

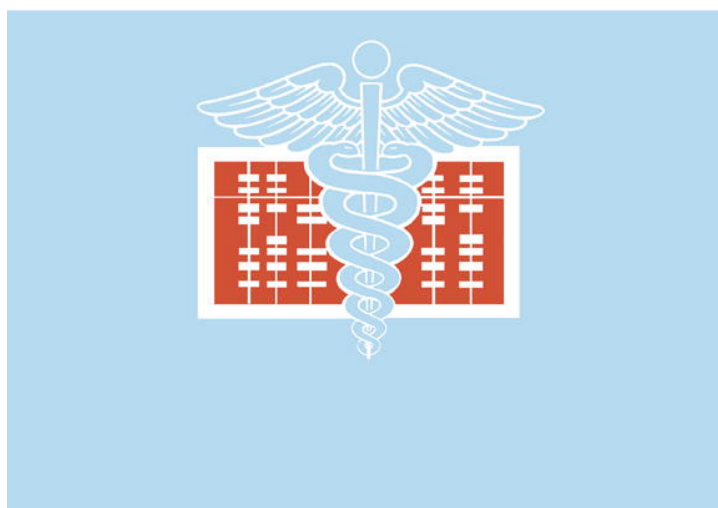
Provided for non-commercial research and educational use only.  
Not for reproduction or distribution or commercial use.



Volume 37 Issue 9 September 2007 ISSN 0010-4825

# Computers in Biology and Medicine

An International Journal



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



ELSEVIER

Computers in Biology and Medicine 37 (2007) 1211–1224

Computers in Biology  
and Medicine

www.intl.elsevierhealth.com/journals/cobm

## Sequence-based protein structure prediction using a reduced state-space hidden Markov model

Christos Lampros<sup>a,b</sup>, Costas Papaloukas<sup>a,c</sup>, Themis P. Exarchos<sup>a,b</sup>, Yorgos Goletsis<sup>a,d</sup>,  
Dimitrios I. Fotiadis<sup>a,e,\*</sup>

<sup>a</sup>Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece

<sup>b</sup>Department of Medical Physics, Medical School, University of Ioannina, GR 45110 Ioannina, Greece

<sup>c</sup>Department of Biological Applications and Technology, University of Ioannina, GR 45110 Ioannina, Greece

<sup>d</sup>Department of Economics, University of Ioannina, GR 45110 Ioannina, Greece

<sup>e</sup>Biomedical Research Institute—FORTH, GR 45110 Ioannina, Greece

Received 24 May 2006; received in revised form 24 October 2006; accepted 30 October 2006

### Abstract

This work describes the use of a hidden Markov model (HMM), with a reduced number of states, which simultaneously learns amino acid sequence and secondary structure for proteins of known three-dimensional structure and it is used for two tasks: protein class prediction and fold recognition. The Protein Data Bank and the annotation of the SCOP database are used for training and evaluation of the proposed HMM for a number of protein classes and folds. Results demonstrate that the reduced state-space HMM performs equivalently, or even better in some cases, on classifying proteins than a HMM trained with the amino acid sequence. The major advantage of the proposed approach is that a small number of states is employed and the training algorithm is of low complexity and thus relatively fast.

© 2006 Elsevier Ltd. All rights reserved.

**Keywords:** Structure prediction; Fold recognition; Hidden Markov models; Protein classification

### 1. Introduction

The number of identified protein sequences has been increased in the last years due to the extensive research in the field. However, the majority of these sequences is not accompanied by any information about their function. One way to understand their function is to link them with known proteins in annotated databases, whose three-dimensional structure (fold) is known. Computer methods for protein analysis address this problem since they study the relations within the amino acids sequence or structure. Since proteins have structural features which define functional similarities, the need for structure

estimation methods is high. Such methods are sequence-based or protein attribute-based. The availability of protein attributes' data is lower than the sequence information, either primary or secondary, so we focus on exploiting sequence data for structure prediction.

Protein computational analysis aims in structure estimation and includes two protein classification tasks: fold recognition and class prediction. Several methods have been proposed in the literature for protein classification. Genetic algorithms (GAs) have been applied for fold recognition [1] in order to exploit sequence and secondary structure information. Artificial neural networks (ANNs) have been used to extract patterns from databases of known structures in fold recognition problems [2]. Lately, ANNs assisted in improving the quality of alignments between sequences and identified structures, which is a crucial step for fold recognition [3]. Support vector machine (SVM) have been used for multi-class protein fold recognition [4]. SVMs have also been used for predicting the three-dimensional

\* Corresponding author. Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, P.O. Box 1186, GR 451 10 Ioannina, Greece. Tel.: +30 26510 98803; fax: +30 26510 98889.

E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

structure of proteins [5]. In this case SVMs were used to transform primary protein sequences into fixed-length feature vectors and then predict their tertiary structure. Moreover, SVMs have been employed in order to improve the accuracy of profile to profile alignment for fold recognition [6].

The prediction of the structural class of a protein is addressed through several methods based mainly on the amino acid composition of the protein. It is well known that the knowledge of protein structural class assists in the determination of the three-dimensional structure of a protein. If the structural class of a protein is known, it can be used to considerably reduce the search space of structure prediction processes, since most of the structure alternatives could be eliminated and therefore the structure prediction task is simplified and the whole process is accelerated [7]. Various statistical methods have been proposed to deal with this problem [8,9].

A review of sequence-based approaches reveals that hidden Markov models (HMMs) are those most commonly used and also demonstrate high performance. HMMs have been applied for multi-class protein fold recognition [10] employing the sequence alignment and modelling (SAM) software [11]. Furthermore, secondary structure information can be incorporated in the HMM and increase the fold recognition performance [12]. Karchin et al. [13] used the same approach and additionally they have evaluated different alphabets for backbone geometry and their effect on the classification performance. However, the main drawback of HMMs is the employment of large model architectures which require large data sets and high computational effort for training. As a consequence, in cases where these data sets are not available, e.g. small classes or folds, their performance deteriorates.

In this work, a novel classification tool for computational protein analysis is proposed. It is based on a HMM with a reduced state-space topology. The model employs an efficient architecture with a small number of states and a low complexity training algorithm. Secondary structure information is introduced to the model to increase its performance and it is used in such a way that allows the use of the low complexity algorithm. The number of states is equal to the number of the different possible formations of secondary structure. The model is trained using the low complexity likelihood maximization algorithm for each candidate class and fold. The problem addressed is the multi-class classification of sequences, so the method employed should classify a query sequence of unknown structural category in one of the candidate categories. The proposed model is evaluated in two different tasks, i.e. class prediction and fold recognition, and in two different data sets, a high homology data set and a low homology data set. For class prediction a Bayesian multi-class classification approach is used while for fold recognition a two-stage classifier is adopted (see also Fig. 2). The obtained results are equivalent or even better than other similar approaches in the high homology data set. However, this does not happen for the low homology data set, since the performance is decreased. The major advantage of the proposed approach is that the computational load of the model is significantly smaller than conventional methods based on full HMM.

