PROTEINS: Structure, Function, and Genetics 25:38–47 (1996)

Predicting Solvent Accessibility: Higher Accuracy Using Bayesian Statistics and Optimized Residue Substitution Classes

Michael J. Thompson[1] and Richard A. Goldstein[1,2]
[1]Biophysics Research Division, [2] Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055

ABSTRACT **We introduce a novel Bayesian probabilistic method for predicting the solvent accessibilities of amino acid residues in globular proteins. Using single sequence data, this method achieves prediction accuracies higher than previously published methods. Substantially improved predictions—comparable to the highest accuracies reported in the literature to date—are obtained by representing alignments of the example proteins and their homologs as strings of residue substitution classes, depending on the side chain types observed at each alignment position. These results demonstrate the applicability of this relatively simple Bayesian approach to structure prediction and illustrate the utility of the classification methodology previously developed to extract information from aligned sets of structurally related proteins.** © 1996 Wiley-Liss, Inc.

Key words: protein structure prediction, Bayesian statistics, amino acid substitution, information theory, solvent accessibility

INTRODUCTION

Knowledge of the detailed three-dimensional structure of a protein is required for developing a full mechanistic understanding of its functionality. It is therefore unfortunate that so few protein structures have been solved compared to the enormous number of proteins that have been sequenced. As this gap grows, so grows the need for reliable and generally applicable structure prediction methods, particularly for those instances where the determination of a protein structure is experimentally infeasible and no related proteins of known structure are available for homology modeling. In addition, successful predictions schemes which help elucidate the relationship between amino acid sequence and protein structure can aid in de novo protein design and engineering.

Many methods have been developed for predicting some quantifiable one-dimensional aspect of three-dimensional protein structure. One common means of subdividing protein structure is to characterize amino acid residues or segments of the polypeptide chain as adopting one of a number of conformational secondary structures, the standard four being α-helix, β-strand, turn, and coil. The implicit hope was that accurate assignment of such segments along the protein chain would facilitate the prediction of protein tertiary structure in parallel to the envisioned process of protein folding as the assembly of preformed secondary structure elements.[1–10] With the growing awareness that non-local interactions and hydrophobically driven chain compaction are major determinants in the concurrent formation of secondary and tertiary structure,[11–15] there is interest in predicting aspects of protein structural organization which directly result from the interplay of these folding forces. One such aspect, which is perhaps the most basic and informative organizational distinction to make, is the degree to which the amino acid residues in the protein structure are capable of interacting with solvent molecules.[16] The one-dimensional descriptor which best captures this division is the static relative solvent accessibility.[17] This descriptor is typically divided into binary categories (buried or exposed) or ternary categories (buried, partially exposed, or exposed) according to some chosen percent solvent-accessibility threshold(s).[18–21] While the prediction of secondary structure has historically dominated this field, the prediction of solvent accessibility has become an increasingly active area of research.[22–25]

The observation that homologous sequences generally adopt the same tertiary fold[26–28] indicated that families of related proteins could provide more information about their common structure than could single sequences. Several secondary structure predictions methods have made use of this information to substantially increase predictive accuracy. (For reviews, see references 29 and 30.) Likewise, approaches to predicting solvent accessibility have found similar benefit from using this type of evolu-

Received July 14, 1995; revision accepted November 20, 1995.

Address reprint requests to Richard A. Goldstein, Biophysics Research Division, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055.

tionarily derived information. Holbrook et al. found they could increase the prediction accuracy by performing a consensus prediction.[22] The solvent accessibility predictions of Wako and Blundell employed substitution tables derived from protein families.[24] The artificial neural networks of Rost and Sander achieved higher accuracy predictions when using information derived from multiple sequence alignments than with single sequence data.[25]

Here, we present a novel method for predicting one-dimensional descriptors of protein structure, specifically solvent accessibility. The first component of our method is the computational formalism for performing the prediction calculations, which confronts one major difficulty met with in attempts to predict the local structure of proteins from amino acid sequences; inter-residue correlations. The structural characteristics (e.g., secondary structure or exposure to solvent) of neighboring residue locations are strongly correlated and impose correlations in the amino acid sequence. These correlations can be uncoupled to a large degree through the use of Bayes' theorem. This theorem allows us to express the conditional probability of a small segment of local protein structure, given the amino acid residues observed in that segment, in terms of the conditional probability of observing those residues given that structural segment. Once the structure of the particular segment is conjectured, the statistics of the individual locations in the segment can be considered independently.

The second component of our method relies on the degenerate nature of the protein folding code—multiple amino acid sequences can adopt equivalent three-dimensional structures. Due to this "structural inertia"[31] on the time scale of protein evolution, it is possible to view the evolutionary process as one in which the protein sequence adapts to the protein structure. The sets of side chains found at positions in alignments of homologous proteins provide more information about the constraints imposed by the local structure than can single sequences. We have developed a methodology for codifying the evolutionary information about structure present in these patterns of amino acid residue substitution.[29] Each position in an alignment of multiple homologous proteins is classified into a particular residue substitution class depending on which side chains are observed at that position. Information theory provides us with a function, "mutual information," which measures how much information is obtained about one random variable (the local structure) by knowledge of another random variable (the substitution classes). The optimal set of residue substitution classes can be obtained by maximizing this function during a stochastic search of the possible sets of substitution classes over a representative database of solved protein structures and their families.[29] Following the Bayesian procedure described in the previous paragraph, we can predict local protein structure by considering the sequence of substitution class designations representing the alignment of the protein and its homologs.

