Protein Science (1997), 6:1963–1975. Cambridge University Press. Printed in the USA.

Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information

MICHAEL J. THOMPSON[1] AND RICHARD A. GOLDSTEIN[1,2]

[1]Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109-1055
[2]Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055

(RECEIVED January 29 1997; ACCEPTED May 19, 1997)

Abstract

We demonstrate the applicability of our previously developed Bayesian probabilistic approach for predicting residue solvent accessibility to the problem of predicting secondary structure. Using only single-sequence data, this method achieves a three-state accuracy of 67% over a database of 473 non-homologous proteins. This approach is more amenable to inspection and less likely to overlearn specifics of a dataset than "black box" methods such as neural networks. It is also conceptually simpler and less computationally costly.

We also introduce a novel method for representing and incorporating multiple-sequence alignment information within the prediction algorithm, achieving 72% accuracy over a dataset of 304 non-homologous proteins. This is accomplished by creating a statistical model of the evolutionarily derived correlations between patterns of amino acid substitution and local protein structure. This model consists of parameter vectors, termed "substitution schemata," which probabilistically encode the structure-based heterogeneity in the distributions of amino acid substitutions found in alignments of homologous proteins. The model is optimized for structure prediction by maximizing the mutual information between the set of schemata and the database of secondary structures.

Unlike "expert heuristic" methods, this approach has been demonstrated to work well over large datasets. Unlike the opaque neural network algorithms, this approach is physicochemically intelligible. Moreover, the model optimization procedure, the formalism for predicting one-dimensional structural features, and our previously developed method for tertiary structure recognition all share a common Bayesian probabilistic basis. This consistency starkly contrasts with the hybrid and ad hoc nature of methods that have dominated this field in recent years.

Keywords: Bayesian statistics; evolutionary information; mutual information; probabilistic schemata; secondary structure prediction

The prediction of protein secondary structure by a number of methods has benefitted from the use of aligned sets of homologous proteins. The patterns of conservation and variation in amino acid residue substitutions at a particular site in a protein convey implicit information about the long-range interactions involved in determining the local structure at that site. Various techniques have been devised to use this information to raise the three-state accuracies to around 70–72%. While this summary statistic is universally reported, a broader evaluation of prediction performance would consider both the practical and scientific value of the prediction scheme in terms of its statistical performance, physicochemical interpretability, and general applicability (reproducibility and robustness). Until now, the success of multiple-sequence alignment-based secondary structure prediction methods has been restricted to one or two of these attributes.

Reprint requests to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055; e-mail: richardg@chem.lsa.umich.edu.

The most common technique for using multiple-sequence alignments is the consensus method. This "signal averaging" approach takes the predictions made for individual members of a protein family and combines them according to some weighting scheme to arrive at a consensus prediction. The two most recent applications of this approach have yielded high accuracies over large datasets (Salamov & Solovyev, 1995; Riis & Krogh, 1996). While this approach is generally applicable and can provide competitive statistical accuracy, it does not model the underlying sequence-to-structure correlations or evolutionary process. Thus, few questions regarding such relationships can be addressed and this technique can be of little use in furthering structure prediction efforts. The widely known artificial neural network method of Rost and Sander (1993, 1994) also performs quite well, but it too is rather opaque. In that approach, position-specific profiles of residue substitutions are fed into an ensemble of complicated neural networks; how to make use of this raw information is left up to the training algorithm and a large number of highly coupled adjustable parameters.

Dominating the realm of transparent methods is the work of Benner and colleagues who focus their efforts on developing prediction heuristics based on expert knowledge of protein chemistry and evolutionary relationships among proteins (Benner, 1989, 1992). Advocates of this approach claim that it allows the predictor to develop an understanding of the principles of protein structure formation. However, given the slow turn-around time for bona fide predictions and the small set of examples, it would be difficult to distinguish between heuristics which reflect general principles and ad hoc "fixes." Moreover, with the pervasive contextuality observed at every level of protein structure, one might wonder at the eventual size of a generally applicable "rule table" for predicting protein structure.

A few efforts have been directed at creating automated prediction methods including explicit models of evolutionarily derived correlations in the form of substitution matrices (Wako & Blundell, 1994a, 1994b; Mehta & Argos, 1995). These algorithms are more physicochemically interpretable than the consensus or neural network methods but achieve lower accuracies. The recent work of Goldman et al. (1996) explores the explicit use of evolutionary trees but the statistical performance of the method was left unquantified. The recent approach of King and Sternberg (1996) is transparent in its linear decomposition but the method achieves 70% accuracy only with the addition of global sequence information, smoothing functions, feedback about predictions, and, finally, filtering rules. These researchers report higher accuracies through judicious substitution of another algorithm for their own.

The picture that arises is that large-scale applicability (automatability) and physicochemical interpretability are somehow incompatible—a view emphasized by some proponents of the more manual methods (Benner & Gerloff, 1993; Benner et al., 1997). We certainly agree with the scientific necessity for using inspectable models to further understanding, and with the criticism of the current state-of-the-art automated schemes in this regard. In this paper, we present a method that is accurate, biophysically intelligible, and generally applicable. It performs comparably to the best of the opaque methods over large datasets of non-homologous proteins. Using only single-sequence information, our basic Bayesian prediction formalism yields three-state accuracies as high as 67% over a dataset of 473 non-homologous proteins. Using our new method for incorporating evolutionary information into prediction algorithms, we obtain accuracies of 72% for a dataset of 304 non-homologous proteins. An additional benefit of our method is that it can be used to predict solvent accessibility (Thompson & Goldstein, 1996b).

Our multiple-sequence alignment-based method employs a probabilistic model of correlations between protein structure and patterns of amino acid residue substitution that have arisen over the course of evolution. The model consists of a set of parametrized distributions (schemata) over the 20 amino acids and gaps. These schemata are constructed to represent the structurally based heterogeneity of substitutions found in multiple-sequence alignments. Based on the substitutions observed at a given location in an alignment, the prediction algorithm assesses the probability that the location belongs to (i.e., was generated according to) each of the schemata and the probability with which each of the schemata is associated with a particular type of local protein structure. This information is then incorporated into our previously developed Bayesian formalism for predicting one-dimensional features of protein structure (Thompson & Goldstein, 1996b).

This work extends from our previous method for classifying sets of amino acid residue substitutions based on the structural information conveyed by the set of classes (Thompson & Goldstein, 1996a). The difference in terminology ("classes" vs. "schemata") emphasizes the fundamental difference between these approaches; whereas amino acid membership in the previous classes was all-or-none, here the representation is probabilistic. The term "schemata" is borrowed from the literature of genetic algorithms where it describes the "building block" patterns from which a solution is constructed (Holland, 1992). As before, our optimization of the schemata is based on maximizing the mutual information between a set of schemata and the corresponding local structures. This work bears some similarity to the work of Sjolander et al. (1996), who derived Dirichlet mixture priors. These two methods, however, differ significantly in purpose. While those researchers sought to construct hidden Markov models of specific families for use in finding remote sequence homologs, we seek to predict the secondary structure of proteins in general. Our schemata, though, could be used in database searches for structural homologs.

Theory I. Basic prediction formalism

We expand on the theoretical basis of our prediction methodology, as introduced by Thompson and Goldstein (1996a).

Structure determines sequence

For each residue site in a protein, we would like to calculate $P(s_i^k|\{a_j\})$, the conditional probability of observing secondary structure s^k (where k indexes the types of secondary structure) given amino acids $\{a_j\}$ in a local window of sequence (indexed by j) around the site of interest, i.

A physical interpretation of this probability implies that the structure of a protein is determined by the amino acid sequence. On the folding time scale this is true, but on the time scale of evolution, it is the relatively fixed protein structures that constrain the evolution of the quickly changing protein sequences—a phenomenon referred to as "structural inertia" (Aronson et al., 1994), which has been observed in the simulated evolution of lattice-model proteins (Govindarajan & Goldstein, 1997a). As the database of protein sequences and structures is a product of this evolution, it would be more natural to model the protein sequence as being probabilistically dependent on the structure, and to write $P(\{a_j\}|s_i^k)$. Application of Bayes' theorem accomplishes this inversion,

$$P(s_i^k|\{a_j\}) = \frac{P(\{a_j\}|s_i^k)P(s^k)}{P(\{a_j\})} \quad (1)$$

where $P(\{a_j\}|s_i^k)$ is the conditional probability of the particular set of amino acid residues in the window given the particular type of secondary structure s_i^k at location i, $P(\{a_j\})$ is the probability of observing the set of residues given no structural information, and $P(s^k)$ is the frequency of occurrence of secondary structure type s^k in the database. Note that $P(\{a_j\})$ is simply a normalization and can be computed as $\sum_{k'} P(\{a_j\}|s^{k'})P(s^{k'})$ where k' indexes all the types of secondary structure. The actual calculation of this denominator is unnecessary in the prediction routine as we take the secondary structure with the maximum conditional probability as the prediction.

