

# Incorporating global information into secondary structure prediction with hidden Markov models of protein folds

Valentina Di Francesco\*, Philip McQueen†, Jean Garnier‡ and Peter J. Munson\*

National Institutes of Health, Bethesda, MD 20892-5626, USA

\*Analytical Biostatistics Section, Laboratory of Structural Biology, †Scientific Computing Resource Center, and ‡Fogarty International Center

Email: {valedf, mcqueenp, munson}@helix.nih.gov, garnier@biotec.jouy.inra.fr

## Abstract

Here we propose an approach to include global structural information in the secondary structure prediction procedure based on hidden Markov models (HMMs) of protein folds. We first identify the correct fold or 'topology' of a protein by means of the HMMs of topology families of proteins. Then the most likely structural model for that protein is used to modify the sequence of secondary structure states previously obtained with a prediction algorithm. Our goal is to investigate the effect on the prediction accuracy of including global structural information in the secondary structure prediction scheme, by means of the HMMs. We find that when the HMM of the predicted topology of a protein is used to adjust the secondary structure sequence, predicted originally with the Quadratic-Logistic method, the cross-validated prediction accuracy (Q3) improves by 3%. The topology is correctly predicted in 68% of the cases. We conclude that this HMM based approach is a promising tool for effectively incorporating global structural information in the secondary structure prediction scheme.

**Keywords:** protein structure prediction, hidden Markov models, fold recognition, secondary structure.

## Introduction

It has been suggested often that the reason for the limited prediction accuracy achieved by the various protein secondary structure prediction algorithms is due to the fact that most of these algorithms do not consider the interactions among residues distant in the chain. Here, we present a new approach for incorporating global structural information based on hidden Markov models (HMMs) of protein folds. We first try to identify the correct fold or "topology" of a protein by fitting to a hidden Markov model of the topology family (Di Francesco, Garnier, and Munson 1997). The most likely structural model is then used to modify the predicted sequence of secondary structure states. Although originally developed to provide a solution to speech recognition problems (Rabiner 1989) hidden Markov models have been applied to such biological problems as gene finding (Krogh, Mian, and Haussler 1994; Kulp et al. 1996), multiple protein and nucleotide sequence alignments (Eddy 1996; Krogh et al. 1994). HMMs have also been applied to secondary structure

prediction (Asai, Hayamizu, and Handa 1993) directly from the primary sequence. In contrast, we use HMMs of the protein topologies together with a predicted secondary structure sequence obtained from a published prediction algorithm. It is necessary only that the predicted state be accompanied by an estimated probability or confidence level of its correctness at each residue. Our main goal is to investigate whether the topology of a predicted secondary structure sequence can be used to improve the secondary structure prediction accuracy. We also estimate the effect of mis-assignment to protein topology

## Methods

### Hidden Markov models of protein folds

A new approach to the fold recognition problem based on sequences of secondary structure states of residues, i.e. helix (H), strand (E) and coil (C), was presented previously (Di Francesco, Garnier, and Munson 1997). With this approach we have shown that protein topology can often be recognized from the sequence of secondary structures alone, using hidden Markov models with the architecture proposed by Krogh and colleagues (1994), to model families of proteins of a specific structural topology (selected from the CATH database, version Jan. 1995 (Orengo et al. 1993)). These models describe the consensus secondary structure of topology groups of proteins, such as globins, TIM barrels or serine proteases. The model parameters (namely the *transition* probability distributions between hidden states (*match*, *insert* and *delete* states at each position in the model) and the *observation symbol* probability distributions for the secondary structures H, E, C given that the residue is in a *match* or an *insert* state) are estimated using well established algorithms (Hughey and Krogh 1995; Rabiner 1989). These algorithms automatically align sequences of crystallographically determined secondary structure states of protein family members as the model parameters are calibrated. A model is then asked to identify those predicted sequences, among a large set of proteins, with the topology described by the model. Each sequence is assigned a log-odds score indicating the relative likelihood that the protein has that particular topology

compared to a generic null model. Based on each available HMM of protein topologies the scores are calculated and ranked so that a high ranking sequence is associated with the model that has the highest likelihood of describing the correct topology for that sequence. For testing purposes, the method is considered successful if the secondary structure sequence is ranked higher by the HMM of its true topology than by HMMs of other topology families.

### Adjusting predictions with the HHMs

Two approaches are tested: The first one modifies the predicted sequence using the most likely secondary structure state in the position of the topology model to which the residue has been aligned. The second one utilizes both the *transition* and *observation symbol* probability distributions of the topology HMM, together with the probability distributions or confidence levels associated with the predicted secondary structure sequence.