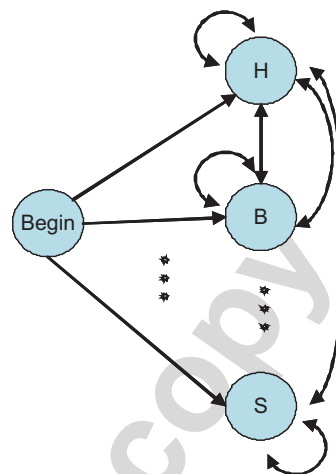


Fig. 1. Topology of the reduced state-space HMM (H, B, ..., S are the letters of DSSP secondary structure alphabet).

In the following paragraphs, the methodology is presented and the training and scoring procedures of the proposed model are explained. The employed data sets are described next, as well as, the implementation of the overall classification process. In the results section the performance of the reduced HMM is demonstrated for the class and fold classification tasks. Finally, the advantages and disadvantages of the proposed method are discussed.

## 2. Methods

HMMs are a highly effective means of modelling a family of unaligned sequences. The trained HMM can then be used for discrimination. The HMMs that have been used for sequence modelling until now consist of a set of positions that correspond to columns in a multiple alignment. In this work a HMM with a smaller topology containing a limited number of states is adopted.

The reduced state-space HMM that is introduced adopts the learning approach used in all protein classification problems, which is to train a model using a set of known sequences (training set). The reduced HMM is used in modelling families of biological sequences and its utilization is an efficient method to implement unsupervised learning in protein sequence data. According to that implementation, the training set sequences are considered to be produced by the model. The aim of the learning procedure is to maximize the likelihood of the model given the training data. This likelihood is maximized when the likelihood of the training data, given the model, is maximized. The likelihood of the training data is the product of all likelihoods of the training sequences. So the probability parameters of the model which are calculated must satisfy the above criterion.

The reduced state-space topology uses the mathematical framework of a typical HMM. It models a series of observations based upon a hypothesized (hidden) process. The model consists of a set of states  $S$  and a set of possible transitions  $T$  between them (Fig. 1). Every state emits a signal based upon a set of emission probabilities and then stochastically transmits

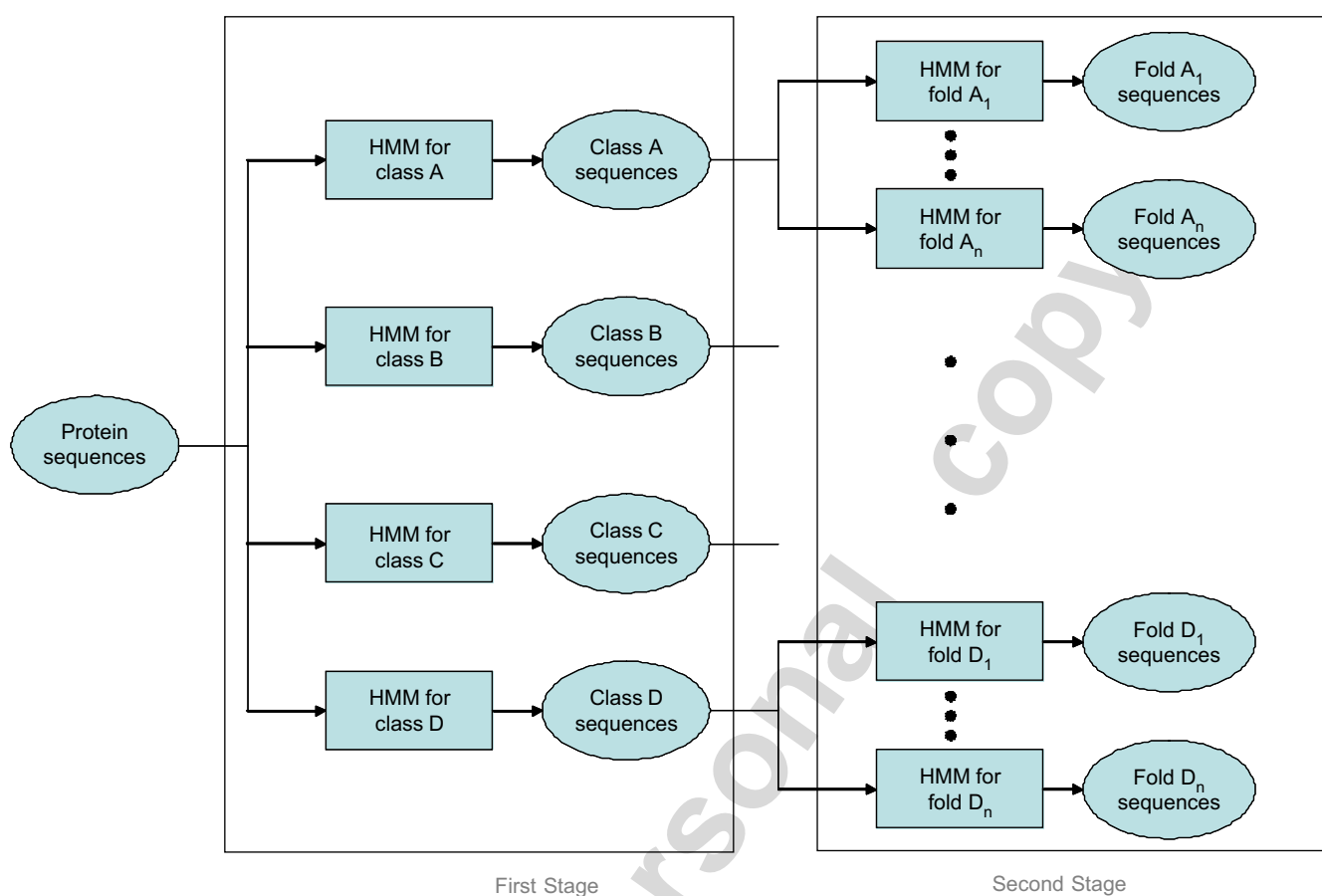


Fig. 2. The two-stage HMM classifier. In the first stage the appropriate class is identified while in the second stage the correct fold is determined.

it to some other state with a probability depending on the previous state. This procedure is repeated until all the observed signals are emitted. There is a starting state where the process starts and transition probabilities also exist from the starting to each possible state. The sum of these probabilities is unity and so is the sum of emission probabilities of possible signals in each state and the sum of transition probabilities from each state to each next state possible. The observer does not know which is the exact state that produces each signal, because the state is hidden. This is a feature that differentiates the model from other stochastic models that do not belong to the HMM category.

Another such characteristic is the Markov property, that is, 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  $t-1$  [14]. The particular feature of the reduced state-space HMM, which differentiates it from other HMMs, is that it uses the secondary structure information in such a way that the states of the model will depict the possible different secondary states. The correct classification among different structural groups demands the use of the secondary structure information and not only that of the primary structure. We use secondary structure sequences which are employed in the context of our HMM as hidden state sequences. This offers the advantage of employing a HMM with a small number of states, equal to the number of

the different letters in the definition of secondary structure of proteins (DSSP) alphabet [15], representing the possible secondary structure formations where each amino acid residue is found. It should be noted that 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 and that fact allows us to use a low complexity training algorithm based on likelihood maximization [14]. So we can avoid complicated iterated learning procedures like the Baum–Welch algorithm [16], which is commonly used in other approaches.