With appropriate "jackknife" tests, this method achieved a 2-state accuracy of 70.7% using single sequence data. Our approach consistently performs as well or better than previously developed methods over the same datasets. Inclusion of information derived from homologous proteins and overall protein length increased this accuracy to 74.9%, comparable with the accuracies reported by other authors.[22,24,25] Similarly high performance predictions were made for three and 10 solvent accessibility states. In contrast to neural networks, the parameters of the model have clear biophysical interpretations, and the assumptions and approximations of the method can be explicitly stated. There is also a lack of ad hoc parameters such as neural network architectures or tuning parameters. This approach combines general applicability with computational affordability.

In addition to providing insight into the organization of three-dimensional protein structure, predictions of solvent accessibility may find use in a number of applications. They could be used to detect amphipathic structures with characteristic periodicities,[32–36] to characterize structural motifs, and to aid in the alignment of sequences in regions where sequence similarity is slight. They could also be used to predict loop regions,[37,38] transmembrane regions,[39] and antigenic determinants[38] in proteins. In addition to predicting solvent accessibility, the Bayesian approach could be applied to the prediction of other one-dimensional descriptors of protein structure.

METHODS

Bayesian Theory

As mentioned in the Introduction, protein structure prediction is complicated by the fact that the amino acid residues in a sequence are highly correlated, so a particular amino acid in one sequence position is indicative of a particular local structure that can influence the amino acids that are likely to be found at nearby positions. It is sometimes possible to decouple correlations by making the statistics depend upon some causal factor, so that the statistics of the dependent factors become independent if the state of the causal factor is known.[40] For instance, rather than considering the structure of the protein to depend upon the amino acid composition, we can instead consider the amino acid composition to depend upon the structure of the protein. By making the amino acids at the various sequence positions a function of the local protein structure, it is plausible to conjecture that once the local structure is selected, the particular choice of residue at one position is relatively uncorrelated to the choice of residue at the other positions. Bayes' theorem gives

us a way to relate the conditional probability of a particular structure given knowledge of the corresponding sequence to the conditional probability of that sequence given the particular structure. The capacity to consider residue locations independently is an underlying aspect of hidden Markov models.[41–43] Other Bayesian-based approaches have been developed for both secondary structure prediction[44] and tertiary structure recognition,[45] although neither of these methods made use of this "decoupling" capability.

We are interested in predicting the solvent accessibility of residue i, ω_i, based on knowledge of the amino acid sequence, $\{A_j\}$, of a "window" of restricted size symmetric about location i. Locations within the window are indexed by j. We write this as the conditional probability for the solvent accessibility given the sequence, $P(\omega_i \mid \{A_j\})$. Bayes' theorem allows us to express this probability in terms of $P(\{A_j\} \mid \omega_i)$, the conditional probability of a particular sequence $\{A_j\}$ given that location i has solvent accessibility ω_i, times $P(\omega_i)$, the probability of that solvent accessibility in the absence of any sequence information, divided by $P(\{A_j\})$, the probability of that sequence in a random structure.

$$P(\omega_i \mid \{A_j\}) = \frac{P(\{A_j\} \mid \omega_i)P(\omega_i)}{P(\{A_j\})} \quad (1)$$

Let us take s_j to designate the local structure at each residue location, j. This descriptor can include the solvent accessibility ω_j alone ($s_j = \omega_j$), or it can include secondary structure information and/or any other subdivision of protein structure which can be reduced to a one-dimensional characteristic (e.g., ω_j = "buried," while s_j = "buried beta strand"). The local structure of the protein chain can be written out as a one-dimensional string of such descriptors. Recalling our consideration of a limited-sized window along the amino acid sequence, we will denote the corresponding substring of local structure descriptors a "structural segment," $S^k = \{s_j^k\}$, where k denotes the particular segment type. We can consider independently all of the possible structural segments, $\{S^k\}$, that have solvent accessibility ω_i at location i. The value of $P(\{A_j\} \mid \omega_i)P(\omega_i)$ is the sum of these probabilities for all of the various possible segments of local structure $\{S_k\}$, multipled by the probability of the sequence given that structural segment.

$$P(\omega_i \mid \{A_j\}) = \sum_k \frac{P(\{A_j\} \mid S^k)}{P(\{A_j\})} P(S^k)\, \delta(s_i^k \in \omega_i) \quad (2)$$

where $\delta(s_k^i \in \omega_i)$ is zero unless the residue in state s_i^k has the solvent accessibility ω_i.

