

Hidden Markov Models That Use Predicted Secondary Structures For Fold Recognition

Jeanette Hargbo and Arne Elofsson*

Department of Biochemistry, Stockholm University, Stockholm, Sweden

ABSTRACT There are many proteins that share the same fold but have no clear sequence similarity. To predict the structure of these proteins, so called “protein fold recognition methods” have been developed. During the last few years, improvements of protein fold recognition methods have been achieved through the use of predicted secondary structures (Rice and Eisenberg, *J Mol Biol* 1997;267:1026–1038), as well as by using multiple sequence alignments in the form of hidden Markov models (HMM) (Karplus et al., *Proteins Suppl* 1997;1:134–139). To test the performance of different fold recognition methods, we have developed a rigorous benchmark where representatives for all proteins of known structure are matched against each other. Using this benchmark, we have compared the performance of automatically-created hidden Markov models with standard-sequence-search methods. Further, we combine the use of predicted secondary structures and multiple sequence alignments into a combined method that performs better than methods that do not use this combination of information. Using only single sequences, the correct fold of a protein was detected for 10% of the test cases in our benchmark. Including multiple sequence information increased this number to 16%, and when predicted secondary structure information was included as well, the fold was correctly identified in 20% of the cases. Moreover, if the correct secondary structure was used, 27% of the proteins could be correctly matched to a fold. For comparison, blast2, fasta, and ssearch identifies the fold correctly in 13–17% of the cases. Thus, standard pairwise sequence search methods perform almost as well as hidden Markov models in our benchmark. This is probably because the automatically-created multiple sequence alignments used in this study do not contain enough diversity and because the current generation of hidden Markov models do not perform very well when built from a few sequences. *Proteins* 1999;36:68–76. © 1999 Wiley-Liss, Inc.

Key words: protein structure; HMM; Scop; HSSP; threading; blast; fasta; ssearch; protein fold recognition

INTRODUCTION

The most promising method for predicting the structure of a protein is to identify a protein with a known structure that shares the same fold. Traditionally, this has been done

by identifying proteins that have similar sequences. However, of late, many examples of structures that have similar folds but no detectable sequence similarity have been found. This has led to the development of methods to detect the fold of a probe sequence from a library of known target folds. These methods are often referred to as fold recognition methods.

Fold recognition methods can roughly be divided into three different types, based on the type of information that they use. Within each category there are many different implementations. The three types of methods are sequence-based methods,^{1,2} structure-based methods,^{3,4} and prediction-based methods.^{5–9,10} In this study, we introduce a new method that combines multiple-sequence-alignment methods with predicted secondary structure information. We also compare the performance of hidden Markov models with standard sequence-based methods. All these comparisons are made with a more rigorous benchmark than those used in most earlier studies.

Sequence-based methods are the oldest methods for fold recognition.¹¹ It seems a bit surprising that sequence-based methods are able to detect a similar fold of proteins that show no sequence similarity, but the amino-acid sequence contains much information about the physical environment at each position in the sequence. Thus, even if there is no detectable sequence similarity between two proteins that have the same fold, the corresponding positions in the proteins will have similar properties. Moreover, there are many examples where there is no obvious sequence similarity, but where two proteins clearly are homologous. Of course, these targets might be detected with improved sequence-based methods. One way to increase the performance of sequence-based methods is to use information from a family of sequences, instead of from just one sequence. With the inclusion of multiple sequence alignment information and modern computational methods, such as hidden Markov models, sequence-based methods have proven to be successful in fold recognition.²

Abbreviations: Scop, a structural classification of proteins database; HMM, Hidden Markov Model; ssHMM, hidden Markov models that use secondary structure information; predHMM, hidden Markov models that use secondary structure information with predicted secondary structures.

*Correspondence to: Arne Elofsson, Department of Biochemistry, Stockholm University, 106 91 Stockholm, Sweden. E-mail: arne@biokemi.su.se

Received 16 September 1998; Accepted 1 March 1999

A hidden Markov model (HMM), or more correctly a profile-HMM, is a generalized version of a profile that is mathematically more consistent. A general description of HMMS (applied in speech recognition, where they were originally used) has been written by Rabiner and Juang.¹² In biology, HMMS have been used in many different areas, such as gene prediction,¹³ membrane protein prediction,¹⁴ and protein sequence comparisons.^{1,2} One major difference between profile-HMMs and a profile is that in a profile the penalty for gaps or insertions are the same in every position of the alignment, even though some regions are more variable than others. Ideally, these regions should have a smaller penalty for gaps than more conserved areas. In the HMM, the penalties are position-dependent, and are learned from the training data.

An alternative type of information has been used in the structure-based fold recognition methods. These methods do not use sequence information to determine if two proteins have the same fold or not. Instead, they use an energy function that describes how well a probe sequence matches a target fold. The energy function is often obtained from a database of known protein structures, and can be used, for instance, to describe the environment of each residue¹⁵ or the probability of finding two residues at a certain distance from each other.^{3,4}

Proteins having a similar fold also have similar secondary structures, so that even though the amino acid sequences may have changed a great deal during evolution, the secondary structure will still be the same for related proteins belonging to the same fold. Today, the secondary structure can be predicted from the amino acid sequence with an accuracy of more than 70%.¹⁶ Several approaches attempt to use this information, in addition to the amino acid sequence, to recognize the correct fold.^{5,6,9} Fischer and Eisenberg⁵ align a probe sequence to known folds and then calculate the probability of the protein having a certain fold. The score for an aligned amino acid normally depends on how likely it is to have that particular amino acid in that position in the fold, but Fischer and Eisenberg also take the predicted secondary structure into account, increasing the score if it fits the secondary structure of the fold and decreasing the score otherwise. The addition of the secondary structure information seems to help significantly in recognizing the correct fold, indicating that, even though the predicted secondary structure is not completely correct, it still contains a lot of useful information that could complement other information.