### 2.1. Model training

There are seven different hidden states in the model corresponding to the underlying secondary structure. In the DSSP approach an eighth letter is also determined, which indicates unknown structure. However, we do not use an eighth state in our method, so the amino acid residues with unknown structure are skipped during the modelling process.<sup>1</sup> The states of the model are fully connected, that is all possible transitions between them are allowed. The topology of the model is shown in Fig. 1. In the training set, there is one to one correspondence

<sup>1</sup> The introduction of an eighth state provided poorer classification results during the validation phase and was therefore omitted.

between the amino acid and the secondary structure residues, thus for each state the distribution over all possible amino acid residues is estimated. 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. The total number of the model parameters is  $7 \times 21$  for the possible emissions,  $7 \times 7$  for the possible transitions between states and  $1 \times 7$  for the transitions from the beginning, yielding a total of 203 parameters.

As already mentioned, the likelihood maximization algorithm was employed to train the reduced state–space HMM. Following this training procedure, the emission and transition parameters are calculated in one step with the use of the maximum likelihood estimators. If  $a_{kl}$  is the transition probability from state  $k$  to another state  $l$  (Fig. 1) and  $e_k(b)$  the emission probability of the residue  $b$  in the state  $k$ , then the estimators are given by the following equations, which are the estimation equations in the HMMs when the state sequences are known [14]:

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

and

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}, \quad (2)$$

where  $A_{kl}$  is the number of times that the transition from  $k$  to  $l$  is observed and  $E_k(b)$  is the number of times the emission of  $b$  from  $k$  is used in the training test of sequences. Whenever there is a state  $k$  never been used in the set of example sequences, then the estimation equations are undefined for that state (numerator and denominator will have zero values). In order to avoid such problems it is preferable to add predetermined pseudocounts to the  $A_{kl}$  and  $E_k(b)$  before using Eqs. (1) and (2), thus we have

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

$$E_k(b) = (\text{number of emissions of } b \text{ from } k) + r_k(b). \quad (4)$$

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  $r_{kl} = r_k(b) = 1$ . This implies that the prior distribution of amino acids in each emission state and the prior distribution of transitions from each state is considered to be the uniform distribution. Thus, the use of pseudocounts is confined here only to avoid overfitting and does not include the incorporation of some specific prior knowledge.

## 2.2. Model scoring

We deal with a multi-class classification problem, and the method used should classify a sequence of unknown structural category in one category among many others. The decision will be based on the Bayes theorem, which claims that the probability of a particular model given a sequence is proportional to the probability of the sequence, given the model. The latter is the likelihood of the sequence and can be calculated since the

parameters of the model are known after the learning process. In order to evaluate the performance of the reduced state–space HMM the posterior probability scores are used. These scores are logarithmic forms of the probability of the sequence, given the model. According to the Bayes theorem, the 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.

The data which is used in the experiments are separated in training and test sets. The test set contains only primary sequences, as all information concerning the structure of a protein is considered unknown. Log-likelihood scores are adopted for evaluating the reduced state–space 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. In addition, 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. Consequently, the log-likelihood score of a sequence against the null model is given as

$$\text{score}(x_p) = \log_z \frac{P_m(x_p)}{P_{\emptyset}(x_p)}, \quad (5)$$

where the quantity  $P_m(x_p)$  represents the probability that a sequence  $x_p$  has been produced by model  $m$  and  $P_{\emptyset}(x_p)$  the probability that the same sequence has been produced by the *null* model. This kind of scoring has been selected so that the effect of differences in length among sequences will be reduced. The criterion for selecting the model, which best classifies a particular protein, is to choose the model with the highest posterior probability for a given part of evidence. Thus, the model selected as the best classification for the protein  $x_p$  would be model  $m_i$  such that  $P_{m_i}(x_p) > P_{m_j}(x_p)$  for all  $i \neq j$ , or equivalently  $\text{score}_{m_i}(x_p) > \text{score}_{m_j}(x_p)$ . The posterior probabilities correspond to the log-likelihood scores against the *null* model. The posterior probabilities are calculated with the use of the forward algorithm [14], which estimates 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 product of probabilities has to be computed. Finally, a Bayesian classification table is constructed for each aspect of the classification problem, fold recognition and class prediction.

## 3. Data set

In order to validate the proposed classifier, an appropriate group of protein sequences, both primary and secondary, have been selected from the protein data bank [17]. The members of this group should correspond to a specific class or fold of the structural classification of proteins (SCOP) database [18], depending on the type of the problem. Then these sequences form the training set that is used in the classification problem. The test set used for each case consists of a group of primary

Table 1

The employed high and low homology data sets. The high homology data set includes data from four SCOP classes and 29 SCOP folds

Fold	Index	High homology data set		Low homology data set	
		Number of sequences in the training set	Number of sequences in the test set	Number of sequences in the training set	Number of sequences in the test set
<i>All alpha proteins</i>		287	282	260	131
Globin-like	a1	48	47	21	11
Cytochrome c	a3	33	32	20	10
DNA-binding 3-helical bundle	a4	100	99	103	52
Four-helical up-and-down bundle	a24	29	28	28	15
EF-hand	a39	46	46	31	15
SAM domain-like	a60	–	–	25	12
Alpha-alpha superelix	a118	31	30	32	16
<i>All beta proteins</i>		743	738	406	203
Immunoglobulin-like beta sandwich	b1	445	444	132	66
Common fold of diphtheria	b2	–	–	20	10
Cupredoxins	b6	33	33	–	–
Galactose-binding domain-like	b18	–	–	21	10
ConA-like lectins/glucanases	b29	41	40	24	12
SH3-like barrel	b34	46	46	44	22
OB-fold	b40	62	61	61	31
Trypsin-like serine proteases	b47	41	40	25	12
PH domain-like	b55	–	–	24	12
Lipocalins	b60	27	26	–	–
Double-stranded beta-helix	b82	–	–	28	14
Nucleoplasmin-like	b121	48	48	27	14
<i>Alpha and beta proteins(a/b)</i>		628	625	658	329
(TIM)-barrel	c1	161	160	143	71
NAD(P)-binding Rossmann fold	c2	93	93	91	46
FAD/NAD(P)-binding domain	c3	–	–	22	11
Flavodoxin-like	c23	48	48	58	29
Adenine nucleotide	c26	25	25	35	17
P-loop containing nucleotide	c37	105	104	91	46
Thioredoxin-like	c47	54	54	39	20
Ribonuclease H-like motif	c55	43	43	31	15
Phosphorylase/hydrolase-like	c56	–	–	20	10
S-adenosyl-L-methionine	c66	29	29	40	20
PLP-dependent transferases	c67	30	29	31	15
Hydrolases	c69	40	40	34	17
Periplasmic binding protein-like II	c94	–	–	23	12
<i>Alpha and beta proteins(a + b)</i>		254	251	189	95
b-grasp	d15	61	60	44	22
Cystatin-like	d17	–	–	20	10
MHC antigen-recognition domain	d19	31	30	–	–
Ferredoxin-like	d58	104	104	102	51
Protein kinase-like (PK-like)	d144	28	27	23	12
C-type lectin-like	d169	30	30	–	–

The sequences come from the ASTRAL95 SCOP 1.67 data set, where no sequences with similarity up to 95% with each other are contained. The folds that contain at least 50 members are used. The low homology data set includes data from four SCOP classes and 34 SCOP folds. The sequences come from the ASTRAL40 SCOP 1.69 data set, where no sequences with similarity up to 40% with each other are contained. The folds that contain at least 30 members are used.

sequences whose class or fold is considered unknown, but is actually known from the SCOP database. Thus, we are able to evaluate the effectiveness of their classification in the correct group after the experiments.

