doi:10.1006/jmbi.2000.3837 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 301, 173–190

JMB

HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins

Christopher Bystroff†[1,2]*, Vesteinn Thorsson†[2,3]* and David Baker[2]*

[1]*Department of Biology Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA*

[2]*Department of Biochemistry University of Washington Seattle, WA 98195-7350, USA*

[3]*Department of Molecular Biotechnology, University of Washington, Seattle 98195-7730, USA*

**Corresponding authors*

We describe a hidden Markov model, HMMSTR, for general protein sequence based on the I-sites library of sequence-structure motifs. Unlike the linear hidden Markov models used to model individual protein families, HMMSTR has a highly branched topology and captures recurrent local features of protein sequences and structures that transcend protein family boundaries. The model extends the I-sites library by describing the adjacencies of different sequence-structure motifs as observed in the protein database and, by representing overlapping motifs in a much more compact form, achieves a great reduction in parameters. The HMM attributes a considerably higher probability to coding sequence than does an equivalent dipeptide model, predicts secondary structure with an accuracy of 74.3 %, backbone torsion angles better than any previously reported method and the structural context of β strands and turns with an accuracy that should be useful for tertiary structure prediction.

© 2000 Academic Press

Keywords: hidden Markov models; I-sites library; sequence patterns; motifs; clustering

Introduction

Proteins have recurrent local sequence patterns that reflect evolutionary selective pressures to fold into stable three-dimensional structures. Many of these local sequence patterns correlate with common structural motifs such as helix caps and beta hairpins. A general model of protein sequence that captures these local features could lead to improved methods for gene finding, protein structure prediction, remote homolog detection and other applications relating to the interpretation of genomic sequence information. Here, we describe the development of such a model, based on the I-sites library of sequence-structure motifs.

The I-sites (invariant or initiation sites) library consists of an extensive set of short sequence motifs, lengths 3 to 19 motifs obtained by exhaustive clustering of sequence segments from a non-redundant database of known structures (Han & Baker, 1996; Bystroff & Baker, 1998). Each sequence pattern correlates strongly with a recurrent local structural motif in proteins. Approximately one-third of all residues in the database are found in an I-sites motif that can be predicted with a high degree of confidence (>70 %). The library is non-redundant in that no motif is completely contained within another, longer motif. However, many of the motifs overlap. For example, the helix cap position may occur in the fourth position of one motif and the eighth position of another. Furthermore, the isolated motif model does not capture higher order relationships, such as the distinctly non-random transition frequencies between the different motifs. For example, sequences characteristic of amphipathic helices are frequently bracketed by N and C-terminal helix-capping motifs (Doig & Baldwin, 1995; Aurora & Rose, 1998).

The redundancy inherent in the I-sites model suggests a better representation that would model the diversity of the motifs and their higher order relationships while condensing features they have in common. A hidden Markov model (HMM; Rabiner, 1989) is well suited to this purpose. An HMM consists of a set of states, each of which is associated with a probability distribution for generating a symbol, such as an amino acid residue or a secondary structure type, and a set of transition probabilities between the states.

†These authors contributed equally to this work.

Abbreviations used: I-sites, invariant or initiation sites; HMM, hidden Markov model; EM, expectation maximization.

E-mail address of the corresponding authors: bystrc@rpi.edu,{thorsson,dabaker}@u.washington.edu

0022-2836/00/010173–18 $35.00/0

© 2000 Academic Press

Previous applications of HMMs to biological sequence data include the problems of finding coding regions in DNA and splice junctions (Kulp *et al.*, 1996, Burge & Karlin, 1998, Lukashin & Borodovsky, 1998, Henderson *et al.*, 1997), finding transmembrane regions in proteins (Sonnhammer *et al.*, 1998), and secondary structure prediction (Asai *et al.*, 1993, Stultz *et al.*, 1993; Di Francesco *et al.*, 1997, Lio *et al.*, 1998). Left-right, or feed-forward HMMs have proven to be very powerful representations of protein sequence families (Krogh *et al.*, 1994; Eddy, 1996; Sonnhammer *et al.*, 1997), able to pick out distant homologs that were undetectable by other sequence alignment methods (Karplus *et al.*, 1999), and to align correctly their sequences.

The HMM we discuss here, HMMSTR, is fundamentally different from previous Markov models of proteins. HMMSTR does not have a pre-defined ''left-right'' topology, as ''profile'' HMMs do (Eddy, 1998), but instead has a highly branched and multi-cyclic topology, discovered from the protein database through a motif clustering process. A single state in our model always represents a single position, i.e. there are no gap or insertion states. Unlike earlier HMMs that model sequence features specific to a single protein family, HMMSTR models local motifs common to all protein families. HMMSTR was trained simultaneously on sequence and structure databases, and shows considerable promise for a variety of applications, including gene finding and prediction of local structure and secondary structural context.

Results

Novel local structure (I-sites) motifs

We first briefly describe novel I-sites motifs found in preparing the construction of HMMSTR. Weak or rare sequence-structure correlations were found for 180 sequence-structure clusters beyond those published (Bystroff & Baker, 1998), including a few previously uncharacterized local structure motifs, such as the α-α corner and the Type-I′ β-hairpin. Additional new clusters contain alternative sequence patterns for previously characterized motifs. Others are new sequence patterns for rare and novel structural motifs, such as a glycine-rich α-helix N-cap, which differs in structure from the well-known N-capping box. Figure 1 shows some of the more interesting new motifs.

The α-α corner has been described by Efimov (1995), but was not fully characterized. A detailed profile of the α-α corner and the other new motifs are available in the latest version of the I-sites library along with a mapping of the position specific scoring matrix score to the confidence of the motif prediction. This prediction service is provided by the I-sites server (isites.bio.rpi.edu).

Description of the HMM

As a starting point for the work described here, each I-sites motif was represented as a chain of Markov states, each of which contains information about the sequence and structure attributes of a single position in the motif (see Figure 2, to be discussed in more detail below). Adjacent positions were represented by transitions from one state to the next. Merging of these linear chains of states, based on sequence and structure similarity (see Methods), resulted in graphs such as the one shown in Figure 3, in this case representing two ways of building a hairpin. The graphs were hierarchically merged, until almost all motifs were contained in a single graph. Branches, bulges (as in Figure 3) and long cycles, but not short cycles, were allowed to form during the merging process.

The merged graph of I-sites motifs comprises a network of states connected by probabilistic transitions, or more specifically, an HMM. Each state can produce, or ''emit'', amino acid residues and structure symbols according to a probability distribution specific to that state. There are four probability distributions defined for the states in our models, *b*, *d*, *r*, and *c* (see Figure 2), which describe the probability of observing a particular amino acid, secondary structure, backbone angle region (see Figure 5), or structural context descriptor, respectively. A context descriptor represents the classification of a secondary structure type according to its context. For example, a hairpin turn is distinguished from a diverging turn, and a β-strand in the middle of a sheet is distinguished from one at the end of a sheet. The conversion of the I-sites library to an HMM is advantageous, as training and application of HMMs is greatly facilitated by existing, powerful algorithms.

Three models are discussed here, denoted with a superscripted λ. Two (λ^D, λ^C) were created by clustering the I-sites motifs using observed adjacencies in the database, then by training the models on sequence plus secondary structure data (λ^D), and on sequence plus structural context data (λ^C). A third model was produced by clustering the I-sites motifs based on hierarchical pairwise alignments, followed by training on sequence plus backbone angle data (λ^R). Due to space considerations, only λ^R is illustrated here, in Figure 4. The differences in model topology were not thoroughly investigated, and are difficult to quantify, though strong similarities are clearly present. The models were improved by modifying the topology of the HMMs and optimizing parameters using a training set of 618 proteins. A considerable reduction in the number of initial parameters was found to improve the predictive power of the model in the applications described below. Thus, the final models are smaller than the initial merged graphs. The name HMMSTR refers to the three models collectively. The specific model to be used depends on the application.