Usually, a HMM only uses the amino acid sequence when modeling a protein family, making very distant homologues difficult to recognize. The aim of this work is to create a HMM that uses the predicted secondary structure in addition to the primary sequence. By combining the information from both sequence and secondary structure, it should be possible to recognize even distant or non-homologous proteins that share a similar fold. The idea of using secondary structure predictions and multiple sequence information HMMS has been proposed earlier but not tested in this type of benchmark.^{8,17} In addition, our implementation of this approach differs from earlier attempts.

MATERIALS AND METHODS

An Implementation of HMMS Using Secondary Structure Information

The program package HMMER, version 1.8.4,¹⁸ was modified to include secondary structure information when building a hidden Markov model (HMM) of a protein family, as well as when matching an amino acid sequence to an HMM. The secondary structure HMMS (ssHMMS) are models of protein families based both on amino acid sequence and on secondary structure information.

Ordinary profile HMMS consist of a sequence of match states, analogous to positions in a multiple sequence alignment, and corresponding insert and delete states. To each insert and match state a probability distribution over all amino acids is associated, these distributions giving the probability of a certain amino acid, given that particular state. The parameters of the model are the probabilities for transitions between states and the amino acid probability distributions, and these are optimized so that all sequences belonging to the modeled family obtain high probabilities and all other sequences low. Thus a sequence $s = x_1 \dots x_L$ following the path $q = q_0 \dots q_{N+1}$ through model μ has the probability

$$P(s|q, \mu) = \prod_{i=1}^{N+1} T(q_i|q_{i-1}) \prod_{i=1}^N P(x_{l(i)}|q_i) \quad (1)$$

where $T(q_i|q_{i-1})$ is the probability for a transition from state q_{i-1} to q_i and $l(i)$ is the index for amino acid x in the sequence in state q_i , $P(x_{l(i)}|q_i)$ is the probability of having amino acid $x_{l(i)}$ in state q_i , and N is the number of states in the path. The lower indexes represent the position in the path. The theory behind HMMS has been described in more detail in earlier work.^{1,18,19} In comparison with sequence profiles, one of the major differences is that for each position there is a correct transition probability for each gap and insertion parameter.

The ssHMM has an extra distribution of probabilities for the secondary structures E, H, and L associated with each insert and match state. In each state, the model emits a probability for the amino acid, as before, but in addition to this it emits another probability for the secondary structure assigned to that position. In this way, the probability for the sequence is higher if the secondary structure is the same as in the modeled family. The total probability for a sequence $s = x_1 \dots x_L$ having the secondary structure $ss = y_1 \dots y_L$ given the path $q = q_0 \dots q_{N+1}$ and model μ is now:

$$P(s,ss|q, \mu) = \prod_{i=1}^{N+1} T(q_i|q_{i-1}) \prod_{i=1}^N P(x_{l(i)}|q_i) \prod_{i=1}^N P(y_{m(i)}|q_i) \quad (2)$$

where $y_{m(i)}$ is the secondary structure emitted in state q_i . The emission probabilities of the secondary structures are found in the same way as the amino acid emission probabilities when training the model. The combined HMM will be referred to as a secondary structure HMM (ssHMM). The

modified HMMER program is available from <http://www.biokemi.su.se/~arne/sshmm/>

As the number of parameters in the model increases, additional information is needed to produce a useful model. To decrease the number of free parameters, the emission probabilities $P(x|i_k)$ for the insert states are set equal or to some background frequency. The problem with having too little information, i.e., too few training sequences, concerns fitting, i.e., a HMM created from this data will be able to recognize only proteins that are very closely related to the proteins used to create the HMM. In this situation, a prior distribution can be used, and the model is not allowed to specialize too much. However, a prior distribution assumes that any change from one amino acid to another is equally probable, which is not the case.¹⁹ A standard HMM could be seen as building a sequence profile using an identity matrix, which certainly is not the most efficient matrix to use. The inclusion of substitution parameters into HMMER can be made through the use of a special prior distribution using a substitution matrix. The inclusion of substitution matrices are made when building the HMM by adding a partial count to all amino acid types when a certain amino acid is found in a position. This partial count is related to the probability of an amino acid having been replaced by another particular amino acid. In this study, we have used the Pam250 substitution matrix, which was included in the HMMER package. For the secondary structure counts, we were not able to create a prior distribution that significantly improved the performance. Therefore we chose not to use any. At the beginning of the training, all secondary structures are assumed to occur at equal probabilities. Thus, even if a position is found in only one secondary structure type, the other secondary structure types will also have a small probability of occurrence.

A library of ssHMMs was built from the sequences and secondary structures of a representative set of all proteins with a known structure. For a given protein, all related proteins in Swissprot were found through the HSSP database,²⁰ and the secondary structure was assumed to be the same for all proteins in a family. The multiple sequence alignment from HSSP, together with the secondary structure, was used to build a ssHMM, as described above. For comparison with the original HMM method, HMMs not using the secondary structure were also created, as were HMMs (and ssHMMs) using substitution matrices. These last will be referred to as HMM-pam and ssHMM-pam. Finally, another set of HMMs, ignoring multiple sequence alignments, were created. These will be referred to as HMM-single, ssHMM-single, etc. For a complete description of all HMMs built see Table I.