We now use our assumption that for a given segment of structure, the probability of each amino acid residue in that segment depends only on the local structure at that point, and not on the identity of the nearby residues. In this case, we can express

$$\frac{P(\{A_j\} \mid S^k)}{P(\{A_j\})}$$

as the product of probabilities for all of the local amino acid residues in the local structural segments:

$$\frac{P(\{A_j\} \mid S^k)}{P(\{A_j\})} = \prod_j \frac{P(A_j \mid s_j^k)}{P(A_j)} \quad (3)$$

where $P(A_j \mid s_j^k)$ is just the probability of residue A_j being at location j in structure segment k given that the local structure at that point is s_j^k. The product is over all the locations in that structural segment. Substituting this result into Eq. (2) yields:

$$P(\omega_i \mid \{A_j\}) = \sum_k \left(\prod_j \frac{P(A_j \mid s_j^k)}{P(A_j)} \right) P(S^k)\, \delta(s_i^k \in \omega_i) \quad (4)$$

As mentioned above, the descriptor for the local structure, s_j, can be as simple as the solvent accessibility at that location, or as rich as we desire, with the assumption of independence increasing with the complexity of the description. For instance, in the work presented here, s_j can be any of $4 \times n$ categories based on combinations of four secondary structure states and n solvent accessibility states, as defined below. The inclusion of secondary structure information makes this approach sensitive to patterns of surface exposure, characteristics of certain secondary structure elements such as surface α-helices and surface turns. The number of adjustable parameters in

$$\frac{P(A_j \mid s_j^k)}{P(A_j)}$$

remains small, even with a more complex description, because of the ability to ignore correlations between the residues in different positions. Although the number of possible segments of local structure grows with the richness of the description, uncommon segments of local structure, where the probabilities are poorly determined, will have negligible effect on the sum of Eq. (2).

Substitution Classes

A richer description of the available sequence information can also be used. Instead of considering just the amino acid sequence of a protein, we can make use of its alignment with a family of structurally related proteins. Based on the side chains observed at a given alignment position, that position is designated as belonging to one of a set of amino acid substitution classes. Using an information theoretic formalism, these substitution classes were constructed to be optimally indicative of protein struc-

ture by maximizing the mutual information between the set of classes and the set of local structures.[29] The set of substitution classes included 20 classes corresponding to conserved examples of the 20 side chain types and eight multiresidue substitution classes representing patterns of structurally indicative residue substitution. The last class contained all residue types and the possibility of a gap so that all alignment positions in the dataset were assigned to one of the classes. While the substitution classes discussed in Thompson and Goldstein[29] were optimized for eight local structure categories based on solvent accessibility and secondary structure, the various sets of 28 substitution classes used in this work were optimized for the set of solvent accessibility states being predicted. Bayesian predictions identical to the one described above were performed using these classes to represent the proteins and their aligned sets of homologs. Taking C_j to represent the substitution class assignment at position j, we rewrite Eq. (4),

$$P(\omega_i \mid \{C_j\}) = \sum_k \left(\prod_j \frac{P(C_j \mid s_j^k)}{P(C_j)} \right) P(S^k)\, \delta(s_i^k \in \omega_i). \quad (5)$$

The propensity for the substitution classes to exist in any structural context can be expressed with log-likelihood ratios. $L(C,s)$, the log-likelihood ratio for class C to be in context s, is defined by

$$L(C, s) = 100 \times \left(\frac{p(C, s)}{p(C) \times p(s)} \right). \quad (6)$$

This ratio quantifies how much more likely it is for an alignment position belonging to a substitution class C to be in local structure s, compared with what would be expected at random.

Databases and Structure Definitions

As used in our earlier work,[29] a set of 111 target protein chains (25,511 residues) (D_{TG}) was selected from the October 1994 PDBselect list of representative structures sharing less than 25% sequence identity between any pair, as compiled by Hobohm and Sander.[46] Alignments of homologs were extracted from the "homology derived structures of proteins" (HSSP) files of Sander and Schneider.[47] Secondary structure information was taken from the "Dictionary of Protein Secondary Structure " (DSSP) file of Kabsch and Sander[48] which were derived from the Protein Data Bank (PDB) files of three-dimensional coordinates for each protein.[49,50] In this work 3-helix and 5-helix locations were assigned as helix, bend locations as turn, and β-bridge locations as β-strand.[29] Percent solvent accessibilities were computed by normalizing the accessible surface area with maximum values obtained by Shrake and Rupley.[51] Accessible surface areas for residues in chains from multimeric proteins were calculated using the multimeric complexes. Various cut-offs were used to define the solvent accessibility states, as explained in the results section.

For the purpose of comparison with the results of other methods, the set of 126 protein chains (23,336 residues) (D_{RS}) compiled by Rost and Sander[52] was used, as were the training set (19 proteins, 3,344 residues) (D_H) and test set (five proteins, 963 residues) ($D_{H'}$) as reported by Holbrook and colleagues.[22] Relative solvent accessibility values for all datasets of other authors were calculated in the same manner as by those authors.[22,25,53]

Prediction Procedures

All of the our single sequence-based predictions followed a single-omission jackknife procedure. Each of the example proteins in the dataset, in turn, was excluded from calculation of the Bayesian statistics. These statistics formed the basis for predicting the solvent accessibilities of that excluded protein. For predictions based on multiple sequence alignments, two separate jackknife procedures were employed. The first procedure was the same as that just described, with each alignment of example protein and homologs represented with a set of residue substitution classes which were optimized over the entire dataset. To avoid any potential memorization of the test proteins by use of these globally optimal residue substitution classes, a second jackknife procedure was followed. Here, only 7/8 of the 111 protein dataset formed the basis for the optimization of substitution classes and calculation of the Bayesian statistics used in predicting the solvent accessibilities of the remaining 1/8 of the dataset. This procedure was repeated until the entire dataset was predicted. This resulted in a near-negligible decrease in accuracy.