Structural segments

The amino acids found in the local window obviously depend upon the structure of the whole window, rather than just a single residue location. We refer to the string of secondary structure identities capturing the structural information about the local window of the protein chain as a "structural segment," $S^\mu = \{s_j^\mu\}$, where μ denotes the particular segment type. These segments are the same length as the window of sequence being considered.

We can evaluate $P(\{a_j\}|s^k)P(s^k)$ as the sum of the probabilities for all of the various possible segments of local structure S^μ that have secondary structure type s^k at the central segment site i (corresponding to the site of interest in the sequence window), multiplied by the probability of the sequence given that structural segment:

$$P(s_i^k|\{a_j\}) = \frac{\sum_\mu P(\{a_j\}|S^\mu)P(S^\mu)\delta(s_i^\mu, s^k)}{\sum_\mu P(\{a_j\}|S^\mu)P(S^\mu)} \quad (2)$$

where $\delta(s_i^\mu, s^k)$ is zero unless s_i^μ, the secondary structure at the central site in the segment, is the same as s^k.

Bayesian decoupling

Unfortunately, due to the relatively small sample of non-homologous proteins that are available and the typically large size (13–17 residues) of sequence windows used, there is an insufficient number of examples to get a good estimate of $P(\{a_j\}|S^\mu)$ in Equation 2. This problem can be overcome by the following considerations. We assume that the amino acid residue at each site in the segment depends only on the structure at that site in the segment, and is independent of the residues at other sites in the segment. This approach views the correlations between neighboring amino acid residues as resulting from underlying structural correlation, which is fully accounted for by the use of our structural segments. In particular, we might use the probabilities $P^\mu(a_j|s_j^\mu)$ for the amino acids given the secondary structure type at each location in the segment. The superscript μ on the probability indicates that it would be estimated from only the instances of the particular segment S^μ. Again, because most structural segments will be observed very few times, the estimations of these parameters will be rather poor. As an approximation, we assume the amino acid residue only depends on the local structure at that site in the segment so that we can estimate these values from the entire dataset. This leads to the series of equations,

$$P(\{a_j\}|S^\mu) \cong \prod_j P(a_j|S^\mu) \quad (3)$$

$$\cong \prod_j P^\mu(a_j|S_j^\mu) \quad (4)$$

$$\cong \prod_j P(a_j|S_j^\mu). \quad (5)$$

Substituting this result into Equation 2 yields

$$P(s_i^k|\{a_j\}) = \frac{\sum_\mu \left(\prod_j P(a_j|s_j^\mu)\right) P(S^\mu)\delta(s_i^\mu, s^k)}{\sum_\mu \left(\prod_j P(a_j|s_j^\mu)\right) P(S^\mu)}. \quad (6)$$

The utility of this "prediction equation" hinges on the decoupling capability resulting from the combined Bayesian and evolutionary perspectives (Thompson & Goldstein 1996b). These key features distinguish our approach from the mathematically similar GOR method (Robson, 1974). In that pioneering method, an explicit consideration of pair-wise dependencies was attempted but was constrained by the size of the datasets available (Gibrat et al., 1987). Other Bayesian statistical approaches toward protein structure prediction have not employed this decoupling concept, either (Maxfield & Scheraga, 1979; Stolorz et al., 1992; Zhang et al., 1992; Goldstein et al., 1994). Such ideas, however, have been used in the probabilistic modeling of inter-residue correlations in the EF-hand motif (Mamitsuka, 1995) and they are an implicit feature of hidden Markov models (Asai et al., 1993; Stultz et al., 1993; Krogh et al., 1994).

Structure descriptors

Equation 6 implies that the identity of an amino acid in the sequence will be determined only by the secondary structure at that location. In fact, other factors, such as surface accessibility, will be major influences. It has been observed that α-helices and β-strands can possess characteristic patterns of exposure to solvent, and this information has been successfully exploited in previous secondary structure predictions (Lim, 1974; Yi & Lander, 1993; Wako & Blundell, 1994b; Salamov & Solovyev, 1995). We can capture these patterns through the use of a richer alphabet for denoting local structure.

We take ϕ_j to be the "descriptor" of the structure at each residue location, j. In this work, we explore the use of four categories of secondary structure, s_j, combined with one, two, or three categories of solvent accessibility, ω_j. Thus, ϕ_j can take on four, eight, or 12 values depending on the use of solvent accessibility information. Where we need to distinguish between the use of four, eight, or 12 structure categories, we will use the notation $\phi_j(4)$, $\phi_j(8)$, and $\phi_j(12)$. Note that in the case where no solvent accessibility information is used (one accessibility category), $\phi_j(4) = s_j$.

Since the secondary structure of a residue location is defined in terms of local bond angles and hydrogen bonding patterns that extend beyond the single location, and since there is statistical evidence that the 20 amino acids' residues have differential preferences for different regions of secondary structural elements (Richardson & Richardson, 1988), we also experimented with the use of a more extended definition of the local structure. This was done by combining the "singlet" descriptors for a residue location with those of its neighbors. We designate these n-tuplets with the superscript n (${}^n\phi_j$). In particular, we considered the use of duplets and triplets of structure descriptors. In both cases, the n-tuplet can be asymmetric about the residue location of interest. Depending on the terminus to which the n-tuplet extends, we add a label of N or C to the superscript. For example, in the case of structural triplets, ${}^{3N}\phi_j = \{\phi_{j-2}, \phi_{j-1}, \phi_j\}$, ${}^3\phi_j = \{\phi_{j-1}, \phi_j, \phi_{j+1}\}$, and ${}^{3C}\phi_j = \{\phi_j, \phi_{j+1}, \phi_{j+2}\}$.

Equation 6 can be easily generalized for these more specific descriptors. Defining $\Phi^\mu = \{\phi_j^\mu\}$;

$$P(s_i^k|\{a_j\}) = \frac{\sum_\mu \left(\prod_j P(a_j|\phi_j^\mu)\right) P(\Phi^\mu)\delta(\phi_i^\mu, s^k)}{\sum_\mu \left(\prod_j P(a_j|\phi_j^\mu)\right) P(\Phi^\mu)} \quad (7)$$

where $\delta(\phi_i^\mu, s^k)$ is zero unless the structure descriptor at site i, ϕ_i^μ, corresponds to the secondary structure s^k combined with any

solvent accessibility category. As the sum is over all possible solvent accessibility categories for a given secondary structure, solvent accessibility information about the target protein is not used.

Inspection of Equation 7 reveals the algorithmic simplicity of this approach. First, we count the number of instances of each type of amino acid residue being associated with each type of structure descriptor and the number of instances of each type of structural segment of a given length. These are converted to probabilities. Then, for each type of secondary structure, given the local window of query sequence, we simply take the product over window positions and sum over the relevant structural segments. Finally, the secondary structure type with the highest probability is taken as the prediction. One clear advantage of this approach is that as the protein datasets increase, all that is necessary is to count new instances and add them to the old ones.

Theory II. Multiple-sequence-based formalism

Substitution count vectors

Rather than use a consensus approach, we seek to develop a model of the evolutionarily derived relationships between protein sequences and structures. The raw data from which we will build our model are the substitutions found at each residue site in the database of multiple-sequence alignments. Each site j can be represented as a vector of counts, $\vec{n}_j = (n_{j_1}, n_{j_2}, \ldots, n_{j_{20}})$, where n_{j_a} is the number of times residues of type $a = 1, \ldots, 20$ are observed. Using this representation, we can simply replace a local amino acid sequence, $\{a_j\}$, with a local string of count-vectors, $\{\vec{n}_j\}$, in the derivation of Equation 7.

$$P(s_i^k|\{\vec{n}_j\}) = \frac{\sum_\mu \left(\prod_j P(\vec{n}_j|\phi_j^\mu)\right) P(\Phi^\mu)\delta(\phi_i^\mu, s^k)}{\sum_\mu \left(\prod_j P(\vec{n}_j|\phi_j^\mu)\right) P(\Phi^\mu)}. \quad (8)$$

This lacks robustness and scientific merit, as it merely treats these count-vectors in a look-up table fashion. A significant fraction of positions in the dataset of alignments have a unique combination of amino acid residues, so this approach is statistically infeasible.