The PDB sequence files include no structural categorization of their data, and this kind of information must be provided

by the SCOP database. There is a hierarchical categorization of proteins with known structure in the SCOP database, where class is the upper level and folds, superfamilies and families follow. Here we deal with the problem of classifying a sequence in one of the categories of the two higher levels, that is in a class or in a fold. This implementation field is chosen because among

members of a superfamily or a family there is a quite high similarity in their primary sequence residues, so the existing methods of protein classification can deal with this aspect of the problem adequately [10].

We used two different data sets, a high homology data set and a low homology data set. The proteins with higher homology come from the ASTRAL95 SCOP 1.67 data set, where no proteins with more than 95% similarity are contained. On the other hand, the proteins with lower homology come from the ASTRAL40 1.69 data set, where only proteins with less than 40% similarity are included. The primary and secondary structure information was extracted from the PDB database. Both databases (PDB and SCOP) contain similar identifiers for the protein sequence information and the data from both databases are combined, so that SCOP will provide the identity and the category of the protein and PDB its primary and secondary content. This is realized for each class and fold to be tested. The data sets used in the current study are shown in Table 1. In the high homology data set, the four most populated classes are used for the class prediction experiment. Moreover, the 29 most populated SCOP folds, and specifically those with at least 50 members, are used to derive the training and test data for the fold recognition experiment. In the low homology data set, the 34 most populated folds (those with at least 30 proteins in this case) of the four major classes are used for the fold recognition experiment.

#### 4. Results

The reduced HMM is compared against SAM which is considered as the most effective method that employs HMMs for protein classification [19]. The same training and test sets were used for both methods.

In the class prediction task, four reduced HMMs are trained which correspond to the relevant SCOP classes. Then the members of the class test sets of the high homology data set are scored against them and the prediction accuracy for each class is determined. The prediction accuracy is the number of test proteins uniquely recognized as belonging to a specific class, divided by the total number of test proteins belonging to that class.

The fold recognition task includes the training of 29 reduced HMMs for each one of the 29 most populated SCOP folds of the high homology data set. A two-stage classifier is adopted here (Fig. 2). In the beginning, the initial test sets for fold recognition are filtered through the class models. The final test sets consist of those test sequences that were classified in the correct class when compared with the four class models. So the class prediction task acts as the first stage of the fold recognition task. In the second stage, the test sequences are scored against all the models of the relevant 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 number of proteins belonging to the initial test of each fold and not to the number of those that have remained after the filtering of the first stage (class prediction). In a similar manner, the SAM performance was estimated.

The posterior probabilities in the case of the SAM models, which are compared with our models, correspond to the negative log-likelihood scores of each sequence. The only difference is that when the negative log-likelihood is decreasing, the posterior probabilities are increasing, so an unknown protein should be assigned to that model which gives the lowest negative log-likelihood score for the specific primary sequence. Relying on the scores themselves to decide which SAM model provides with the best classification is not a reasonable choice because the scores assigned by different models are not comparable as they depend on the model length. The model length changes when a different training set is used, unlike when using the reduced HMM. Di Francesco et al. [20] showed that the decision in that case is based upon ranked scores, which means that the higher the rank assigned by a model to a query sequence, the more we believe that the model produced that sequence. In the output files containing the scores given by SAM for a test set against a model these scores are already ranked. So each test sequence will be assigned to that SAM model which gives the highest rank at that sequence among all others in the same test set.

In the low homology data set, two experiments take place. In the first experiment, 34 reduced HMMs are trained for each one of the 34 most populated folds. Here only one stage was adopted, because results in class prediction get worse. The test sequences are scored against all folds of every 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 number of test proteins belonging to that fold. In the second experiment, 1513 reduced HMMs are trained for each one of the training sequences of all folds. In that case the prediction accuracy is calculated in the same way, but the test sequences are first scored against the models of all proteins of every fold. Then each protein is classified to the fold whose member is the protein the model of which gave the maximum probability score among all 1513 models of sequences used for training.

The results of the applications in the high homology data set are shown in Tables 2–4. Specifically Table 2 demonstrates the performance of the two-stage reduced HMM classifier compared to that of SAM for class classification while Table 3 for fold recognition. Table 4 shows the results of the second stage only, so, in this case, the denominator of the prediction accuracy corresponds to the number of those proteins that have remained after the filtering procedure occurred in the first stage. Consequently, the numerical results presented in Table 3 are the aggregate results from Tables 2 and 4. The results of the experiments in the low homology data set are shown in Tables 5 and 6. Table 5 shows the performance of the reduced HMM classifier when different models were trained for each fold and Table 6 its performance when different models were trained for each sequence of the training set of every fold.

In Tables 7 and 8, the classification results of our method for each fold separately in the low homology data set are presented in terms of Top 1–Top 5 sensitivity [21] for both experiments. Top 1–Top 5 sensitivity is computed by considering a classification as correct even if the actual (true) fold receives a score

Table 2  
Comparison in terms of prediction accuracy of the proposed model with SAM for the four SCOP classes (high homology data set)

Class index	Reduced HMM prediction accuracy		SAM prediction accuracy	
A	200/282	70.9%	182/282	64.5%
B	475/738	64.4%	438/738	59.4%
C	327/625	52.3%	340/625	54.4%
D	94/251	37.5%	126/251	50.2%
Overall	1096/1896	57.8%	1086/1896	57.3%

Table 3  
Comparison of the proposed model with SAM for the 29 SCOP folds (overall classification in the high homology data set)