**Method 1.** The secondary structure prediction algorithm provides a sequence of predicted states  $S = s_1s_2\dots s_L$ , where  $s_i \in \{H,E,C\}$  and  $L$  is the length of the sequence and  $1 \leq i \leq L$ . When the sequence  $S$  of predicted states is aligned to an HMM, it chooses a series of hidden states  $q_1q_2q_3\dots q_L$  forming a path  $Q$  through the model from beginning to end. One way to obtain a new secondary structure prediction  $S' = s'_1s'_2\dots s'_L$  is by simply replacing the predicted state  $s_i$  with the most likely observation symbol  $s'_i$  for the corresponding state  $q_i$  in the HMM, that is

$$s'_i = \underset{s_i \in \{H,E,C\}}{\operatorname{argmax}} \left[ P(s_i | q_i) \right].$$

If the residue is aligned to an insert state then its predicted secondary structure state is not changed. Note that this method is highly dependent on the specific alignment  $Q$  of the predicted sequence  $S$  to the model  $\lambda$ .

**Method 2.** For each amino acid in a query sequence  $S$ , some prediction algorithms provide a probability distribution  $P' = (P'(H), P'(E), P'(C))$  describing the likelihood of that residue having one of the three secondary structure states. The residue's predicted state  $s$  is set to be the  $\operatorname{argmax}_{s \in \{H,E,C\}} [P'(s)]$ . It often happens that a secondary structure

state, say H, is chosen instead of another, say E, but that  $P'(H)$  is not much larger than  $P'(E)$ , so that there is no strong indication that the helical state is to be preferred to the extended state. Another way to modify the secondary structure prediction while taking into account the most likely protein topology model  $\lambda$ , and the probability distributions  $P'_i$  associated with each residue is to find a new sequence of secondary structure states  $S' = s'_1s'_2\dots s'_L$  and a new sequence of hidden states  $Q' = q'_1q'_2\dots q'_L$ , that maximizes

$$\prod_{i=1}^L P(q_i | q_{i-1}) P(s_i | q_i) P'_i(s_i) P(q_{L+1} | q_L)$$

where the maximum of this product is taken over all the possible

paths through the model  $Q = q_1q_2q_3\dots q_L$  and all possible secondary structure sequences  $S = s_1s_2\dots s_L$ . Note that in the new predicted sequence  $S'$  each secondary structure state is obtained by weighing for the observation symbol probability at the position to which the residue was aligned to the model. Moreover, the new predicted state depends also on the original probability distribution  $P'_i$ , so that residue states predicted with high probability values are more likely not to be changed by this procedure. A modification of the Viterbi algorithm (Rabiner 1989) was used to obtain the new predicted sequence  $S'$  and its associated new alignment  $Q'$  to the model  $\lambda$ . Details of the algorithm will be described elsewhere; the code is available from the authors upon request.

### Test proteins

Twenty-eight predicted secondary structure sequences (4,604 residues; 2041, 758, 1805 in respectively H, E and C conformation) were used to test whether the knowledge of the topology of a protein improves the secondary structure prediction quality. We realize that the test set is too small to draw definitive conclusions, but the small size is due to the limited number of HMMs of protein folds available so far. Moreover, the database of 112 predicted secondary structure sequences (obtained with the Quadratic-Logistic algorithm (Munson, Di Francesco, and Porrelli 1994), average cross-validated Q3=68%) used as a 'control' database in the fold recognition experiments with HMMs (Di Francesco, Garnier, and Munson 1997) contains only few representative sequences for protein topology family of each available fold HMM. The protein topology families and member protein PDB identifiers for which an HMM was available are: In the  $\alpha$  class: *Globin*: 1eca, 2lhb, 1sdhA, 2lh4, 1colA; *Cytochrome C*: 1cc5, 5cytR; *Cytochrome b562*: 2hmzA, 2ccyA, 256bA, 2tmvP; *EF-Hand*: 4cpv, 3icb, 3cln; In the  $\alpha$ - $\beta$  class: *TIM Barrel*: 3timA, 4xiaA, 1wsyA, 3dhq; *Ras P21*: 1etu, 1s0l, 4fxn; *Plait ptf*: 2fxb; *OB*: pns1; In the  $\beta$  class: *Orthogonal Barrel*: 1bbpA, 1rbp, 1lib; *Serine Protease*: 2alp, exta. Three proteins in this test set are part of a group of five *bona fide* fold recognition predictions submitted to the 2nd Critical Assessment of techniques for protein Structure Prediction meeting (CASP2). At the time of the prediction submission, we only had available the HMM of the true fold for those three target proteins: T0004 (pns1), T0014 (3dhq) and T0020 (exta). None of these five target sequences had sequence identity higher than 25% with any protein of known structure. The twenty-eight test sequences are chosen so as not to have pairwise sequence identity higher than 25%. Moreover, careful cross-validation of the model training sets was performed by removing from each training set all the amino acid sequences homologous to the query sequence (Di Francesco, Garnier, and Munson 1997). For example, since the test set has five globin sequences, five different HMMs for the Globin fold were trained. In all, twenty-eight cross-