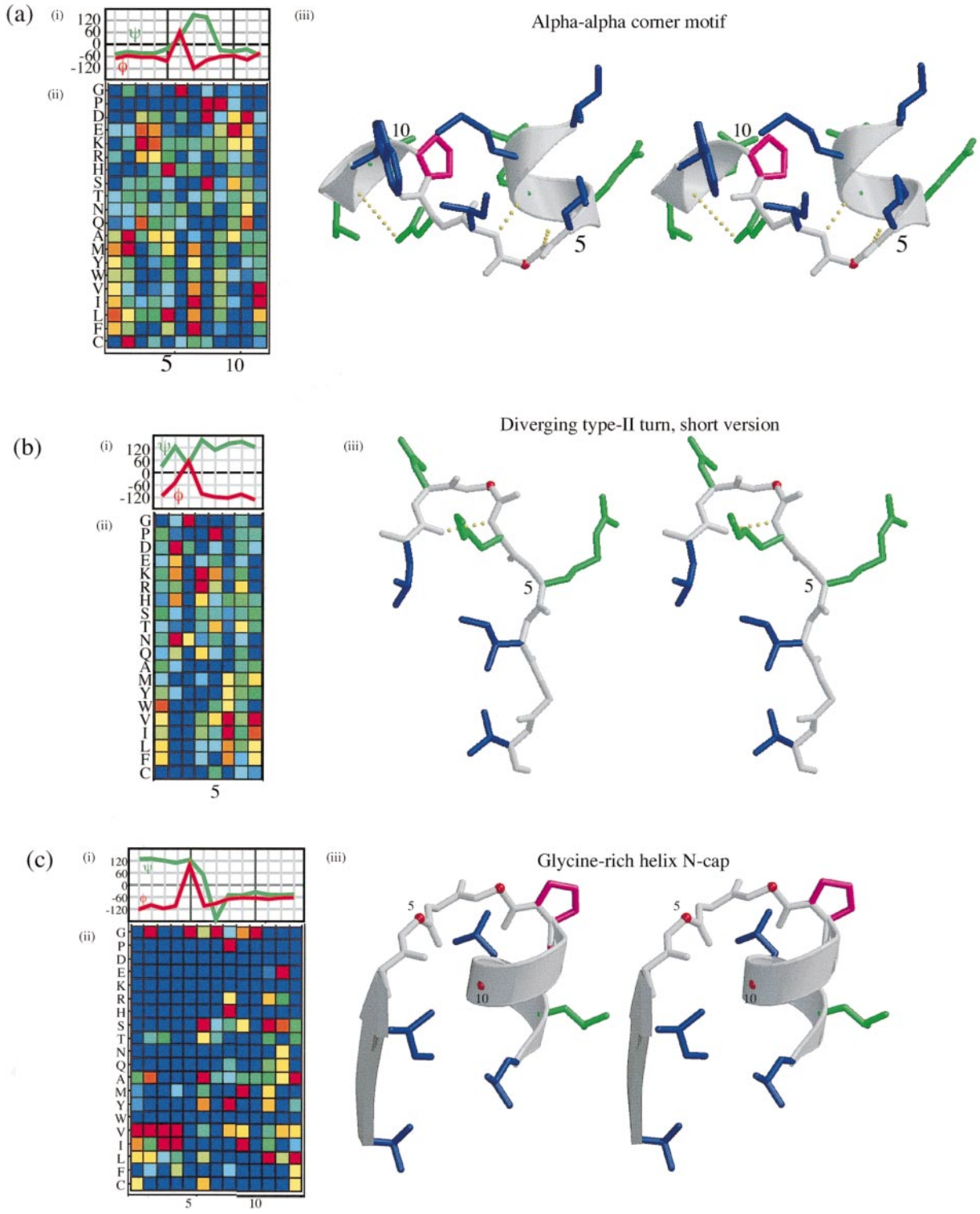

Figure 1. Rare or weak I-sites motifs. Three of the new weak or rare sequence-structure correlations found in the growing sequence-structure database since the publication of the I-sites library (Bystroff & Baker, 1998). (a) The α-α corner, first described by Efimov (1993). (b) A shorter version of the diverging β-turn described previously. (c) The glycine-rich helix N-cap, which appears to be novel. Shown are (i) backbone angles; (ii) sequence profiles, using a color scale of red (greater than three-times background amino acid frequency) to blue (less than one-third of background amino acid frequency), for representative structures (iii): (a) 1phg_163-174, (b) 1eny_5-12, and (c) 1pbe_5-15. Blue side-chains are conserved non-polar positions, green: conserved polar, magenta: conserved proline residue, and red dots: conserved glycine residue (Motifs were found by a clustering method, prior to development of the HMM.).

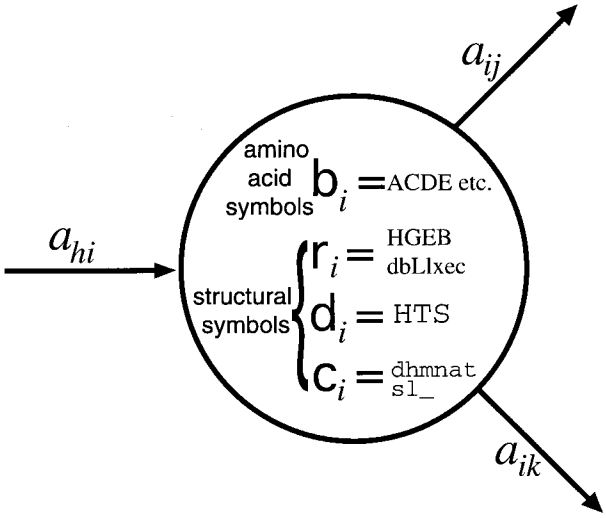

Figure 2. A Markov state. A hidden Markov model consists of Markov states connected by directed transitions. Each state emits an output symbol, representing sequence or structure. There are four categories of emission symbols in our model: *b*, *d*, *r*, and *c*, corresponding to amino acid residues, three-state secondary structure, backbone angles (discretized into regions of phi-psi space) and structural context (e.g. hairpin *versus* diverging turn, middle *versus* end-strand), respectively.

Clear statistical interactions between local structure motifs are encoded in HMMSTR. For example, as seen in Figure 4, different α-helix C-caps branch from amphipathic helices at different points relative to the amphipathic periodicity and are followed by different loop or strand motifs. The type-3 α-C-cap is often followed by a hydrophobic β-strand, while the α-C-caps types 1 and 2 are followed by amphipathic β-strands or a coil, and then another helix. The reason for the statistical interactions between motifs lies in the chemical interactions between conserved side-chains and in the geometric constraints imposed by the local structure. Favorable energetic interactions between adjacent motifs lead to greater abundance of those adjacencies in the database. Because this is an overview of what is clearly a complex system of interactions, no more will be said here about the underlying chemistry. Instead, we concentrate on establishing the validity of the model.

Applications of the HMM

The HMM captures recurrent features of both protein sequences and protein structures and thus has application to a wide range of problems of interest. Potential applications are summarized in Table 1 in terms of the input, output, and the probability evaluated for each application. The input is a sequence of known attributes of a protein for each position, such as an amino acid sequence or profile (*O*), a sequence of secondary structure symbols (*D*), symbols denoting the context in which the secondary structure symbols are found (*C*), or dihedral angle regions (*R*, see Figure 5).

Gene finding

Presenting the model with a potential coding sequence *O*, we may determine the likelihood that it is a genuine coding sequence by evaluating the probability of the sequence according to the model, $P(O)$. $P(O)$ is the sum, over all possible paths through the network, of the probability of that path multiplied by the probability that the path emits the sequence *O* (equation (1), in Methods).

Secondary structure prediction

A prediction of the three-state secondary structure symbol for each position in the sequence *O*, may be obtained by summing the state-specific probability distribution *d* (see Figure 2) multiplied by the position-specific state probability, over all states (see Methods). The predicted secondary structure state is the one with the highest value in the summed distribution.

Local and super-secondary structure prediction

Structural context symbols and dihedral angle regions are predicted much the same way as secondary structure, but using probability distributions *c* and *r*, respectively.

Sequence design

Since sequence and structure are formulated in a dual and symmetric fashion in our model, design proceeds analogously to structure prediction, i.e. a sequence of structure descriptors (i.e. *D*, *R* or *C*) leads to a sequence of position-specific probabilities for all states, which may be used to predict preferred residues.

Sequence alignment

The final entries in Table 1 concern the alignment of two amino acid sequences O_1 and O_2, as is commonly done in homology searches. One method to increase the sensitivity of such alignments is to derive amino acid substitution matrices specific to the local environment at each position from the HMM. An alternative is to compute the probability the two sequences take the same path through the HMM by the sum over all paths of the product of single path contributions to $P(O_1)$ and $P(O_2)$†.

† An efficient algorithm is needed to reduce the computational cost for the direct sum, as the standard forward-backward dynamic programming algorithm does for $P(O)$

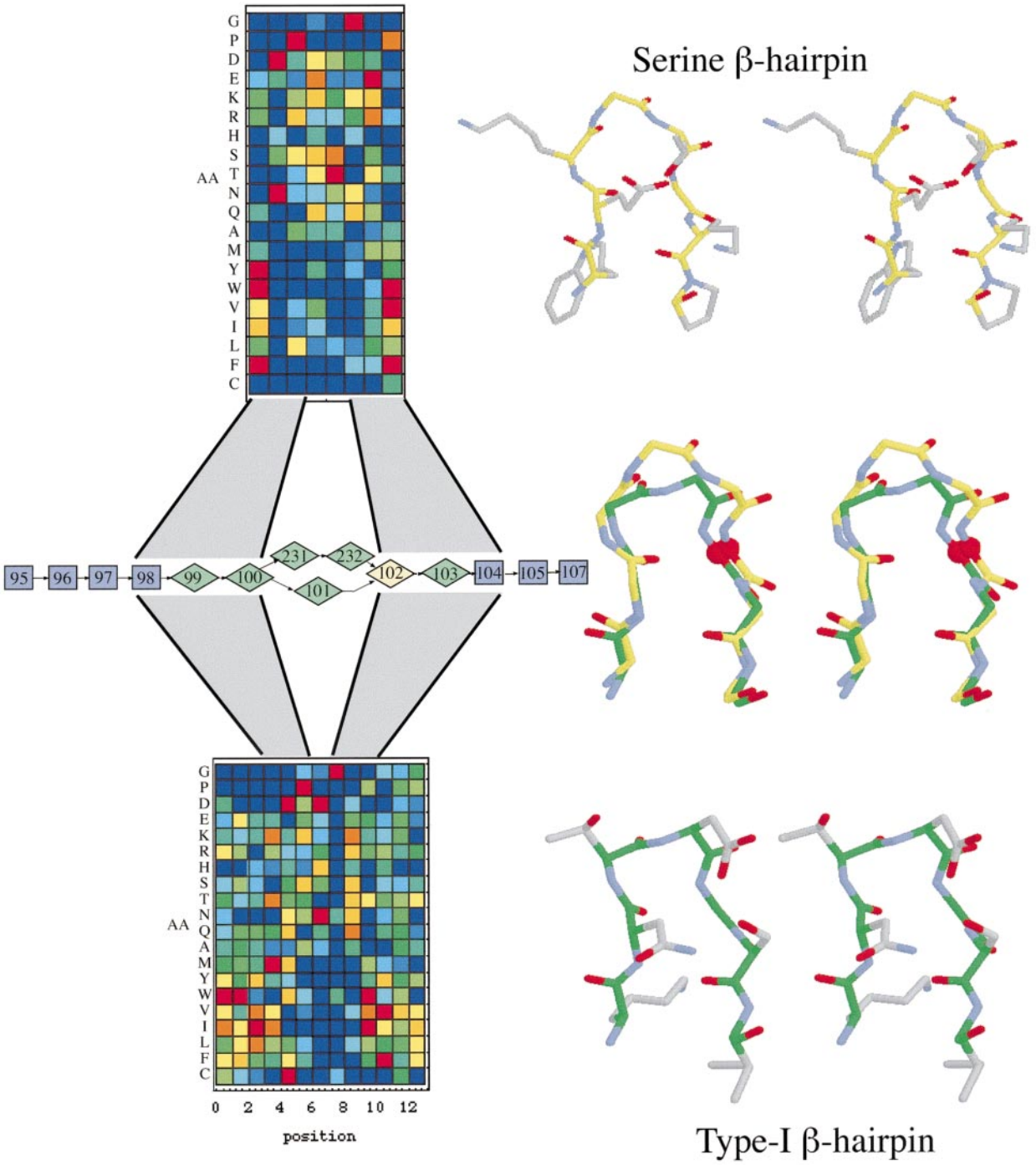

Figure 3. Merging of two I-sites motifs to form an HMM. The extended Type-I hairpin motif and the Serine hairpin align in 3D space and in sequence with a short mismatch in the turn. The color scheme is as described in legend to Figure 1. The resulting HMM topology is shown. The shaped icons represent Markov states, with probabilistic emission properties (see Figure 2). Rectangles are predominantly β-strand states, and diamonds are predominantly turns. The color of the icon indicates a sequence preference as follows: blue, hydrophobic; green, polar; yellow, glycine residue. Numbers in icons are arbitrary (unique) state identifiers.