Fold index	Reduced HMM prediction accuracy		SAM prediction accuracy	
a1	32/47	68.1%	30/47	63.8%
a3	17/32	53.1%	23/32	71.9%
a4	39/99	39.4%	30/99	30.3%
a24	10/28	35.7%	8/28	28.6%
a39	33/46	71.7%	33/46	71.7%
a118	8/30	26.7%	7/30	23.3%
Overall class A	139/282	49.3%	131/282	46.5%
b1	301/444	67.8%	79/444	17.8%
b6	5/33	15.2%	23/33	69.7%
b29	23/40	57.5%	17/40	42.5%
b34	1/46	2.2%	17/46	37%
b40	0/61	0%	5/61	8.2%
b47	10/40	25%	19/40	47.5%
b60	3/26	11.5%	6/26	23.1%
b121	38/48	79.2%	14/48	29.2%
Overall class B	381/738	51.6%	180/738	24.4%
c1	10/160	6.3%	24/160	15%
c2	42/93	45.2%	24/93	25.8%
c23	1/48	2.1%	3/48	6.3%
c26	2/25	8%	4/25	16%
c37	8/104	7.7%	7/104	6.7%
c47	5/54	9.3%	10/54	18.5%
c55	6/43	14%	8/43	18.6%
c66	1/29	3.4%	2/29	6.7%
c67	8/29	27.6%	9/29	31%
c69	15/40	37.5%	12/40	30%
Overall class C	98/625	15.7%	103/625	16.5%
d15	12/60	20%	25/60	41.7%
d19	19/30	63.3%	22/30	73.3%
d58	8/104	7.7%	20/104	19.2%
d144	14/27	51.2%	7/27	25.9%
d169	8/30	26.7%	22/30	73.3%
Overall class D	61/251	24.3%	96/251	38.2%
Overall	679/1896	35.8%	510/1896	26.9%

between the first and fifth highest ones. For example, the Top 5 sensitivity provides the five most probable folds that the unknown protein belongs to. In our case this sensitivity reached up to 47.6% when different models are trained for each one of the folds and 47.9% when different models are trained for every sequence in the training set. We have not calculated the Top 1–Top 5 sensitivity in the high homology data set due to the use of the two-stage classifier.

Finally, receiver operating characteristic (ROC) analysis [22] was performed for our method and for all experiments for high

and low homology data set. The analysis followed the class reference formulation, where each class is considered separately against all others. The ROC curves with the corresponding areas under curves (AUC) for all folds are shown in Fig. 3 for the first experiment in low homology data set. The multi-class AUC is calculated by using the following formula [23]:

$$AUC_{\text{total}} = \sum_{c_i \in C} AUC(c_i) p(c_i), \quad (6)$$

Table 4  
Comparison of the proposed model with SAM for the 29 SCOP folds (second stage only, in the high homology data set)

Fold index	Reduced HMM prediction accuracy		SAM prediction accuracy	
a1	32/33	97%	30/38	78.9%
a3	17/20	85%	23/26	88.5%
a4	39/69	56.5%	30/53	56.6%
a24	10/18	55.6%	8/13	61.5%
a39	33/40	82.5%	33/39	84.6%
a118	8/20	40%	7/13	53.8%
Overall class A	139/200	69.5%	131/182	72%
b1	301/363	82.9%	79/309	25.6%
b6	5/11	45.5%	23/30	76.7%
b29	23/28	82.1%	17/19	89.5%
b34	1/7	14.3%	17/23	73.9%
b40	0/5	0%	5/12	41.7%
b47	10/15	66.7%	19/21	90.5%
b60	3/6	50%	6/6	100%
b121	38/40	95%	14/18	77.8%
Overall class B	381/475	80.2%	180/438	41.1%
c1	10/82	12.2%	24/80	30%
c2	42/72	58.3%	24/48	50%
c23	1/25	4%	3/25	12%
c26	2/13	15.4%	4/15	26.7%
c37	8/50	16%	7/68	10.3%
c47	5/16	31.3%	10/40	25%
c55	6/20	30%	8/20	40%
c66	1/11	9.1%	2/13	15.4%
c67	8/16	50%	9/9	100%
c69	15/22	68.2%	12/22	54.5%
Overall class C	98/327	30%	103/340	30.3%
d15	12/20	60%	25/30	83.3%
d19	19/23	82.6%	22/22	100%
d58	8/26	30.8%	20/41	48.8%
d144	14/14	100%	7/7	100%
d169	8/11	72.7%	22/26	84.6%
Overall class D	61/94	64.9%	96/126	76.2%
Overall	679/1096	62%	510/1086	47%

where  $AUC(c_i)$  is the area under the class reference ROC curve for the class (fold)  $c_i$ , ( $i = 1, 2, \dots, 34$ ),  $C$  is the number of classes (folds) and  $p(c_i)$  is the prevalence of class  $c_i$  in the data. Using Eq. (6) the  $AUC_{total}$  is equal to 0.80 for the first stage (class prediction) and 0.96 for the second stage (fold recognition) in the high homology data set. For the low homology data set, the  $AUC_{total}$  is equal to 0.76 for the first experiment and to 0.73 for the second experiment. All these values indicate an effective classifier.

## 5. Discussion

We developed a model that employs a HMM with a reduced state–space architecture for protein classification. The implementation of our model is based on the concept of its simultaneous training using both primary and secondary sequence for class and fold modelling. 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. For each state a probability distribution over all possible amino acids is estimated. The model employs a fast learning algorithm based on the calculation of the maximum likelihood estimators of all parameters in a single step. After training, the probability score of unknown sequences against the created models is calculated with the use of the forward algorithm while 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.

In the case of the high homology data set, the classification takes place in two stages. In the first stage the test sequences are classified in the appropriate class and the results are validated. In the second stage the correctly assigned sequences of the previous step are classified in the appropriate fold of their corresponding class.

In the case of the low homology data set, there is only one step and the test sequences are assigned in the appropriate fold

Table 5

Comparison of the proposed model with SAM for the 34 SCOP folds (overall classification in the low homology data set). Different models were trained for each fold

Fold index	Reduced HMM prediction accuracy		SAM prediction accuracy	
a1	5/11	45.5%	9/11	81.8%
a3	3/10	30%	6/10	60%
a4	5/52	9.6%	1/52	1.9%
a24	1/15	6.7%	2/15	13.3%
a39	6/15	40%	11/15	73.3%
a60	5/12	41.7%	2/12	16.7%
a118	6/16	37.5%	0/16	0%
Overall class A	31/131	23.7%	31/131	23.7%
b1	27/66	40.9%	22/66	33.3%
b2	0/10	0%	1/10	10%
b18	2/10	20%	3/10	30%
b29	2/12	16.7%	3/12	25%
b34	4/22	18.2%	8/22	36.4%
b40	4/31	12.9%	1/31	3.2%
b47	4/12	33.3%	7/12	58.3%
b55	5/12	41.7%	4/12	33.3%
b82	1/14	7.1%	1/14	7.1%
b121	12/14	85.7%	0/14	0%
Overall class B	61/203	30.1%	50/203	24.6%
c1	1/71	1.4%	7/71	9.9%
c2	10/46	21.7%	8/46	17.4%
c3	2/11	18.2%	9/11	81.8%
c23	0/29	0%	8/29	27.6%
c26	1/17	5.9%	3/17	17.6%
c37	1/46	2.2%	21/46	45.7%
c47	2/20	10%	6/20	30%
c55	4/15	26.7%	2/15	13.3%
c56	0/10	0%	1/10	10%
c66	0/20	0%	4/20	20%
c67	1/15	6.7%	10/15	66.6%
c69	7/17	41.2%	2/17	11.8%
c94	5/12	41.7%	4/12	33.3%
Overall class C	34/329	10.3%	86/329	26.1%
d15	4/22	18.2%	0/22	0%
d17	1/10	10%	0/10	0%
d58	0/51	0%	2/51	3.9%
d144	5/12	41.7%	9/12	75%
Overall class D	10/95	10.5%	11/95	11.6%
Overall	136/758	17.9%	178/758	23.5%

after they have been scored against all candidate models. There are two experiments in the low homology data set: in the first one different models are trained for every candidate fold and in the second one different models are trained for every protein in each candidate fold.

