a protein's structure from the sequence information alone is potentially useful.

Although there are over 600 structures in the Brookhaven protein structural database (Bernstein et al., 1977;

Abola et al., 1987), comprising over 100 nonhomologous structures (Hobohm et al., 1992), it is well known that a number of structural themes recur. One example of such a recurring structure is the Rossman fold (Rossman et al., 1974), found in such proteins as lactate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, and alcohol dehydrogenase (Branden & Tooze, 1991). Although these proteins have no significant sequence similarity outside of the nucleotide-binding region, they have very similar α/β tertiary structures: they are composed of a parallel six-stranded central β -sheet with mostly α -helical returns.

A major attempt to classify the known structures was undertaken by Jane Richardson in her 1981 review (Richardson, 1981), in which she systematically identified common structural patterns and developed a taxonomy of structural domain classes. Examples of such structural classes include the eight-stranded β -barrel, as exemplified by triose phosphate isomerase, and the four-helix antipar-

Reprint requests to: Temple F. Smith, BMERC, Boston University, 36 Cummington Street, Boston, Massachusetts 02215.

allel bundle, as exemplified by myohemerythrin. In this paper, we address the following questions: Given such a comprehensive set of predefined tertiary structural classes and the amino acid sequence of a protein of unknown structure, to which of these structural classes does this sequence most probably belong? Furthermore, given the most probable tertiary structural class, what is the secondary structure of the sequence?

Previous work on structural templates and profiles has addressed a related problem: Given a known tertiary structure, what sequences are compatible with that structure (Ponder & Richards, 1987; Bowie et al., 1991)? In particular, the method of Ponder and Richards combines the known tertiary structure with stereochemical rules to determine the probability that a sequence will fold into a known core structure. However, these methods do not include the generalizations of the core secondary structure and their connectivities, as is done in this study.

There have been a number of methods developed to identify the most likely structural class of a protein given only its amino acid sequence. Most of these methods have focused on determining the gross overall structural class (i.e., all α , all β , α/β , $\alpha + \beta$) based only on the analysis of average sequence properties such as amino acid composition (Nishikawa et al., 1983; Chou, 1989; Zhang & Chou, 1992). When this relatively coarse information is utilized to predict secondary structure, the accuracy of the predictions is increased (Deleage & Roux, 1989). In addition, previous work by Cohen and coworkers claims that turns (defined as those irregular regions between regular secondary structure) may be predicted with greater than 90% accuracy when the overall structural class is known (Cohen & Kuntz, 1989). Therefore, we expect that

a more detailed classification scheme should yield even more accurate secondary structure predictions. We have used an expanded predefined set of structural classes that closely follow the classifications of Jane Richardson (see Table 1). For each class considered, a discrete state-space model (DSM) is constructed as previously described by White et al. (1992).¹ Each model is an idealized representation of a particular tertiary structural class: it is intended to represent the important constraints imposed on primary and secondary structure by that tertiary structure. The models may be viewed as mathematical generators of amino acid sequences. The models are stochastic, and their most probable sequences are most compatible with the modeled tertiary structures.

Each DSM describes probabilistically the allowed secondary structural elements, their types, lengths, connectivity, and amino acid compositions for each class. The structural elements are modeled by a hierarchy of states in a Markov chain, with the transitions between structural elements defined by a set of transition probabilities. The structural elements considered include grouped secondary structures, single secondary structures, and relative amino acid positions within those secondary structures (see Fig. 1). The timelike parameter of the Markov chain is identified with the position of each residue in the primary sequence. Fifteen domain structural classes are currently modeled (see Table 1).

We use particular groupings of secondary structural elements, such as the β plex in Figure 1, as a convenient way to organize the secondary structural elements in the

¹ As discussed in White et al. (1992), the structural profiles of Bowie et al. (1991) are a limiting case of similar models.

