# PREDICTION OF PROTEIN STRUCTURE USING A HIDDEN MARKOV MODEL WITH LIMITED NUMBER OF STATES

Christos Lampros[*], Costas Papaloukas[*,**], Yorgos Goletsis[*,***], Dimitrios I. Fotiadis[*]

[*] Unit of Medical Technology and Intelligent Information Systems,
Dept. of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece
[**] Department of Biological Applications and Technology, University of Ioannina, GR 45110 Ioannina, Greece
[***] Department of Economics, University of Ioannina, GR 45110 Ioannina, Greece

fotiadis@cs.uoi.gr

**Abstract: Computational analysis of proteins can be used for structure prediction of newly identified protein sequences. In this way, valuable information related to their function can be derived. Hidden Markov Models (HMMs) have been largely applied for this task. However, due to their topology they suffer from lack of computational simplicity and the need for complex training algorithms. In this work, a Hidden Markov Model with a reduced state-space topology is proposed to serve as a protein classification tool. The model employs an efficient architecture and a low complexity training algorithm based on likelihood maximization. In addition, secondary structure information is introduced to the model to increase its performance. The proposed model was tested in two different tasks, i.e. class prediction and fold recognition. The dataset, used for the evaluation of the proposed methodology, comes from SCOP and PDB databases. The classification performance of our model was tested against the SAM approach, which is considered as a benchmark in sequence based protein classification. The proposed classifier performed better than SAM in the overall prediction accuracy.**

## Introduction

The large number of genome projects during the last years led to an exponential increase in the number of identified protein sequences. Nevertheless, for the majority of these sequences there is no information available concerning their function or their structure. Understanding the structure of these sequences is a way to define their function, as proteins with similar structure have in general similar function. A newly identified protein can be related with proteins in annotated databases whose structure is known. In computational analysis of proteins this can be considered as a classification problem which can be divided in two tasks: fold recognition and class prediction.

Several machine learning methods have been proposed in the literature for these tasks. Genetic algorithms have been applied for fold recognition [1], as well as artificial neural networks [2] and support vector machines [3,4]. For the prediction of the structural class of a protein several methods have been also suggested. These methods are mostly based on the amino acid composition of the protein [5]. Various statistical approaches have been adopted to deal with that problem [5,6].

Among the sequence-based approaches that use hidden Markov models (HMMs) Sequence Alignment and Modelling (SAM) method is considered as the most prominent and more representative [7,8]. Furthermore, other surveys have shown that secondary structure information can be incorporated in the HMM and increase the fold recognition performance [9]. Recently the same approach was extended with the additional application of different alphabets for backbone geometry [10]. However, the main disadvantage of HMMs is the employment of large model architectures which demand large datasets and high computational effort for training. As a consequence, in cases where the available datasets are inadequate, e.g. small classes or folds, their performance deteriorates.

In this work, a HMM with a limited number of states is proposed to serve as a classification tool for structure prediction. Due to its simplified architecture the proposed model is easy to be trained. The model's architecture includes a small number of states while a low complexity training algorithm is used. Secondary structure information is introduced to the model to increase its performance. The model addresses the problem of multi-class classification of sequences, meaning that the method employed classifies a query sequence of unknown structural category in one of the candidate categories, classes or folds. The proposed model is used for two tasks: class prediction and fold recognition.

Folds from two different major structural classes were used. Class prediction is employed at the first stage. The proteins correctly classified are assigned to the folds of the specific class at the second stage. A Bayesian multi-class classification approach is used for the classification of proteins in the appropriate category of each stage. The obtained results are equivalent or even better than other similar techniques, whereas the computational load is significantly smaller.