validated HMMs of protein folds were used to test the protein fold recognition capabilities of the HMMs and evaluate their contribution to improving the secondary structure prediction quality.

## Results

### The knowledge of a protein topology improves the secondary structure prediction quality

Table I shows the results of using the protein topology HMMs to modify the QL-predicted secondary structure sequences. The application of method 1 decreased the prediction accuracy by 0.4%, even when the right topology for a query sequence is known. However, using method 2, the knowledge of the protein topology can improve the predicted secondary structure sequence, up to 5.6% if the correct protein topology is known. With method 2, 78% (22 out of 28) of the sequences showed an increased prediction accuracy.

Table I. Difference in Q3 values due to the use of fold HMMs.

	Using the correct model		Using the predicted model	
	QL	QL Adjusted Method 1	QL Adjusted Method 2	QL Adjusted Method 2
Q3* (%)	69.6	69.2	75.2	72.6
difference	(0.0)	(-0.4)	(+5.6)	(+3.0)

\* Percentage of correctly predicted residues in three secondary structure states. The DSSP assignments were considered as the true secondary structure state.

### Adjusting predictions with predicted protein fold models

In real protein structure prediction experiments one does not know the correct fold in advance, therefore the correct HMM for adjusting the predicted sequence is unknown. Thus, we also adjusted the predicted sequence using the highest ranking model for the original predicted sequence. Table II summarizes the results obtained using method 2. With this procedure, 19 out of 28 cases (or 67.8%) found the correct topology model confirming the previously determined success rate of protein topology HMMs to recognize protein folds. Of those 19 cases, 14 proteins showed an improvement in prediction accuracy, so 50% of the 28 proteins were recognized correctly and achieved an improvement (5.4%) in secondary structure prediction accuracy.

Table II: Summary of protein fold recognition results versus changes in QL prediction accuracy, due to the use of method 2.

	# Proteins with increased Q3	# Proteins with decreased Q3	Total
Correctly identified fold	14	5	19
Incorrectly identified fold	5	4	9
Total	19	9	28

On the whole set of 28 proteins, Q3 was improved by 3% (Table I). This average includes the case of the tobacco mosaic virus protein 2tmvP, that was incorrectly recognized by the orthogonal beta barrel model, while its correct topology is 4 helix bundle. The protein 2tmvP was originally predicted with 60.4% accuracy and, after adjustment with the wrong model, accuracy dropped to 46.1%. Surprisingly, 5 of the 9 proteins for which the wrong fold model was utilized to adjust the prediction sequence, still showed an improvement in prediction accuracy. For example, the globin 2lh4 showed a prediction improvement (~10%) even though the prediction was adjusted with the EF-Hand model. Figure 1 also shows that in most cases the decrease in prediction accuracy is limited to about 3%, excluding 2tmvP, while, not surprisingly, bigger increases in Q3 values are more common and associated with the use of the correct fold. Table III shows that the use of the HMM based approach produces a large improvement in the correct QL prediction of beta strands (+9.5%) mainly due to the reduction of extended residues predicted as helices (-10.9%). Similarly a decrease in the number of helical residues predicted as strands is observed (-2.2%).

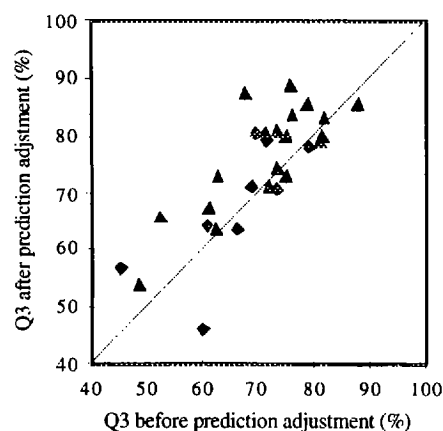


Figure 1. Plot of the Q3 prediction accuracy values for QL predicted secondary structure sequences before and after adjustment with method 2, using the HMM chosen by the fold recognition procedure. Q3 values of proteins which were correctly associated to the true HMM are indicated by ▲; Q3 values of those proteins that were mis-classified to the wrong HMM are indicated by ◆.