Table 1. Discrete state-space models (DSMs) for tertiary structural classes

Model no.	Super class	Macro class	Example protein	Mean no. of elements		
				Helices	Strands	Turns
1	α	α -Box	Hemoglobin	4-7	0	3.42
2	α	Antiparallel bundle	Myohemerythrin	4	0	2.68
3	α	Mixed helix lengths	λ repressor (one domain)	5.57	0	2.95
4	α/β	Central β -sheet	Glutathione peroxidase	5.75	8-10	7.65
5	α/β	Central β -sheet	Flavodoxin	3.75	5-7	5.12
6	α/β	Barrel	Triose phosphate isomerase	7.47	8	9.45
7	α/β	Mixed helix and strand lengths	Lactate dehydrogenase (domain 2)	4.53	4-12	6.97
8	β	β -Sheets	Proteinase complex	0.91	11-13	6.21
9	β	β -Sheets	Tomato bushy stunt virus	0.73	8-10	4.70
10	β	β -Sheets	Concanavalin A	0.55	5-7	3.19
11	β	Mixed strand lengths	Rhizopuspepsin	0.72	4-12	4.65
12	Irregular	Cysteine-rich	Insulin-like growth factor	0	0	1.32
13	Irregular	α -Helices	Ferrocyclochrome	1.35	0	0
14	Irregular	β -Strands and β -turns	fd phage gene 5	0	2.10	3.40
15	Irregular	α -Helices and β -strands	Ferrodioxin	1.5	3.0	1.35

Fig. 1. Schematic of model for the α/β structural class having a central β -sheet containing five, six, or seven β -strands. **A:** The DSM organization at the β plex level of detail. **B:** The β plex is composed of secondary structural elements. **C:** One of these elements, the β -strand submodel, which has a chain of discrete states at the amino acid position level of detail.

construction and representation of our models. In Figure 1A, the transition probabilities between $\alpha\beta$ plexes are given for our central-sheet α/β class model. The model contains seven $\alpha\beta$ plexes, and each $\alpha\beta$ plex contains exactly one β -strand submodel. Because of this and the choice of transition probabilities, this model generates, with equal probability, only sequences compatible with five, six, or seven strands. Figure 1B portrays the contents of a β plex. Each plex contains a β -strand submodel, which is followed by either a loop or an amphipathic α -helix submodel. Note that strands are connected to helices by either a four-residue turn or a short loop. Figure 1C describes the contents of the β -strand submodel. In this particular case, it models a strand containing three to nine residues.

Given a sequence of unknown structure, we analytically determine the probability that each model has generated it, using an optimal Bayesian filtering algorithm (White, 1988; White et al., 1992). Once the most probable struc-

tural model is determined, the most probable secondary structure for each residue is calculated for the sequence using an optimal smoothing algorithm (see Methods) on the sequence with the most probable DSM (White, 1992).

Results

To demonstrate the utility of the approach, two sequences from the Brookhaven Protein Data Bank are analyzed: flavodoxin (locus 3fxn) and thioredoxin (locus 2trx). They have similar tertiary structures (both have central β -sheets with mostly α -helical returns), but no significant sequence similarity according to the algorithm of Smith and Waterman (1981).

Figure 2A shows the probabilities that the flavodoxin sequence was generated by each of 15 candidate DSMs. Model 5, in Table 1, has the highest probability and represents the class of α/β domains having a central β -sheet containing five, six, or seven strands. This model represents the true class to which flavodoxin belongs. Figure 3A shows the secondary structure probability map for flavodoxin as computed by the smoothing algorithm using model 5. This figure is a contour map of the marginal probability distribution of the 11 secondary struc-

Fig. 2. Probabilities of DSMs for (A) flavodoxin sequence and (B) thioredoxin sequence. In both cases, the most probable model is shown hatched.

Fig. 3. Secondary structure probability maps (smoothing probabilities of secondary structural elements using most-probable DSMs) for (A) flavodoxin sequence and (B) thioredoxin sequence. Probabilities are computed as described in the text. The contours are drawn at 0.2, 0.3, 0.4, and 0.5. The solid areas correspond to regions with probability of at least 0.5. The y-axis denotes the secondary structural elements and the x-axis denotes the positions along the amino acid chain. The atomic coordinates from the Brookhaven Data Bank Release 58, 1992, were used to identify the known β -strand and α -helical regions (Kabsch & Sander, 1983). Helices and strands in the known structure are denoted with solid lines. The dashed lines indicate an irregular helical region and the dotted regions indicate regions with a hydrogen bond pattern that is consistent with 3_{10} -helices. Regions labeled with + denote short regions with hydrogen patterns consistent with 3_{10} -helical structure (a 3-turn in Kabsch-Sander terminology) and regions labeled with o denote a Kabsch-Sander 4-turn. The matrix used to calculate the contour plot may be found on the Diskette Appendix.

tural element types for each residue of the sequence. The actual locations of the secondary structural elements, as determined from X-ray coordinates, are also shown (Smith et al., 1977; Kabsch & Sander, 1983). The regions of high probability for all the β -strands correspond well with their actual locations. Helices located at residues 10–25 and 94–105 correspond to regions of high proba-

bility, but the other two helices show a poorer correspondence with the probability maps. However, all of the actual helical and strand residues within the structure have probability above 0.50.