The classification performance of the reduced HMM model is tested by comparing it to a SAM model trained with the same data sets. The SAM model is linear and its length is equal to that of the multiple alignment which the SAM method gives for the specific set. In the high homology data set, the comparison shows that for the class prediction problem the reduced HMM outperforms SAM in classes *A* and *B* and performs worse in classes *C* and *D* (Table 2). As far as the fold recognition problem is concerned, the reduced HMM approach is

again more efficient in classifying correctly those test proteins whose folds are members of the classes *A* and *B*. Thus, for the folds belonging to classes *A* and *B* the reduced HMM is more accurate than SAM while the opposite happens for the folds belonging to classes *C* and *D*. The total precision though of our classification model based on the reduced HMM is better than that of the model based on SAM (Table 3). The same conclusion is drawn from the comparison of the performance of both methods in the second stage, where the reduced HMM is able to classify more sequences correctly in folds among those already correctly assigned to classes (Table 4). On the other hand, in the low homology data set, the reduced HMM performs equivalently with SAM in the folds of classes *A* and *D*, better in the folds of class *B*, but it performs worse in the

Table 6  
Comparison of the proposed model with SAM for the 34 SCOP folds (overall classification in the low homology data set) when different models were trained for each sequence in the training set

Fold index	Reduced HMM prediction accuracy		SAM prediction accuracy	
a1	2/11	18.2%	5/11	45.5%
a3	0/10	0%	1/10	10%
a4	7/52	13.5%	13/52	25%
a24	1/15	6.7%	0/15	0%
a39	3/15	20%	1/15	6.7%
a60	1/12	8.3%	0/12	0%
a118	4/16	25%	0/16	0%
Overall class A	18/131	13.7%	20/131	15.3%
b1	18/66	27.3%	14/66	21.2%
b2	3/10	30%	0/10	0%
b18	2/10	20%	0/10	0%
b29	3/12	25%	1/12	8.3%
b34	0/22	0%	0/22	0%
b40	4/31	12.9%	3/31	9.7%
b47	1/12	8.3%	1/12	8.3%
b55	0/12	0%	1/12	8.3%
b82	0/14	0%	0/14	0%
b121	2/14	14.3%	1/14	7.1%
Overall class B	33/203	16.3%	21/203	10.3%
c1	21/71	29.6%	11/71	15.5%
c2	9/46	19.6%	5/46	10.9%
c3	1/11	9.1%	0/11	0%
c23	4/29	13.8%	5/29	17.2%
c26	4/17	23.5%	0/17	0%
c37	10/46	21.7%	9/46	19.6%
c47	2/20	10%	2/20	10%
c55	0/15	0%	0/15	0%
c56	0/10	0%	1/10	10%
c66	1/20	5%	2/20	10%
c67	2/15	13.3%	1/15	6.7%
c69	4/17	23.5%	0/17	0%
c94	3/12	25%	0/12	0%
Overall class C	61/329	18.5%	36/329	10.9%
d15	1/22	4.6%	3/22	13.6%
d17	1/10	10%	1/10	10%
d58	8/51	15.7%	8/51	15.7%
d144	0/12	0%	1/12	8.3%
Overall class D	10/95	10.5%	13/95	13.7%
Overall	122/758	16.1%	90/758	11.9%

folds of class C, which is the most populated one. In terms of total precision the reduced HMM performs worse than SAM in the first experiment (Table 5). But when different models are trained for every sequence of the training set, the reduced HMM performs better than SAM in terms of overall accuracy, even in the low homology data set (Table 6). Moreover, the ROC analysis that is applied in all cases for our method gives  $AUC_{total}$  values 0.8 for class prediction in high homology data set, 0.96 for fold recognition in high homology data set and more than 0.7 for both experiments in low homology data set, which indicate good classification competence. Finally, the Top 1–Top 5 sensitivity results (Tables 7 and 8) indicate good performance for both experiments in the low homology data set.

The results of other methods, apart from SAM, that use HMM in protein classification, are not directly comparable with the proposed reduced HMM. They either use different data sets [10,12,13] or the application field is different and refers to the superfamily level of SCOP structural categorization [24]. Moreover, in all those cases it is necessary to include the corresponding secondary sequences in the test set, which is not the case in our model. In other methods that do not employ HMMs, only [4] refers to classification in SCOP folds and uses a different data set coming from an older SCOP version.

The main advantage of the reduced HMM implementation for classifying proteins in the appropriate class or fold is the avoidance of iterative procedures that demand huge computational effort in training. A multi-class approach for structural

Table 7

Classification results of the proposed method in terms of Top 1–Top 5 accuracy for every fold when different models were trained for each fold (low homology data set)

Fold index	Top 1 (%)	Top 2 (%)	Top 3 (%)	Top 4 (%)	Top 5 (%)
a1	45.5	63.6	72.7	81.8	90.9
a3	30	60	80	80	100
a4	9.6	23.1	28.9	44.2	53.9
a24	6.7	13.3	13.3	13.3	13.3
a39	40	46.7	53.5	66.7	80
a60	41.7	41.7	50	66.7	66.7
a118	37.5	56.3	62.5	68.8	68.8
Overall class A	23.7	36.6	43.5	54.2	61.8
b1	40.9	53	59.1	66.7	69.7
b2	0	0	10	30	50
b18	20	50	60	60	60
b29	16.7	50	66.7	66.7	66.7
b34	18.2	27.3	27.3	36.4	40.9
b40	12.9	25.8	38.7	48.4	48.4
b47	33.3	33.3	41.7	50	50
b55	41.7	58.3	58.3	75	83.3
b82	7.1	14.3	21.4	21.4	21.4
b121	85.7	92.9	92.9	92.9	92.9
Overall class B	30.1	42.4	49.3	56.7	59.6
c1	1.4	2.8	7	19.7	25.4
c2	21.7	34.8	43.5	47.8	50
c3	18.2	36.4	45.5	45.5	45.5
c23	0	0	3.5	20.7	20.7
c26	5.9	23.5	23.5	29.4	35.3
c37	2.2	6.5	13	15.2	21.7
c47	10	40	55	65	65
c55	26.7	40	40	40	53.3
c56	0	20	20	30	40
c66	0	5	10	25	30
c67	6.7	26.7	53.3	60	73.3
c69	41.2	70.6	82.4	82.4	88.2
c94	41.7	66.7	66.7	66.7	66.7
Overall class C	10.3	21.3	28	35.6	40.4
d15	18.2	22.7	27.3	27.3	31.8
d17	10	10	20	30	30
d58	0	3.9	5.9	9.8	19.6
d144	41.7	41.7	50	50	50
Overall class D	10.5	13.7	17.9	21.1	27.4
Overall	17.9	28.6	35.1	42.6	47.6

classification with the use of HMMs was also adopted [24], but there the complexity was quite high, as the model used in [12] was adopted. Moreover, it is the only method, among those using secondary sequence information for fold recognition [12,13,24], where the knowledge of the secondary sequence of the target protein is not needed during the validation process, due to the nature of the model's architecture.