Schemata

The heterogeneity observed in the number and type of amino acid substitutions from position to position in a database of alignments arises from a mixture of biophysical generative processes, stochastic evolutionary mechanisms, and statistical biases. This suggests a probabilistic treatment. There are easily detectable patterns among the sets of count-vectors derived from multiple-sequence alignments. There is the well-known division between sites with preferentially hydrophobic substitutions and sites with preferentially hydrophilic substitutions corresponding to the characteristic relative degrees of solvent exposure of those two types of sites. Residue sites within the same type of secondary structure share common structural constraints, giving rise to specific patterns of substitutions. Likewise, the counts of substitutions at similarly functional sites in various proteins could represent samples from an underlying distribution characteristic of that particular function.

To capture this statistical and structurally based variation in patterns of substitution, we postulate the existence of a number of probability distributions (much smaller in number than the number of substitution count vectors) from which the count vectors have been generated. Each schema consists of a probability vector, $\vec{p}^\eta = (p_1^\eta, p_2^\eta, \ldots, p_{20}^\eta, p^\eta)$, where the parameters, p_a^η, represent the probabilities of "drawing" each of the amino acid types a according to the particular probability distribution η. The parameter p^η denotes the a priori probability of the schema itself existing at any site in a protein (i.e., without reference to count-vectors or secondary structure information).

Assuming each amino acid substitution occurs independently, these parameters will allow us to calculate the conditional probability that a particular count-vector would be drawn from a particular η. This is done by taking the product over the probabilities of the counts of the amino acids multiplied by the number of ways of selecting the particular set of amino acid counts. According to the combinatorics of the problem, the number of ways of generating the vector, $\vec{n}_j$, is $|\vec{n}_j|!/(n_{j_1}!n_{j_2}!\ldots n_{j_{20}}!)$ where $|\vec{n}_j|$ is the sum total of the number of amino acids observed at the particular alignment position. Thus,

$$P(\vec{n}_j|\eta) = \frac{|\vec{n}_j|!}{n_{j_1}!n_{j_2}!\ldots n_{j_{20}}!} \prod_{a=1}^{20} (p_a^\eta)^{n_{j_a}}. \quad (9)$$

In contrast to our earlier substitution classes, a location in the multiple alignment of proteins can only be assigned to schemata in a *probabilistic* manner. Again using Bayes' law, the probability that a location with vector $\vec{n}_j$ was generated by schemata η defined by $\vec{p}^\eta$ is given by

$$P(\eta|\vec{n}_j) = \frac{P(\vec{n}_j|\eta)p^\eta}{\sum_{\eta'} P(\vec{n}_j|\eta')p^{\eta'}}. \quad (10)$$

Predicting with schemata

As in our earlier work, we replace the multiple sequence alignment with a single sequence characterized by the underlying schemata, and consider the correlations between these schemata and the local structure in making our predictions. The indeterminacy in assigning locations to schemata, however, must be included in every part of the prediction scheme. Assuming we know the schemata from which nature has assembled the proteins in our databases, we can calculate $P(\vec{n}_j|\phi_j^\mu)$ in our substitution count vector-based "prediction equation" (Equation 8) by explicitly summing over all η, or

$$P(\vec{n}_j|\phi_j^\mu) = \sum_\eta P(\vec{n}_j|\eta, \phi_j^\mu)P(\eta|\phi_j^\mu) \quad (11)$$

$$= \sum_\eta P(\vec{n}_j|\eta)P(\eta|\phi_j^\mu) \quad (12)$$

where we have taken advantage of the fact that the probability of a count vector only depends on the schemata and not the local structure.

The probabilistic nature of the schemata must also be included in the accumulating of statistics, specifically in the calculation of $P(\eta|\phi_j^\mu)$ above, as we can not count instances of the various schemata in an unambiguous way. We approach this problem by noting that $P(\eta|\phi_j^\mu) = P(\eta, \phi_j^\mu)/P(\phi_j^\mu)$. The joint probability $P(\eta, \phi_j^\mu)$ can be written as the probability that location j with its corresponding count vector $\vec{n}_j$ can be assigned to schemata η, given by Equation 10, summed over all positions in the database that have the type of local structure ϕ_j^μ, and normalized by N, the total number of positions in the database.

$$P(n|\phi_j^\mu) = \frac{1}{N}\sum_{j'} P(\eta|\vec{n}_{j'})\delta(\phi_{j'}, \phi_j^\mu) \quad (13)$$

where $\delta(\phi_{j'}, \phi_j^{\mu})$ is 1 if the structure indicated by $\phi_{j'}$ is the same as that indicated by ϕ_j^{μ} and 0 otherwise. The problem remains how to obtain the schemata that would allow us to calculate these expressions.

Optimizing schemata

Our purpose is to predict protein secondary structure. Therefore, we would like our probabilistic knowledge of the different schemata that may have generated the count vector at a given site in a protein to provide us with as much information as possible about the secondary structure at that position. More generally, we would like a set of schemata that represents the entire database of substitution count vectors to provide the maximum possible information about the secondary structure identities at all of the corresponding residue sites. In the section below, we describe an optimization procedure that allows us to adjust the parameters for some number of schemata so as achieve this.

Mutual information

As in our previous work in constructing binary-state classes of amino acid residue substitution, we have chosen mutual information as our optimization function (Thompson & Goldstein, 1996a). This function, taken from information theory, is based on the Shannon entropy function, which quantifies the "uncertainty" with regard to the state of a random variable (Shannon & Weaver, 1949). It is calculated over the probability distribution of states of the random variable and behaves as follows: When the probability distribution is uniform (all possible states are equally likely) the entropy is maximized, and when the probability of one particular state is unity (no uncertainty) then the entropy is zero.

For our purposes, we can calculate an entropy over our probabilistic schemata,

$$H_\eta = -\sum_\eta P(\eta) \ln P(\eta) \tag{14}$$

where η is an index over the schemata. Likewise, we calculate H_ϕ as the entropy over local protein structures.

$$H_\phi = -\sum_k P(\phi^k) \ln P(\phi^k) \tag{15}$$

where k is an index over types of local protein structure.

It is also possible to calculate a joint entropy over the joint probability distribution of two random variables (e.g., a set of schemata and local protein structure),

$$H_{\eta,\phi} = -\sum_n \sum_k P(\eta, \phi^k) \ln P(\eta, \phi^k). \tag{16}$$

It is natural to equate a gain in "information" with a reduction in "uncertainty." More generally, information can be defined as a difference between entropies. For two random variables, the amount of information about the state of one variable conveyed by knowledge of the state of the other variable is quantified by the mutual information (Cover & Thomas, 1991). This function is expressed as the difference between the sum of the independent entropies of the two random variables and their joint entropy. We write

$$M_{\eta,\phi} = H_\eta + H_\phi - H_{\eta,\phi}. \tag{17}$$

Inspection of this equation reveals that if a set of schemata has no specific correspondence with local protein structure, then $H_{\eta,\phi} \to H_\eta + H_\phi$ and $M \to 0$. Conversely, if the correspondence between the two sets of variables is highly specific, then M is maximized.

To compute the entropies of this mutual information, we need to calculate the quantities $P(\eta)$, $P(\phi^k)$, and $P(\eta, \phi^k)$. The probabilities, $P(\phi^k)$, are taken as frequencies of the local structure types denoted by the descriptors ϕ^k. We have already seen how to calculate $P(\eta, \phi^k)$ in Equation 13 from the previous section. The terms, $P(\eta)$, are calculated in a similar manner, except that there is no restriction to positions of a particular structural type,

$$P(\eta) = \frac{1}{N} \sum_{j'} P(\eta | \vec{n}_{j'}) \tag{18}$$

where j' is an index over all positions.

The mutual information is calculated based on the parameter values that define the set of schemata. These parameters can be iteratively updated using a gradient descent algorithm so as to maximize the mutual information. This procedure produces a set of schemata that represent a structurally optimal compression of the count-vector data. This may, however, not be exactly what we desire. This could correspond to memorization of specific patterns found in the dataset over which the optimization is performed. Rather, we seek compression of these data into schemata that will be useful in predicting the secondary structures of proteins in general. To achieve this, cross-validation techniques are used, as discussed in the results section.

For our secondary structure prediction application, we optimize the schemata based on secondary structure information only (no solvent accessibility information). As found in our previous work optimizing binary-state substitution classes, the solvent accessibility information tends to dominate the results of the search (Thompson & Goldstein, 1996b). Since it is imperfectly correlated with secondary structure, this drives the search away from what would be optimal for secondary structure prediction. However, schemata optimized based on only secondary structure information can be used within a prediction setting which makes use of solvent accessibility information.