As with flavodoxin, the most probable generator of the thioredoxin sequence is model 5 (see Fig. 2B). However, while the flavodoxin sequence was correctly classified with a probability of 0.60, the most probable model for the thioredoxin sequence has the much smaller probability of 0.20. Thus, it is not surprising that the secondary structure probability map for thioredoxin in Figure 3B shows a slightly weaker correspondence with the crystallographically determined secondary structure assignments (Kabsch & Sander, 1983; Katti et al., 1990). Just as with flavodoxin, the strands in the actual structure correspond to regions of high probability in the probability map. Here most of the residues assigned a single secondary structure probability of over 0.50 are in fact found to be in such a structure.

To quantitate the information contained in Figure 3A,B the secondary structure probability maps were combined with a simple rule for predicting secondary structure. Let $p(s_k|A)$ denote the conditional probability that the k th amino acid residue a_k is in conformational state s_k , given all amino acids in the primary sequence A . An amino acid a_k is said to be in a particular secondary structural state s_k if $p(s_k|A) \geq \frac{1}{2}$. If $p(s_k|A) < \frac{1}{2}$ for all of the secondary structural elements considered, then the secondary structure of a_k is said to be indeterminate. For simplicity, the 13 states associated with model 5 were collapsed to 3 states: α -helix, β -strand, and all remaining states are combined into a "coil state." Residues whose structures are indeterminate were treated as coil residues. Using this rule, the overall accuracy is 77% for flavodoxin and 66% for thioredoxin. Furthermore, different secondary structural elements are predicted with different accuracies as shown in Table 2. Strands are predicted with a sensitivity of 76% and a specificity of 95% for flavodoxin. For thioredoxin, strands are predicted with a sensitivity of 70% and a specificity of 92%. Helices, on the other hand, are predicted well only for the flavodoxin sequence using this rule.

A secondary structural element (i.e., an α -helix or a β -strand) is said to be correctly predicted if more than 50% of the residues within that structure are correctly predicted. All of the helical structures are correctly predicted, and four of the six strands in this structure are predicted. For thioredoxin, only the second and longest of the four helices is correctly predicted while four of the five strands are predicted (see Table 2).

All α -helices in this model are amphipathic: they are modeled as having one side packed against β -strands (buried residues) and one side exposed to solvent (exposed residues). Buried states are distinguishable from exposed states only by the fact that exposed states have probability distributions that favor hydrophilic residues (see Table 3).

Table 2. Results on secondary structure prediction^a

	α -Helix	β -Strand
Flavodoxin		
Sensitivity	0.87	0.76
Specificity	0.86	0.95
Overall accuracy = 0.77		
Thioredoxin		
Sensitivity	0.47	0.70
Specificity	0.89	0.92
Overall accuracy = 0.66		

^a The results of the secondary structure probability maps are combined with a simple rule for predicting secondary structure (see text). The sensitivity of α -helical predictions is defined as the fraction of helical residues that are predicted to be in an α -helical conformation. The specificity of α -helical predictions is the fraction of nonhelical residues that are predicted to be nonhelical. Similar definitions apply for β -strand predictions.

Implicit in this assumption is the fact that hydrophobic residues tend to be buried and hydrophilic residues tend to be exposed (Chothia, 1974). For those helical regions within the structure, the actual pattern of exposed and

buried residues in the structure is shown in Figures 4 and 5. With this pattern the probability of each residue being in an exposed region of the helix is also shown. Peaks in the probability plots correspond to peaks in the solvent-accessible area plots for both flavodoxin and thioredoxin.

Discussion

Figure 3A,B displays a secondary structure marginal probability distribution for each residue. In several instances one residue has significant probability for more than one secondary structural element. In the case of flavodoxin, all of the actual β -strands within the structure correspond to regions with high probability for β -strands, but residue 60 has significant probability in both the strand and helical conformations. In this instance, the algorithm conveys conflicting information about the structure of this residue. Furthermore from residue 39 to residue 44 and 58 through 64, there is significant probability for amphipathic helix. Although there is no helix per se in this region of the structure, the Kabsch-Sander algorithm places residues 41–43 in a 4-turn: an H-bonded turn three residues long. The phi-psi values of residues 42 and 43 are