To match a protein against a library of HMMs, a query sequence is matched against all HMMs. We examined the four different alignment algorithms included in HMMER local, global, endsfree, and fragmentary matches. However, in all cases, the hmms program that uses a global alignment algorithm performed best, and only results using this algorithm were evaluated in this study. When a

TABLE I. Description of Information Used in Methods Studied[†]

Name	SS in HMM	Query True SS	Query Pred SS	Substitution matrix	MSA
HMM					X
predHMM	X		X		X
ssHMM	X	X			X
HMM-single					
predHMM-single	X		X		
ssHMM-single	X	X			
HMM-pam				X	X
predHMM-pam	X		X	X	X
ssHMM-pam	X	X		X	X
HMM-pam-single				X	
predHMM-pam-single	X		X	X	
ssHMM-pam-single	X	X		X	
blast2				X	
fasta				X	
ssearch				X	

[†]SS in HMM, secondary structure in the HMM; Query True SS, correct secondary structure in query sequence; Query Pred SS, predicted secondary structure in query; MSA, multiple sequence alignment.

protein is matched against a ssHMM it is necessary to assume the secondary structure of the protein; this was done in two different ways. First, the correct secondary structure was used. Second, the secondary structure predicted by predator²¹ was used. The tests using the predicted secondary structures are referred to as predHMM etc. (see Table I). The rather mediocre performance of 68% was probably due to the fact that 45% of the sequences in our database had 10 or fewer homologous sequences in HSSP. For comparison with the standard sequence search methods we have used blast2,²² fasta,²³ and ssearch²³ on our benchmark. These methods were used with default parameters, and the scoring has been done by using the expectation-values.

Measuring the Performance

To compare the performance of different fold recognition methods, it is of great importance to use a large and well-crafted benchmark. Several recent studies^{6,24,25} have shown that a useful benchmark can be created using Scop²⁶ as a standard for classifying proteins into families of similar fold or of evolutionary relationship. Scop is a database in which all known protein structures are classified into a hierarchical classification: class, fold, superfamily, and family. In this study we have focused on proteins that have the same fold but belong to different families, according to Scop. Two proteins that are classified into the same fold have the same secondary structure elements in a similar topological arrangement, while two proteins that belong to the same family have a clear common evolutionary origin. Two proteins classified into the same fold but to different families might belong to the same superfamily or they might not.

We created a benchmark from the pdb40 dataset of Scop version 1.37. This dataset contains a subset of Scop where

no proteins have more than 40% sequence identity to any other member of the dataset.²⁵ However, this dataset did not completely match the latest release of HSSP in that (1) HSSP was created from another subset of pdb and (2) the proteins in Scop are divided into domains, whereas proteins in HSSP are not. To overcome this problem, we matched each sequence in pdb40 to the HSSP database and replaced the sequence with the HSSP sequence if the match had a significance better than $1.e-5$ using fasta, and if the alignment produced was of the same length as the original sequence. Using this procedure, 1,130 out of 1,272 sequences in pdb40 were retained. This procedure removed all Scop entries of the “non-proteins” class and many of the peptides, as they were not present in HSSP. For each of the 1,130 sequences, the multiple sequence alignment and the secondary structure were read from HSSP. On average, 26 sequences were included in a sequence family. However, many of these sequences were identical or almost identical to the original sequence. This dataset of sequences and multiple sequence alignments is available from <http://www.biokemi.su.se/~arne/sshmm/>

In our benchmark, see Figure 1, all proteins were matched to the HMMs of all other proteins, and for each pair the folds and families (according to Scop) were recorded. As the family classification in Scop is a subclassification of a fold, two proteins can belong either to the same family, to two different families but to the same fold, or to two different folds. If the two proteins belong to the same family, we have eliminated them from further consideration, because this indicates that they are homologous and thereby not a good test of fold recognition methods. If the fold, but not the family, of the two entries is the same,

TABLE II. Description of the Benchmark

Data	Number of data points
Protein domains in pdb40	1,272
Protein domains both in HSSP and in pdb40	1,130
Protein domains with at least one true match (another domain from the same fold but from another family)	730
Number of pairwise comparisons	1,273,618
True matches (protein domains from the same fold but different families)	8,312
False matches (protein domains from different folds and families)	1,265,306
Number of different protein families	666
Number of different protein folds	359

the match was considered to be a *true* match, while if the two entries belong to different folds they were considered to be a *false* match. To create a good benchmark it is necessary to have a large and complete set of proteins; in our benchmark set there are 730 proteins that have at least one true match, i.e., there are 400 proteins in the database that do not have any true match. These 400 entries were retained, because they provided potentially important information about false matches. The total number of true hits is 8,312, and there are more than 1.2 million false hits in the benchmark (see Table II). The benchmark includes proteins from 359 different folds and 666 different families in Scop. We believe that this benchmark contains a significant fraction of all possible targets for fold-recognition.