Materials and methods

Single-sequence datasets

Two datasets of proteins were used in our single-sequence-based predictions. The first dataset, comprising 473 protein chains, was taken from the March 1996 PDBselect list of representative structures with less than 25% sequence identity between any pair of chains (Hobohm & Sander, 1994). The second dataset consists of 126 protein chains, also with less than 25% pairwise sequence identity, compiled by Rost and Sander (1994).

Multiple-sequence datasets

The construction of our dataset of proteins with homologs also began with the March 1996 PDBselect list. We extracted multiple-sequence alignment data from the HSSP files for these proteins (Sander & Schneider, 1991).

Table 1. *Summary of size, in residues (N_{res}) and secondary structural characteristics for the various subsets of proteins used in this work*[a]

Dataset[b]	N_{res}	$\%\alpha$	$\%\beta$	$\%T$[c]	$\%C$[d]
473	121750	30	22	27	21
126	23348	28	22	28	22
151A	37760	29	22	27	21
150B	37327	29	22	27	21
102A	25086	30	22	28	21
101B	24977	29	22	27	21
101C	25024	29	23	27	22

[a]The percentages of structure types may not sum to 100% due to round-off error.
[b]The number denoting the dataset indicates the number of protein chains in that dataset.
[c]*T* denotes turn.
[d]*C* denotes coil.

Two modifications were made to these multiple-sequence alignments to maximize the usefulness of the information they contain. First, we eliminated all homologs with ≤40% identity with the protein of known structure. In earlier work, we found that a 40% sequence identity cut-off in the homologs used provided the greatest amount of information about local structure in terms of patterns of residue substitution (Thompson & Goldstein, 1996a). The second step was to eliminate redundant sequences in the alignments. If two members of a given alignment are nearly identical then the addition of one of those members after the other member is already present provides no additional useful information. The presence of these sequences gives rise to "apparent" conservation that would mislead the prediction algorithm. Alignments were examined for pairs of homologs with >90% identity and one of members of the pair was excluded from the alignment. After these modifications, we took each protein of known structure that had at least five homologs for a minimum at 80% of its residue sites. This resulted in a dataset of 304 proteins. For cross-validation purposes, this dataset was divided into either two or three subsets.

Summary statistical information about the various datasets used is shown in Table 1. Lists of all sets of proteins used in the single-sequence predictions and schemata-based predictions and optimization are available by anonymous ftp at chem.lsa.umich.edu in directory pub/goldstein/.

Structure information

Information about secondary structure was extracted from the "Dictionary of Protein Secondary Structure" (DSSP) files of Kabsch and Sander (1983), which were derived from the Protein Data Bank (PDB) files of three-dimensional coordinates for each protein (Bernstein et al., 1977; Abola et al., 1987). In addition to the standard four types of secondary structure—α-helix, β-strand, turn, and coil—the DSSP files contain four other types that we assigned to the standard four according to the following mapping: Five-helix to helix, and 3_{10}-helix, bend, and β-bridge to turn. The probabilities for the turn and coil categories were combined. Solvent accessibility values were also taken from the DSSP files and were normalized with maximum values obtained by Shrake and Rupley (1973). Solvent accessibility thresholds are needed for defining the solvent accessibility states of the residue sites. Thresholds were chosen such that equal numbers of residue sites were assigned to each of the states. For two-solvent accessibility states the threshold for the 126-protein dataset is 23% and for the 473-protein dataset it is 19%. To define three solvent accessibility states for the 126-protein dataset and the 473-protein dataset, we set thresholds at 9% and 36% and at 6% and 36%, respectively. For all datasets in the multiple-sequence alignment-based work, a two-state threshold of 20% was used. In order to use a window-based scheme, virtual residue locations were added to the *N*- and *C*-termini of the protein chains. These locations were all taken to be in the fully exposed coil state.

Results and discussion

Memorization

Regardless of methodology, one way to improve prediction performance is to include more information relevant to the sequence-to-structure correlations that the prediction algorithm seeks to learn and exploit. This can be done by increasing the specificity of the structural description at the resolution of the "structural segments" (increase the window size) or at the resolution of individual residue sites (increase the alphabet of structure descriptors). Due to the inherent statistical nature of most efforts in this field, the size of the local window or the number of structural descriptors cannot be increased without bound. For the machine-learning approaches, there is not enough data from which to learn, and for the statistical schemes, the probabilities become ill-defined.

While the literature of secondary structure prediction stresses the importance of cross-validation or jackknifing in the optimization of neural network synaptic weights or in the estimation of parameters such as $P(a_j|s_k)$ in our model, the selection of more global parameters such as the structural descriptors, window sizes, number of nearest neighbors, and neural architectures is often left uncritiqued. These parameters are frequently selected through multiple prediction trials beyond the cross-validation protocols. As such, the values of these parameters are potentially specific to the dataset being used.

The most successful methods (in terms of reported accuracies), like neural networks and nearest neighbor algorithms, make use of "black box" tuning parameters, such as the number of nearest neighbors and the neural architectures, in addition to selecting descriptor alphabets and window sizes. Moreover, the most recent of these two types of approaches have both employed a jury-decision scheme over the prediction outputs of multiple variations of their algorithms (Rost & Sander, 1993, 1994; Salamov & Solovyev, 1995). Unlike the descriptor alphabet or the window size, however, these methodological parameters do not clearly have anything to do with proteins in a physical sense. There is no a priori reason to believe that 50 nearest neighbors, or 15 hidden units, or a jury decision taken over algorithms using windows of 11, 17, and 23 residues will give the best results regardless of dataset. The fact that the jury-decision procedure (signal averaging) works for these schemes is evidence that each of the algorithmic variations of these authors is making systematic errors, possibly due to overlearning.

In contrast, the only parameters that are selected over the entire dataset in our method are the alphabet of structure descriptors and the window size. In both cases, these parameters control the "spec-

ificity" of information being used in the predictions. The predictions that returned the highest accuracies correspond to the maximum specificity of descriptions that can be statistically supported by the datasets used. In this sense, it is unlikely that the prediction accuracies for these parameter choices is an overestimation of the predictive capability of our method. For larger future datasets, better estimates can be calculated for the various parameters in the model. Moreover, larger datasets will support the estimation of additional parameters and/or more specific parameterizations of the prediction problem.

By using parameter values obtained from our current set of proteins, though, this method will not produce accuracies equivalent to the mean accuracies reported here for new proteins that do not match the average characteristics of our dataset. The question, then, becomes how likely new proteins are to match the characteristics of the current database. While this is a complicated issue, various researchers have estimated the number of protein folds to be in the low thousands, and it is commonly known that a small number of protein folds are overwhelmingly populated relative to the majority of observed structures. One explanation of this biased distribution of sequences among the various protein tertiary structures was provided by a recent lattice-model protein study (Govindarajan & Goldstein, 1996).

Evaluation I: Single-sequence-based performance

A summary of prediction highlights can be found in Table 2, for both the 473-protein and 126-protein datasets. We report Q_3 scores, the percentage of residues correctly predicted in three states, and the Matthew's correlation coefficients for α-helical (C_α) and β-strand structures (C_β) (Matthews, 1975). All evaluations reported here were obtained using a single-chain-exclusion jackknife procedure; each protein in the dataset was, in turn, left out from the calculation of the probabilities used to predict the structure of that protein.

Table 2. *Best prediction results for our Bayesian method (Bayes-TG) over the set of 126 non-homologous proteins compiled by Rost and Sander (1994), and a set of 473 non-homologous proteins*[a]

Method	N_{chains}	Q_3	$C\alpha$	$C\beta$
Bayes-TG $^{2N}\phi(8)$	126	66.2	0.47	0.35
Bayes-TG $^{3N}\phi(8)$	473	67.5	0.50	0.39
Bayes-TG $^{3N}\phi(8)$	8	66.5	—	—
PHD	126	62.1	0.40	0.35
eNN	126	66.3	0.48	0.41
Homol.	126	67.6	—	—
Bayes-SLX	14	61.1	0.33	0.27

[a]We also report the results of "blind predictions" made for eight proteins in the CASP2 prediction experiment, as explained in the text. For comparison we show single-sequence-based prediction results over the same 126-protein dataset reported by other methods, including the neural network (PHD) of Rost and Sander (1994), the ensemble of neural networks (eNN) of Riis and Krogh (1996), and the nearest-neighbor method (Homol.) of Salamov and Solovyev (1995). We also include results obtained over a 14-protein dataset using a Bayesian statistical method developed by Stolorz et al. (1992). Q_3 is the three-state percentage correct predictions. The Matthew's correlation coefficients for α-helix and β-strand are C_α and C_β, respectively (Matthews, 1975).