Table 3. Conditional amino acid probabilities in DSMs^a

Amino acid	Secondary structural element												
	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0.100	0.112	0.099	0.087	0.104	0.075	0.088	0.079	0.074	0.080	0.143	0.061	0.084
C	0.013	0.012	0.011	0.002	0.004	0.005	0.018	0.015	0.013	0.009	0.016	0.011	0.015
D	0.102	0.053	0.074	0.077	0.078	0.131	0.071	0.059	0.029	0.086	0.021	0.054	0.010
E	0.087	0.077	0.072	0.087	0.091	0.075	0.071	0.054	0.038	0.126	0.030	0.072	0.013
F	0.031	0.047	0.036	0.019	0.009	0.010	0.018	0.033	0.053	0.019	0.074	0.025	0.073
G	0.059	0.043	0.067	0.068	0.091	0.176	0.079	0.064	0.039	0.044	0.042	0.046	0.034
H	0.015	0.028	0.034	0.010	0.026	0.035	0.035	0.024	0.027	0.037	0.020	0.041	0.017
I	0.028	0.062	0.056	0.048	0.009	0.020	0.062	0.045	0.111	0.025	0.098	0.052	0.155
J	0.046	0.004	0.023	0.019	0.013	0.025	0.003	0.043	0.006	0.004	0.003	0.007	0.005
K	0.074	0.077	0.088	0.077	0.082	0.085	0.071	0.064	0.044	0.125	0.030	0.082	0.015
L	0.046	0.078	0.061	0.048	0.026	0.025	0.044	0.057	0.093	0.032	0.123	0.044	0.130
M	0.025	0.038	0.027	0.002	0.002	0.015	0.009	0.020	0.021	0.016	0.060	0.010	0.029
N	0.056	0.039	0.056	0.068	0.065	0.080	0.035	0.059	0.035	0.063	0.015	0.066	0.012
P	0.033	0.006	0.018	0.068	0.091	0.025	0.018	0.053	0.019	0.063	0.006	0.022	0.017
Q	0.043	0.048	0.047	0.058	0.052	0.035	0.079	0.047	0.025	0.079	0.019	0.046	0.009
R	0.023	0.038	0.040	0.019	0.060	0.050	0.018	0.033	0.037	0.062	0.015	0.069	0.013
S	0.082	0.055	0.067	0.077	0.095	0.065	0.079	0.077	0.073	0.056	0.054	0.086	0.064
T	0.069	0.068	0.045	0.068	0.060	0.045	0.053	0.075	0.075	0.070	0.067	0.088	0.065
V	0.051	0.081	0.056	0.058	0.022	0.015	0.106	0.064	0.137	0.033	0.128	0.064	0.191
W	0.003	0.012	0.007	0.010	0.004	0.005	0.009	0.010	0.018	0.008	0.015	0.014	0.020
Y	0.015	0.021	0.016	0.029	0.017	0.000	0.035	0.024	0.034	0.022	0.021	0.039	0.029

^a Structural elements are N-cap (1), average α -helix (2), C-cap (3), β -turn residues (4, 5, 6, 7), loop or coil (8), buried β -strand (9), exposed and buried residues of amphipathic α -helix (10, 11), and exposed and buried residues of amphipathic β -strand (12, 13). Probabilities are based on the following 16 example proteins from the Brookhaven Protein Data Bank and their homologs in the SWISS-PROT database: carbonic anhydrase, ferrocyclochrome, hemoglobin, insulin-like growth factor (somatomedin), glutathione peroxidase, proteinase complex, triose phosphate isomerase, cytochrome c', concanavalin A, myohemerythrin, Bence Jones λ chain (variable region), flavodoxin, lactate dehydrogenase (domain 1), lactate dehydrogenase (domain 2), plastocyanin, and Ca-binding parvalbumin (Abola et al., 1987; PDB loci: 1CA2, 1CYC, 1ECD, 1GF1, 1GP1, 1HVP, 1TIM, 2CCY, 2CNA, 2MHR, 2RHE, 3FXN, 3LDH, 3PCY, 4CPV). Note that the letter J corresponds to a "gap" character, which is present in aligned sequences. Because aligned sequences were not used in the current study, this row is not essential.

Fig. 4. Plots of solvent-accessible surface area and the probability of being in an exposed helical position for flavodoxin. **A:** Probability of being in an exposed helical state. (Only the probability maps for those regions that would be predicted to be helical by the method described in the text are shown.) **B:** The fraction of each residue within the protein that is exposed to solvent, with helical regions darkened. The solvent-accessible surface area of each residue was taken from the program of Kabsch and Sander (1983). This number was divided by the value of the accessible surface area of the amino acid in extended polypeptide chains (Eisenberg & McLachlan, 1986). For completeness, nonhelical regions are shown with dotted lines.

significantly different from the ideal α -helical value of -60° and -40° .