Performance summary

Table 2 gives a summary of the overall performance of the model for a number of possible applications, and below we discuss the results in more detail. The performance is summarized for the training set and for a randomly selected, independent test set of 61 proteins. In nearly all cases, the scores of the test and training sets are comparable; indicating that the process of refining the model has not resulted in overfitting of the data. Performance is evaluated on models λ^D, λ^C, and λ^R. The sequence score $L(S; \lambda^D - \lambda^{bg})$ is a measure of the probability of sequences occurring under the

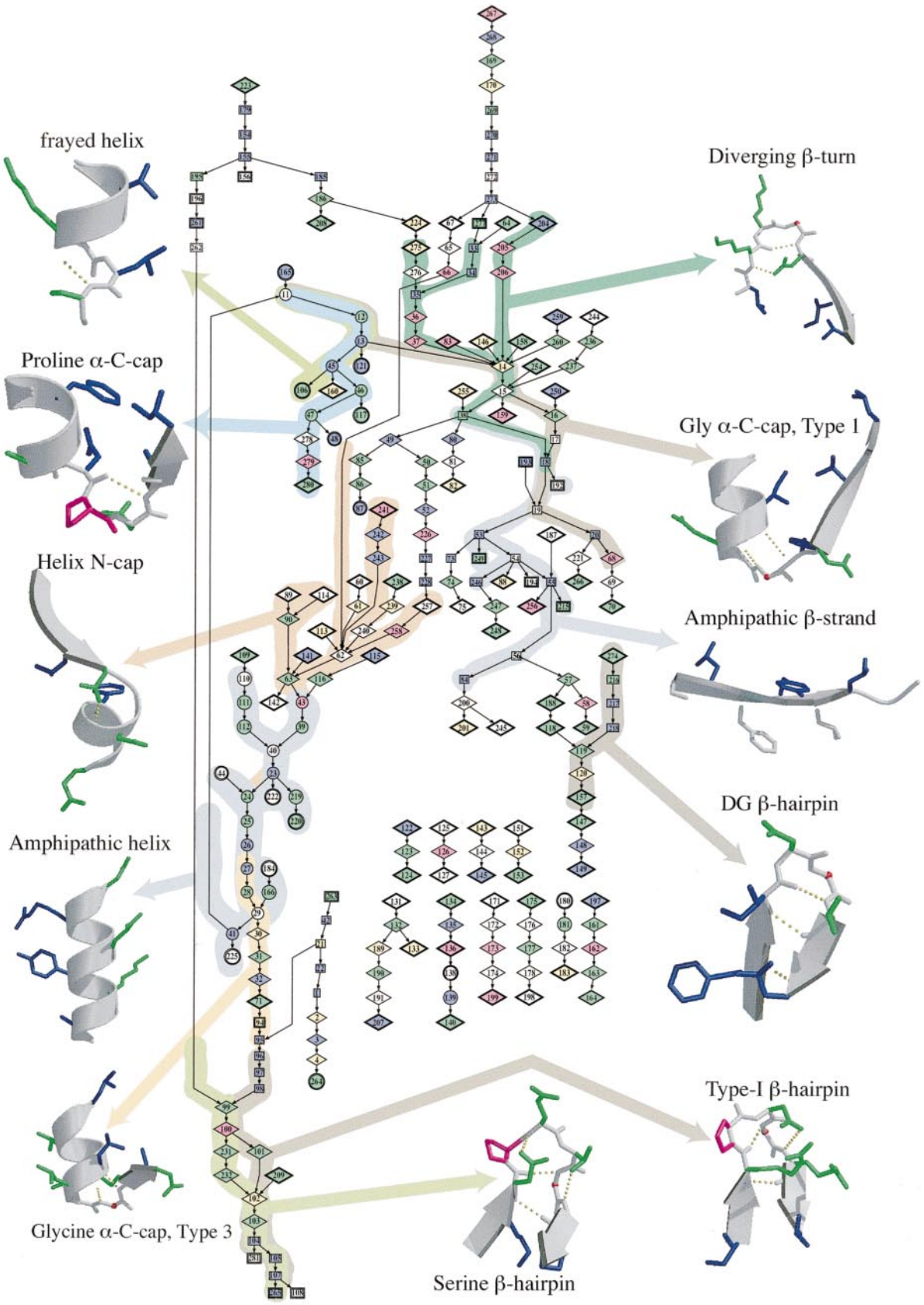

Figure 4 (*legend opposite*)

Table 1. Applications of the HMM

Application	Input	Output	Measure
Gene finding	Single sequences (O)	Likelihood for coding	$P(O)$
Secondary structure prediction	Sequence profile (O)	Secondary structure (D)	$P(D\|O)$
Structural context prediction	Sequence profile (O)	Context (C)	$P(C\|O)$
Dihedral angle region prediction	Sequence profile (O)	Dihedral angle region (R)	$P(R\|O)$
Protein design	Structure (D, C, R)	Sequence (O)	$P(O\|D, C, R)$
Sequence comparison	Sequence 1 (O_1) Sequence 2(O_2)	Likelihood for alignment	$P(O_1 \sim O_2)$

Applications of the model are given in terms of input for the application, output, and the corresponding measure or statistic that may be evaluated with the HMM. The measures are probabilities P, where a vertical bar specifies a conditional probability. O denotes a sequence of amino acid residues, D a sequence of secondary structure symbols, C a sequence of structural context symbols, and R a sequence of dihedral angle region symbols. In the final entry, $\sim$ denotes a similarity measure.

model λ^D, relative to the probability under a simpler model which takes into account only amino acid composition (λ^{bg}). In the first row of Table 2 this score is evaluated on single sequences, as is relevant for the case of novel sequences with no known homologs. The value on the test set, 0.024, indicates that a protein of length 350 residues generally has probability $2^{0.024 \times 350} = 338$ times greater with λ^D than with the simpler model. For profile-based evaluation, our sequence score is 0.226 on the test set. The *Q3* score is the fraction of positions correctly assigned to one of the three states: helix, strand, or turn. The value on the test set $Q3 = 74.3\,\%$, is an indication that local structure at the level of three-state prediction is accurately reproduced by λ^D. The next two lines in Table 2 show the accuracy of high-confidence (>70 %) predictions of structural ''context'', β-hairpin *versus* β-diverging turn on the one hand and middle β-strand *versus* end β-strand on the other, for model λ^C. The *MDA* score is the fraction of residues that are found in correctly predicted eight-residue segments, i.e. segments in which no predicted backbone torsion angle differs by more than 120° from the true structure. The overall MDA score, 59 % on the test set, for model λ^R, is substantially higher than the 48 % overall accuracy reported in our earlier work (Bystroff & Baker, 1998). The success of the secondary structure, context and torsion angle predictions indicates that these predictions may give useful constraints for tertiary structure prediction (Simons *et al.*, 1998). For the applications above, scores evaluated from the two models other than the one quoted were only slightly lower.

Identification of coding regions

The desired test for the potential of our HMM to aid in the identification of coding regions involves integration into one of the several excellent gene identification systems now in use (see e.g. Li, linkage.rockefeller.edu/wli/gene/), and evaluation by comparison with existing methods. In lieu of a comprehensive test of this nature, we compare our sequence scores to those obtained from simple but relevant models of protein sequence. We also evaluate scores on randomized sequences.

The sequence score $L(S;\ \lambda - \lambda^{bg})$ can be used to gauge whether or not a given string of amino acid residues or a profile is likely to be a protein sequence, as modeled by the HMM λ. The score is given by the logarithm of the probability of the sequence given λ, minus the corresponding quantity given the very simple ''background'' model λ^{bg}. This is equivalent to a likelihood ratio test, which assesses which of the two models is best supported by observed protein sequences. In the background model, the probability of a sequence is a product of independent contributions from each position, where the probability of a residue is given by the frequency of that residue in the training set as a whole. To facilitate comparisons in the context of information theory we use the base 2 logarithm and average over all positions to obtain $L(S;\ \lambda - \lambda^{bg})$ in bits per position. In terms of information content, $L(S;\ \lambda - \lambda^{bg})$ is the reduction of uncertainty, or information gain, in using λ as a model of sequence rather than λ^{bg}. (See Schneider, T., http://www.lecb.ncifcrf.gov/ ~ toms/paper/primer/for a primer on information theory in biology.)

HMMSTR significantly improved in general performance when profiles were used for training

Figure 4. The topology of the model λ^R and the locations of selected I-sites motifs. Icon colors and shapes as in Figure 3, and with circles predominantly helix; magenta, proline; white, no sequence preference. All forward transitions with probability greater than 0.001 and all Markov states with posterior probabilities (obtained by summing over all database positions) greater than 0.0001 are shown. Icons with bold borders represent sinks and sources. All sinks are connected to all sources through transitions to the non-emitting ''nought'' state (not shown). Disjoint graphs (lower right) connect to the main graph only through the nought state. Backshaded pathways are mapped to I-sites motifs, whose structures are drawn in the margins. In some cases, there are multiple pathways that map to a single structural motif (e.g. helix N-cap) and in other cases, multiple local structure motifs share a common segment of the model (e.g. states 14 and 15).

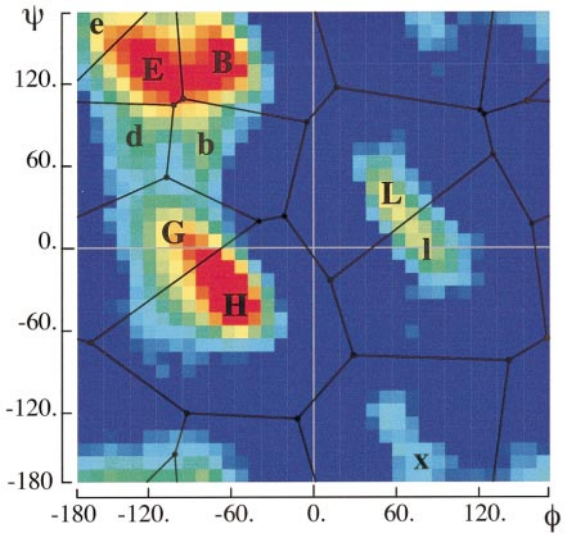

Figure 5. Backbone angle regions. The color scale indicates relative frequency of occurrence in the database. Frequently occurring $\phi\psi$ angles are shown in red and infrequent ones in blue. Not shown is the *cis*-peptide region (c).

instead of single sequences. For gene identification, scores on single sequences are relevant. The distribution of the single sequence scores for individual proteins in the training set (0.026 bits/residue overall) and the test set (0.024 bits/residue) is given in Figure 6. There is no clear correlation between the per-residue score and the length of the sequence (data not shown). A preliminary study indicates that training on a properly weighted set of single sequences rather than profiles can give rise to a per-position score of 0.05 or more, but further evaluation is required (see Discussion).