All reported Q_3 scores are averaged over residues. In the following, we discuss the performance of the method in terms of Q_3 only, as the Matthews' correlation coefficients for our various prediction runs followed similar trends.

We find that the use of increased specificity of the structural description at the residue level via the use of solvent accessibility information is beneficial and relatively robust to the size of the window. Over the 126-protein dataset, using no solvent accessibility information ($^1\phi_j(4)$), the Bayesian method yields a peak accuracy of 62.7%, whereas if two or three categories of solvent accessibility are used ($^1\phi_j(8)$ or $^1\phi_j(12)$), respective peak accuracies of 65.1% and 65.7% are obtained. These results are shown in Figure 1A. All these peaks occur for window sizes in the range of 19 to 23 residues and memorization (decline in the jackknifed

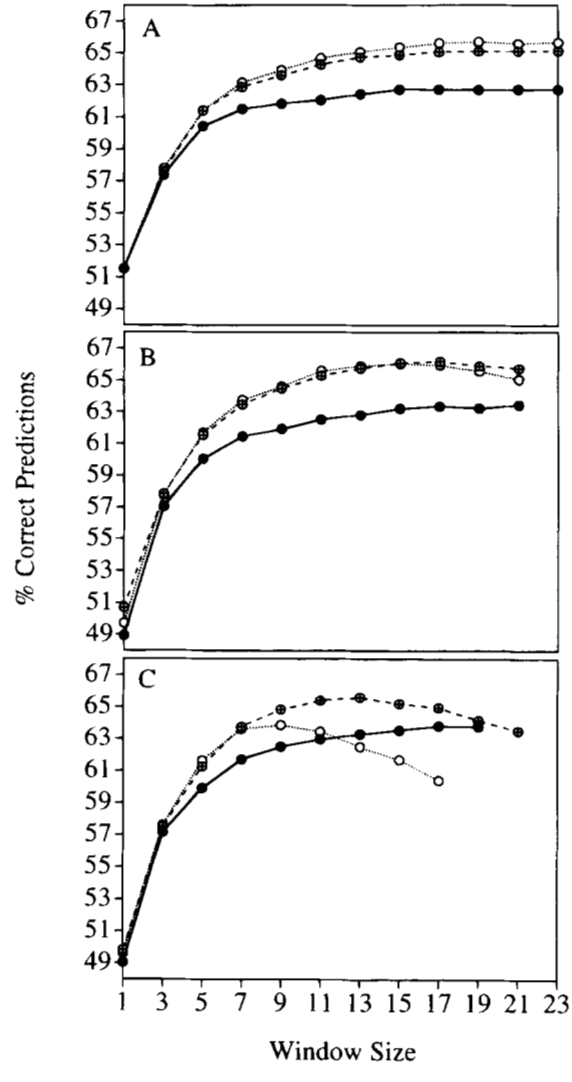

Fig. 1. A: Prediction accuracies (Q_3) over the 126-protein dataset as a function of window size for three types of singlet descriptors $^1\phi_j(4)$ (—●—), $^1\phi_j(8)$ (– – ⊕ – –), and $^1\phi_j(12)$ (···○···), as explained in Materials and methods. **B:** Same as the previous plot, but for the duplet descriptors, $^{2N}\phi_j(4)$ (—●—), $^{2N}\phi_j(8)$ (– – ⊕ – –), and $^{2N}\phi_j(12)$ (···○···). **C:** Same as the previous plots but for the triplet descriptors, $^{3N}\phi_j(4)$ (—●—), $^{3N}\phi_j(8)$ (– – ⊕ – –), and $^{3N}\phi_j(12)$ (···○···).

accuracies) does not occur until larger window sizes are attempted (data not shown).

With even greater specificity of structural description at the residue level, the phenomenon of memorization becomes increasingly apparent, as shown in Figures 1B and 1C for the use of structural n-tuplets, ${}^{2N}\phi_j$ and ${}^{3N}\phi_j$ over the 126-protein dataset. The peaks in accuracy for all curves are shifted to smaller window sizes relative to results obtained with singlet descriptors. While the use of two solvent accessibility categories is beneficial for both duplets and triplets (${}^{2N}\phi_j(8)$ and ${}^{3N}\phi_j(8)$), the use of three solvent accessibility categories in combination with n-tuplets provides overly specific information. Overall, the best accuracy (66.2%) was achieved with duplets (${}^{2N}\phi_j(8)$) for a window of 17 residues.

In general, the problem of memorization can most easily be addressed through the use of larger datasets, and so we would expect that larger datasets could support the use of richer structural descriptions. The highest accuracy (67.4%) for the 473-protein dataset was obtained using structural triplets (${}^{3N}\phi_j(8)$) rather than duplets. For this dataset, as with the 126-protein dataset, the combination of n-tuplets and three solvent accessibility categories showed a decline in accuracy (data not shown).

In the above discussion and accompanying figures, all accuracies have been reported for the N-terminal asymmetric n-tuplets because these consistently yield higher accuracies than the symmetric or C-terminal asymmetric n-tuplets. For instance, compared to the accuracy of 66.2% obtained over the 126-protein dataset with ${}^{2N}\phi_j(8)$, the accuracy for the corresponding C-terminal asymmetric duplet, ${}^{2C}\phi_j(8)$ was 65.9%. With the triplet descriptors over the 473-protein dataset, the accuracies were 67.4%, 67.1%, and 66.7% for ${}^{3N}\phi_j(8)$, ${}^{3}\phi_j(8)$, and ${}^{3C}\phi_j(8)$, respectively. Although these results may not be statistically significant, it is possible that the amino acid residues of a protein have a propensity to interact more strongly with their neighbors to one side rather than to the other. If this is the case, then according to the asymmetry observed in the accuracies using the n-tuplets, the structure to the N-terminal side of a residue location more strongly influences the amino acid identities at that location than does the structure to the C-terminal side. This implies that the inverse relationship should be true from the point of view of the sequence; the amino acid residues to the C-terminal side of a residue location should provide more information about the structure of that location than the residues to the N-terminal side. To explore this possibility, we made predictions at each of the positions in the window, rather than at just the central site. This was done over the 126-protein for two test cases—with ${}^{1}\phi_j(8)$ and a 13-residue window and with ${}^{1}\phi_j(12)$ and a 19-residue window. The results of this are shown in Figure 2A. The asymmetric distribution of accuracies over the positions in the window is quite clear, with the highest values occurring at positions in the N-terminal side of the the window. As expected, then, regarding the structural state of a given residue location, there is more useful information to be found in the neighboring residues extending toward the C-terminus of the protein.

Finally, we examined the accuracies as a function of window position for the parameter settings where the best performance for each dataset was observed. For the 126-protein dataset, using ${}^{2N}\phi_j(8)$, the best accuracy (66.2%) was found at the eighth position in a 19-residue window. For the 473-protein dataset, using ${}^{3N}\phi_j(8)$, the best accuracy (67.5%) was found at the ninth position in a 21-residue window. In both cases, the asymmetric distribution of accuracies is clear (Fig. 2B), but the differences in the performances between the peaks and the central positions are small.

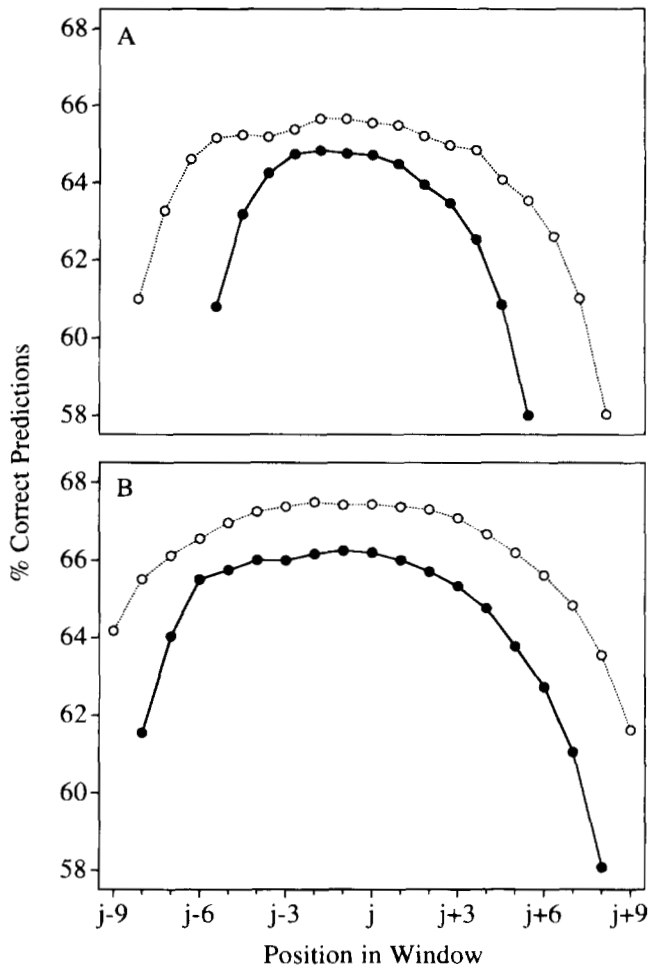

Fig. 2. A: Prediction accuracies (Q_3) over the 126-protein dataset as a function of window position for ${}^{1}\phi_j(8)$ and a 13-residue window (—●—), and for ${}^{1}\phi_j(12)$ and a 19-residue window (···○···). **B:** Prediction accuracies as a function of window position using asymmetric duplet descriptors, ${}^{2N}\phi_j(8)$, and a window of 17 residues over the 126 protein dataset (—●—) and using asymmetric triplet descriptors, ${}^{3N}\phi_j(8)$, and a window of 21 residues over the 473-protein dataset (···○···). j is a general position index over the local window of sequence, with $j = 0$ indicating the central position.