Benner et al. (1992) recently noted the difficulty of properly predicting the conformation of proteins from amino acid sequences. They noted that most methods predicted the folding class of a recently determined Src-homology 3 (SH3) domain (Musacchio et al., 1992) incorrectly as being mostly helical, while their multiple sequence approach correctly predicted a mostly β structure. In a test of our procedure, we correctly predict this domain to be in the β superclass (see Table 1) with probability 0.60. Given the most probable model from the filtering, our smoothing algorithm correctly predicts the secondary

Fig. 5. Plots of solvent-accessible surface area and the probability of being in an exposed helical position for thioredoxin. See legend of Figure 4.

structure of 68% of the residues (among α -helices, β -strands, and loops). This accuracy is to be compared to only 52% by Benner et al. (1992), whose prediction includes one helix incorrectly.

Our results are reasonable when it is noted that (1) the smoother can compute probabilities only for those secondary structure elements that are included in the DSMs, and (2) 3_{10} -helices and other types of turns were not included in the DSMs used for these examples. This highlights one limitation of the smoothing analysis in its current implementation. The secondary structure probabilities for a sequence can reflect only the information encoded in the chosen DSMs. In the thioredoxin case, there was apparently no model in our set that properly allows for some of the characteristics of its structure (e.g., the two-residue turn between β -strands 4 and 5 is not incorporated within our current set of DSMs, all of which require a minimum of four residues in a turn).

An additional limitation is that the approach may give misleading results when the set of models does not adequately cover the space of realizable structures. When the given protein represents a structure that is not represented in the current set of DSMs, the secondary structure analysis may be inaccurate. A possible example of this is our analysis of thioredoxin. For this sequence, all candidate DSMs had low probabilities. The probability of model 5 was only 0.20, suggesting that a fully appropriate model may not exist within the current set. In such a case, there is little justification for choosing one particular model as the basis of a secondary structure analysis, and thus there is little reason to expect the secondary structure assignments to be in agreement with the crystallographic data. If there is external prior information that supports a particular domain class DSM assignment (e.g., CD data, sequence similarity with a known structure, or a diagnostic signature pattern such as the mononucleotide-binding fold), then the DSM(s) closest to that class would be considered a priori more probable than those classes contradicting the external information. This type of prior-driven model selection fits naturally into our approach, because relatively low prior probabilities can be assigned to those DSMs that are inconsistent with any given prior information.

The various domain models created so far are not equivalent to each other in their generality and/or complexity. Some DSMs place more constraints on the set of possible sequences of secondary structural elements than others. For example, the eight-stranded α/β -barrel, model 7, can only produce eight β -strands, while model 5 can generate five, six, or seven, and of course the all- α model cannot generate any β -strands. Some of this variation in model information content may be exploited when assigning model prior probabilities (see Methods).

In another paper (White et al., 1992), results on the accuracy of our structural class assignments are reported for study set clusters of homologous sequences (see Table 3

footnote) rather than individual sequences. These results were also obtained with a previous set of DSMs that did not include amphipathic helices or strands. By classifying each cluster based on its most probable model (given all sequences in the cluster and modeling them as independent realizations of a single unknown DSM), the correct structural class was obtained for just over 80% of the clusters at the macroclass level of detail. In contrast, the classification of single sequences is expected to be a harder problem. This is partially due to the fact that information on allowed amino acid substitutions within given secondary structural states is no longer available. On the other hand, similar classification accuracy may be possible with our current DSMs for only single sequences, because our current DSMs have been expanded to include amphipathic helices and strands. These structures allow the DSM to include at least indirectly information as to what residue states are expected to be on the surface of the modeled structure. However, a thorough statistical analysis of our classification method on a large independent set has not yet been conducted.

The approach presented here has the advantage that the entire sequence is used to compute the secondary structural probabilities for each residue. This is different from the use of residues contained in a sliding local "data window" used by some traditional secondary prediction methods (Chou & Fasman, 1974; Garnier et al., 1978). However, the more important advantage is that the wealth of knowledge about the common structural themes among protein domains can be incorporated into an explicit mathematical representation, i.e., the DSMs. Proteins with very different primary sequences and patterns of secondary structure connectivity can have the same fundamental domain structure. The DSM approach can directly include such variation in connectivity, as well as other properties that can be common to topologically similar, but nonhomologous, domains. Thus this approach may prove useful in classifying domains and identifying the secondary structure of proteins with little sequence similarity but similar tertiary structures.

Methods

In this section, our methods are outlined with a minimum of mathematical notation. The mathematical details of our modeling, filtering, and smoothing techniques are presented elsewhere (White, 1992; White et al., 1992).

