Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 41 (2008) 165–179

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Mining sequential patterns for protein fold recognition

Themis P. Exarchos [a,b], Costas Papaloukas [b,c], Christos Lampros [a,b], Dimitrios I. Fotiadis [b,d,*]

[a] Department of Medical Physics, Medical School, University of Ioannina, GR 451 10 Ioannina, Greece
[b] Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, P.O. Box 1186, 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

Received 20 November 2006
Available online 17 May 2007

Abstract

Protein data contain discriminative patterns that can be used in many beneficial applications if they are defined correctly. In this work sequential pattern mining (SPM) is utilized for sequence-based fold recognition. Protein classification in terms of fold recognition plays an important role in computational protein analysis, since it can contribute to the determination of the function of a protein whose structure is unknown. Specifically, one of the most efficient SPM algorithms, cSPADE, is employed for the analysis of protein sequence. A classifier uses the extracted sequential patterns to classify proteins in the appropriate fold category. For training and evaluating the proposed method we used the protein sequences from the Protein Data Bank and the annotation of the SCOP database. The method exhibited an overall accuracy of 25% in a classification problem with 36 candidate categories. The classification performance reaches up to 56% when the five most probable protein folds are considered.

Keywords: Data mining; Sequential patterns; Fold recognition

1. Introduction

Structure prediction is a challenging field strongly related with function determination which is of high interest for the biologists and the pharmaceutical industry. As the genome projects worldwide progress, we are presented with an exponentially increasing number of protein sequences which are not accompanied by any knowledge concerning their structure or biochemical function. Proteins have structural features which define functional similarities, so the need for structure estimation methods is high. One way to define their structure is to link them with proteins in annotated databases, whose three-dimensional structure (fold) is known. Determining how amino acid sequences are related to those of proteins with known structure, helps us make predictions for their structural, functional and evolutionary attributes [1].

The proteins that share the same fold category have considerable structural similarities even when no evolutionary relationship (homology) of their sequences can be detected [2,3]. Various methods have been developed to identify the fold category where a protein of unknown structure belongs (fold recognition). These methods are divided into two methodological approaches: (a) the informatics based methods that involve the sequence-based methods [4–15] and the structure based methods [16–19] and (b) the biophysics based methods [20–22]. Sequence based methods use protein sequence or predicted secondary structure information to perform sequence comparison and detect whether two proteins share a fold or not. Structure based

[*] Corresponding author. Address: Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, P.O. Box 1186, GR 45110 Ioannina, Greece. Fax: +30 26510 97092.
E-mail address: fotiadis@cs.uoi.gr (D.I. Fotiadis).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.05.004

or threading methods create an energy function describing how well a probe sequence matches a target fold. In fold recognition by threading, we must take the amino acid sequence of a protein and evaluate how well it fits into one of the known three-dimensional (3D) protein structures. Besides purely sequence based or structure based methods, a combination of them is also possible [23]. On the other hand, methods based on biophysics perform *ab initio* structure prediction. They detect a native conformation or ensemble of conformations of the protein that are at or near the global free-energy minimum [24].

Sequence-based methods are very common in fold recognition. Machine learning techniques, such as genetic algorithms [8], support vector machines [9,10], hidden Markov models [11,13,14] and segmentation conditional random fields [15], have been adopted to exploit protein sequence or secondary structure information. The amino acid composition (protein sequence), in specific, has been employed in many areas of bioinformatics, like protein structural class prediction [25–27], discrimination of DNA binding proteins [28] and discrimination of outer membrane proteins [29]. However, although significant improvement has been made in the field of fold recognition, the accuracy of the existing methods remains limited and there is a need to develop new methods.

In this study, a novel classification method for biological data is proposed. The method uses sequential patterns that are extracted with data mining techniques and is validated in the common problem of protein fold recognition. Previous studies that utilized data mining techniques for biological data analysis [30] and proteins in specific [31], provided very promising results revealing that data mining can play a vital role in the field of bioinformatics. Currently, data mining is employed in the form of sequential pattern mining (SPM) [32] which is a technique appropriate for analyzing sequential data, like time series, texts and biosequences (e.g., proteins, DNA). Our approach extracts a large number of sequential patterns in order to characterize each class (protein fold in our case). The patterns discovered using the protein data could also assist the domain experts by providing them with previously unknown knowledge.

Sequential patterns can match significant combinations of amino acids that may correspond to functionally or structurally important regions in the proteins, like, for example, dipeptide combinations [33]. They follow the notion of deterministic motifs and are able to allow flexible length for the pattern due to the variable insertion of gaps between the amino acids of the patterns. A motif is defined as the occurrence in the protein's sequence of a particular cluster of residue types [34]. However, determining consistent motifs, requires first multiple alignment of the input protein sequences, which is not needed in mining sequential patterns.

The proposed method was applied in automated protein fold recognition, by classifying an unknown protein to the corresponding fold. In the training phase, sequential patterns are extracted from the training data with the use of the cSPADE algorithm [35]. During testing, a classifier uses the extracted sequential patterns and classifies the unknown proteins. Our method introduces several novelties. The employment of SPM for protein structure analysis offers the potential of discovering new knowledge in the form of patterns. Furthermore, the method uses only the protein's sequence for classification, which is easier to be acquired, whereas other similar approaches make use of the secondary structure [6], as well as other features [36]. For training and testing we employed a dataset with low similarity between proteins. The classification results indicate that our method performs well in terms of accuracy (considering a 36-class classification problem where the accuracy of the random prediction is 2.8%) and compares favorably with the Sequence Alignment and Modeling (SAM) approach (version 3.3.1) [11,12], which is an effective and widely used tool for sequence-based classification of proteins in structural/functional categories and thus for fold recognition [37,38].

In the following paragraphs, the adopted method is presented and the training and testing procedures are explained. The employed dataset and the results of the classification method are described then. The advantages and disadvantages of our approach are given in the discussion section, where possible further improvements are also discussed.

2. Materials and methods

The formulation of SPM can cover almost any categorical sequential domain [33,39,40]. In order to apply SPM to a specific domain, the following notions are required: a database of sequences *D*, a set of items (alphabet) *I*, a definition of the transaction *id* (*tid*) and a definition of an itemset. In what concerns our problem, protein sequences form the database *D*. The set of items *I* is the 20 amino acids that compose the protein sequences plus one for the unknown amino acid. *tid* currently denotes the position of the amino acid in the protein sequence and an itemset consists only of a single item (one of the 21 letters), since only one amino acid exists in a specific position of the protein sequence. The SPM procedure can be favored by incorporating constraints that allow for flexible gap of the extracted sequential patterns. More details are provided in the Appendix A.

Several algorithms have been reported in the literature which implement the above described SPM procedure [32,41,42]. However, limited work has been done in constrained SPM [35,40,43,44]. An algorithm that performs constrained SPM is the cSPADE algorithm [35]. cSPADE finds the set of all frequent sequences with constraints, such as minimum and maximum gap between sequence items, based on the SPADE algorithm [45]. In what concerns the performance and the computational effort required, cSPADE is considered superior, compared to other constrained SPM approaches [40,44].