Table 2. Summary of performance measures

Statistic (model)	Training set	Test set
$L(S; \lambda^D - \lambda^{bg})$, single sequences (bits)	0.026	0.024
$L(S; \lambda^D - \lambda^{bg})$, profiles (bits)	0.236	0.226
Q3 secondary structure (λ^D) (%)	74.9	74.3
Hairpin *vs.* diverging turn (λ^C) (%)	75.1	85.2
End *vs.* middle strand (λ^C) (%)	77.3	78.4
MDA ratio (λ^R) (%)	57.1	59.1

Performance is evaluated on models λ^D, λ^C, and λ^R optimized for prediction of secondary structure, structural context, and dihedral angle regions, respectively. $L(S; \lambda^D - \lambda^{bg})$ is the sequence score per position, in bits, evaluated on single sequences or on profiles. *Q3* is the fraction of positions correctly predicted to be helix, strand, or turn. For turns separating two strands, we give the fraction of positions that are correctly predicted in a hairpin or diverging configuration. For strands, we give the fraction of positions correctly predicted as middle strands (e.g. interior of a β-sheet) or end strands (e.g. boundary of a β-sheet). The latter two measures are shown for predictions with greater than 70% confidence. The *MDA* ratio is the fraction of residues in segments of length 8 with no greater than 120° maximum deviation between observed and predicted backbone torsion angles.

Table 3 summarizes result for scores on single sequences for one of our models, λ^D, and a ''dipeptide model'', λ^{dip}, evaluated on natural protein sequences and on two randomly generated sequence sets. The dipeptide model, λ^{dip}, is a 20-state Markov chain model that encodes observed dipeptide frequencies in its transition probabilities. The sequence sets, S^{bg} and S^{dip}, were generated stochastically using the overall amino acid frequencies and the dipeptide frequencies, respectively. Dipeptide content is related to dicodon content, used successfully as a component in gene recognition methods (Fickett & Tung, 1992).

We first note that the score for the model λ^D on the database, 0.0260, is considerably higher than the corresponding score for the dipeptide model, 0.0072. To evaluate the significance of this score difference, we compared the distributions of these scores over all proteins. Since the scores are paired (one score from λ^D and one score from the dipeptide model for each protein), we carried out a paired *t*-test with the null hypothesis being that the mean of the distribution of the differences between the scores of λ^D and the dipeptide model, taken for each protein, is zero. For the test set, the paired *t*-test rejects the null hypothesis (*P*-value $P = 2 \times 10^{-5}$, *t*-statistic $t = 4.5$, 95% confidence interval $CI = (0.009, 0.022)$). For the training set, the hypothesis is rejected with $P = 2 \times 10^{-71}$, $t = 17.8$, $CI = (0.020, 0.025)$. This suggests that the score $L(S; \lambda^D - \lambda^{bg})$ is likely to be a useful addition to gene identification methods. Secondly, there is a trend toward smaller scores going from right to left in Table 3. For λ^D, this is consistent with the fact that λ^D is optimized on S^{db} and would expect to fare less well on sequences lacking correlations found in natural sequences. For scores using λ^{dip}, the downward trend and the interesting relations among the magnitude of the scores may be confirmed independently by relatively simple calculations (see Appendix).

Three-state secondary structure prediction

The *Q3* score is the fraction of positions correctly assigned to one of the three states: helix, strand, or

Table 3. Single sequence score

Model	S^{bg}	S^{dip}	S^{db}
λ^{dip}	−0.0073	0.0073	0.0072
λ^D	−0.0105	−0.0071	0.0260

$L(S; \lambda^D - \lambda^{bg})$ in bits per position, for selected models and sequence sets. Models; λ^{dip}, dipeptide model; λ^D, HMMSTR. Sequence sets: S^{bg}, random amino acids with background distribution; S^{dip}, random with dipeptide distribution; S^{db} training set.

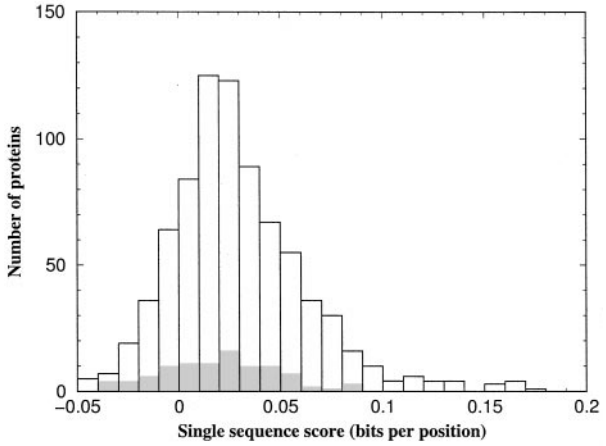

Figure 6. Distribution of single sequence scores. The number of proteins with single sequence scores within intervals of width 0.01 bits per position is shown. White bars, training set; Gray bars, test set.

turn. Evaluating the prediction of λ^D on the test set, we obtain $Q3 = 74.3\,\%$, an indication that local structure at the level of three-state prediction is accurately reproduced. Our initial models gave $Q3 = 55\,\%$, and the score was greatly increased by factors such as training, pruning, and the use of profiles instead of single sequences. Significant improvements were also obtained through the use of a voting scheme, as a means of combining contributions of multiple paths through the HMM, instead of basing predictions on the single most likely (Viterbi) path. The early statistical prediction methods of Chou & Fasman (1978) gave $Q3 = 60\,\%$, while recent work on neural net models report $Q3 = 74\,\%$ (Rost, 1997) and $Q3 = 77\,\%$ (Jones, insulin.brunel.ac.uk/psipred/).

As shown in Table 4, there are some imbalances in the secondary structure prediction. For example, the model overpredicts turns, and underpredicts helices and strands. However, as indicated in Figure 7, the length distributions of secondary structures are reproduced fairly well. The true structure of segments incorrectly predicted to be short helical segments of length 1-3 is highly variable. Naïve conversion of such segments to a different secondary structure simply tends to introduce new types of errors, such that there is no net gain in *Q3* score. However, some improvement in prediction quality might be produced by a post-processing step analogous to the second layer in the PHD secondary structure prediction method, which reassigns the predicted secondary structure at a position based on the predictions at surrounding positions (Rost & Sander, 1994). The complex topology of HMMSTR clearly provides the flexibility to accurately model secondary structure length distributions.

Backbone angle prediction

Backbone angle predictions were assessed using the *MDA* score (Bystroff & Baker, 1998), which is an appropriate measure for low-resolution backbone angle prediction, and is more precise than the *Q3* score. The *MDA* score is the percentage of all residues that are found in correctly predicted eight-residue segments, meaning 8-mers in which no backbone angle deviates by more than 120° from the true value (Bystroff & Baker, 1998). Pairs of 8-mers with a maximum backbone angle deviation of 120° or less tend to have backbone *rmsds* (root mean square deviations of superimposed backbone atom positions) of less than 1.4 Å.

Backbone angles were predicted using a voting procedure to combine contributions from multiple Markov states. A hierarchical voting procedure gave an overall *MDA* score of 58.8 % on training data, and outperformed a non-hierarchical one, which gave 56.7 %. The hierarchical voting method overpredicts the most heavily populated angle regions: *H*, *E* and *B* (Figure 5) and under-predicts most of the sparsely-populated regions, especially regions *b*, *d*, and *e* (Table 5), although the distribution is even more skewed using the non-hierarchical procedure. States representing *b*, *d*, and *e*

Table 4. Three-state secondary structure prediction

	H^{pred}	S^{pred}	T^{pred}	Total
A. *Training set*				
H^{obs}	35,943	1794	8983	46,720
S^{obs}	1484	18,154	10,454	30,092
T^{obs}	7115	6406	54306	67,827
Total	44,542	25,354	73,743	144,639
B. *Test set*				
H^{obs}	4252	276	1160	5688
S^{obs}	178	1972	1129	3279
T^{obs}	706	619	5547	6872
Total	5136	2867	7836	15,839

Training and test set positions are categorized by observed and predicted three-state secondary structure for the model λ^D. The right margin shows the summed observed observations, and the bottom margin shows the summed predictions for each secondary structure. *H*, helix; *S*, strand; *T*, turn.

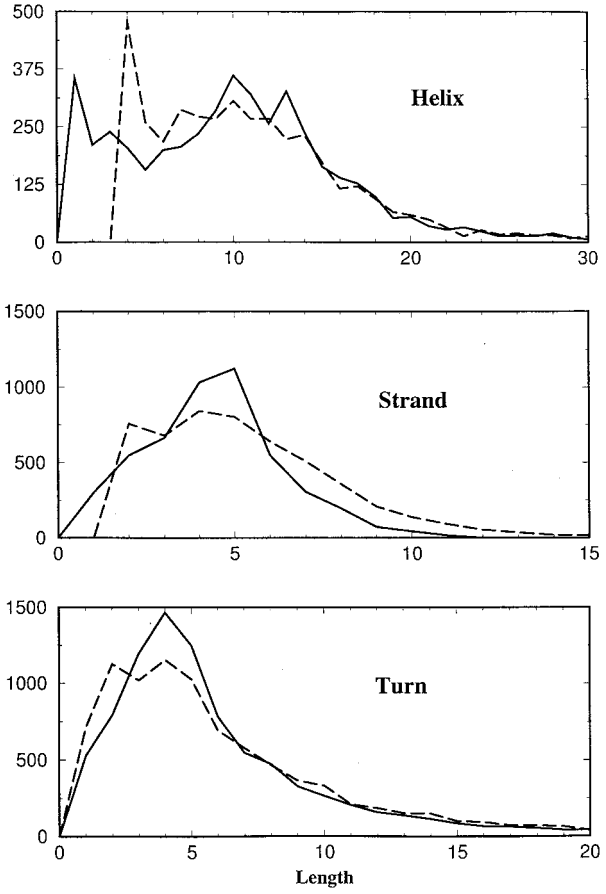

Figure 7. Length distributions for three-state secondary structures. For a given three-state secondary structure, the ordinate represents the number of occurrences of contiguous regions of that state, of length given on the abscissa. Solid line, prediction by the model λ^D, broken line, true distribution. Uppermost panel, helix; middle panel, strand; Lowest panel, turn.

often represent them weakly, having diffuse backbone angle emission probabilities (*r*), while most states representing regions H, E and B have sharply peaked *r*-vectors, sometimes to the exclusion of all but one region. It may be that the rarer backbone conformations, corresponding to the sparsely populated *r* regions, may be determined more by non-local strains and stresses than by strong local sequence signals, in which case they would not be accurately modeled by the HMM.