ISSN: 1727-1983 © 2005 IFMBE

## Materials and Methods

HMMs are widely used in modelling families of biological sequences. Each HMM consists of a set of states *S* and a set of possible transitions *T* between them. Each state stochastically emits a signal and then the procedure is transmitted to another state with a probability depending on the previous state. The procedure continues until the total of each sequence is emitted. There is also a beginning state where the process starts and a set of transition probabilities from the beginning of each possible state. That set of probabilities sums to unity and so does the set of emissions of possible signals in each state and the set of transitions from each state. The observer does not know which state produced each specific signal, because that state is hidden from him. This is the first main characteristic of a HMM, which differentiates it from other stochastic models. The second is the Markov property, which means that given the value of the previous state $S_{t-1}$ the current state $S_t$ and all future states are independent of all the states prior to *t-1* [11].

A HMM is trained using a set of sequences called training set. The aim of the learning procedure is to maximize the likelihood of the model given the training data.

The current methods which employ HMMs for protein classification adopt the above approach but they have some limitations. First they use a very big number of states, as their topology corresponds to a multiple sequence alignment among sequences and the length of the model is proportional to the length of the alignment. That leads to a huge number of parameters to be calculated. The next problem is the training algorithm adopted (e.g. Baum – Welch) which is very complex and demands a vast amount of calculations in order to locate a local maximum of the likelihood [12].

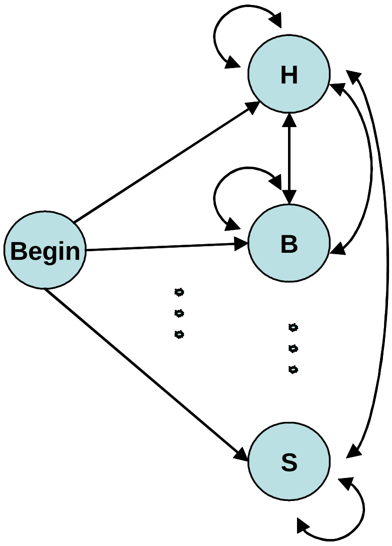

Figure 1: Topology of the HMM with a limited number of states (H, B,…, S are the letters of the DSSP alphabet).

*Model description*

The proposed HMM, whose topology is shown in Fig. 1, can overcome the above mentioned limitations. The key for that is the additional use of secondary structure information in such a way that the states of the model will depict the possible different secondary states and not different positions in a multiple alignment. The correct classification among different structural groups demands the use of the secondary structure information and not only that of the primary structure. It has been proved [9] that it is more effective to use such information which is confirmed and published in databases like the Protein Data Bank (PDB) [13], than to use predicted secondary structure which inevitably contains errors in the sequence of the residues.

We use secondary structure sequences taken from the PDB, which are used in the context of our HMM as hidden state sequences. This offers the advantage of employing a HMM with a small number of states, which is equal to the number of the different letters in the Definition of Secondary Structure of Proteins (DSSP) alphabet [14] representing the possible secondary structure formations where each amino acid residue is found. The set of letters in the DSSP alphabet is {H,B,E,G,I,T,S}. Moreover, the state sequence of each primary sequence produced by the model is known, thus this fact enables us to use a low complexity training algorithm based on likelihood maximization and avoid complicated learning schemes.

There are seven different hidden states in the model corresponding to the underlying secondary structure. In the training set, there is one to one correspondence between the amino acid and the secondary structure residues. It should be noted that in the DSSP method an eighth state is also determined, which indicates unknown structure, but it is not taken into consideration in our method, so the amino acid residues with unknown structure are skipped during the modelling process. The states of the model are fully connected, that is all possible transitions between them are allowed. In each state a distribution over all possible amino acid residues is found. There are 21 possible residues which are the variables in each distribution, the 20 different amino acids and one more residue indicating amino acids of unknown origin. So the total number of the model parameters is 7x21 for the possible emissions, 7x7 for the possible transitions between states and 1x7 for the transitions from the beginning. The sum of all this is 203 parameters, which is much less than the number of parameters which is needed to be calculated by other current HMM methods for protein modelling.