CASP2 performance

Using our single-sequence based method with a window of 19 residues and structural triplets of eight local structure categories (${}^{3N}\phi_j(8)$), we submitted predictions for a number of protein targets in the recent CASP2 prediction experiment. The purpose of this experiment was to gather predictions from various researchers in the field of protein structure prediction for several types of structure prediction. These predictions were made on proteins whose structures were not solved at the time of prediction, thus making them "blind predictions." The intent of the experiment and subsequent conference was to make objective comparisons of the various methods available. Results from CASP2 can be examined at http://predictioncenter.llnl.gov. For the eight protein targets for which we submitted predictions, we achieved a 66.5% accuracy. While this is a small sample of proteins, these results indicate that our method can perform at levels reported in the section above for proteins not included in our datasets.

Table 3. *Summary of the statistical performance of our method, Bayes-TG ($^{3N}\phi_j(8)$ and a window of 17 residues), over the three subsets of the 304-protein database*[a]

Dataset	Q_3	C_α	C_β
102A	71.6	0.58	0.47
101B	73.0	0.61	0.50
101C	70.2	0.56	0.44
304	71.6	0.58	0.47

[a]The last row gives the combined results over the subsets. Performance measures are as in Table 2.

Evaluation II: Schemata-based performance

As we are using multiple-sequence alignments, our method requires two types of cross-validation. The actual prediction calculations based either on amino acids or on substitution schemata are jackknifed in a single-chain exclusion procedure; all summations over residue sites in the calculations of the theory section are performed over all the proteins in the dataset except the one that is being predicted.

The cross-validation of the schemata, however, cannot be performed in a single-chain exclusion procedure because the optimization of the schemata is computationally intensive. Instead, similarly to the cross-validated training of neural networks, we divide the dataset into larger subsets for training and testing. In this work, we employed two variations of this idea.

Threefold cross-validation

First, we split our dataset of 304 proteins into thirds labeled 102A, 101B, and 101C to indicate the number of chains in each set. The optimization routine searched for sets of 44 schemata using structural duplets and no solvent accessibility information, $^{2N}\phi_j(4)$. The choice of 44 for the number of schemata is somewhat arbitrary. We used duplet structure descriptors in the optimization of these schemata.

The cross-validation was performed as in this example: The optimization of schemata was performed over 102A, with the set of schemata at each step of the search then used to predict the structures in 101B. Memorization occurs when the mutual information continues to increase over 102A, but the prediction accuracy declines over 101B. Taking the set of schemata prior to the onset of overlearning, we find the set of descriptors and the window size that maximizes the prediction accuracy over 101B. Finally, with this set of schemata derived from the dataset 102A, and with descriptor and window choices made for dataset 101B, we make predictions for the proteins in 101C. Thus, absolutely no information about the 101C dataset has been used in predicting the structures in that dataset, either through the statistical parameters or through the window size and descriptor alphabet choices. Permuting this procedure over the three subsets of proteins gives an estimate of 71.6% for the accuracy over the entire 304 proteins. These results are shown in Table 3. While the optimization of the schemata was performed using structural duplets ($^{2N}\phi_j(4)$), the highest prediction accuracies were achieved with structural triplets and two solvent accessibility categories ($^{3N}\phi_j(8)$) and a window of 17 residues. This was true for each of the three permutations of the cross-validation.

Table 4. *Summary of the statistical performance of our method, Bayes-TG ($^{3N}\phi_j(8)$ and a window of 17 residues), over the over the two subsets of the 304 protein database*[a]

Dataset	Q_3	C_α	C_β
152A	72.0	0.60	0.48
152B	72.6	0.61	0.49
304	72.3	0.60	0.49

[a]The last row is their combined results. Performance measures are as in Table 2.

Table 5. *Best prediction results for our method, Bayes-TG ($^{3N}\phi_j(8)$ and a window of 17 residues), using the twofold and threefold cross-validated schemata over 304 proteins*[a]

Method	N_{chains}	Q_3	$C\alpha$	$C\beta$
Matrix-WB	13	69.0[b]	—	—
Matrix-MA[c]	36	70.9[b]	—	—
PHD[c,d]	126	71.6	0.61	0.52
eNN[c]	126	71.3	0.59	0.52
Homol.[c]	126	72.2	0.64	0.50
DSC[c,d]	126	70.1	0.58	0.51
Bayes-TG 3-fold	304	71.6	0.58	0.47
Bayes-TG 2-fold	304	72.3	0.60	0.49

[a]For comparison we show prediction results reported by other methods, including the substitution matrix-based methods (Matrix) of Wako and Blundell (1994b) and of Mehta et al. (1995), the neural network (PHD) of Rost and Sander (1994), the ensemble of neural networks (eNN) of Riis and Krogh (1996), the local homology method (Homol.) of Salamov and Solovyev (1995), and the linear discriminant analysis (DSC) of King and Sternberg (1996). Performance measures are as in previous tables.

[b]Reported accuracies are averages over per-chain accuracies instead of per-residue accuracy.

[c]Employs post-prediction filtering.

[d]Includes global information, such as the fractions of residue types, fractions of predicted secondary structure types, and distances to N and C termini.

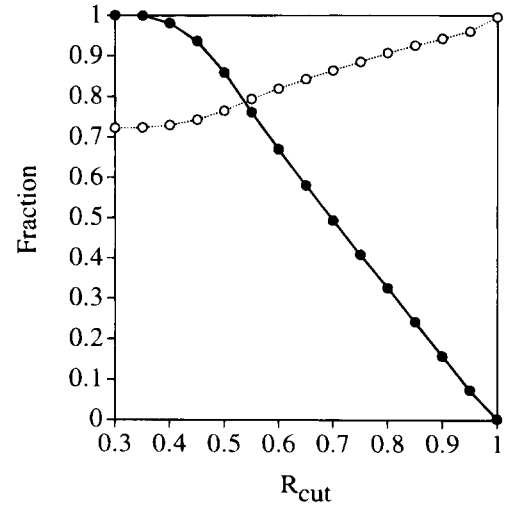

Fig. 3. (—●—) denotes the fraction of the dataset predicted with a reliability score, R, greater than R_{cut}. (···○···) denotes the fraction of these $R > R_{cut}$ predictions which were correct.

Twofold cross-validation

It would be beneficial to use larger datasets for the purpose of getting improved estimates of the various statistical parameters in our model. Therefore, we developed another cross-validation technique that divides the data into larger subsets. What we need is an estimate of the mutual information value that could, in general, indicate the onset of memorization. This value has to be obtained over a dataset different from the one over which we will estimate our prediction accuracy. To achieve this, we did the following.

We divided the 304-protein dataset into halves labeled 152A and 152B. Here we searched for sets of 40 schemata using triplets of secondary structure and no solvent accessibility information, ${}^{3_N}\phi_j(4)$. First, we optimized a set of schemata over the 152A dataset. For each set of schemata along the search pathway we obtained a prediction accuracy over the 152B dataset. Prior to the onset of memorization of the 152A dataset, we noted the mutual information value of the optimization over 152A. We then optimized a set of schemata over the 152B dataset. From this search, we selected the set of schemata whose mutual information value was nearest to that of the previous search. This set of schemata was used to predict the structures of the 152A dataset. Thus, this was done without knowledge of the prediction accuracies obtainable over the 152A dataset using schemata optimized over the 152B. There is no reason, a priori, to expect that the estimated "best" mutual information value obtained by the optimization over 152A should be able to pick out the schemata optimized over 152B that maximize the prediction accuracy over 152A.