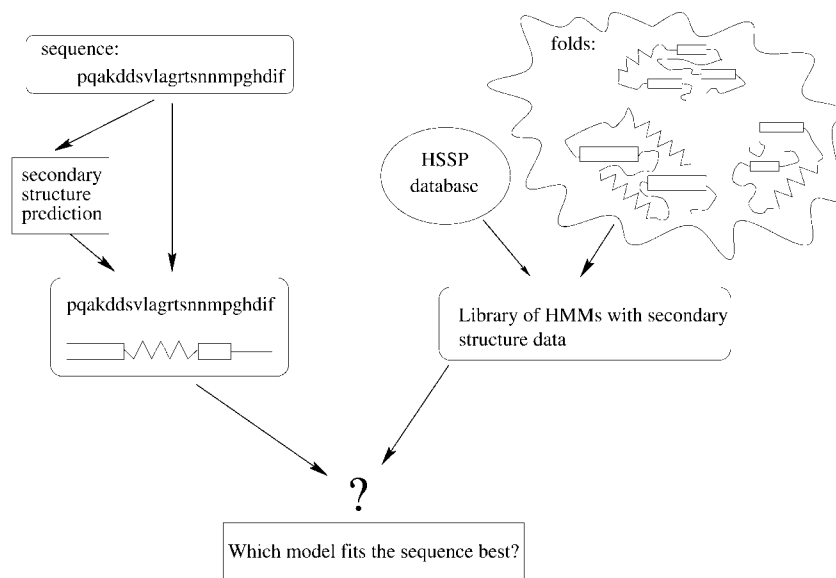


Fig. 1. A schematic description of the ssHMMs. First a library of representative folds is created, second, all homologous sequences of these proteins are found. These multiple sequence alignments, together with the secondary structures of the representative proteins, are used to

construct the library. For the probe sequence, a secondary structure prediction is performed. Finally, the sequence with the predicted secondary structure is probed against all folds in the fold library.

We have used two different criteria to analyze the performance of a fold recognition method on our benchmark. First, we simply examined at what rank the first true hit was found. This is a very intuitive measure, however, and it does not measure the reliability of a match of a certain score. For some proteins there are several possible correct hits and with this measure the first match could be to any one of these proteins, while for others there is only a single match. Second, as a complementary measure, we have used specificity-sensitivity plots, or spec-sens plots, as in Rice and Eisenberg.⁶ The main advantage of this method is that it describes the ability of a method to find all pairwise matches in the benchmark. The sensitivity is based on the model's ability to find all members of the same fold. In other words:

$$\text{SENS}(score) = \text{TP}(score) / (\text{TP}(score) + \text{TN}(score)) \quad (3)$$

where $\text{TP}(score)$ is the number of true hits that have a score above $score$, and $\text{TN}(score)$ is the number of true hits with a score less than $score$. The specificity measures the probability that a pair of sequences with a score greater than a certain threshold really belong to the same fold. The specificity is defined as:

$$\text{SPEC}(score) = \text{TP}(score) / (\text{TP}(score) + \text{FP}(score)) \quad (4)$$

where $\text{FP}(score)$ is the number of false hits that have a score above $score$ and TP is defined as above. The sensitivity is plotted as a function of specificity, each point in the plot corresponding to a certain score. One difference between our two measures is that the spec-sens curves represent a method's ability to recognize all proteins from the same fold (but from different families), while the simple counting method measures the ability of a method to identify any member of the same fold (but from another family).

RESULTS AND DISCUSSION

Every two years there is a community-wide effort, CASP, to analyze protein structure prediction methods by blind predictions, allowing predictors to "guess" the structure of soon-to-be solved protein structures.²⁷ At the second CASP process in 1996, five groups were selected for the best performance in the threading category. One of these groups used predicted secondary structures,⁷ another group used hidden Markov models (HMM),² a third group used a hidden Markov model that only used secondary structure and matched a predicted secondary structure against this model.⁸ The last two groups^{4,28} used either human expert knowledge or a physical energy function in their threading studies. The success of using HMMs and the idea of using predicted secondary structures makes it a natural step to try to combine these two methods, as we have done in this study.

This study is based on matching all proteins in our test set against all other proteins of the test set. Each protein is classified as belonging to a protein family and as having a

certain fold, according to Scop.²⁶ The Scop classification is hierarchical, i.e., a fold is a superset of one or several families, and thus two proteins might belong to the same fold but to different families. Two proteins from the same fold, but from different families, are not assumed to be homologous but still have a similar structure. A match between two proteins is ignored if the two proteins belong to the same family, it is considered as a true match if the proteins belong to different families but to the same fold, and it is considered to be a false match if the proteins belong to different folds. Using this benchmark, we have compared the performance of the newly developed ssHMMs, standard HMMs, and pairwise sequence comparisons methods.