The proposed method provides equivalent or even better results than SAM, which, however, is of higher computational complexity in the training of the model. In the reduced HMM we have to train 203 parameters in total and the calculation for each parameter takes place in just one step. On the other hand, SAM employs a much higher number of parameters which are

calculated iteratively. SAM is currently considered as a very effective method for protein classification based on HMMs. However, this method has some specific limitations, that the reduced HMM overcomes. First, it uses a large number of states, since its topology corresponds to a multiple sequence alignment among sequences and also, the length of the model is similar to the length of the alignment. This leads to a huge number of parameters to be calculated, unlike the reduced HMM, where there are only seven states and there is no need for sequence alignment. Second, SAM employs the Baum–Welch algorithm [16] for training the model, which is an iterative procedure. Baum–Welch algorithm, given  $M$  training sequences of length  $L$  and a model consisting of  $S$  states, has complexity

Table 8  
Classification results of the proposed method in terms of Top 1–Top 5 accuracy for every fold when different models were trained for each sequence in the training set (low homology data set)

Fold index	Top 1 (%)	Top 2 (%)	Top 3 (%)	Top 4 (%)	Top 5 (%)
a1	18.2	18.2	18.2	27.3	27.3
a3	0	10	30	30	40
a4	13.5	21.2	32.7	40.4	46.2
a24	6.7	13.3	13.3	13.3	20
a39	20	26.7	26.7	26.7	40
a60	8.3	8.3	16.7	16.7	25
a118	25	37.5	56.3	62.5	62.5
Overall class A	13.7	20.6	29.8	34.4	40.5
b1	27.3	45.5	53	62.1	68.2
b2	30	30	30	40	40
b18	20	20	40	50	50
b29	25	50	50	50	50
b34	0	0	0	0	4.6
b40	12.9	22.6	25.8	29	45.2
b47	8.3	25	33.3	33.3	33.3
b55	0	8.3	16.7	25	25
b82	0	0	0	7.1	7.1
b121	14.3	28.6	50	64.3	71.4
Overall class B	16.3	27.6	34	40.4	45.8
c1	29.6	49.3	62	71.8	80.3
c2	19.6	39.1	58.7	65.2	76.1
c3	9.1	27.3	36.4	54.6	72.7
c23	13.8	17.2	20.7	24.1	27.6
c26	23.5	35.3	35.3	52.9	52.9
c37	21.7	39.1	43.5	47.8	58.7
c47	10	15	20	20	20
c55	0	6.7	13.3	13.3	26.7
c56	0	0	10	10	10
c66	5	20	40	40	55
c67	13.3	26.7	46.7	46.7	60
c69	23.5	29.4	35.3	47.1	47.1
c94	25	41.7	58.3	66.7	66.7
Overall class C	18.5	32.5	43.2	49.5	57.5
d15	4.6	4.6	4.6	13.6	13.6
d17	10	10	20	20	20
d58	15.7	25.5	29.4	35.3	39.2
d144	0	8.3	16.7	25	25
Overall class D	10.5	16.8	21.1	27.4	29.5
Overall	16.1	27.2	35.6	41.7	47.9

of  $O(MS^2L)$  for each iteration. Moreover, many iterations are required until the algorithm converges. The training algorithm we have employed, which is based on likelihood maximization, has complexity of  $O(1)$ , which is independent of the size of the training set. In addition, Baum–Welch locates a local maximum of the likelihood, while in the reduced HMM there is an easily found global maximum, determined in only one step. Finally, the employment of large models, generated by SAM, is problematic when the size of the training set for some folds is quite small.

On the other hand, the proposed model is not flexible. More specifically, the number of states is always fixed and equal to the number of possible secondary sequence states, so alternate topologies cannot be tested, unless a different secondary struc-

ture alphabet is used. The model is simplified, so every possible future attempt to improve its precision should inevitably be followed by increase of its size and complexity. Moreover, the model provided lower classification results in data sets with low homology which are considered as a more challenging task in fold recognition.

As a possible future use the current model can also be applied to secondary structure prediction tasks. Decoding the state sequence of a protein of unknown structure can reveal a highly probable secondary sequence for that protein. Moreover, additional structural features can be incorporated like residue solvent accessibility, for example. These features will add more states in the model without significant increase in the complexity. The ability of distinguishing folds among each other can be