The emission and transition parameters of the model are calculated in a single step with the use of maximum likelihood estimators. If $a_{kl}$ is the transition probability from state *k* to another state *l* and $e_k(b)$ the emission probability of the residue *b* in the state *k*, then the estimators are given by the following equations:

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}, \qquad (1)$$

and

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}. \qquad (2)$$

Where $l'$ all the states where the procedure can go after state $k$ and $b'$ all the symbols that can be emitted from state $k$. These are the estimation equations in the HMMs when the state sequences are known [11]. $A_{kl}$ is the number of times that the transition from $k$ to $l$ is used and $E_k(b)$ is the number of times the emission of $b$ from $k$ is used in the training test of sequences. Nevertheless, maximum likelihood estimators are vulnerable to overfitting when there is insufficient data. Whenever there is a state $k$ that is never used in the set of example sequences, then the estimation equations cannot be defined for that state, because both the numerator and the denominator will have zero value. In order to avoid such problems it is preferable to add predetermined pseudocounts to the $A_{kl}$ and $E_k(b)$ before using equations (1) and (2). Their values are given as:

$A_{kl}$ = number of transitions $k$ to $l$ in training data + $r_{kl}$ (3)

$E_k(b)$ = number of emissions of $b$ from $k$ + $r_k(b)$ . (4)

The pseudocounts $r_{kl}$ and $r_k(b)$ should reflect our prior biases about the probability values. In our case there is actually not prior belief, which means that the value of each pseudocount equals unity. It is as the prior distribution of amino acids in each emission state and the prior distribution of transitions from each state has been the uniform distribution. So here the pseudocounts are used only for avoiding overfitting and do not incorporate any prior knowledge.

*Implementation*

The proposed HMM uses the posterior probability scores. These scores are logarithmic forms of the probability of the sequence, given the model. According to the Bayes theorem, a test sequence is classified to that group whose model gives the maximum probability compared with the probabilities produced from all the other models of the candidate groups. Bayes theorem claims that the probability of a particular model given a sequence is proportional to the probability of the sequence, given the model. The later quantity is the likelihood of the sequence and can be calculated since the parameters of the model are known after the learning process.

The posterior probabilities are calculated with the use of the forward algorithm. The forward algorithm gives the probability that a sequence has been produced by a HMM by adding the probability of all possible paths of the sequence through the model. It is necessary to use logarithms in order to avoid underflow problems appearing when a very long product of probabilities has to be computed. Then a Bayesian classification table is constructed for the classification of the test sequences in the appropriate category, class or fold.

The data available is separated to make the training and test sets. The test sets contain only primary sequences, as all information concerning the structure of a protein is considered unknown. Log–likelihood scores are adopted for evaluating the proposed HMM. These scores will be calculated for each class model in the class prediction case and for each fold model in the fold recognition case. The likelihood score for a sequence against a model is divided with the score of that sequence against the so called null model. The null model assumes that the amino acid symbols are independent at each position, and assigns fixed emission probabilities based on the uniform distribution over the possible amino acids. The criterion for selecting the model which best classifies a particular protein is to choose the model with the highest posterior probability, and posterior probabilities correspond to the log-likelihood scores against the null model.

A group of protein sequences, both primary and secondary, is taken from the PDB. The members of this group correspond to specific classes or folds of the SCOP database [15]. A part of these sequences forms the training set used in the classification problem. The test set used for each case consists of the rest of the primary sequences and their class or fold is considered unknown. Actually, the structural categories of all sequences are known from the SCOP database. So we are able to evaluate the effectiveness of their classification in the correct group after the experiments.

The PDB sequence files include no organization of their data in structural groups, so that kind of information should be found in the SCOP database. There is a hierarchical categorization of proteins with known structure in the SCOP database, where class is the highest level and fold is the level that follows. The SCOP database contains files with categorization of the primary sequences, indicating the structural class and fold where they belong. The data from both databases should be combined, so that SCOP will provide the correct categorization and PDB will provide the relevant sequences, both primary and secondary. That happens for each class and fold to be tested in the classification tasks. The sequence identifiers come from the ASTRAL SCOP 1.67 dataset, where no proteins with more than 95% similarity are contained. The dataset used in our experiments is shown in Table 1. The most populated SCOP folds of classes A and B, most specifically those who have at least 50 members, are used to derive the training and test data for the experiments.

