Computers in Biology and Medicine 39 (2009) 907–914

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Improving the protein fold recognition accuracy of a reduced state-space hidden Markov model

Christos Lampros[a,b], Costas Papaloukas[a,c], Kostas Exarchos[a,b], Dimitrios I. Fotiadis[a,d,*], Dimitrios Tsalikakis[a,e]

[a] Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, PO Box 1186, GR 45110 Ioannina, Greece
[b] Department of Medical Physics, Medical School, University of Ioannina, GR 45110 Ioannina, Greece
[c] Department of Biological Applications and Technology, University of Ioannina, GR 45110 Ioannina, Greece
[d] Biomedical Research Institute - FORTH, GR 45110 Ioannina, Greece
[e] Engineering Informatics and Telecommunications Department, University of Western Macedonia, Kozani, Greece

ARTICLE INFO

Article history:
Received 27 October 2008
Accepted 13 July 2009

Keywords:
Fold recognition
Hidden Markov models
Protein classification
Secondary structure prediction
Secondary structure alphabet

ABSTRACT

Fold recognition is a challenging field strongly associated with protein function determination, which is crucial for biologists and the pharmaceutical industry. Hidden Markov models (HMMs) have been widely used for this purpose. In this paper we demonstrate how the fold recognition performance of a recently introduced HMM with a reduced state-space topology can be improved. Our method employs an efficient architecture and a low complexity training algorithm based on likelihood maximization. The fold recognition performance of the model is further improved in two steps. In the first step we use a smaller model architecture based on the {E,H,L} alphabet instead of the DSSP secondary structure alphabet. In the second step secondary structure information (predicted or true) is additionally used in scoring the test set sequences. The Protein Data Bank and the annotation of the SCOP database are used for the training and evaluation of the proposed methodology. The results show that the fold recognition accuracy is substantially improved in both steps. Specifically, it is increased by 2.9% in the first step to 22%. In the second step it further increases and reaches up to 30% when predicted secondary structure information is additionally used and it increases even more and reaches up to 34.7% when we use the true secondary structure. The major advantage of the proposed improvements is that the fold recognition performance is substantially increased while the size of the model and the computational complexity of scoring are decreased.

1. Introduction

In recent years there has been a vast increase in the number of proteins whose primary sequences have been identified. For most of these proteins, structure and function remains to be defined. The proteins with similar structure usually have similar function, so finding the structure can lead to the determination of function. Therefore, managing to relate amino acid sequences of unknown structure, with those of proteins with known structure, provides an indirect way to make predictions for both their structural and functional attributes [1]. Proteins which have major structural similarities are considered to share the same fold category. Finding the fold category of a protein with unknown structure is very important, as it reveals its three-dimensional structure. Different proteins can be found in the same fold category even when there is a very low sequence similarity among them [2,3]. So, as the fold of a protein is more evolutionarily conserved than its amino acid sequence, a target sequence can be modelled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment [4]. The task of identifying the fold category to which a protein of unknown structure belongs is called fold recognition.

A variety of methods have been proposed in the literature to address the problem of fold recognition. There are two main categories, the informatics-based methods and the biophysics-based methods. The informatics-based methods are divided into two subcategories, the sequence-based methods and the structure-based methods. The sequence-based methods use the primary sequence and/or the predicted secondary sequence information of the protein with unknown structure to perform sequence comparison with proteins of known structure [5–16]. Various machine

* Corresponding author at: Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, PO Box 1186, GR 45110 Ioannina, Greece. Tel.: +30 2651098803; fax: +30 2651098889.
E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

0010-4825/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2009.07.007

learning methods have been used for that comparison, such as hidden Markov models (HMMs) [5–11], genetic algorithms (GAs) [12], support vector machines (SVMs) [13–15], artificial neural networks (ANNs) [14] and segmentation conditional random fields (SCRFs) [16]. The structure-based or threading methods scan the amino acid sequence of an unknown structure against a database of known structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models [17–20]. This type of method is also known as three-dimensional–one-dimensional fold recognition due to its compatibility analysis between three-dimensional structures and linear protein sequences. Unlike the informatics-based methods, the biophysics-based methods perform *ab initio* structure prediction. More specifically, they do not perform any comparison of the primary sequence of the query protein with known folds. Instead they try to find or approach the specific three-dimensional structure of the protein which minimizes its energy [21–23].