All results reported for the Bayesian predictions were based on a window of length 13 residues, symmetric about the residue being predicted. In order to predict residue locations near the N and C terminals of the proteins with this window-based scheme, virtual residue locations were added to the ends of the chains, all taken to be in the exposed coil state. For the assessment of prediction performance, two numbers are generally computed. One is the percentage of correctly predicted residues (%-correct). The other is a correlation coefficient between the observed, o_i, and predicted, p_i solvent accessibility states for a dataset of N residue locations, as given by

$$\text{Correlation coefficient} \equiv \frac{N \Sigma o_i p_i - \Sigma o_i \Sigma p_i}{[N \Sigma o_i^2 - (\Sigma o_i)^2]^{1/2} [N \Sigma p_i^2 - (\Sigma p_i)^2]^{1/2}} \quad (7)$$

TABLE I. Prediction Results Based on the Use of Single Sequence Data*

	Single sequence data		
Method	Dataset	% Correct†	Correlation
2-States			
Bayes	$D_{TG_{20}}$	70.7	0.414
Bayes	$D_{TG_{20}}$ (mono)	70.9	0.417
Bayes	$D_{H_{20}}/D_{H'_{20}}$	70.3	0.408
2-State NN[16]	$D_{H_{20}}/D_{H'_{20}}$	72.0	0.44
Bayes	$D_{TG‡20}/D_{H'_{20}}$	72.3	0.445
Bayes	D_{RS_9}	72.8	0.370
3-State NN[19]	D_{RS_9}	71.4	—
Bayes	$D_{RS_{16}}$	71.1	0.404
10-State NN[19]	$D_{RS_{16}}$	70.0	—
Bayes	$D_{RS_{23}}$	70.0	0.401
3-States			
Bayes	$D_{RS_{9,64}}$	60.8	0.358
3-State NN[19]	$D_{RS_{9,64}}$	55.1	0.356
Bayes	$D_{TG_{9,64}}$	61.4	0.369
Bayes	$D_{RS_{9,36}}$	54.2	0.437
10-State NN[19]	$D_{RS_{9,36}}$	52.4	—
Bayes	$D_{TG_{9,36}}$	54.1	0.449
Bayes	$D_{H_{5,40}}/D_{H'_{5,40}}$	56.8	0.487
3-State NN[16]	$D_{H_{5,40}}/D_{H'_{5,40}}$	52.0	—
Bayes	$D_{TG‡5,40}/D_{H'_{5,40}}$	56.2	0.510
10-States			
Bayes	$D_{RS_{1,2,9,16,25,36,49,64,81}}$	21.6	0.432
10-State NN[19]	$D_{RS_{1,2,9,16,25,36,49,64,81}}$	21.6	0.432
Bayes	$D_{RS_{0,3,8,15,24,34,44,55,69}}$	20.6	0.443
Bayes	$D_{TG_{1,2,9,16,25,36,49,64,81}}$	22.3	0.459
Bayes	$D_{TG_{0,3,8,15,23,32,42,53,67}}$	21.3	0.460

*Correlation, correlation coefficient as defined in Eq. (7); Bayes, Bayesian probabilistic predictions as described in this work; k-State NN, Neural network predictions for k solvent accessibility states; D_{TG}, 111 protein chain dataset of Thompson and Goldstein[29]; D_{RS}, 126 protein chain dataset of Rost and Sander[52]; D_H, 19 protein training set of Holbrook et al.[22]; $D_{H'}$, 5 protein test set of Holbrook et al.[22]; D_X/D_Y, X dataset of proteins used to train neural network or generate statistics for predicting solvent accessibilities of proteins in Y dataset; $D_{X_{1,2,\ldots,k}}$, subscripts of X denote the solvent accessibility cutoffs used to define the k solvent accessibility states for residues in protein dataset X; mono, subset of 60 monomeric proteins from D_{TG} dataset.

†Because the %-correct measure can be artifically elevated by an uneven splitting of the dataset, accuracies for different solvent accessibility cut-offs cannot be directly compared.