As far as the prediction of classes is concerned, two hidden Markov models with a limited number of states were trained. These models correspond to the SCOP classes A and B, which are two of the most populated ones. The training set for each class is the sum of the training sets of the folds of each class and the same happens with the test set. Then the members of the class test sets are scored against the models of each class and the test sequences are assigned to that class having the maximum probability. The class prediction task is the pre-stage of the fold recognition task and the Table 2 summarizes the results. The two stages of classification are shown in Fig. 2.

Table 1: The dataset used (2 SCOP classes and the 12 SCOP folds).

| Fold index | Number of sequences in the training set | Number of sequences in in the test set |
|---|---|---|
| **A** | **287** | **282** |
| a1 | 48 | 47 |
| a3 | 33 | 32 |
| a4 | 100 | 99 |
| a24 | 29 | 28 |
| a39 | 46 | 46 |
| a118 | 31 | 30 |
| **B** | **743** | **738** |
| b1 | 445 | 444 |
| b6 | 33 | 33 |
| b29 | 41 | 40 |
| b34 | 46 | 46 |
| b40 | 62 | 61 |
| b47 | 41 | 40 |
| b60 | 27 | 26 |
| b121 | 48 | 48 |

The fold recognition task includes the training of 12 hidden Markov models with limited number of states of the most populated SCOP folds. The final test sets used in this task consist of those test sequences that can be classified in the correct class when compared with the two class models. So the initial test sets for fold recognition are first filtered through the class models for the limited HMM. Then the remaining test sequences are scored against all fold models of the specific class and the prediction accuracy is calculated for all folds. The prediction accuracy is the number of test proteins uniquely recognized as belonging to a specific fold divided by the total numbers of test proteins belonging to that fold. The denominator corresponds to the total number of the proteins belonging to the initial test of each fold and not to the number of those which remaine after the filtering of the first stage. Finally, the total number of protein assigned correctly from all folds of a class divided by the total number of test proteins belonging to the class provides the class prediction accuracy.

## Results

The HMM with a limited number of states is compared against SAM [7] which is considered the most effective current method that employs HMMs for protein classification [16]. In the case of the SAM models, which are compared with our models, the posterior probabilities correspond to the negative log–likelihood scores of each sequence. So when the negative log - likelihood scores decrease, the posterior probabilities increase and an unknown protein must be assigned to that model which gives the lowest negative log – likelihood score for its primary sequence. Nevertheless, the decision taken for the classification of a sequence to a structural group is based upon ranked scores, because the comparison of log-likelihood scores of a sequence against different models is not trustworthy as those scores depend on model length [17]. It should be noted that model length varies in SAM and depends on the training set, unlike our model.