The prediction accuracy of the I-sites library was $MDA = 48\,\%$, and 54 % when combined with PHD secondary structure predictions (Rost, 1997). The current results, $MDA = 59\,\%$, represent a considerable improvement in accuracy for the prediction of backbone angles. A paired *t*-test establishes the significance of this difference at $P = 10^{-4}$ (*t*-statistic $t = 4.1$, 95 % confidence interval $CI = (0.04,\ 0.11)$) for the test set, and $P = 2 \times 10^{-44}$ ($t = 14.1$, CI = (0.06, 0.08)) for the training set.

Context prediction

We examined the ability of the HMM to predict local structural context, in addition to secondary structure and backbone angles. Short turns (<eight residues) between secondary structure were classified according to the types of secondary structure that bracket them: helix-turn-helix, helix-turn-strand, strand-turn-strand, etc. Turns between two strands were designated to be ''hairpins'' (*h*) if the strands were paired, and ''diverging turns'' (*d*) if they were not. Table 6 shows results of predictions from model λ^C. Short turns that were correctly predicted to be between strands were binned according to the confidence of the prediction that the turn is diverging, rather than hairpin, $P_d/(P_d + P_h)$. Results for the training set and test set are quite comparable. Note that hairpin and diverging turn predictions of high confidence are generally correct.

The context of a strand unit is defined as ''middle'' (*m*) if there are hydrogen-bonding β-strands on both sides of it, otherwise it was an ''end'' strand (*n*) (we are using the DSSP definition of β-strand residues, in which a residue is part of a strand if it participates in strand-strand H bonds). The context is defined residue by residue, thus β-strands may be partly ''middle'' and partly ''end''. In Table 7, positions that were correctly predicted to be strand positions using λ^C were binned according to the value of $P_n/(P_u + P_m)$, the confidence that the position is found in an end strand. Evaluation on both training and test sets shows the desired behavior for high-confidence predictions.

Table 2 summarizes results for confidence above 70 %. The high scores indicate that there is in some cases strong and reliable sequence signal for structural context. We are unaware of any precedent for this type of prediction, which may be considered super secondary structure prediction, and therefore cannot compare our results with existing methods. We are presently investigating the utility of these context predictions for tertiary structure prediction algorithms.

Discussion

The HMM developed in this paper captures simultaneously the recurrent local features of protein sequences and protein structures in a single compact form. In contrast to the more familiar family-specific HMMs (Sonnhammer *et al.*, 1997, Eddy, 1998), which have shown considerable power in remote homolog detection, HMMSTR models features common to protein sequences and structures generally. The models λ^D, λ^C, and λ^R, and c++ source code for model training and evaluation, are available at isites.bio.rpi.edu/hmmstr/.

Comparison to the I-sites library model

HMMSTR extends and generalizes the I-sites library of sequence-structure motifs. First, the

Table 5. Backbone angle region prediction

	H^{pred}	G^{pred}	B^{pred}	E^{pred}	d^{pred}	b^{pred}	e^{pred}	L^{pred}	l^{pred}	x^{pred}	c^{pred}	Total
H^{obs}	6909	42	593	589	2	0	13	0	15	37	0	8200
G^{obs}	887	55	266	198	0	0	3	0	10	6	0	1425
B^{obs}	904	19	1429	747	1	2	19	0	6	31	0	3158
E^{obs}	677	8	598	1737	1	0	13	1	6	10	0	3051
d^{obs}	107	3	129	127	10	0	0	0	1	2	0	379
b^{obs}	174	4	165	174	0	1	1	0	1	0	0	520
e^{obs}	188	5	162	201	0	0	35	0	16	34	0	641
L^{obs}	214	14	83	52	0	0	12	0	21	19	0	415
l^{obs}	140	2	35	29	0	0	33	2	107	74	0	422
x^{obs}	62	3	38	16	0	0	17	0	18	45	0	199
c^{obs}	6	0	33	5	0	0	1	0	1	1	0	47
Total	10,269	154	3531	3875	14	3	147	3	202	259	0	18,640

Position-wise predicted backbone angle regions are tabulated according to their true values. Letters in the margins refer to the regions of phi/psi space mapped out in Figure 5. c denotes residues N-terminal to a *cis*-peptide bond. The right margin shows the summed observations, and the bottom margin show the summed predictions in each region.

non-random transitions between different sequence motifs are described. For example, a helix followed by a certain type of helix cap tends to transition into a strand further downstream, whereas a helix followed by a different type of cap preferentially transitions into another helix. These regularities in the order of occurrence of motifs are not treated in the original I-sites model, but are compactly represented by transitions between motifs in the HMM. In this sense, the model captures the grammatical structure of protein sequence. Second, overlapping motifs are condensed, resulting in a more complete description with fewer parameters. The I-sites library, from which the HMM was built, may be viewed as a mixture model (Bailey & Elkan, 1994) where the I-sites motifs are the mixture components. Many of the mixture components contain overlapping portions of the same sub-segment pattern, both in sequence and in structure (Figure 3); the overlapping regions are represented by a single set of states in the HMM. The added information and condensed representation are likely to account for the considerably improved local structure prediction capabilities of the HMM compared to the original I-sites model, which predicted secondary structure with only 64 % accuracy and local structure (MDA) with 48 % accuracy.

Table 6. Hairpin *versus* diverging turn

$P_d/(P_h + P_d)$	Hairpin	Diverging
A. *Training set*		
0.0-0.1	74	36
0.1-0.2	235	68
0.2-0.3	318	108
0.3-0.4	232	178
0.4-0.5	252	136
0.5-0.6	278	175
0.6-0.7	116	195
0.7-0.8	101	138
0.8-0.9	49	117
0.9-1.0	18	262
Total	1673	1413
B. *Test set*		
0.0-0.1	2	0
0.1-0.2	16	4
0.2-0.3	15	4
0.3-0.4	18	18
0.4-0.5	39	50
0.5-0.6	15	30
0.6-0.7	17	11
0.7-0.8	5	29
0.8-0.9	4	12
0.9-1.0	0	24
Total	131	182

Each position in turn regions between strands is categorized by the confidence that the turn is a diverging turn rather than a hairpin turn, $P_d/(P_d + P_h)$ (column 1), and by its true context (columns 2 and 3).

One drawback to the model, as with the library model, is the assumption of positional independence. When multiply aligned sequences are condensed to form a profile, the information contained in pairwise statistical interactions (sequence covariance) is lost. One place where covariance is significant is in the region of the helix N-capping box (Doig *et al.*, 1997), another is in polar residues in helices (Lacroix *et al.*, 1998). Covariance can result from conserved salt bridges or from packing interactions. The problem is partially overcome by allowing multiple paths for each motif, each path containing a different covariant pair. To a certain extent this has occurred in the case of the helix N-cap, for which there are several independent paths in the HMM (Figure 4) Use of the full multiple sequence alignment rather than profiles in the training process could improve the retention of this type of covariance between nearby positions.

What is the origin of the non-random connectivities between the different I-sites motifs in HMMSTR? The presence of a motif pattern has a selective effect on the allowable sequence-structure patterns upstream and downstream. This is one manner in which structure may propagate through the chain during the very early stages of folding. Alternatively, the connectivities between motifs may arise from global structural constraints on protein structures and recurring super-secondary structure motifs, for example the β-α-β motif of many α-β-proteins.

Table 7. Middle *versus* end strand

$P_n/(P_m + P_n)$	Middle	End
A. *Training set*		
0.0-0.1	596	92
0.1-0.2	2346	533
0.2-0.3	2545	873
0.3-0.4	2103	1104
0.4-0.5	1285	988
0.5-0.6	724	735
0.6-0.7	418	572
0.7-0.8	131	333
0.8-0.9	94	156
0.9-1.0	38	25
Total	10,280	5411
B. *Test set*		
0.0-0.1	79	5
0.1-0.2	266	69
0.2-0.3	236	93
0.3-0.4	194	128
0.4-0.5	171	117
0.5-0.6	80	68
0.6-0.7	52	62
0.7-0.8	9	24
0.8-0.9	6	19
0.9-1.0	1	1
Total	1094	586

Each position in a correctly predicted strand is categorized by the confidence that the strand is an end strand rather than a middle strand, $P_n/(P_n + P_m)$ (column 1), and by its true context (columns 2 and 3).

As with the I-sites model, low complexity regions and particular types of proteins (such as membrane-associated, see Methods), were excluded. Our HMM is therefore less suitable for gene or structure prediction for such proteins. Treatment of such sequences would necessitate the addition of modules for membrane spanning and low complexity/disordered region segments to the HMM.

Applications

The local structure predictions made by the HMM may be useful for both *ab initio* structure prediction and homology modeling. In the case of *ab initio* structure prediction, discrimination between internal and external strands and between hairpin and diverging turns should facilitate identification of native-like topologies of beta sheet proteins, which tend to be particularly challenging for *ab initio* folding methods. In the case of homology modeling, the challenging loop building problem may be facilitated by the backbone angle predictions produced by the HMM, which can serve as prior distributions to guide loop closing/building procedures.

The ''inverse'' folding problem is the design of protein sequences that have a desired structure. In our model, the dual nature in which sequential and structural information is incorporated enables the prediction of sequence from structure to be performed analogously to the prediction of structure from sequence. A sequence profile consistent with a given structure can be obtained from the HMM simply by threading the desired secondary structure, backbone angle, and structural context strings simultaneously through the HMM and recording the amino acid residues emitted from the visited states. Such sequence profiles could be useful as prior distributions for sequence searching methods based on explicit side-chain modeling (Dahiyat *et al*., 1996), particularly in turns and on the surface where steric repulsion plays a less important role. Such sequence predictions could potentially be made more effective by incorporating additional structural attributes, such as the number of residues within a certain distance from the position of interest.