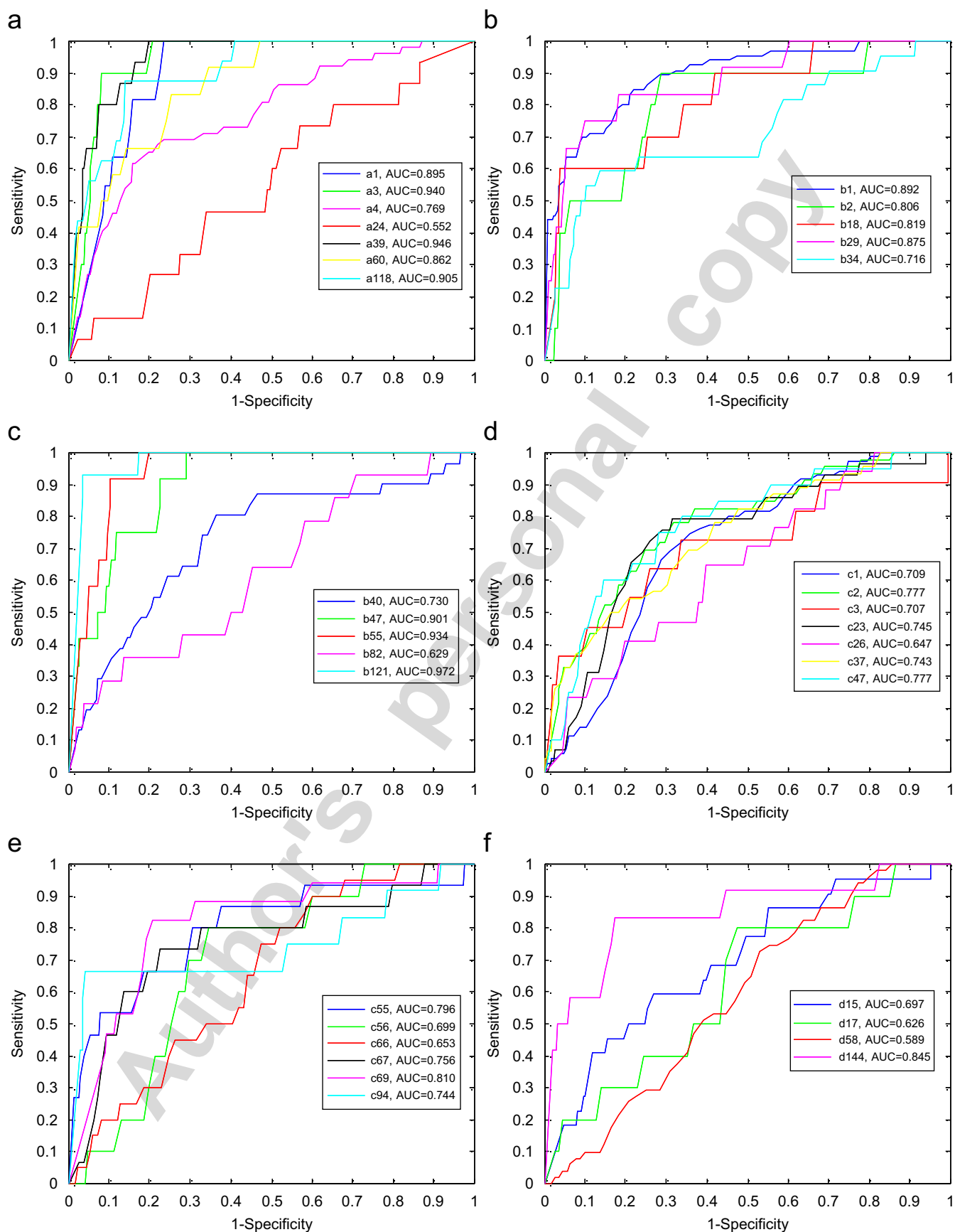


Fig. 3. ROC curves and areas under ROC curves (AUC) for all 34 folds for the experiment in which different models were trained for each fold in the low homology data set. The multiclass AUC was 0.76.

enhanced in that way with a small cost. But the incorporation of a significant amount of information in an efficient way is a difficult task. In that case, a large number of calculations will probably be required while considerable amelioration of the results will remain doubtful. Nevertheless, the reduced state–space HMM implementation is a simplified, fast and efficient way to address the problem of modelling structural categories.

## References

- [1] T. Dandekar, P. Argos, 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 (1996) 645–660.
- [2] B. Rost, PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol.* 266 (1996) 525–539.
- [3] A. Reinhardt, D. Eisenberg, DPANN: improved sequence to structure alignments following fold recognition, *Proteins* 56 (2004) 528–538.
- [4] C. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics* 17 (2001) 349–358.
- [5] R. Karchin, K. Karplus, D. Haussler, Classifying G-protein coupled receptors with support vector machines, *Bioinformatics* 18 (2002) 147–159.
- [6] S. Han, B.C. Lee, S.T. Yu, C.S. Jeong, S. Lee, D. Kim, Fold recognition by combining profile-profile alignment and support vector machine, *Bioinformatics* 21 (2005) 2667–2673.
- [7] Z. Isik, B. Yanikoglu, U. Sezerman, Protein structural class determination using support vector machines. In: Aykanat C, Dayar T, Korpeoglu I (Eds.), *Lecture Notes in Computer Science*, vol. 3280, Computer and Information Sciences, Springer, New York, 2004, pp. 82–89.
- [8] Z.X. Wang, Z. Yuan, How good is the prediction of protein structural class by the component-coupled method?, *Proteins* 38 (2000) 165–175.
- [9] R.Y. Luo, Z.P. Feng, J.K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, *Eur. J. Biochem.* 269 (2002) 4219–4225.
- [10] E. Lindahl, A. Elofsson, Identification of related proteins on family, superfamily and fold level, *J. Mol. Biol.* 295 (2000) 613–625.
- [11] R. Hughey, A. Krogh, Hidden Markov models for sequence analysis: extension and analysis of the basic method, *CABIOS* 12 (2) (1996) 95–107.
- [12] J. Hargbo, A. Elofsson, Hidden Markov models that use predicted secondary structures for fold recognition, *Proteins* 36 (1999) 68–76.
- [13] R. Karchin, M. Cline, Y. Mandel-Gutfreund, K. Karplus, Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry, *Proteins* 51 (2003) 504–514.
- [14] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, New York, 1998.
- [15] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* 22 (1983) 2577–2637.
- [16] L.E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, *Inequalities* 3 (1972) 1–8.
- [17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242.
- [18] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247 (1995) 536–540.
- [19] S. Payne, Classification of Protein Sequences into Homogenous Families. Master Thesis in Software Engineering, University of Frankfurt, Germany, 2001.
- [20] V. Di Francesco, J. Garnier, P.J. Munson, Protein topology recognition from secondary structure sequences: applications of the Hidden Markov Models to the alpha class proteins, *J. Mol. Biol.* 267 (1997) 446–463.
- [21] J. Xu, Fold recognition by predicted alignment accuracy, *IEEE/ACM Trans. on Comput. Biol. Bioinform.* 2 (2) (2005) 157–165.
- [22] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, Reading, MA, 2005.
- [23] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers, Technical Report HPL-2003-4, HP Labs (2003).
- [24] A. Raval, Z. Ghahramani, D.L. Wild, A Bayesian network model for protein fold and remote homologue recognition, *Bioinformatics* 18 (2002) 788–801.

**Christos Lampros** was born in Ioannina, Greece, in 1978. He received his Diploma degree in Electrical and Computer Engineering from the Democritus University of Thrace, in 2001. He is currently working toward the Ph.D. degree in Medical Physics at the University of Ioannina. His research interests include bioinformatics, protein engineering and artificial intelligence.

**Costas Papaloukas** was born in Ioannina, Greece, in 1974. He received his Diploma degree in Computer Science and the Ph.D. degree in biomedical technology from the University of Ioannina, Ioannina, Greece, in 1997 and 2001, respectively. He is a Lecturer of Bioinformatics with the Department of Biological Applications and Technology, University of Ioannina. His research interests include biomedical engineering and bioinformatics.

**Themis P. Exarchos** was born in Ioannina, Greece, in 1980. He received his Diploma degree in Computer Engineering and Informatics from the University of Patras, in 2003. He is currently working toward the Ph.D. degree in Medical Physics at the University of Ioannina. His research interests include medical data mining, decision support systems in healthcare and biomedical applications.

**Yorgos Goletsis** was born in Ioannina, Greece in 1972. He received his Diploma degree in Electrical Engineering and the Ph.D. degree in Electrical and Computer Engineering both from the National Technical University of Athens, Athens, Greece. Since 2006 he is Lecturer in the Department of Economics, University of Ioannina, Ioannina, Greece. His research interests include operational research, decision support systems and evolutionary computation.

**Dimitrios I. Fotiadis** was born in Ioannina, Greece, in 1961. He received his Diploma degree in chemical engineering from National Technical University of Athens, Greece, and the Ph.D. degree in chemical engineering from the University of Minnesota, Twin Cities. Since 1995, he has been with the Department of Computer Science, University of Ioannina, Greece, where he currently is an Associate Professor. He is the director of the Unit of Medical Technology and Intelligent Information Systems. His research interests include biomedical technology, biomechanics, scientific computing, and intelligent information systems.