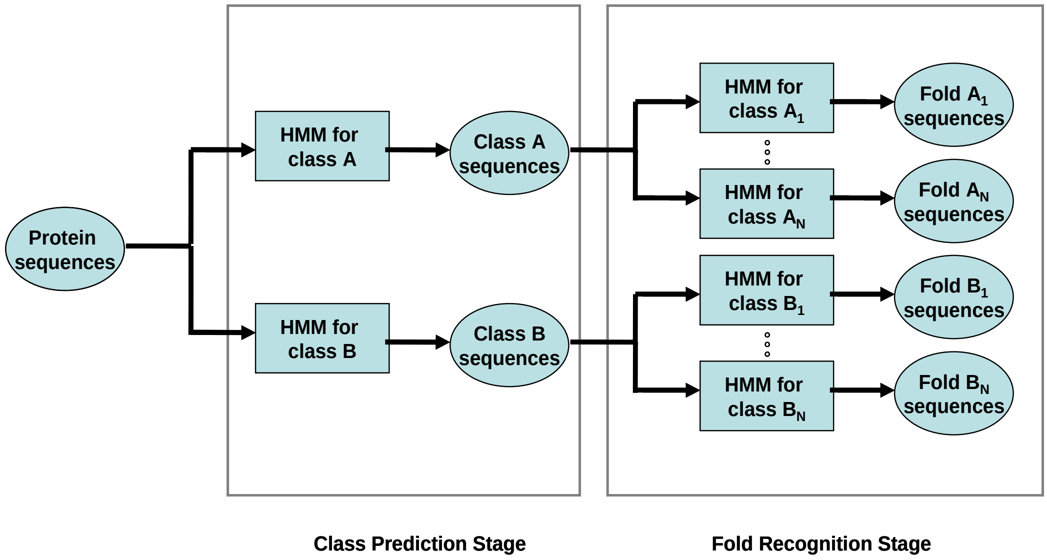

Figure 2: In the first stage of the classification procedure the appropriate class is identified while in the second stage the correct fold is determined**.**

The same training and test sets were used for both methods and the results are shown in Table 2. As it can be seen, the HMM with limited number of states outperforms SAM in the overall prediction accuracy for the folds of classes A and B.

Table 2: Comparison of the proposed model accuracy against SAM.

| Fold Index | Proposed HMM prediction accuracy | | SAM prediction accuracy | |
|---|---|---|---|---|
| a1 | 43/47 | 91.5% | 42/47 | 89.4% |
| a3 | 22/34 | 64.7% | 30/34 | 88.2% |
| a4 | 47/88 | 53.4% | 46/88 | 52.3% |
| a39 | 32/46 | 69.6% | 38/46 | 82.6% |
| a118 | 19/31 | 61.3% | 15/31 | 48.4% |
| **Class A** | **163/246** | **66.3%** | **171/246** | **69.5%** |
| b1 | 264/400 | 66% | 152/400 | 38% |
| b6 | 18/39 | 46.2% | 33/39 | 84.6% |
| b10 | 50/53 | 94.3% | 37/53 | 69.8% |
| b29 | 26/37 | 70.3% | 19/37 | 51.4% |
| b34 | 5/43 | 11.6% | 24/43 | 55.8% |
| b40 | 6/62 | 9.7% | 24/62 | 38.7% |
| b47 | 21/37 | 56.8% | 29/37 | 78.4% |
| **Class B** | **390/671** | **58.1%** | **318/671** | **47.4%** |
| **Overall** | **553/917** | **60.3%** | **489/917** | **53.3%** |

## Discussion

The HMM with a limited number of states that has been presented here is based on the concept of training with both primary and secondary sequence for class and fold modeling. Each hidden state of the model corresponds to a possible secondary state an amino acid can adopt, so the number of states is equal to the number of all possible versions of secondary structure. In each state a probability distribution over all possible amino acids is found. The state sequence is known during training, as the secondary sequences of the correspondent primary ones are given, so the learning algorithm is very fast and based on the calculation of maximum likelihood estimators of all parameters in a single step. After training, the probability score of unknown sequences against the created models are calculated with the use of the forward algorithm and Bayesian classification tables are constructed for assigning the test sequences to that category, either class or fold, whose model gave the maximum probability score. In all cases, only the primary sequences of proteins are needed in the test set. The classification takes place in two stages. In the first stage the test sequences are assigned to the appropriate class. In the second stage those sequences which are correctly assigned in the previous step are classified in the appropriate fold and the results are validated.

The classification performance of the HMM with a limited number of states is tested by comparing it to a SAM model trained with the same datasets. This model is linear and its length is equal to that of the multiple alignment which the SAM method gives for the specific test. Experiments indicate that the HMM with limited number of states is more accurate than SAM in the overall classification rate.