Sequence comparison methods are among the most widely used computational methods in biology and can provide considerable insight into both the structure and function of a novel sequence. Scoring any two positions in sequence alignments of proteins typically involves substitution matrices, such as those of the BLOSUM series (Henikoff & Henikoff, 1992). Most scoring matrices in current use are position independent. Context dependent substitution tables can be computed for each of the states in the HMM, and subsequently computed for any query sequence from the paths it takes through the HMM in a manner analogous to that described here for secondary structure prediction, etc. These context dependent substitution tables could then be used to increase the sensitivity and specificity of standard sequence searching methods, perhaps achieving some of the improved performance of methods, such as PSI-BLAST (Altschul *et al*., 1997), which use multiple sequence information, but using only a single sequence.

Methods

Identification of weak sequence-structure correlations

A procedure was described previously for identifying increasingly weak sequence-structure correlations by ''peak removal'' and re-clustering (Bystroff & Baker, 1998). In short, the I-sites motifs were used to mask locations in the database that had strong correlations with structure, and we then clustered the remaining sequence segments. Segments were clustered at each length from 3 to 19 residues, and cluster correlations were refined by supervised learning, as before. Confidence curves were calculated for the new clusters, mapping sequence score to the probability of being within approximately 1.4 Å rmsd of the cluster's paradigm (most common) three-dimensional structure. Weak motifs are classified as those sequence clusters whose highest-scoring 30 members were at least 40% but less than 70% correctly predicted. Previously, the I-sites motifs had been defined as clusters where at least 70% of the 30 highest-scoring members were correctly predicted. By masking out all sequence segments that would be more confidently predicted by one of the strong I-sites motifs, we obtained only new sequence-patterns. The entire set of I-sites clusters can be viewed at: http://isites.bio.rpi.edu/Isites/by_motif.html.

Formulation of the hidden Markov model

Here, we describe the parameters of the model and the general equations for training and prediction. We refer the reader to the tutorial by Rabiner (1989) for a full treatment of the methods (expectation-maximization, the forward-backward algorithm, and the Viterbi algorithm), which is not included here.

A hidden Markov model is a network of Markov states connected by directed transitions. Each state emits symbols representing sequence and structure. More specifically, the components of the model, collectively denoted by the symbol λ, are as follows. The HMM consists of a network of N states denoted $q_i (1 \leqslant i \leqslant N)$. The probability of initiating a sequence at state j is given by π_j. The probability of a transition from state i to state j is given by a_{ij}. For a given state i, there are a set of emission probabilities collectively called B_i. Here, we use four in this collection, denoted b, d, r, and c (see Figure 2). The values $b_i(m)$ $(1 \leqslant m \leqslant 20)$ are associated with probabilities for the emission of amino acid residues. The values $d_i(m)$ $(1 \leqslant m \leqslant 3)$, are the probabilities of emitting helix (H), strand (S) or loop (T), respectively. (In Bayesian notation: $d_i(m) = P(m = \{\mathrm{H,S,T}\}|q_i)$.) The values $r_i(m)$ $(1 \leqslant m \leqslant 11)$ are the probabilities of emitting one of the 11 dihedral angle symbols (Figure 5). Finally, $c_i(m)(1 \leqslant m \leqslant 10)$ is the probability of emitting one of ten structural context symbols (described below).

The database is encoded as a linear sequence of amino acid and structural observables. The amino acid sequence data consists of a ''parent'' amino acid sequence of known three-dimensional structure, and an amino acid profile obtained by alignments to the parent sequence (Bystroff & Baker, 1998). The amino acid of the parent sequence is denoted by O_t, and the profile by $\{O_t^m\}(1 \leqslant m \leqslant 20)$. For the structural identifiers at each position t, the following nomenclature is used: three-state secondary structure, D_t; discrete backbone angle region, R_t; and the context symbol, C_t. A sequence s of length T is given by the values of the attributes at all positions $s_t = \{O_t, \{O_t^m\}, D_t, R_t, C_t\}$ $(1 \leqslant t \leqslant T)$.

The utility of the HMM to model protein sequences is based on the notion of a path. A path is a sequence of states through the HMM, denoted $Q = q_1 q_2 \ \ldots \ q_T$. Thus, the probability of a sequence s given the model λ, $P(s|\lambda)$, is obtained by summing the relevant contributions from all possible paths Q:

$$P(s|\lambda) = \sum_{allQ} \pi_{q_1} B_{q_1}(s_1) a_{q_1 q_2} B_{q_1}(s_2) \ldots a_{q_{T-1} q_T} B(s_T) \qquad (1)$$

Here, $B_i(s_t)$ is the probability of observing s_t at state i, which for observation of a single sequence is given by:

$$B_i(s_t) = \begin{pmatrix} d_i(D_t) \\ r_i(R_t) \\ c_i(C_t) \end{pmatrix} b_i(O_t) \qquad (2)$$

Usually, only one of the structural emission symbols d, r, or c is included in B_i in any given training run. However, in principle, any combination could be used. Our HMMs showed significant improvements in performance when we used amino acid profiles instead of single amino acid sequences for training and for subsequent predictions. For the probability of observing a given profile $\{O_t^m\}$ position t in a sequence, we use the multinomial distribution, and the expression for B becomes

$$B_i(s_t) = \begin{pmatrix} d_i(D_t) \\ r_i(R_t) \\ c_i(C_t) \end{pmatrix} \sum_{m=1}^{20} (b_i(m))^{N_{count} \times O_t^m} \qquad (3)$$

To give equal weight to the information in sequence families of different depths, N_{count} was taken to be a global parameter. To each term in B_i we added a small pseudocount, ε (not shown), to prevent machine errors and to allow state paths Q to be introduced in the training process which are otherwise excluded when terms in B_i are zero.

Protein database

For training, evaluation and testing of the HMM we used a non-redundant database of proteins of known structure, PDBselect: December 1998 (Hobohm & Sander, 1994) containing 691 proteins and their sequence families. Entries in the database were selectively removed if the structure was solved by NMR, had a large number of disulfide bridges or *cis*-peptide bonds, or if it was a membrane-associated protein according to the header records. Disordered or missing coordinates in the middle of a sequence were addressed by dividing the sequence at that point. Contiguous segments of length less than 20 were ignored. Multiple sequence alignments were generated from each sequence using PSI-BLAST (Altschul *et al*., 1997) after filtering the query sequences for low-complexity regions using SEG (Wootten & Federhen, 1996). Data for training the HMM included the sequence profile, computed from the multiple sequence alignment as described (Bystroff & Baker, 1998), the DSSP secondary structure assignments (Kabsch & Sander, 1983), the backbone angles, and a structural ''context'' symbol.

All *trans* phi/psi pairs from the protein database were partitioned using k-means clustering ($k = 10$). Boundaries enclosing the cluster centroids were found using a Voronoi method, yielding ten regions in phi/psi space (Figure 5). All *cis* peptides form an 11th region.

A randomly selected set of 73 of the 691 proteins (19,000 positions) was then set aside and not used for training, but only for the final cross-validation. Before cross-validation, a test for true independence was applied to each member of the test set as described below, and 12 members were removed. The final test set thus contained 61 proteins and 16,000 positions.

The remaining set of 618 parent sequences (145,000 positions) was used for training, and divided into a large set of 564 sequences (133,000 positions), used for optimization *via* the EM algorithm, and a small set of 54 sequences (12,000 positions) used to evaluate of the predictive ability of the model during training. Note that the small set of 54 sequences is used only for evaluation of the performance of a model and may thus appear to be a test set. However, decisions regarding the modification of the model are based on results of those evaluations. The set of 54 sequences is therefore not a test set, but a training set. For the final round of training we recombined the large and small training sets, to a total of 618 sequence families. After the final round of training, the models were frozen.

The final evaluations (the only ones reported here) were done on the independent test set of 61 proteins. This set of proteins was not used in any way prior to the final evaluations (in particular, the set was not used to evaluate any earlier models).

Cross-validation

Here, we are looking for local sequence-structure correlations that transcend protein sequence families. We must be careful that the information contained in the model is not biased toward any one molecular ancestry and that the test set is completely unrelated to the training set. While the PDBselect list used in this work does have most homologous pairs of proteins already removed, for this study it was important to take further precautions.

To eliminate redundancy in the multiple sequence alignments used in the training process, we used PSI-BLAST (Altschul *et al*., 1997) to detect remote homologs. First, a file was created of pointers from each sequence in the ''nr'' database to each homologous protein on the PDBselect list. If a sequence was found in a PSI-BLAST search using one of the PDB select parent sequences, then that parent, with annotations specifying the overlap region and the percent identity, was recorded in the file. In a second pass through the PDBselect list, a ''nr'' sequence was omitted from the multiple alignment if among its pointers there was another parent with a higher percent ID in an overlapping region. This guaranteed that no sequence was used more than once in the resulting database.

To remove sequences from the test set that could be related to proteins in the training set, a list was made of sets of PDB select sequences that mutually matched any target sequence in ''nr'' in the same overapping region of that target with a BLAST *E*-value less than 0.1. Parent sequences that both matched the same target were considered *remote homologs*, even if they did not recognize each other directly. Test set members that were found to be remote homologs of training set members were removed from the test set. For example, 1alvA (calcium binding domain Vi) and 1wdcB (scallop myosin) both hit the target sequence caltractin (GI 729051) with *E*-values of less than 0.1 (and in fact the two parent proteins share a superimposable calcium binding domain). 1wdcB is a member of the training set, therefore 1alvA was removed from the test set.

Can training set bias originate from the I-sites library?

One might be concerned that care was not taken to exclude the test set proteins during the discovery of the *I*-sites motifs and subsequent initialization of the HMM topology based on these motifs. Even after training the parameters and topology of the HMM on the training set, there is a remote chance that one or more members of the test set could be favored (i.e. give artificially high scores) purely as an artifact of this initialization. The existence of a such a bias would require the ''survival'' of long, non-local, sequence-structure correlations, or paths, specific to ''test-set'' protein families (a sort of profile-HMM (Eddy, 1998) within our HMM) through the processes of initialization and training of the HMM. Below, we argue that such paths are unlikely to exist, even at HMM initialization. Subsequent training cannot enhance any such bias, since the training and test sets are truly independent. Ultimately, only *bona fide* blind predictions for new structures solved since the publication of this work can remove residual doubts concerning this potential source of bias.

Method 1 of initiating the HMM (based on co-occurrence, see below) measured the adjacencies of I-sites motifs in the training dataset. However, method 1 produced a model with no long pathways. The length of a Markov state pathway over which there is significant ''memory'' of past states can be computed as the product of N adjacent transition probabilities p. The longest sequence of states with $p > 0.1$ was $N = 12$ in λ^D and λ^C. Method 2 of initializing the HMM (based on motif alignment, see below) did not use the I-sites training set, but instead used only local profile similarity between I-sites motifs. Incorporation of non-local bias toward specific proteins used to test the I-sites library is therefore not likely for λ^R.