Secondary Structure Increases the Performance of HMMs

Earlier studies showed that including predicted secondary structure sequence into single sequence-based search methods increased the performance significantly.^{5,6,9} Therefore, we believed that the same would be true for hidden Markov models. In Figure 2 it can be seen that our assumption are apparently correct, as the sensitivity of a hidden Markov model is increased when the secondary structure is included. For instance, at a specificity of 5%, the sensitivity increases from 2% to 30% if the true

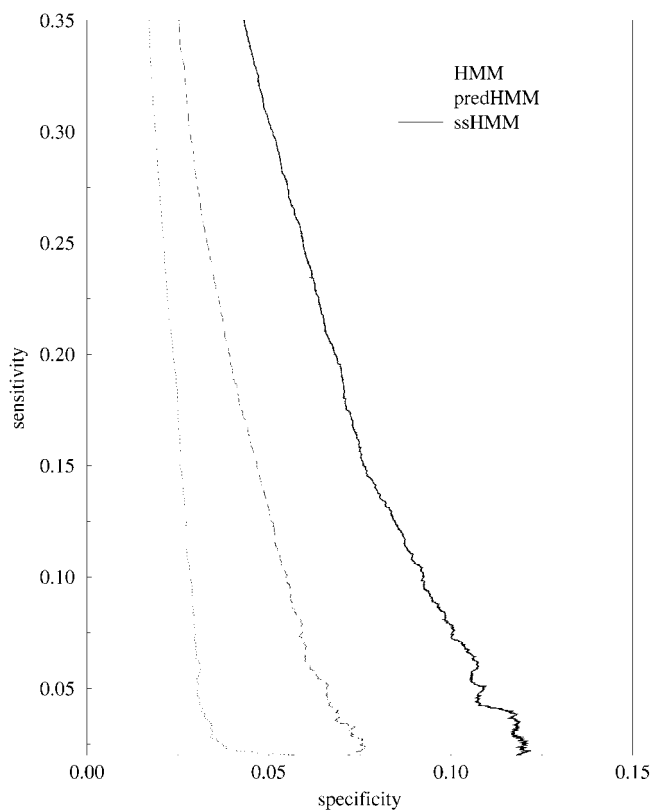


Fig. 2. A specificity versus sensitivity plot of HMM, predHMM, and ssHMM. It can be seen that the sensitivity increases when predicted or true secondary structure information is included.

TABLE III. Sensitivity of Methods at Specificity = 5% and 10%

Name	Spec = 5%	Spec = 10%
HMM	2%	1%
predHMM	13%	1%
ssHMM	30%	8%
HMM-single	0%	0%
predHMM-single	6%	0%
ssHMM-single	17%	0%
HMM-pam	11%	6%
predHMM-pam	11%	2%
ssHMM-pam	26%	7%
HMM-pam-single	8%	2%
predHMM-pam-single	17%	5%
ssHMM-pam-single	24%	11%
Blast2	3%	2%
Fasta	5%	3%
ssearch	13%	6%

TABLE IV. Fraction of Possible True Hits Placed at Ranks 1, 5, 10, and 25

Name	#1	#5	#10	#25
HMM	12%	24%	32%	45%
predHMM	19%	38%	47%	59%
ssHMM	30%	49%	59%	69%
HMM-single	4%	15%	24%	38%
predHMM-single	10%	29%	38%	51%
ssHMM-single	14%	34%	44%	56%
HMM-pam	16%	30%	40%	51%
predHMM-pam	20%	36%	45%	57%
ssHMM-pam	27%	48%	56%	67%
HMM-pam-single	10%	22%	31%	44%
predHMM-pam-single	17%	35%	44%	55%
ssHMM-pam-single	21%	39%	48%	60%
Blast2	17%	30%	37%	48%
Fasta	13%	25%	37%	43%
ssearch	17%	25%	30%	40%

secondary structure is used and to 13% if the predicted secondary structure is used (Table III). The fraction of the possible hits that were ranked in first place is increased as well, from 12% to 30% when using the secondary structure, and to 19% if the predicted secondary structure is used (Table IV). The increase in performance is similar to that reported for single sequence-based methods; for instance, Fischer and Eisenberg increased the fraction of hits found in first rank from 54% to 65% by using predicted secondary structures and the BLOSUM62 matrix.²⁹ In the study by Rice and Eisenberg, the sensitivity increased from approximately 15% to 30% when predicted secondary structures were used at 5% specificity.

It should also be noted that our benchmark seems significantly more difficult than the benchmark used by Fisher and Eisenberg, as they were able to detect 54% of the proteins in first place using sequence alignment methods, while we were able to detect only 17%. The difficulty of the benchmark used by Rice and Eisenberg seems to be similar to the difficulty of ours.

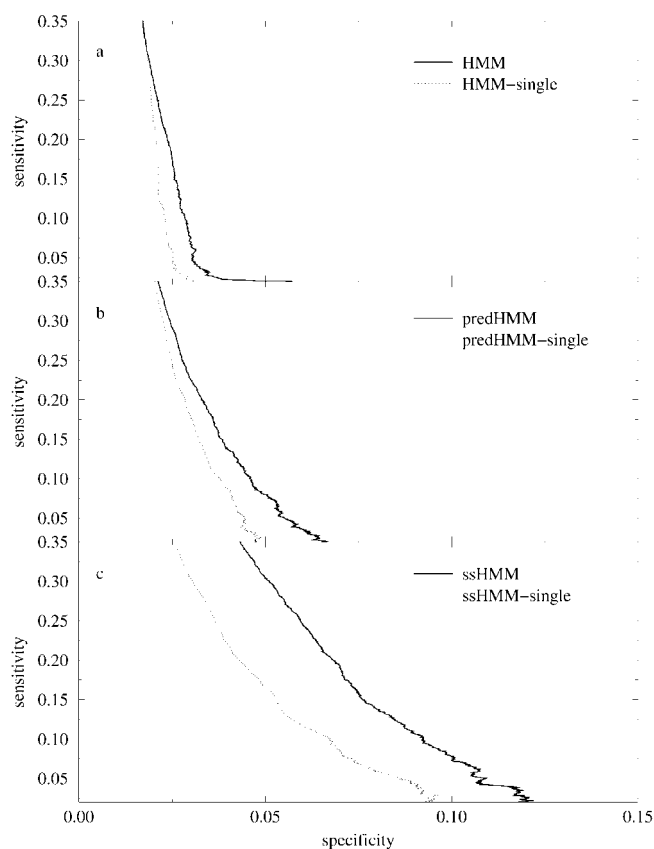


Fig. 3. When multiple sequence alignment is used (bold lines) the sensitivity of the hidden Markov models is increased, compared to using only single sequence alignments. In (a) standard HMMs are used, in (b) predHMMs, and in (c) ssHMMs.