The proposed HMM implementation for classifying proteins in the appropriate class or fold is an approach which avoids iterative procedures demanding huge computational effort in training. Moreover, it is the only method, among those using secondary sequence information for fold recognition, where the knowledge of the secondary sequence of the target protein is not needed during the validation process, due to the nature of model’s structure. It provides equivalent or even better results than SAM implementation which demands extremely higher computational complexity in model’s training and larger number of states.

Additional structural features could be incorporated in the future to improve the performance, like residue solvent accessibility, for example. These features will add more states in the model without significant increase in the complexity, thus the low complexity training algorithm will be again appropriate for training. The ability of classifying proteins in the correct fold can be improved in that way with a small increase at computational cost.

## Conclusions

A HMM with a limited number of states was implemented to address the problem of modelling structural categories. The proposed HMM provides equivalent or even better results than other methods (SAM). It demands extremely lower computational complexity in model’s training and employs much smaller model architecture. Moreover, it is the only approach where the knowledge of the secondary sequence of the target protein is not needed during the testing phase. As a consequence, there is no need to predict the secondary sequence of the protein which is considered unknown.

## References

[1] DANDEKAR T. and ARGOS, P. (2004): ‘Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions’, *J. Mol. Biol.*, **256**, pp. 645-60

[2] ROST B. (1996): ‘PHD: predicting one-dimensional protein structure by profile-based neural networks’, *Methods Enzymology*, **266**, pp. 525-39

[3] BHASIN M., RAGHAVA GP. (2004): ‘GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors’, Nucleic Acids Research,32, pp. 383-89

[4] HAN S., LEE BC., YU ST., JEONG CS., LEE S. and KIM D. (2005): ‘Fold recognition by

combining profile–profile alignment and support vector machine', Bioinformatics, 21, pp. 2667 – 73

[5] WANG ZX and YUAN Z. (2000): 'How good is the prediction of protein structural class by the component-coupled method?' Proteins, 38, pp. 165-75

[6] LUO RY., FENG ZP. and LIU JK. (2002): 'Prediction of protein structural class by amino acid and polypeptide composition', Eur. J Biochem., 269, pp. 4219-25

[7] HUGHEY R. and KROGH. A. (1996): 'HMMs for sequence analysis: Extension and analysis of the basic method', CABIOS, 12, pp. 95-107

[8] LINDAHL E. and ELOFSSON A. (2000): 'Identification of related proteins on family, superfamily and fold level', J. Mol. Biol., 295, pp. 613-25

[9] HARGBO J. and ELOFSSON A. (1999): 'Hidden Markov Models That Use Predicted Secondary Structures for Fold Recognition', Proteins, 36, pp. 68-76

[10] KARCHIN R., CLINE M., MANDEL-GUTFREUND Y. and KARPLUS, K. (2003): 'Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry', Proteins, 51, pp. 504-14

[11] DURBIN, R., EDDY, S., KROGH, A. and MITCHISON, G. (1998): 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press, New York

[12] BAUM L. E. (1972): 'An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes' Inequalities, 3, pp. 1-8

[13] BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T.N., WEISSIG H., SHINDYALOV I. N. and BOURNE P. E. (2000): 'The Protein Data Bank', Nucleic Acids Res., 28, pp. 235-42

[14] KABSCH W. and SANDER C. (1983): 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', Biopolymers, 22, pp. 2577-637

[15] MURZIN A. G., BRENNER S. E., HUBBARD T. and CHOTHIA C. (1995): 'SCOP: a structural classification of proteins database for the investigation of sequences and structures', J. Mol. Biol., 247, pp. 536-40

[16] PAYNE, S. (2001): 'Classification of Protein Sequences into Homogenous Families', Master thesis in Software Engineering, University of Frankfurt, Germany

[17] DI FRANCESKO V., GARNIER J. and MUNSON P.J. (1997): 'Protein topology recognition from secondary structure sequences: applications of the Hidden Markov Models to the alpha class proteins', J. Mol. Biol., 267, pp. 446-63