To obtain a performance evaluation over the entire 304-protein dataset, we also performed the inverse of the above procedure. These results are given in Table 4. Using structural triplets and two

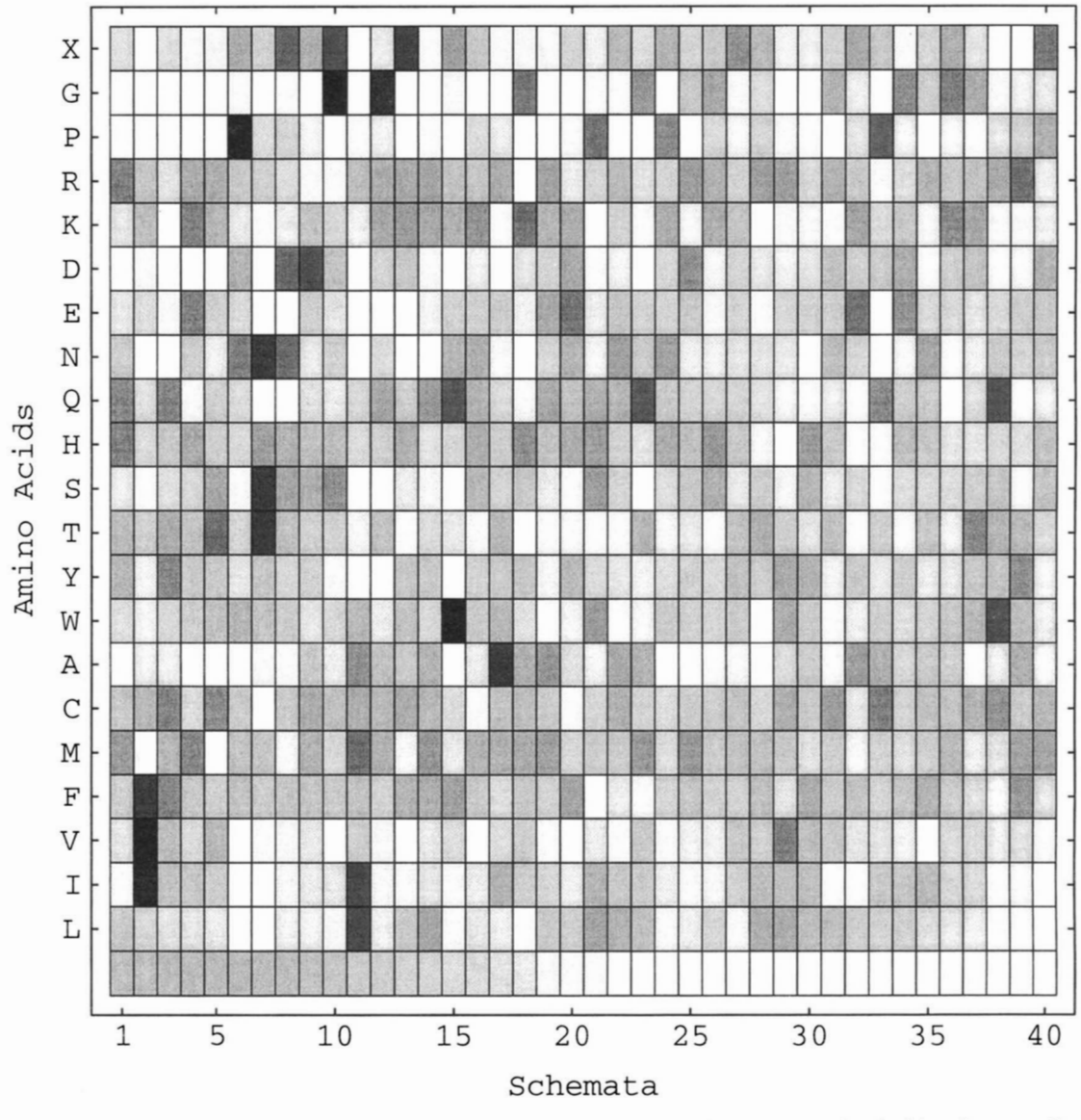

Fig. 4. Density plot representing the joint parameter values, $P(a_i, \eta)$, for each of the amino acids (a_i) and each of the schemata, η. Rows are labeled by the single-letter code of the 20 amino acids. X denotes gap. The first row (unlabeled) denotes the a priori probabilities (p^{η}) of the schemata. The parameter values range from 0 (white) to 1 (black). These schemata were optimized over the dataset 152A and used to predict 152B.

solvent accessibility categories (${}^{3N}\phi_j(8)$) and a window of 17 residues, an accuracy of 72.3% was achieved. In Table 5, we report prediction summaries of the two cross-validation protocols along with the best results reported by other authors.

Reliability

We note that with these probabilistic formalisms there is an easy means for calculating a confidence measure, R, for the predictions.

$$R \equiv \frac{\max_k [P(s^k|\{\vec{n}_j\})]}{\sum_k P(s^k|\{\vec{n}_j\})}. \tag{19}$$

In Figure 3 we plot the fraction of predictions made with an R value above a cut-off value ranging from 0 to 1 and the fraction of these subsets of the predictions that were correct. This measure is monotonic with prediction accuracy.

Physicochemical interpretability

The advantage of our method is the physicochemically transparent nature of the models we employ. Our prediction formalism is based on a simple evolutionary perspective of sequence dependence on structure, and our schemata represent a structurally meaningful condensation of the information derived from multiple-sequence alignments. Unlike the optimized weights of a neural network, the parameters of our model have a biophysical interpretation. Although it bears some mathematical similarity to the recent work of Goldman et al., our approach is diametrically opposed. Whereas they employ evolutionary information (phylogenetic trees) (Goldman et al., 1996), we use evolutionarily derived information (substitution schemata). Whereas they seek to model the structural information conveyed by "phylogenetic inertia" (Harvey & Pagel, 1991), we concentrate on modeling the correlations between patterns of substitution and local protein structure result-