Sequence classification

To support our modeling activities and provide verification for some of our modeling assumptions, 16 distinctly different proteins were chosen from the Brookhaven database. (These proteins are listed in the caption of Table 3.) They, along with Jane Richardson's taxonomy, provided a study set of examples containing prototypical domains

belonging to most of the structural classes we modeled (Table 1). This study set contains one of the two proteins analyzed in this study. All of the models were constructed as generalizations of the study-set example structures in terms of allowed connectivities and surface loop/turn sizes, for example. In addition, the study-set proteins with their homologs provided amino acid sequences for testing our classification technique.

Different amino acid conditional probabilities are associated with the 13 distinct types of secondary structural elements we considered. These elements are: the α -helical elements consisting of N-cap, C-cap, average-interior, and exposed and buried residues (in amphipathic helices²); the four residues of β -turns; the β -strand average, exposed and buried residues; and loop or coil residues. The conditional probabilities of observing any amino acid in each of these types of elements were obtained from the study-set proteins and their homologs (see Table 3). The sequences of the example proteins, which have known structures, were first aligned (Smith & Smith, 1990) with up to 20 of their close homologs from the SWISS-PROT Database (release 18) (Bairoch et al., 1992). Then the distinct amino acids observed at each aligned position were enumerated and expressed as relative frequencies. The use of distinct amino acids prevents the amino acid repetitions between very close homologs from dominating the counts. These distinct amino acid frequencies were then used in our DSMs as the conditional probabilities that any of the 20 amino acids occurs in each of the 13 types of structural elements.

Each DSM is conveniently represented by three matrices: Φ , H , and x_1 , as explained more fully elsewhere (White, 1988). The state transition matrix Φ holds the conditional probabilities $\phi(s|s')$ of transitioning from any secondary structural state s' to any other state s . The amino acid matrix H holds the conditional probabilities $h(a|s)$ of amino acid residue a , given that the residue is in state s . The initial state-distribution matrix x_1 is a vector holding the probabilities $x_1(s)$ that the Markov chain starts out in any state s at the beginning of the sequence.

Table 1 summarizes the DSMs employed in this study. In this table, we use superclass and macroclass designations for convenience in distinguishing between structural classes at different levels of detail. The DSMs used in this study differ from those used in our previous study (White et al., 1992) by the inclusion of state-space submodels for the amphipathic helix and strand, described below. For each macroclass, from 4 to 25 size-scaled candidate models are initially constructed, each capable of generating sequences ranging in length from a particular minimum to a particular maximum. These multiple models for each macroclass are designed to model proteins over the range of sizes believed to be compatible with these classes. Given a sequence of length L , a set of models is selected that has

² Our amphipathic helix models do not contain N-caps or C-caps.

that length within the central 90% of the model's cumulative distribution of generatable lengths. From this set, the model with the highest likelihood of generating the given sequence is selected to represent that macroclass. Therefore, at most one of these candidate models is ultimately selected to represent each macroclass for the purpose of analyzing a particular primary sequence $\mathbf{A}$ of length L .

The submodel for amphipathic helices is constructed by concatenating the basic building block shown in Figure 6 with itself. The number of building blocks is selected to model helices of different lengths. The amino acid probabilities for exposed and buried residues in this model are listed in columns 12 and 13 of Table 3.

Given a model $M = (+, H, \mathbf{x}_1)$, and a sequence $\mathbf{A}$, the model likelihood is computed using an optimal filtering algorithm (White, 1988; White et al., 1992). The likelihood of model M is the probability $p(A|M)$ of generating sequence $\mathbf{A}$, given that it came from model M . To compute the posterior probability $p(M|\mathbf{A})$ that model M generated the sequence $\mathbf{A}$ (assuming that $\mathbf{A}$ was generated by one of the models being considered), the model likelihood is combined with the prior probability $p(M)$ of the model, using Bayes' rule

$$p(M|A) = \frac{p(A|M)p(M)}{p(A)}; \quad (1)$$

$$p(A) = \sum_M p(A|M)p(M). \quad (2)$$

The prior probability for each model is selected to model our uncertainty about the tertiary class of the sequence prior to analyzing the sequence. To model maximum prior uncertainty (and to determine the amount of information in the sequence that favors one model over another), the prior probabilities of the competing macroclass models are set equal to each other. In this way, the

Fig. 6. Schematic of the repeating unit used to construct models of amphipathic helices. Open circles represent the exposed or surface residue states, and the solid circles represent buried residue states. Parameter P was chosen such that a generated sequence of infinite length has a hydrophobicity profile of periodicity 3.6.

posterior probability favors one model over another based exclusively on the information in the sequence. However, this is not a requirement of the method: if prior information is available (e.g., helical content based on CD data), then the priors can be selected to favor or disqualify specific models.