Using Multiple Sequence Information Increases the Performance of HMMs

It has been assumed that using multiple sequences improves the performance of sequence-based search methods. However to our knowledge, there has been no studies showing that this is in fact true, using as complete benchmark as the one we have used here. Figure 3 shows that the sensitivity at a given specificity is increased for models built from multiple sequences compared to models built from just one sequence. This is most obvious for the ssHMMs, where at a specificity of 5%, the sensitivity increases from 17% to 30% when using multiple sequences to build the ssHMMs, compared to single sequences. A clear increase can also be seen for ordinary HMMs, and when using predicted secondary structures. The number of sequences placed at rank one is more than doubled when building models from multiple sequence alignments. They increase from 14% to 30% for the ssHMMs, from 10% to 19% using predHMMs, and from 4% to 12% for the ordinary HMMs (Table IV). It should be remembered that when using multiple sequence alignments we have used only automatically-created alignments from HSSP, and for many proteins these alignments do not contain enough

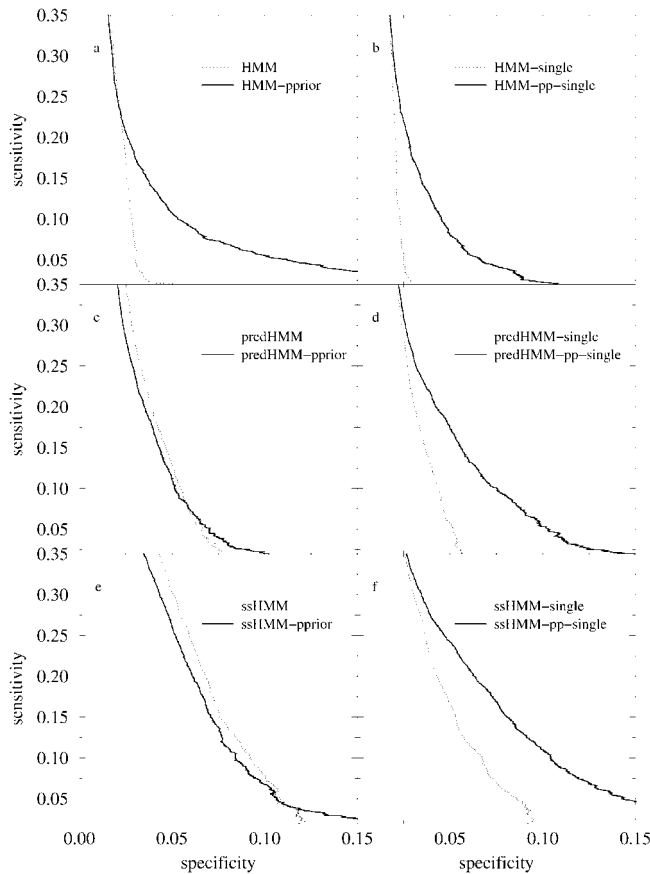


Fig. 4. The specificity is increased when a substitution matrix is used (bold lines). In (b,d,f) HMMs created from single sequences are used, while in (a,c,e) multiple sequence HMMs are used. In (a,b) standard HMMs are used, in (c,d) predHMMs, and in (e,f) ssHMMs.

diversity to perform as well as HMMs created from a more diverse set of sequences.

Using a Substitution Matrix Increases the Performance of HMMs

A standard hidden Markov model does not include any information about which substitutions are most likely, i.e., a substitution matrix is not used. If the protein family is large enough and diverse enough this should not be a problem. However, in our benchmark, we have many small families with low diversity. By including a substitution matrix we attempted to overcome this problem. As can be seen in Figure 4a,b, the use of a substitution matrix when building the models increased the sensitivity significantly. For hidden Markov models built from multiple sequence alignments, the sensitivity increases from 2% to 11%, at a specificity of 5%, when using the substitution matrix. When comparing Figures 4a and 3a, and Figures 4a and 4c, it can be seen that the use of a substitution matrix helps more than the use of multiple sequence alignments.

In Figure 4d,f, it can be seen that the ssHMMs and predHMMs built from single sequences have higher sensitivities when using a substitution matrix than when not.