Among the sequence-based methods, HMMs are commonly used for fold recognition and have proved to be very effective [5–11]. The main drawback of HMMs is the employment of large model architectures which require large datasets and high computational effort for training. Recently, we introduced a reduced state-space HMM with a much smaller architecture which also adopts a low complexity training algorithm for training, that proved to be equally effective in fold recognition [11]. In the current work we introduce certain improvements to that model that further improve its fold recognition performance without increasing its complexity. These improvements take place in two steps.

More specifically, in the first step we decrease the number of states by adopting the simple DSSP-EHL alphabet for the secondary structure. This reduction leads to an even smaller number of parameters that needs to be calculated in the training phase and simultaneously to better results. In the second step, we additionally use the predicted or the true secondary structure sequences in scoring the test set sequences. Thus, we avoid the use of the complex forward algorithm [11] for scoring and also we exploit the secondary structure information of the test set proteins.

In the following, the initial model is briefly described followed by the proposed improvements, and the training and the scoring procedures are also explained. The employed dataset is described next, as well as the experiments that we performed in order to evaluate the proposed methodology. Finally, the advantages and the disadvantages of the proposed approach are discussed.

2. Methods

HMMs are widely used in modelling families of biological sequences. A HMM is trained using a set of sequences called a training set and then it can be used for discrimination. The aim of the learning procedure is to maximize the likelihood of the model given the training data. 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 S_{t-1} [25].

The reduced state-space HMM [11], which was recently introduced, uses the mathematical framework of a typical HMM though it adopts a much smaller architecture containing a limited number of states. It consists of a set of states S and a set of possible transitions T among them. Each state stochastically emits a signal and then the procedure is moved forward 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.

Table 1
Correspondence between letters of the DSSP alphabet and the letters of the DSSP-EHL alphabet.

DSSP	Type	Corresponding letter of the DSSP-EHL alphabet
H	Alpha-helix	H
G	3_{10}-helix	H
I	Π-helix	H
E	Extended (β-strand)	E
B	Residue in isolated β-bridge	E
T	Turn	L
S	Bend	L

The reduced state-space topology that was used in our recent model addressed the main disadvantage of the previous HMMs, which is the employment of large model architectures which demanded large datasets and consequently high computational effort for training. The reduced HMM could be trained and then used for classification of proteins into fold categories. It contains a small number of states, because it incorporates the secondary structure in such a way that each state of the model corresponds to every possible different secondary structure state. This fact enables us not only to use the secondary structure information to train the model, which is necessary for more accurate classification of proteins into fold categories, but also to drastically reduce the number of states employed in the model and, thus, the number of parameters which must be estimated. Moreover, the model is trained with the use of a low complexity algorithm based on likelihood maximization, because the sequence of states during the training phase is known, so we can avoid complicated iterative procedures.

In the current work we introduce specific improvements in the reduced state-space model which lead to a substantial increase in its ability to classify proteins in the correct fold category. These improvements are presented in two steps. In the first step we change the topology of the model while in the second step we change the way that the test proteins are scored against the improved model.

More specifically, in the first step we use a different alphabet in order to encode the secondary structure of the proteins. The different possible secondary structure formations of each amino acid are represented by the 3-letter alphabet DSSP-EHL instead of the 7-letter DSSP alphabet. As it is shown in [10], the DSSP-EHL alphabet is a reduced representation of the DSSP alphabet. Specifically, H, G and I correspond to H, E and B–E, while T and S–L. Those amino acids which were considered of unknown structure in the DSSP representation are considered as loop (L) in the DSSP-EHL representation. The correspondence of letters between the two alphabets is presented in Table 1. So, now we employ three states in the model, corresponding to the three different possible formations of underlying secondary structure that each amino acid may have according to the DSSP-EHL alphabet, as it is shown in Fig. 1.

There is one to one correspondence between the amino acid residues and the secondary structure residues in the training set. The states of the model are fully connected, which means that every possible transition between them is allowed. In each state there is a distribution over all possible amino acids. There are 21 possible residues in each distribution. Twenty of them correspond to the 20 different amino acids and there is also one residue which represents amino acids of unknown origin (due to experimental limitations). More specifically, in the model there are 3×21 emission parameters, 3×3 transition parameters between the states and three parameters for the transitions from the starting state. Therefore, the total