The likelihood of model M , $p(A|M)$, is computed without approximation by using the following optimal filtering algorithm. The filter computes the incremental likelihood of M , $p(a_k|A_1^{k-1}, M)$, for every residue a_k in the sequence. These numbers are the conditional one-step-ahead probabilities according to model M for the k th amino acid residue, a_k , given the previous amino acids in the subsequence $A_1^{k-1} = (a_1 a_2 \dots a_{k-1})$. These numbers are called incremental likelihoods, because the model likelihood is computed by multiplying them together: $p(A|M) = \prod_{k=1}^L p(a_k|A_1^{k-1}, M)$. To avoid underflow in numerical computations, the log-likelihood, $\log p(A|M)$, is computed in practice.

The optimal filtering algorithm for a specific model M can be described with the following two formulas. For simplicity the dependency on M is omitted from all of the equations below. For example, $p(a_k|A_1^{k-1}, M)$ becomes $p(a_k|A_1^{k-1})$.

The incremental likelihoods are computed sequentially for residues $k = 1$ to L , using

$$p(a_k|A_1^{k-1}) = \sum_{s_k} h(a_k|s_k) p(s_k|A_1^{k-1}), \quad (3)$$

and the next state probability distribution is computed using

$$p(s_{k+1}|A_1^k) = \frac{\sum_{s_k} \phi(s_{k+1}|s_k) h(a_k|s_k) p(s_k|A_1^{k-1})}{p(a_k|A_1^{k-1})}. \quad (4)$$

Equation 3 expresses the incremental likelihood in terms of the amino acid probabilities $h(a_k|s_k)$ for every state s_k . These probabilities are in that row of the model's H-matrix that corresponds to the observed amino acid a_k . The state probability distribution $p(s_k|A_1^{k-1})$ is computed by the filtering algorithm based on the previous $k-1$ amino acids. For the first amino acid, the probability state distribution is given by the initial state distribution, $\mathbf{x}_1$. For the remaining residues, the state distribution is propagated from one residue to the next in accordance with Equation 4, in which the transition probabilities, $\phi(s_{k+1}|s_k)$, are taken from the model's $+$ -matrix.

Figure 7 is a block diagram showing the calculations involved in computing the posterior probabilities of the models for a single sequence. Note that all the model likelihoods may be computed independently of one another. Therefore, they may be implemented as concurrent computations via a suitable parallel computer architecture.

Fig. 7. Block diagram of probability computations.

Secondary structure analysis

We compute the probability of each residue being in each of the possible secondary structural elements. These structure probabilities are determined by the probability distribution $p(s_k|A)$ of the k th residue being in state s_k , given the entire primary sequence. This is the posterior marginal distribution of the state of the k th residue. These distributions can be visualized as a secondary structure probability map for an entire sequence. This map is constructed from the matrix, $P = [p(s_1|A), p(s_2|A), \dots, p(s_L|A)]$, of smoothing distributions for each residue. In each column, the probability of a strand is computed by summing the probabilities of all strand states. The same is done for each of the other secondary structural elements. This results in a smaller matrix with as many rows as there are types of structural elements and as many columns as there are residues in the sequence. A contouring algorithm is then used to plot a contour map of this matrix. Examples of these probability maps are presented in the Results.

For a given model, the smoothing distribution $p(s_k|A)$ is recursively computed for all residue positions by using the following optimal smoothing algorithm. The term "smoothing" originated in the theory of optimal estimation (Wiener, 1949; Gelb, 1974). The name is appropriate, because smoothing is a model-based generalization of the old idea of interpolation.

It can be shown that the smoothing distribution for the state of the k th residue can be computed using two optimal filters: a forward filter that computes $p(s_k|A_1^{k-1})$ in

accordance with Equation 4, and a reverse filter that computes $p(s_k|A_{k+1}^L)$ in reverse order by starting at the last residue and working back to the first residue. These two conditional distributions are combined by the smoother to compute the smoothing distribution

$$p(s_k|A) = C_k \frac{p(s_k|A_1^{k-1})h(a_k|s_k)p(s_k|A_{k+1}^L)}{x_k(s_k)}. \quad (5)$$

In the numerator, $p(s_k|A_1^{k-1})$ is the state distribution from the forward filter. The probability $h(a_k|s_k)$ is available in the model's H -matrix. The probability $p(s_k|A_{k+1}^L)$ is the state distribution from the reverse filter, and $x_k(s_k)$ is the unconditional state probability, which is computed from the model's initial state distribution $\mathbf{x}_1$ using the recursion $x_{k+1}(s) = \sum_{s'} \phi(s|s')x_k(s')$. The normalizing factor C_k is chosen so that $\sum_{s_k} p(s_k|A) = 1$. The details of computing the reverse state distributions and a proof of Equation 5 are to be presented elsewhere. An alternative smoothing algorithm is presented as the solution to "problem 2" in Rabiner (1989).