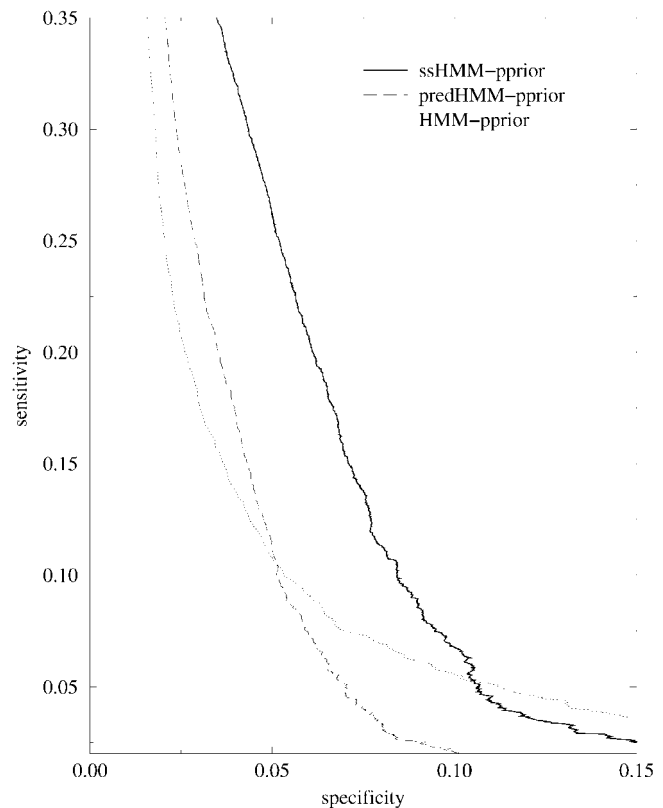


Fig. 5. A final comparison of HMMs that use multiple sequence information as well as substitution matrices. It can be noted that at higher specificities the sensitivity is lower for the ssHMMs and predHMMs than for the HMMs that do not use secondary structure information.

However, for the ssHMMs built from multiple sequence alignments, using a substitution matrix does not seem to improve the performance. On the contrary, the sensitivity decreases from 30% to 26% when the substitution matrix is added to the ssHMMs (Fig. 4e, Table III). This indicates that the prior distribution might not be optimized for the secondary structure HMMs. For these, another prior, where the secondary structure is included, could be used.

When Creating a Hidden Markov Model It Is Best to Use Multiple Sequence Alignments and Substitution Matrices

From the previous results it was concluded that the use of multiple sequence alignments and substitution matrices give the best results. A comparison between the HMM-pam methods with or without secondary structure information can be seen in Figure 5. At a low specificity (<5% for predHMM and <10% for ssHMM), the secondary structure HMMs have a higher sensitivity than the ordinary HMMs. For higher specificities, however, the ordinary HMMs have a higher sensitivity. One possible explanation of this is that the ssHMMs give very high scores to some false matches. When studying false matches with high scores for predHMM-pams, we found that there were a few families that caused a very large part of these false positives. The majority of these matches were between

different families that all consisted of a various number of alpha helices. From this data, it seems plausible that the contribution from the secondary structure was ranked too high in comparison with the contribution from the sequence. The secondary-structure-based HMMS still place more correct sequences at high ranks than the ordinary HMMS (Table IV). For example, the number of sequences correctly ranked as number one is increased from 16% to 27% when adding the secondary structure, and to 20% when using predicted structures.

In Tables III and IV and in Figure 5, a summary of all methods is shown. The ranks clearly support the conclusions that using multiple sequence alignment, predicted secondary structures, and a substitution matrix improves the performance of HMMS. For instance, when using a single sequence HMM, only 4% of the probe sequences recognize a correct target. This figure increases to 12% when using a multiple sequence alignment, and to 10% when using either a predicted secondary structure or a substitution matrix. When using a combination of all three methods, the number of probe sequences that recognize a correct target is increased further to 20%. The number of probes that recognize a correct target among the top 10 hits is increased from 24% to 31–38% when using multiple sequence alignments, predicted secondary structures, or substitution matrix, and to 45% when using all three.

The sensitivity shows a pattern similar to that of the ranks, although there are also some notable differences. First, it can be seen that predHMM-pam-single performs better than the HMMS that use multiple sequence alignments. This might indicate that the use of substitution matrices is not the optimal choice with the ssHMMS, as discussed above. Second, the standard HMMS that use substitution matrices perform better at higher specificity than the predHMMS. This might be due to the occurrence of a few false positives that have very high scores, as described above.

HMMS Perform as Well as but not Better Than Single-Sequence-Based Methods

The performances of all these methods were compared with the performance of single-sequence-based methods—*fasta*, *blast*, and *ssearch*. It could be assumed that the performance of *ssearch* should be similar to the performance of single sequence HMMS using a substitution matrix. However, *ssearch* performs better than HMM-single, as can be seen in Tables III and IV. Actually, all the single-sequence-based methods perform significantly better than HMM-pam-single and when it comes to ranks, they actually perform as well as standard HMMS. When studying the spec-sens curves it can be seen that the performance of *blast* and *fasta* are not superior to HMM-pam-single. However, *ssearch* still performs as well as standard (multiple sequence) HMM methods.

The reason why the multiple sequence information does not improve the performance further is probably due to the following. (1) In our benchmark, 45% of the HMMS are built from sequences with less than 10 sequences and HMMER is not optimized for small families. Furthermore,

even in the case where there are several sequences they are often very similar, and thus still fail to provide the necessary diversity. (2) The gap penalties in a HMM are calculated individually for each position in the model. However, when an HMM is created from a family with low diversity, and thus few gaps, the gap penalties will not be optimal for recognizing a distant member of the family. (3) *Blast*, *fasta*, and *ssearch* use an extreme value distribution to fit the scores. This method has been included in HMMER-2.0, and consequently the performance has improved (data not shown). (4) When a hidden Markov model is created, it includes a process of optimizing the transition probabilities. Ideally, one should make several tries and create several hidden Markov models for a given sequence family and then use the one that performs best. However, this was not possible in this study, due to computational limitations. All these points show some of the limitations of the current generation of HMMS, but also indicate some easy methods to improve the performance of HMMS.