‡Solvent accessibilities for both datasets calculated as in Holbrook et al.[22]

RESULTS AND DISCUSSION

Single Sequences

Table I lists the results of predictions based on single sequence data. The single-omission jackknife procedure, applied to our dataset of 111 proteins with a 2-state-defining solvent accessibility threshold of 20% ($D_{TG_{20}}$), achieved an accuracy of 70.7% (0.41 correlation coefficient). Predictions over the subset of 60 chains from monomeric proteins gave 70.9% correct predictions (0.42 correlation coefficient)—a slight improvement also observed by other authors.[22,24,25]

In order to obtain a direct comparison with the results of Holbrook and coworkers, Bayesian statistics were generated based on two different sets of proteins: the 19 training set proteins ($D_{H_{20}}$) listed by Holbrook et al. and our 111 protein dataset with fractional solvent exposures calculated as in their work ($D_{TG^*_{20}}$).[22] Predictions on the five test proteins ($D_{H'_{20}}$) using $D_{H_{20}}$ gave 70.3% correct predictions (0.41 correlation coefficient) while 72.3% correct predictions (0.45 correlation coefficient) were obtained using $D_{TG^*_{20}}$. These results compare favorably with the 72.0% accuracy (0.44 correlation coefficient) reported by Holbrook et al., especially considering the problematic nature of comparing performance on such an extremely small test set.[25,54] Rost and Sander reported a 75.7% accuracy for these same five test proteins using their neural network method and a 2-state solvent accessibility

threshold of 16%. However, in addition to using this threshold that assigned a disproportionate number of locations to the "exposed" state, their neural networks were trained on a dataset ($D_{RS_{16}}$) which contained members of the $D_{H'_{16}}$ test set. In order to provide a direct comparison with the neural network approach of Rost and Sander,[25] the Bayesian prediction method was performed over their set of 126 proteins with their alternate solvent accessibility cut-offs of 9 and 16% (D_{RS_9} and $D_{RS_{16}}$) both of which split that dataset unevenly. As shown in Table I, results in both cases were superior to those achieved by their much more complicated scheme.

As noted by Rost and Sander, the choice of solvent accessibility cut-offs is problematic.[25] To explore the issue of threshold-dependent prediction accuracy more fully, 2-state predictions were performed for solvent accessibility cut-offs ranging from 5 to 25% over the 126 proteins compiled by Rost and Sander (D_{RS_5}-$D_{RS_{25}}$). These results are shown in Figure 1. Both common measures of prediction performance—percentage of predictions correct and correlation coefficient—were sensitive to the choice of solvent accessibility thresholds. This phenomenon is easily understood. A simple prediction scheme which assigns the most likely state will produce increasingly accurate results as the solvent accessibility is varied to produce increasingly biased partitioning of the dataset. While the percentage correct measure was highest in the range of cut-offs which severely favor one state over the other, the correlation coefficient plateaued around the cut-offs which produced more evenly populated solvent accessibility states. We propose that the selection of thresholds which partition a dataset uniformly among the solvent accessibility states (e.g., $D_{RS_{23}}$ and $D_{TG_{20}}$) would provide a standard which would facilitate performance comparison among prediction methodologies; else, as demonstrated above, the various quantitative assessments of prediction performance—particularly the percentage correct measure—are of dubious value.

In order to obtain more detailed information, solvent accessibility has sometimes been classified into 3 or 10 states.[22,25] The 3-state predictions were made based on datasets constructed using two pairs cut-offs proposed by Rost and Sander.[25] For both sets, the buried and the partially exposed states were separated by a 9% cut-off, while the exposed state was distinguished from the partially exposed state by alternate cut-offs of 36% ($D_{RS_{9,36}}$ and $D_{TG_{9,36}}$), or 64% ($D_{RS_{9,64}}$, and $D_{TG_{9,64}}$). For comparisons to the results of Holbrook et al., the pair of cut-offs 5 and 40% were used ($D_{H_{5,40}}$, $D_{H'_{5,40}}$, and $D_{TG^*_{5,40}}$). As shown in Table I, results using the Bayesian scheme were significantly and consistently superior to results obtained with neural networks. Again, we see that an uneven partitioning of the dataset (9%,64%) can produce misleadingly high

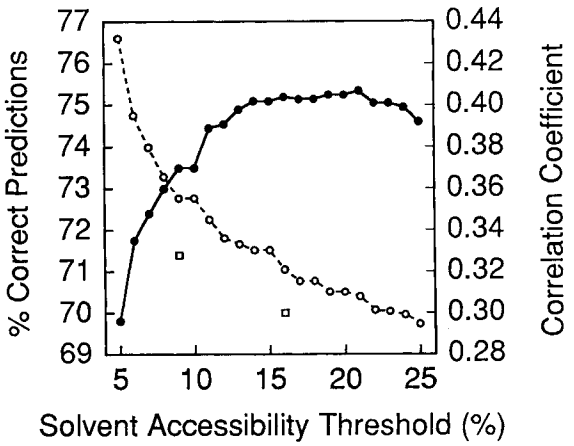

Fig. 1. Dependence of 2-state prediction accuracies and correlation coefficients on solvent accessibility threshold. Curves represent results obtained using the Bayesian scheme over the 126 protein dataset. (•) Denotes the correlation coefficients; (○) denotes %-correct prediction accuracies; (□) designates the %-correct values reported for the recent neural network method.[25]

prediction accuracies, in contrast to those obtained with cut-offs which distribute the dataset more evenly among the ternary states (9%,36%).