Models are available via anonymous ftp from darwin.bu.edu.

Supplementary material

On the Diskette Appendix are (1) the Φ , H , and $\mathbf{x}_1$ matrices that define the central β -sheet model used to smooth

the flavodoxin sequence and (2) the smoothing distributions plotted in Figure 3.

Acknowledgments

This research was supported by TASC, NSF grant DIR-8715633, a Ford Foundation Fellowship, and a Pfeifer Fund Fellowship. The authors thank Mathworks, Inc. for making a β version of Matlab, version 4, available for developing sparse-matrix software and Nancy Sands for a critical reading of the manuscript.

References

- Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F.H., Bergerhoff, G., & Sievers, R., Eds.), pp. 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.
- Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* 20, 2019–2022.
- Benner, S.A., Cohen, M.A., & Gerloff, D. (1992). Correct structure prediction. *Nature* 359, 781.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, T. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.
- Bowie, J.U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–169.
- Branden, C. & Tooze, J. (1991). *Introduction to Protein Structure*, Chapter 10. Garland Publishing, Inc., New York/London.
- Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* 248, 338–339.
- Chou, P. (1989). Prediction of protein structural classes from amino acid compositions. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), Chapter 12. Plenum Press, New York/London.
- Chou, P.Y. & Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry* 13, 222–245.
- Cohen, F. & Kuntz, I.D. (1989). Tertiary structure prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), Chapter 17. Plenum Press, New York/London.
- Deleage, G. & Roux, B. (1989). Use of class prediction to improve protein secondary structure prediction joint prediction with methods based on sequence homology. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), Chapter 13. Plenum Press, New York/London.
- Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* 319, 199–203. [Values of the solvent-accessible surface area for each of the amino acids were taken from the computer program made available by the authors.]
- Garnier, J., Osguthorpe, D.J., & Robson, B. (1978). Analysis of the accuracy and implications for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120, 97–120.
- Gelb, A. (1974). *Applied Optimal Estimation*. M.I.T. Press, Cambridge, Massachusetts.
- Hobohm, U., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* 1, 409–417.
- Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 22, 2577–2637.
- Katti, S.K., LeMaster, D.M., & Eklund, H. (1990). Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* 212, 167–184.
- Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., & Saraste, M. (1992). Crystal structure of a Src-homology 3 (SH3) domain. *Nature* 359, 851–855.
- Nishikawa, K., Kubota, Y., & Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characteristics. *J. Biochem.* 94, 981–995.
- Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193, 775–791.
- Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. Inst. Electrical Electronics Eng.* 77, 257–286.
- Richardson, J.S. (1981). The anatomy and taxonomy of protein structures. *Adv. Protein Chem.* 34, 167–339.
- Rossman, M.G., Moras, D., & Olsen, K.W. (1974). Chemical and biological evolution of a nucleotide binding protein. *Nature* 250, 194–199.
- Smith, R.F. & Smith, T.F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA* 87, 118–122.
- Smith, T.F. & Waterman, M.S. (1981). Comparison of biosequences. *Adv. Appl. Math.* 2, 482–489.
- Smith, W.W., Burnett, R.M., Daring, G.D., & Ludwig, M.L. (1977). Structure of the semiquinone form of flavodoxin from *Clostridium mp.* extension of 1.8 Å resolution and some comparisons with the oxidized state. *J. Mol. Biol.* 117, 195–228.
- White, J.V. (1988). Modeling and filtering for discretely valued time series. In *Bayesian Analysis of Time Series and Dynamic Models* (Spall, J.C., Ed.), Chapter 10. Marcel Dekker, New York.
- White, J.V., Stultz, C.M., & Smith, T.F. (1992). Protein classification by stochastic modeling and optimal filtering of amino acid sequences. *Math. Biosci.*, in press.
- Wiener, N. (1949). *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Wiley, New York.
- Zhang, C. & Chou, K. (1992). An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.* 1, 401–408.