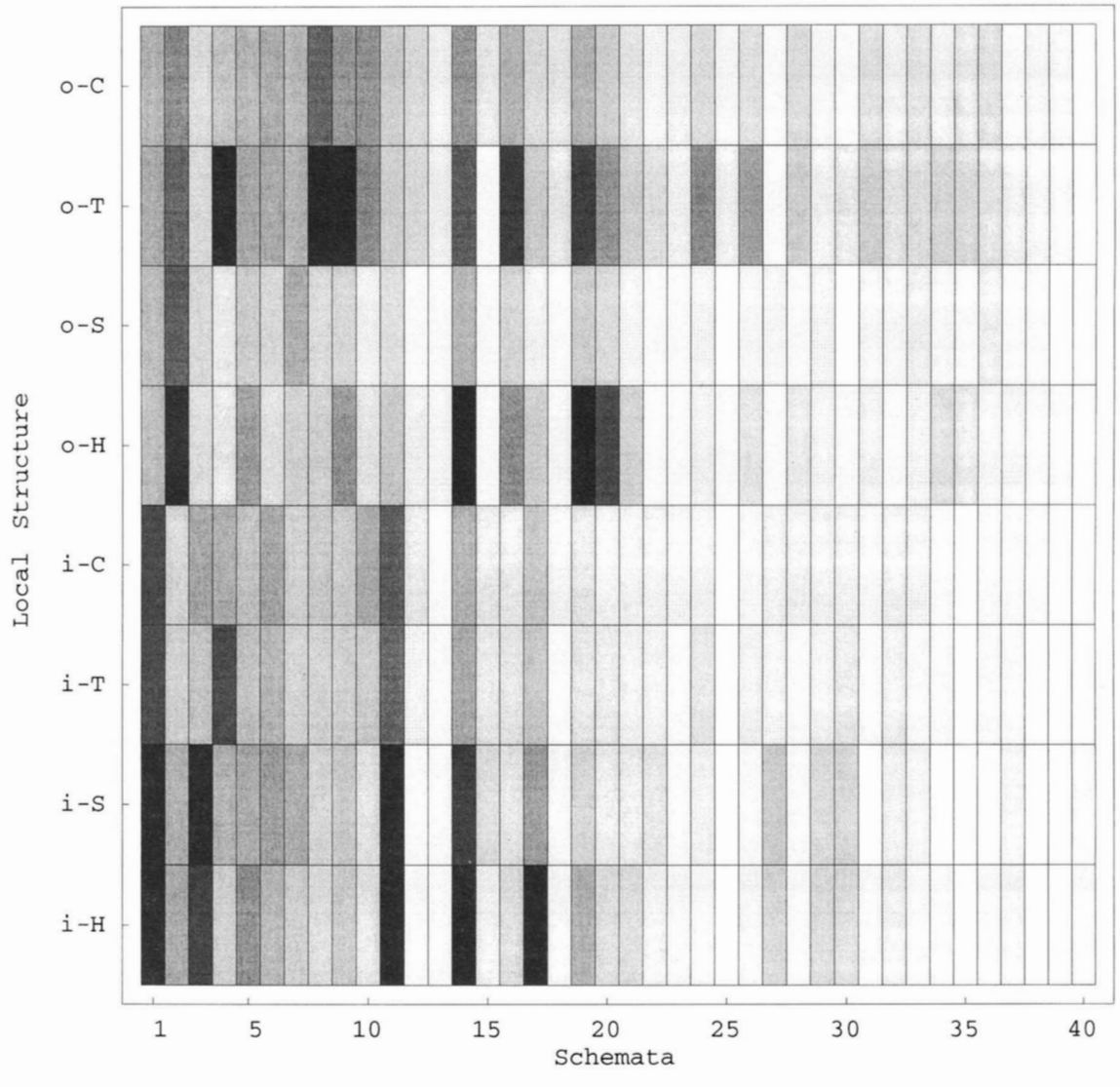

Fig. 5. Density plot representing the probability with which each schema (same set and order of schemata as in Fig. 2) is associated with each of eight types of local structure, $P(\eta, s^k)$. H denotes α-helix, S denotes β-strand, T denotes turn, and C denotes coil. The prefixes i- and o- denote "inside" and "outside" ($\leq$20% and >20% solvent accessible surface area), respectively. The parameter values range from 0 (white) to 1 (black).

ing from the "structural inertia" (Aronson et al., 1994) of protein evolution.

The parameters of our schemata can be taken as estimates of the probabilities that a particular amino acid would be "generated" according to each of the schemata. These probability estimates for the set of schemata derived from dataset 152B and used to predict 152A are shown graphically in Figure 4. In Figure 5, we display the probabilities with which each schemata is associated with each type of triplet structure. While we claim our method is transparent, we make no claims about the simplicity of the result. It is possible to pick out some patterns which fit with general intuition about the physicochemical identity of residues and their membership in these schemata. However, the information depicted in Figures 4 and 5 is rather complex. This is testimony to the pervasive contextuality of relationships between protein sequences and structures. This contextuality and the diversity of protein structures suggest that finding a general-purpose set of "expert heuristic" structure prediction rules of manageable size is unlikely. However, the pursuit of such rules leads to insights into protein structure formation that can be incorporated into probabilistic models, particularly within simple and mathematically rigorous formalisms such as ours.

Summary

We have demonstrated our simple Bayesian prediction formalism to the problem of predicting secondary structure. The advantages of our basic approach include its conceptual simplicity, ease of implementation, lack of ad hoc parameters, and low computational cost. With our method, overlearning or memorization is simply a problem with ill-defined probabilities resulting from overly specific structural descriptions—either the window size or alphabet of structure descriptors is too large. Moreover, as the database of proteins increases, this method requires no retraining like a neural network, or reconfiguring of the algorithmic architecture and rechoosing of the various parameter settings as in a more ad hoc approach like the nearest neighbor (local alignment) scheme. It is enough to merely add the new probabilities to the pre-existing ones.

We have also introduced a novel method for including evolutionary-derived sequence-to-structure correlations within our prediction method. This extended schemata-based formalism performs comparably to the best of methods using multiple-sequence alignment information. The use of a biophysically interpretable model makes this approach superior to neural network algorithms for the development of biophysical insight. The statistical performance and reproducibility of this method give it greater practical value than that of the expert heuristic methods. Thus, our method occupies a new niche in the field of secondary structure prediction—possessing some of the transparent qualities of the expert heuristic methods while having a demonstrated ability to perform well over large datasets. Lastly, this approach fits into a larger Bayesian framework which has already provided successful applications to solvent accessibility prediction and tertiary fold recognition (Goldstein et al., 1992, 1994; Thompson & Goldstein, 1996b).

Acknowledgments

We would like to thank Matthew Shtrahman and Jeffrey Koshi for helpful discussions and Kurt Hillig for computational assistance. We extend a general thanks to those who compile 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, NIH grant LM05770, and NSF equipment grant BIR9512955.

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein data bank. In: Allen FH, Bergerhoff G, Sievers R, eds. Crystallographic databases—Information content, software systems, scientific applications. Bonn: Data Commission of the International Union of Crystallography. pp 107–132.

Aronson HEG, Royer WE, Hendrickson WA. 1994. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci 3*:1706–1711.

Asai K, Haymizu S, Handa K. 1993. Prediction of protein secondary structure by the hidden Markov model. *CABIOS 2*:141–146.

Benner SA. 1989. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv Enz Regul 28*:219–236.

Benner SA. 1992. Predicting de novo the folded structure of proteins. *Curr Opin Struc Biol 2*:402–412.

Benner SA, Chelvanayagam G, Turcotte M. 1997. Bona fide predictions of protein secondary structure using transparent analyses of multiple-sequence alignments. *Chem Rev*. Forthcoming.

Benner SA, Gerloff DL. 1993. Predicting the conformation of proteins: Man versus machine. *FEBS Lett 325*:29–33.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Cover TM, Thomas JA. 1991. *Elements of information theory*. New York: John Wiley & Sons, Inc.

Gibrat JF, Garnier J, Robson B. 1987. Further developments of protein secondary structure prediction using information theory. *J Mol Biol 198*:425–443.

Goldman N, Thorne JL, Jones DT. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol 263*:196–208.

Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc Natl Acad Sci USA 89*:9029–9033.

Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1994. A Bayesian approach to sequence alignment algorithms for protein structure recognition. In: *Proc. 27th Annual Hawaii International Conference on System Sciences*. Los Alamitos: IEEE Computer Society Press. pp 306–315.

Govindarajan S, Goldstein RA. 1996. Why are some protein structures so common? Proc Natl Acad Sci. USA 93:3341–3345.

Govindarajan S, Goldstein RA. 1997. Evolution of model proteins. *Proteins*. Forthcoming.

Harvey PH, Pagel MD. 1991. Comparative methods for explaining adaptation. *Nature (London) 351*:619–624.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Holland J. 1992. *Adaptation in natural and artificial systems*. Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

King RD, Sternberg MJE. 1996. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci 5*:2298–2310.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. *J Mol Biol 235*:1501–1531.

Lim V. 1974. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J Mol Biol 88*:873–894.

Mamitsuka H. 1995. Representing inter-residue dependencies in protein sequences with probabilistic networks. *CABIOS 11*:413–422.

Matthews BW. 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biocim Biophys Acta 405*:402–451.

Maxfield FR, Scheraga HA. 1979. Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochem 18*:697–704.

Mehta PK, Heringa J, Argos P. 1995. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci 4*:2517–2525.

Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of α-helices. *Science 240*:1648–1652.

Riis SK, Krogh A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple-sequence alignments. *J Comput Biol 3*:163–183.

Robson B. 1974. Analysis of the code relating sequences to conformation in globular proteins. Theory and application of expected information. *Biochem J 141*:853–867.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins 19*:55–72.

Salamov A, Solovyev V. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple-sequence alignments. *J Mol Biol 247*:11–15.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*:56–68.

Shannon CE, Weaver W. 1949. The mathematical theory of communication. Urbana, Illinois: University of Illinois Press.

Shrake A, Rupley JA. 1973. Environment and exposure to solvents of protein atoms: Lysozyme and insulin. *J Mol Biol 79*:351–371.

Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian I, Haussler D. 1996. Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *CABIOS 12*:327–345.

Stolorz P, Lapedes A, Xia Y. 1992. Predicting protein secondary structure using neural nets and statistical methods. *J Mol Biol 225*:363–377.

Stultz CM, White JV, Smith TF. 1993. Structural analysis based on state-space modeling. *Protein Sci 2*:305–314.

Thompson MJ, Goldstein RA. 1996a. Constructing amino acid residue substitution classes maximally indicative of local protein structure. *Proteins 25*:28–37.

Thompson MJ, Goldstein RA. 1996b. Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins 25*:38–47.

Wako H, Blundell T. 1994a. 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.

Wako H, Blundell T. 1994b. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J Mol Biol 238*:693–708.

Yi TM, Lander ES. 1993. Protein secondary structure prediction using nearest-neighbor methods. *J Mol Biol 232*:1117–1129.

Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. *J Mol Biol 225*:1049–1063.