Initialization of HMM state profiles and connectivity

An I-sites motif may be represented as a simple HMM in which each position defines a state and each state, except the last, has a single transition to the next state. To generate the initial HMM, we treat the I-sites motifs (262 in number, including strong and weak motifs) as individual linear HMMs, then anneal them together using a measure of similarity. In this case, ''annealing'' refers to the merging of two or more single-residue positions to make one, retaining all of the connectivity. The resulting HMM has a single state for each set of merged single-residue positions, and a forward transition for each position merged, unless that position was the last one in the motif (Figure 3). Two measures of similarity were used to define which states to merge.

Initialization based on co-occurence of I-sites motifs

The first measure of state similarity was derived by scoring a large sequence database for matches to each of the I-sites patterns, and then assessing the correlations in the scores (over all positions in the database) of each pair of positions in the I-sites library. S, a similarity matrix with indices corresponding to individual positions in the 262 I-sites motifs was initialized to

$$S_{ij} = \frac{\sum_t cf(t,i) \times cf(t,j)}{\min\left(\sum_t cf(t,i), \sum_t cf(t,j)\right)} \qquad (4)$$

where $cf(t, i)$ is the confidence associated with the prediction of an I-sites motif i at position t in the database. If the predicted structure was wrong using the criteria described (Bystroff & Baker, 1998), cf was set to zero.

A greedy algorithm was employed to choose which motif positions to merge: the highest value of S_{ij} was identified, and the ''states'' i and j were merged. The matrix elements in columns i and j are reset to the value $\min(S_{ik}, S_{jk})$ for all rows k. and S_{ij} was set to zero. The process is repeated until all elements of S are equal to zero. By this process, 2169 initial states in the I-sites library were condensed to 208 states. To initialize the forward transition probabilities, each position in the database was assigned a state according to the motif at which it scored the highest. The probability of a transition from state i to state j was initialized to the frequency of ij pairs in the database divided by the frequency of state i. The total number of states in the starting model, generated by this method, was 209 (including the unknown state) and the total number of non-zero forward transitions was 7329. This initial model led to models λ^D and λ^C.

Initialization based on pairwise alignment of I-sites motifs

The second approach to initializing the HMM topology employed pairwise all-against-all dynamic programming alignment of the 262 I-sites motifs using the correlation between the amino acid frequency profiles as the similarity measure. To force aligned positions to have the same backbone conformation, a large negative value was added to the pairwise score if there was a backbone angle deviation of more than 90° between aligned positions. In addition, the average distance matrix error (DME) over the aligned positions was computed during the traceback step, and the part of the DME exceeding 0.3 Å was scaled and subtracted from the score. Each pair of the aligned positions was merged as described above, starting with the highest-scoring alignment and proceeding down the sorted list. However, a motif-motif alignment was ignored if: (i) the merging of the current pair of motifs would create a self-loop in the HMM; (ii) the merging of the current pair of motifs would create a cyclic graph, and the alignment score was below a cut-off; or (iii) any ungapped subsegment of the alignment did not contain at least three contiguous positions.

Some subsets of the I-sites motifs did not merge with other subsets under these constraints, resulting in a fragmented HMM. Disjoint graphs within the HMM were connected using a non-emitting state (see below). The three rules above improved the prediction of local and secondary structure over models made without them.

The forward transition probabilities a_{ij} were set to one if states i and j were adjacent in at least one I-sites motif, and zero otherwise. The a_{ij} matrix was then normalized by rows. This relatively crude initialization was acceptable because of the robustness of the expectation-maximization (EM) algorithm. EM optimization of the model initialized in this manner converged in about five cycles when the a_{ij} values were the only parameters allowed to vary. The total number of states in the starting model, initialized by this method, was 281, and the total number of non-zero transitions was 371. Of these transition probabilities, 254 were exactly one and not variable, leaving 117 variable transitions. This initial model led to model λ^R.

An example of how the two methods differ is in the gapped alignment of the two β-hairpins shown in Figure 3. The two hairpins never co-occur in the database, since they have different turn lengths, therefore they would not be merged by the first method. However their sequences and structures can be aligned with a one-residue gap, so the second method merges them.

The non-emitting state ("nought" state)

A single, non-emitting (*nought*) state was used to connect all ''sink'' states (states with no forward transitions) with all ''source'' states (states with no backward transitions). All transitions to and from the *nought* state were initially set equal and normalized to one, and were optimized along with the other a_{ij} values. The formalism for the use of a non-emitting state was found to be a simple extension of that for a standard HMM. By replacing a_{ij} with $(a_{ij} + a_{i0}a_{0j})$ in equation (3), where ''0'' represents the *nought* state, we have the expression for the probability of a sequence given two alternative paths between each pair of states. In practice, the second term in the parentheses is non-zero only if the first term is zero. The new equation satisfies the mathematical requirement that the sum of $P(s|\lambda)$ over all sequences s is equal to one. New update rules for each of the model parameters were derived from the new equation.

The alternative model for connecting sinks and sources would be to provide an explicit transition for each sink to each source. In the initial model there were 53 sinks and 47 sources, so this would require $53 \times 47 = 2491$ independent variable parameters. The use of a centralized, non-emitting connector requires only $53 + 47 = 100$ variable parameters, with some loss of generality.

Expectation maximization

The parameters of the model were iteratively optimized to best reflect correlated sequence-structure patterns of the database. This involved maximizing the probability of observing the training set given the model, equation (1). Expectation maximization (Lawrence & Reilly, 1990) was performed with a generalization of the algorithm described by Rabiner (1989). The expectation step involves computing $\gamma_t(i)$, the probability of being in state i at time t, and $\zeta_t(i, j)$, the probability of being in state i at time t and state j at time $t + 1$. The maximization step involves re-evaluating the parameters of λ based on $\gamma_t(i)$ and $\zeta_t(i, j)$. For any given training run, we included in B_i the b_i term and only one of the symbols d_i, r_i, or c_i. Choosing two or more of these symbols typically resulted in loss of performance. This is to be expected, since there is a high degree of interdependence among variables of these sets. Whether or not they were used in the expectation step, all variables in λ were re-evaluated (updated) in the maximization step.

Modifications of HMM topology

Here, we describe a number of auxiliary functions that were developed to allow flexible exploration of different topologies. We also briefly describe the role of these functions in creating the models λ^D, λ^C and λ^R, which were optimized for the prediction of three-state secondary structure, context, and backbone angles, respectively. Expectation maximization can modify the topology of an HMM in the sense that transitions may vanish that are not consistent with the training set, but the topology is otherwise limited to that of the initial HMM. Three auxiliary functions were created to reduce the complexity of models; one to remove weak transitions (both forward and backward), one to remove seldom-visited states, and one to merge states. We also developed three routines that add complexity to the model: one that redistributes a portion of the transition probability associated with selected strong transitions among the transitions a_{ij} which are currently frozen at zero, a function that splits states into two (nearly identical) copies, and a function to link two states by a new transition.

In general, changes in topology were followed by running EM to convergence, followed by evaluation of sequence, *Q3* and *MDA* scores on the small test set. The smaller training set was not used during EM optimization but only as a tool for accepting or rejecting modifications of the model. If the scores on the smaller training set improved we kept the changes, otherwise we scaled back or undid the modification.

The common ancestor of models λ^D and λ^C was a model based on motif co-occurrences (see above). The

initial graph was highly complex, with 209 states and 7329 transitions, whereas the final model λ^D has 107 states and 263 variable transitions. This model underwent an initial smearing, followed by alternate cycles of removing weak transitions and states. The resulting model underwent final rounds of training using emission vectors d and c in equation (3), separately; creating models λ^D and λ^C, respectively.

Model λ^R originated from a model initialized based on pairwise motif alignments (see above). This initial graph was less complex, having 281 states and only 117 variable transitions. In the final model λ^R there are 247 states and 149 variable transitions. The model underwent automatic removal of weak transitions and states. Subsequently, new transitions were added to mend the fragmented nature of the initial model, and states with high priors (frequently visited states) or high connectivity were targeted for splitting. Splitting a state usually resulted in the two copies taking on different properties and connections. If the two copies remained nearly identical to each other, we undid the change. Occasionally one of the two states would subsequently disappear during EM. Unconnected state pairs ij that exhibited high values of:

$$\xi_{ij} = \sum_t \alpha(i,t)\beta(j,t+1) \tag{5}$$

were targeted for new transitions (splicing). Here, α and β are the forward and backward variables (Rabiner, 1989). New transitions were initialized to very small values, which either went back to zero or increased during subsequent cycles of EM. In all, 80 new transitions were added and 48 transitions were removed.

A trend we observed for all three models was that the probability distributions for structural attributes often sharpened during training, to the point that most states can be identified with a single structural attribute. If the probability of any attribute went to zero during training, it reduced the number of variable parameters in our model by one. In this way, EM served as an unbiased pruning device.

Identification of coding regions

Given a set S of K sequences of variable lengths T_k, the sequence score L is defined as the base 2 logarithm of the probability of all sequences in S given the model λ, relative to the probability given the background distribution:

$$L(S;\lambda-\lambda^{bg}) = \frac{1}{T}\sum_{k=1}^{K}(\log_2 P(s^k|\lambda) - \log_2 P(s^k|\lambda^{bg})) \tag{6}$$

Here, T is the total number of positions in the set, obtained by summing all lengths T_k. The score is thus represented in bits per position, which is convenient for assessing the potential of our model to discriminate actual protein sequences from non-coding sequences through the use of information theory. In this sense, $L(S;\lambda-\lambda^{bg})$ is the information gain, or entropy loss; in using λ in comparison to the simple sequence model λ^{bg}. In scoring models on sets of amino acid sequences or profiles, no structural attributes were used to evaluate $P(s^k|\lambda)$.

Our HMM models showed significant improvements in performance when we used profiles for training and scoring in the place of single sequences. However, for the purpose of comparing our models to simpler models of protein sequence, more direct relations can be made considering only parent sequences of the training set. It is also worth bearing in mind that the parameters $b_i(m)$ in our model are parameters of a multinomial distribution, whereas in this test they are used as ''emission'' probabilities for single amino acids. Two randomized sets of amino acid sequences were constructed for assessing the ability of our models to discriminate protein sequence from non-coding sequence. The first set, S^{bg}, contains sequences with amino acid frequencies given by the training set, in random order. The second set, S^{dip}, contains sequences having the same dipeptide frequencies as the training set, but otherwise unconstrained. The random sequences consist of 1000 ''proteins'' of length 900. Results are insensitive to the length and number of sequences as long as sufficiently many positions are generated to accurately reflect the relevant statistical distributions. For single sequences the reference score $P(s^k|\lambda^{bg})$ is particularly simple, being the product of $b^{bg}(s_t)$ for all positions t in the sequence s^k.