For 10-state predictions, our 111-protein dataset and the 126-protein dataset of Rost and Sander were divided into 10 roughly-equally populated accessibility states ($D_{TG_{0,3,8,15,23,32,42,53,67}}$ and $D_{RS_{0,3,8,15,24,34,44,55,69}}$). In addition, for comparative purposes, both datasets were divided using the thresholds proposed by Rost and Sander ($D_{TG_{1,2,9,16,25,36,49,64,81}}$ and $D_{RS_{1,2,9,16,25,36,49,64,81}}$), which strongly emphasized buried locations over exposed locations. Accuracies for the various methods were roughly equivalent, with slightly higher accuracies obtained by all methods with the more uneven distribution of accessibilities. Interestingly, for $D_{RS_{1\ldots81}}$, the Bayesian method and the neural network method achieved essentially identical accuracy.

Multiple Sequence Alignments

Information was extracted from sets of aligned homologous proteins for each of the example proteins in the 111 protein dataset, $D_{TG_{20}}$, by representing the alignments with residue substitution classes. As explained in the methods section, the set of substitution classes was optimized by maximizing the amount of information about the solvent accessibility provided by knowledge of membership in a residue substitution class. The 28 substitution classes optimized over the 111 protein dataset for two solvent accessibility states ($D_{TG_{20}}$) are shown in Table II along with their log likelihoods for the buried and exposed states. The sets of 28 substitution classes optimized for the 7/8 subsets of the 111 protein dataset and for the 3-state and 10-state predictions were similar.

TABLE II. A Listing of Residue Memberships and Log Likelihood Ratios for the Set of 28 Residue Substitution Classes Optimized Over the Dataset of 111 Protein Chains to Provide Solvent Accessibility Information*

Class		Buried	Exposed
L		53	−122
I		52	−117
V		50	−109
F		52	−121
M		37	−60
C		40	−68
A		42	−75
W		54	−129
Y		34	−53
T		26	−36
S		28	40
H		34	−53
Q		9	−10
N		21	−27
E		−7	6
D		11	−12
K		−24	19
R		3	−3
P		13	−16
G		23	−31
LIVFMCAW	—	54	−131
LIVFMCAWYTSH G	—	37	−60
LIVFMCA Y HQ PG	—	17	−21
L VF A YTSHQNE R G	—	−35	26
LIVFMCAWYTSHQNE R G	—	8	−9
LI A YTSH DKRPG	—	−63	38
LIVFMCAWYTSH DKRPG	—	−10	9
LIVFMCAWYTSHQNEDKRPG	—	−124	54

*(—) Indicates gap.

Using the residue substitution class methodology significantly improved prediction accuracies, as shown in Table III. Compared to the 70.7% correct predictions (0.41 correlation coefficient) achieved for $D_{TG_{20}}$ based on single example sequences, use of the residue substitution classes yielded 74.7% correct predictions (0.50 correlation coefficient). As shown in Table III, these values were only slightly affected by the 7/8 jackknife procedure.

We also experimented with the inclusion of chain length information. In the Bayesian step of the prediction procedure, the example proteins were separately considered as "short" chains (< 250 residues) or as "long" chains (> 250 residues). This gave an increased accuracy of 74.9% (0.50 correlation coefficient). Thus, protein chain length provided information about both the likelihood of various structures and about the residues likely to be found in those structures. These numbers compare with the highest 2-state accuracies reported by the more recent neural network scheme which took advantage of additional sources of information, such as profiles of insertions and deletions, amino acid compositions, and residue locations relative to the N and C termini of the proteins.[25] Again, the 7/8 jackknife procedure did not significantly decrease the prediction performance.

Predictions were also performed for the 3-state and 10-state cases using the residue substitution class methodology. For the 3-state predictions, the 9%,36% pair of cut-offs was used to define the solvent accessibility states ($D_{TG_{9,36}}$). Use of the residue substitution classes gave a 3.5% increase to 57.5% correct predictions (0.54 correlation coefficient) over the 54.1% (0.45 correlation coefficient) obtained using only single sequences from the same dataset. Inclusion of protein length information, as described in the paragraph above, produced a further improvement to 57.9% correct predictions (0.55 correlation coefficient), equivalent to the highest 3-state accuracies obtained with neural networks.[25]

In the 10-state predictions, the solvent accessibility states were defined with the two sets of thresholds described in the previous section for the 111 protein dataset ($D_{TG_{0\ldots67}}$ and $D_{TG_{1\ldots81}}$). Inclusion of information from homologous sequences caused an increase in accuracy of approximately 2.3% over that obtained with single sequences, for both sets of cut-offs. Our accuracy was, again, equivalent to that reported by Rost and Sander for their dataset using the same cut-off values, though our predictions showed a slightly stronger correlation with the observed accessibilities (0.56 vs. 0.54).

It should be noted that the comparisons made above between the Bayesian approach and neural network method of Rost and Sander were based on results obtained over different datasets (D_{TG} and D_{RS}, respectively). In order to make direct comparisons with the neural network scheme it would be necessary to apply the residue substitution class-based method to their 126 protein dataset. This was not done due to the fact that the multiple sequence alignment data available for many of the proteins in that dataset was inadequate for making accurate residue substitution class assignments.