Table III: Summary of prediction accuracies (%) for the three predicted states.

Observed	QL Predicted			QL Adjusted* predicted			Total
	H	E	C	H	E	C	
H	75.2	6.2	18.6	77.4	4.2	18.4	100
E	16.8	45.8	37.4	5.9	55.3	38.8	100
C	18.7	8.1	73.2	14.4	10.8	74.8	100

\* QL predictions adjusted with method 2. Each number in the main diagonal represents the percentage of correctly predicted residues in one of the three secondary structure states. The numbers off diagonal are the percentage of mis-assigned residues.

## Discussion

We have shown here that the hidden Markov models of protein folds are a promising tool to effectively include global structural information in the secondary structure prediction scheme. The present approach can be readily applied to predicted sequences obtained with other prediction algorithms, as long as each predicted state is accompanied by an estimated probability or confidence of its correctness.

Given that the global structural information is explicitly included by means of the HMM, one question to ask is why the prediction accuracy improvement obtained with method 2 is not higher than that achieved when using the model of the actual fold of a predicted sequence (75.2% with QL). One possible reason is that these HMMs, as much as many other protein fold recognition techniques, do not produce good quality sequence-to-structure alignments, as we have noted previously (Di Francesco, Garnier, and Munson 1997). The alignment quality is probably the cause of the lack of performance of method 1, and it also influences how the HMM parameters and the QL's probability distributions are combined by the modified Viterbi algorithm used in method 2 to adjust the predicted sequence. The reason why method 2 is less affected than method 1 is that the second approach provides a new alignment, while also adjusting the predicted sequence. It is reasonable to expect that with a better predicted sequence, such as those adjusted with the correct model, a better sequence-to-structure alignment is obtained.

Another possible reason for the limited overall prediction accuracy value is related to the nature of a topology model itself. A model describes the secondary structural features common to the proteins in its training set, so it can potentially adjust the prediction only in those regions aligned to the common structural features. For example, a protein whose core structure consists of a small percentage of its chain length, such the case of TIM Barrels which often have additional structural domains, will only achieve a limited benefit from this procedure. It has been shown by Russell and Barton (1994) that protein pairs having similar 3D structures, but sequence identity less than 20%, can have as low as 30% of their residues in structurally equivalent regions forming the core. They have also shown that the same protein pairs with low sequence identity have rarely more than 80%-85% of residues in identical secondary structure states. These values are not too far from what we obtained here with QL predictions (Q3=75.2%) when using the model of the correct fold for the predicted sequences, especially considering that the proteins in the HMM training sets have low sequence identity to the query proteins.

In conclusion, the major limitations of the approach presented here are due to the alignment quality of the predicted sequence to the model and the intrinsic nature of a structural model itself, which depends on the level of conservation of the secondary structure of proteins having a similar fold. More tests with a larger database of test proteins are in progress to obtain a more statistically robust proof of the usefulness of this approach.

## References

- Asai, K., Hayamizu, S., and Handa, K. 1993. Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.* 9(2): 141-146.
- Di Francesco, V., Garnier, J., and Munson, P. J. 1997. Protein topology recognition from secondary structure sequences - Application of the hidden Markov models to the alpha class proteins. *J. Mol. Biol.* 267(2): 446-463.
- Eddy, S. R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* 6(3): 361-365.
- Hughey, R., and Krogh, A. 1995. SAM: Sequence alignment and modeling software system. Technical Report. UCSC-CRL-95-7. UC Santa Cruz.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235: 1501 - 1531.
- Krogh, A., Mian, I. S., and Haussler, D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res* 22(22): 4768-4778.
- Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. In Proc. 4th Int. Conf. Intel. Sys. Mol. Biol. : 134-142. St. Louis, MO, U.S.A.
- Munson, P. J., Di Francesco, V., and Porrelli, R. 1994. Protein secondary structure prediction using periodic-quadratic-logistic models: statistical and technical issues. In Proc. 27th Hawaii Int. Conf. on System Sciences. V: 375 - 384.
- Orengo, C. A., Flores, T. P., Taylor, W. R., and Thornton, J. M. 1993. Identification and classification of protein fold families. *Protein Eng.* 6(5): 485 - 500.
- Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE. 77: 257 - 286.
- Russell, R. B., and Barton, G. J. 1994. Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* 244: 332 - 350.