Structure prediction by voting

The predictions of three-state (helix/strand/turn), context, and backbone angles, given a sequence $\{O_t^m\}$, were made using a voting procedure. For three-state secondary structure prediction, the predicted state, H, S or T, is given by the largest of the sum of the prior-weighted emission symbols over all states:

$$\begin{aligned} D_t &= \underset{i=\{H,S,T\}}{\operatorname{argmax}}\ P_t(i) \\ &= \underset{i=\{H,S,T\}}{\operatorname{argmax}}\left[\sum_{n=1}^{N} P(d_i|q_n)P(q_n|s_t)\right] \\ &= \underset{i=\{H,S,T\}}{\operatorname{argmax}}\left[\sum_{n=1}^{N} d_n(i)\gamma_t(n)\right] \end{aligned} \tag{7}$$

A two-tiered voting scheme was found to be better when choosing backbone angles. We first vote for a broadly defined region (regions *BEbde*, *GH*, *Ll* and *ex* in Figure 5), and then choose a specific region within it. Similar to the case of parliamentary elections, underpopulated regions are better represented in the elected distribution when their votes are consolidated. In our case, the four-β-sheet backbone angle regions were first consolidated. If the consolidated region ''won'', then ''run-off'' elections were held between the sub-regions. The most highly populated backbone angle region is the one corresponding to α-helix; these angles also occur in about one-fourth of all loop positions. Using hierarchical voting decreased the overprediction of these angles.

The prediction of context was based on the secondary structure prediction. Sequence segments predicted to be a strand were assigned a probability of being either an end strand or a middle strand based on the position-specific probabilities $P_t(n)$ and $P_t(m)$, respectively, obtained by the voting scheme described above. Similarly, $P_t(d)$ and $P_t(h)$ were computed for all turns flanked

by strands to assess the likelihood of finding diverging *versus* hairpin turns.

The summand in equation (1), $P(s, Q|\lambda)$, represents the joint probability of the observation sequence s and the state sequence Q given the model λ. The prediction of structural attributes is also possible by determining the state sequence that maximizes $P(s, Q|\lambda)$ for a given sequence of amino acids or profile, known as the Viterbi path (see Rabiner, 1989). Structural prediction based on voting consistently outperformed predictions based on the Viterbi path in our models.

Acknowledgments

We wish to thank Phil Green, Richard M. Karp, Anders Krogh, Chip Lawrence, Ingo Ruczinski, and Ed Thayer for helpful discussions, and a referee for pointing out an error in an earlier version of this manuscript. This work was supported by a University of Washington Training Grant in Interdisciplinary Genome Sciences (V.T.), by a Sloan Foundation/Department of Energy Fellowship in Computational Molecular Biology(V.T.), by the NSF (STC cooperative agreement BIR-9214821, D.B.), and by a Packard Fellowship in Science and Engineering (D.B.) and a grant from the Howard Hughes Medical Institute to the Rensselaer Bioinformatics Program (C.B.).

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids. Res.* **25**, 3389-3402.

Asai, K., Hayamizu, S. & Handa, K. (1993). Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.* **9**, 141-146.

Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci.* **7**, 21-38.

Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (Altman, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D., eds), pp. 28-36, AAAI Press, Menlo Park, USA.

Burge, C. B. & Karlin, S. (1998). Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346-354.

Bystroff, C. & Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281**, 565-577.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol.* **47**, 45-148.

Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895-903.

Di Francesco, V., McQueen, P., Garnier, J. & Munson, P. J. (1997). Incorporating global information into secondary structure prediction with hidden Markov models of protein folds. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 100-103, AAAI Press, Menlo Park, USA.

Doig, A. J. & Baldwin, R. L. (1995). N and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci.* **4**, 1325-1336.

Doig, A. J., MacArthur, M. W., Stapley, B. J. & Thornton, J. M. (1997). Structures of N-termini of helices in proteins. *Protein Sci.* **6**, 147-155.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361-365.

Eddy, S. R. (1998). Profile Hidden Markov models. *Bioinformatics,* **14**, 755-763.

Efimov, A. V. (1993). Standard structures in proteins. *Prog. Biophys. Mol. Biol.* **60**, 201-239.

Fickett, J. W. & Tung, C. S. (1992). Assessment of protein coding measures. *Nucl. Acids Res.* **20**, 6441-6450.

Han, K. F. & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl Acad. Sci. USA,* **93**, 5814-5818.

Henderson, J., Salzberg, S. & Fasman, K. H. (1997). Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.* **4**, 127-141.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Karplus, K., Barrett, C. & Hughey, R. (19??). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846-856.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.

Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology* (States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. & Smith, R. F., eds), pp. 134-142, AAAI Press, Menlo Park, USA.

Lacroix, E., Viguera, A. R. & Serrano, L. (1998). Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **284**, 173-191.

Lawrence, C. E. & Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct. Funct. Genet.* **7**, 41-51.

Lio, P., Goldman, N., Thorne, J. L. & Jones, D. T. (1998). PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics,* **14**, 726-733.

Lukashin, A. V. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucl. Acids Res.* **26**, 1107-1115.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE,* **77**, 257-286.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55-72.

Rost, B. (1997). Better 1D predictions by experts with machines. *Proteins: Struct. Funct. Genet.* **Suppl. 1**, 192-197.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1999). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Proteins: Struct. Funct. Genet.* **34**, 82-95.

Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405-420.

Sonnhammer, E. L. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds), pp. 175-182, AAAI Press, Menlo Park, USA.

Stultz, C. M., White, J. V. & Smith, T. F. (1993). Structural analysis based on state-space modeling. *Protein Sci.* **2**, 305-314.

Wootton, J. C. & , Federhen S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.

Appendix

There are several relations among the results in Table 3 that are noteworthy. For the sequence scores evaluated with λ^{dip} we find that: (i) the score on the random sequences S^{dip} is equal to that of the database sequences S^{db}; and (ii) the score on the random sequences S^{bg} is the negative of the score on S^{db}, up to the precision with which these values were computed. Here, we show how relations (i) and (ii) may be confirmed independently by relatively simple calculations.

We first show that the sequence score of λ^{dip} on the database S^{db} is equal to the mutual information for the probability distribution of adjacent amino acids relative to the background amino acid frequencies. Let N_i be the number of positions in the database S^{db} with amino acid i, and N_{ij} be the number of positions where amino acid j is preceded by amino acid i. The amino acid frequency in the database is $p_i = N_i/T$, and the dipeptide frequency is $p_{ij} = N_{ij}/(T-K)$, where K is the number of sequences. The mutual information $M(p_{ij})$, which reflects the information in the joint distribution p_{ij} in comparison to that of the marginal distribution p_i, is defined as:

$$M(p_{ij}) = \sum_{ij} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j} \qquad \text{(A1)}$$

The dipeptide model λ^{dip} is a Markov chain with 20 states, one corresponding to each amino acid (a state in a Markov chain, unlike that in a Hidden Markov Model, can ''emit'' only one symbol). The transition probabilities are: $a_{ij} = p_{ij}/p_i$. The probability of a sequence for λ^{dip} is given by the product of the a_{ij} for adjacent pairs of amino acids. The background model reference term in equation (4) is the same for all sequence sets considered here and is given by: $\Sigma_i\ (p_i \log_2 p_i)$. To evaluate $L(S^{db}; \lambda^{dip} - \lambda^{bg})$, we collect adjacent ij positions in the database to obtain

$$\frac{1}{T}\sum_{k=1}^{K} \log_2 P(s^k \in S^{db} | \lambda^{dip}) = \sum_{ij} \log_2 a_{ij} \qquad \text{(A2)}$$

Subtracting the reference term, one finds $L(S^{db}; \lambda^{dip} - \lambda^{bg}) = M(p_{ij})$. We are now ready to address statements (i) and (ii) above.

Statement (i) is straightforward, as the dipeptide model is sensitive only to adjacent residues, which obey identical distributions in the sequence sets S^{dip} and S^{db}. Therefore, $L(S^{dip}; \lambda^{dip} - \lambda^{bg}) = L(S^{db}; \lambda^{dip} - \lambda^{bg}) = M(p_{ij})$.

(ii) For sequences in S^{bg}, the joint probability on the right hand side of equation (A2) is replaced by the product of the marginal probabilities $p_i p_j$, resulting in:

$$L(S^{bg}; \lambda^{dip} - \lambda^{bg}) = \sum_{ij} p_i p_j \log_2 \frac{p_{ij}}{p_i p_j} \qquad \text{(A3)}$$

Surprisingly, this differs from the negative of equation (A1) by only 0.4 %, for the p_{ij} of S^{db}, which is consistent with the results of Table 3. To verify this, we define the small deviations ε_{ij} by $p_{ij} = p_i p_j(1+\varepsilon_{ij})$ and perform expansions in ε_{ij}. Expanding the logarithm gives $\log(1+\varepsilon_{ij}) = \varepsilon_{ij} - \varepsilon_{ij}^2/2 + O(\varepsilon_{ij}^3)$, where $O(\varepsilon_{ij}^3)$ denotes terms of order ε_{ij}^3. The sum of equation (A1) and (A3) is (we omit the constant log 2):

$$\begin{aligned}\sum_{ij}(p_{ij}+p_i p_j)\log\frac{p_{ij}}{p_i p_j} &= \sum_{ij} p_i p_j(2+\varepsilon_{ij}) \\ &\quad \times \left(\varepsilon_{ij} - \frac{1}{2}\varepsilon_{ij}^2 + O(\varepsilon_{ij}^3)\right) \\ &= \sum_{ij} p_i p_j O(\varepsilon_{ij}^3) \end{aligned} \qquad \text{(A4)}$$

The leading, $O(\varepsilon_{ij})$, term vanishes due to the constraint $\Sigma_j(\varepsilon_{ij} p_j) = 0$, for each i, which follows from $p_i = \Sigma_j p_{ij}$. Exact cancellation takes place for the subleading, $O(\varepsilon_{ij}^2)$, terms. Repeating this expansion for $M(p_{ij})$ itself, there is no such cancellation and $M(p_{ij})$ is of order $(\varepsilon_{ij})^2$. This accounts for relation (ii).

Edited by J. Thornton

(Received 27 March 2000; accepted 2 May 2000)