There are a number of serious problems in comparing our results to those obtained by Wako and Blundell.[24] They report the use of a 20% solvent accessibility threshold when the threshold which evenly partitions their dataset is 26%, indicating that their reported accuracies could be inflated by the statistical phenomenon discussed above. Their method also includes an ad hoc parameter which "tunes" the relative weighting of residue substitution information vs. structural propensity information in the prediction calculations. As this parameter was adjusted for optimal performance for the test set, it is possible that poorer performance would be achieved for other protein data sets. Most critically, these authors reported 2-state prediction accuracies calculated by averaging over individual proteins in each of 13 families and then over the set of 13 families[24] in order to achieve their reported 76.5% ac-

TABLE III. Prediction Results Based on the Use of Multiple Sequence Data

	Multiple sequence data		
Method	Dataset	% Correct	Correlation
2-States			
Bayes	$D_{TG_{20}}$	74.7	0.496
Bayes[7/8]	$D_{TG_{20}}$	74.2	0.485
Bayes*	$D_{TG_{20}}$	74.9	0.499
Bayes*,[7/8]	$D_{TG_{20}}$	74.7	0.495
3-State NN[19]	D_{RS_9}	74.6	—
10-State NN†,[19]	$D_{RS_{16}}$	74.2	—
10-State NN*,†,‡,[19]	$D_{RS_{16}}$	75.0	—
3-States			
Bayes	$D_{TG_{9,36}}$	57.5	0.537
Bayes*	$D_{TG_{9,36}}$	57.9	0.547
3-State NN[19]	$D_{RS_{9,64}}$	58.0	0.450
10-State NN*,†,‡,[19]	$D_{RS_{9,36}}$	57.9	—
10-States			
Bayes	$D_{TG_{0,3,8,15,23,32,42,53,67}}$	22.9	0.561
Bayes	$D_{TG_{1,2,9,16,25,36,49,64,81}}$	24.7	0.565
10-State NN*,†,‡,[19]	$D_{RS_{1,2,9,16,25,36,49,64,81}}$	24.4	0.541
10-State NN*,†,‡,[19]	$D_{RS'_{1,2,9,16,25,36,49,64,81}}$	25.3	0.544

[7/8] Prediction based on jacknife procedure using 7/8 subsets of the D_{TG} dataset, as described in the text.
*Used protein chain length information.
†Used conservation weights.
‡Jury-averaged multi-network system using various combinations of additional inputs such as amino acid composition, insertion/deletion profiles and distances to N and C terminals.
$D_{RS'}$, an alternate set of 112 protein chains compiled by Rost and Sander.[25]

curacy, rather than averaging over residues as done by other researchers. Since their method performs better for shorter proteins than for longer proteins, the type of averaging they use is biased in favor of their method. The method presented in this article performs in a more balanced way over proteins of different length. When we ran our Bayesian prediction method over the same set of proteins as reported by Wako and Blundell, excepting those members of the "immunoglobulin constant domain" family whose HSSP files contained constant and variable domains, we achieved 75.0% using the standard type of averaging and 75.8% using the averaging method of Wako and Blundell.

CONCLUSION

As is evident from the preceding section, quantitative assessment of the performance of our Bayesian prediction method indicates that it consistently performs better than existing methods over single sequence data and comparably with those methods which use multiple sequence data. In addition, this approach bears some significant advantages compared to these previous methods. One disadvantage of neural network-based methodologies is the relative difficulty with which these methods can be inspected to gain an understanding of the biophysical basis of the predictions.[24,30] Holbrook et al. stated that neural networks have the advantage of "not needing a preconceived model."[22] One might argue that without a model to test, there is less possibility of expanding the insights learned beyond the realm of structure prediction or generating understanding that would suggest further improvements in the prediction methodology. Another disadvantage of neural networks is that periodic updating and generalizing of the system of networks for new proteins requires complete retraining of the networks and retuning of their individual architectures and input formats—a computationally daunting procedure. On the other hand, using the Bayesian scheme, the relevant statistics for new proteins can be rapidly calculated and assimilated into the statistics stored for previous datasets.

This new Bayesian approach to solvent accessibility prediction provides superior prediction performance with a relatively simple and inspectable formalism which is more computationally affordable and more easily generalizable to larger datasets. This same formalism could be applied to the prediction of other one-dimensional descriptors of protein structure. The methodology for representing alignments of multiple homologous proteins with optimal residue substitution classes, which lends itself with ease to the Bayesian approach, conveys a competitive increase in prediction accuracy while retaining

a conceptual simplicity advantageous for the development of biophysical insight.

ACKNOWLEDGMENTS

We thank Kurt Hillig for computational assistance. We extend a general thanks to those who solve protein structures and make this information available, and to those who construct and maintain databases of protein sequences and structures. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies at the University of Michigan, and NIH grant R29 LM05770.