In fold recognition it is not enough to identify the correct fold of a protein, it is also necessary to make the correct alignment between the two proteins to obtain three-dimensional studies. In the alignments obtained for ssHMM and the other methods from our benchmark, however, most pairs in our benchmark contained proteins that were very distantly related, or not homologous at all, and these proteins are extremely difficult to align correctly. We were, unfortunately, not able to detect any significant improvement of the alignments using ssHMM (data not shown). In a future study we plan to create an alignment benchmark using a set of less difficult proteins to align and examine whether ssHMMS, or standard HMMS, are able to align proteins better than standard pairwise sequence methods.

Use of ssHMM in CASP3

The ssHMM method, together with other methods and manual judgment, were used for blind predictions in the CASP3 process.²⁷ Three successful fold predictions were made of CASP3 targets T0046, T0053, and T0071a. T0046 (gamma-adaptin, ear domain) is an IG-like fold, and several methods (ssHMM, standard HMMS, and *threader*³) consistently scored high for IG-like domains. For T0053 (CbiK protein), we mainly focussed on the *threader* results. Our best prediction was T0071 (Alpha adaptin ear domain), in which, using ssHMM, we were able to identify the first 125 residues as an IG-like fold. We were also able to produce a rather good alignment, with 21 out of 125 residues correctly aligned.

Summary

The program package HMMER was modified to allow the construction of hidden Markov models (HMMS) that use the secondary structure, in addition to the amino acid sequence, to model protein families. This was accomplished by adding a distribution over emission probabilities for secondary structures to each match and insert state in the model. It was shown that the resulting secondary structure HMMS perform better than the ordi-

nary HMMs, with both the true and the predicted secondary structures used to recognize proteins having the same fold as the modeled sequences. We have also analyzed the performance of automatically-created HMMs, using a rigorous benchmark. It was shown that using a substitution matrix improved the performance of HMMs. Finally, it was shown that the automatically-created HMMs did not perform significantly better than single sequence based methods.

REFERENCES

1. Krogh A, Brown M, Mian I, Sjölander K, Haussler D. Hidden Markov models in computational biology: application to protein modeling. *J Mol Biol* 1994;235:1501–1531.
2. Karplus K, Sjölander K, Barrett C, et al. Predicting structures using hidden Markov models. *Proteins Suppl* 1997;1:134–139.
3. Jones D, Taylor W, Thornton J. A new approach to protein fold recognition. *Nature* 1992;358:86–89.
4. Flöckner H, Domingues F, Sippl M. Proteins folds from pair interactions: a blind test in fold recognition. *Proteins Suppl* 1997;1:129–133.
5. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. *Protein Sci* 1996;5:947–955.
6. Rice D, Eisenberg D. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol* 1997;267:1026–1038.
7. Rice D, Fischer D, Weiss R, Eisenberg D. Fold assignments for amino acid sequences of the CASP2 experiment. *Proteins Suppl* 1997;1:113–122.
8. Di Francesco V, Geetha V, Garnier J, Munson P. Fold recognition using predicted secondary structure sequences and hidden Markov models of proteins folds. *Proteins Suppl* 1997;1:123–128.
9. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. *J Mol Biol* 1997;270:471–480.
10. Elofsson A, Fischer D, Rice D, Le Grand SDE. A study of combined structure/sequence profiles. *Fold Des* 1996;1:451–461.
11. Dayhoff M, Barker W, Hunt L. Establishing homologies in protein sequences. *Methods Enzymol* 1983;91:254.
12. Rabiner L, Juang B. An introduction to hidden Markov models. Los Alamitos CA: IEEE ASSP Magazine. Jan 4–15, 1986.
13. Krogh A. Two methods for improving performance of an HMM and their application for gene finding. *ismb* 1997;5:179–186.
14. Sonnhammer E, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *ismb* 1998;6:175–182.
15. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;253:164–170.
16. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584–599.
17. Hubbard JT, Park J. Fold recognition and ab initio structure predictions using hidden Markov models and β -strand pair potentials. *Proteins* 1995;23:398–402.
18. Eddy SR. HMMER—hidden Markov model software. <http://www.genome.wustl.edu/eddy/hmmer.html>
19. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological sequence analysis*. Cambridge, UK: Cambridge University Press; 1998.
20. Sander C, Schneider R. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins* 1991;9:56–68.
21. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 1997;27:329–335.
22. Altschul S, Madden T, Schaffer A, et al. Gapped blast and ψ -blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
23. Pearson W. Comparison of methods for searching protein sequence databases. *Protein Sci* 1995;4:1145–1160.
24. Abagyan R, Batalov S. Do aligned sequences share the same fold? *J Mol Biol* 1997;273:355–368.
25. Brenner S, Chothia C, Hubbard T. Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. *Proc Natl Acad Sci USA* 1998;95:6073–6078.
26. Murzin AG, Breener SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
27. Moult J, Hubbard T, Bryant S, Fidelis K, Pedersen J. Critical assessment of methods of proteins structure predictions (CASP): round II. *Proteins Suppl* 1997;1:2–6.
28. Murzin A, Bateman A. Disant homology recognition using structural classification of proteins. *Proteins Suppl* 1997;1:105–112.
29. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:101915–10919.