REFERENCES

1. Ptitsyn, O.B., Rashin, A.A. Model of myoglobin self-organization. Biophys. Chem. 3:1–20, 1975.
2. Cohen, F.E., Richmond, T.J., Richards, F.M. Protein folding—evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J. Mol. Biol. 132:275–288, 1979.
3. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of protein beta-sheet structures by a combinatorial approach. Nature 285:378–382, 1980.
4. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis of the structure of protein beta-sheet sandwiches. J. Mol. Biol. 148:253–272, 1981.
5. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. J. Mol. Biol. 156:821–862, 1982.
6. Taylor, W.R. Towards protein tertiary fold prediction using distance and motif constraints. Protein Eng. 4:853–870, 1991.
7. Smith-Brown, M.J., Kominos, D., Levy, R.M. Global folding of proteins from a limited number of distance constraints. Protein Eng. 6:605–614, 1993.
8. Gunn, J.R., Monge, A., Friesner, R.A., Marshall, C.H. Hierarchical algorithm for computer modeling of protein tertiary structure: Folding of myoglobin to 6.2Å resolution. J. Phys. Chem. 98:702–711, 1994.
9. Monge, A., Friesner, R.A., Honig, B. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. Proc. Natl. Acad. Sci. U.S.A. 91:5027–5029, 1994.
10. Monge, A., Lathrop, E.J.P., Gunn, J.R., Shenkin, P.S., Friesner, R.A. Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. J. Mol. Biol. 247:995–1012, 1995.
11. Chan, H.S., Dill, K.A. Compact polymers. Macromolecules 22:4559–4573, 1989.
12. Chan, H.S., Dill, K.A. The effects of internal constraints on the configurations of chain molecules. J. Chem. Phys. 92:3118–3135, 1990.
13. Chan, H.S., Dill, K.A. Origins of structure in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87:6388–6392, 1990.
14. Hunt, N.G., Gregoret, L.M., Cohen, F.E. The origins of protein secondary structure. J. Mol. Biol. 241:214–225, 1994.
15. Yee, D.P., Chan, H.S., Havel, T.F., Dill, K.A. Does compactness induce secondary structure in proteins? J. Mol. Biol. 241:557–573, 1994.
16. Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., Sauer, R.T. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. Science 247:1306–1310, 1990.
17. Lee, B.K., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
18. Janin, J. Surface and inside volumes in globular proteins. Nature 277:491–492, 1979.
19. Hubbard, T.J.P., Blundell, T.L. Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. Protein Eng. 1:159–171, 1987.
20. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.
21. Lawrence, C., Auger, I., Mannella, C. Distribution of accessible surfaces of amino acids in globular proteins. Proteins 2:153–161, 1987.
22. Holbrook, S.R., Muskal, S.M., Kim, S.-H. Predicting surface exposure of amino acids from protein sequences. Protein Eng. 3:659–665, 1990.
23. Bohr, H., Goldstein, R.A., Wolynes, P.G. Predicting surface structures of proteins by neural networks. AMSE Periodicals C 31:53–56, 1992.
24. Wako, H., Blundell, T. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J. Mol. Biol. 238:682–692, 1994.
25. Rost, B., Sander, C. Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226, 1994.
26. Chothia, C. Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537–572, 1984.
27. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
28. Pastore, A., Lesk, A.M. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. Proteins 8:133–155, 1990.
29. Thompson, M.J., Goldstein, R.A. Constructing amino acid residue substitution classes maximally indicative of local protein structure. Proteins 25:28–37, 1996.
30. Eisenhaber, F., Persson, B., Argos, P. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit. Rev. Biochem. Molec. Biol. 30:1–94, 1995.
31. Aronson, H.E.G., Royer, W.E., Jr., Hendrickson, W.A. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci. 3:1706–1711, 1994.
32. Schiffer, M., Edmundson, A.B. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. Biophys. J. 7:121–135, 1967.
33. Kuntz, I.D. Protein folding. J. Am. Chem. Soc. 94:4009–4012, 1972.
34. Lim, V. Algorithms for prediction of α-helical and β-structural regions in globular proteins. J. Mol. Biol. 88:873–894, 1974.
35. Eisenberg, D., Weiss, R.M., Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. U.S.A. 81:140–144, 1984.
36. Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., Delisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. 195:659–685, 1987.
37. Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272:586–590, 1978.
38. Hopp, T.P., Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78:3824–3828, 1981.
39. Kyte, J., Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
40. Pearl, J. "Probabilistic Reasoning in Intelligent Systems." San Mateo, CA: Morgan Kauffmann Publishers, Inc., 1988.
41. Stultz, C.M., White, J.V., Smith, T.F. Structural analysis based on state-space modeling. Protein Sci. 2:305–314, 1993.
42. Asai, K., Haymizu, S., Handa, K. Prediction of protein secondary structure by the hidden Markov model. CABIOS 2:141–146, 1993.
43. Krogh, A., Brown, M., Mian, I.S., Sjölander, K., Haussler, D. Hidden Markov models in computational biology. J. Mol. Biol. 235:1501–1531, 1994.
44. Stolorz, P., Lapedes, A., Xia, Y. Predicting protein secondary structure using neural nets and statistical methods. J. Mol. Biol. 225:363–377, 1992.
45. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. A Bayesian approach to sequence alignment algorithms for protein structure recognition. In: "Proceedings of the 27th

Annual Hawaii International Conference on System Sciences." Los Alamitos, CA: IEEE Computer Society Press, 1994.
46. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
47. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
48. Kabsch, W., Sander, C. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
49. 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, M. Protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
50. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In: "Crystallographic Databases—Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn: Data Commission of the International Union of Crystallography. 1987:107–132.
51. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
52. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
53. Rose, G., Geselowitz, A., Lesser, G., Lee, R., Zehfus, M. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
54. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.