Protein fold recognition with a protein feature extraction technique blending evolutionary and secondary structural information through the use of metric learning

Pooya Zakeri[1,2] , **Forough Amini**[3] , **Mehdi Sadeghi**[3,4] ,**Yves Moreau**[1,2]

[1]**Department of Electrical Engineering, STADIUS, KU Leuven, Leuven 3001, Belgium**

[2]**iMinds Medical IT, Leuven 3001, Belgium**

[3]**School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran**

[4]**National Institute of Genetic Engineering and Biotechnology; Tehran, Iran**

Abstract

In this work, we address protein fold recognition (PFR) which is among the most challenging problem in cell and molecular biology. The main features of our work are: (1) We propose a protein feature extraction technique based on the mixture of evolutionary-based and secondary structural-based information. (2) We design a predictor for PFR based on the large margin nearest neighbor (LMNN) algorithm [1]. (3) In order to evaluate our model, we develop an unbiased and prospective benchmark based on SCOP database [2].

Introduction

Protein fold recognition is a challenging problem in computational biology. Tertiary structural information of proteins can provide new knowledge on their function. In addition, understanding the three-dimensional structure of proteins can be facilitated through the knowledge of protein folds; hence, determining this structure is among the most essential objectives in molecular and cell biology. Various approaches based on features extracted from protein sequences have been used in the prediction of protein folds [3, 4]. It has been shown that the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins [3]. In this work, we propose a protein feature extraction technique based on the blending of evolutionary and secondary structural information.

Methods

A protein sequence P with L amino acid residues can be represented by its evolutionary information through position-specific scoring matrix (PSSM), which have L columns and 20 row. Each row of PSSM represents the log-likelihood of the residue substitution at the corresponding position in the protein sequence. A protein sequence P can also be represented by its secondary structure elements (SSEs) such as alpha-helix, beta-sheet and coiled-coil. In this study, the predicted secondary structure results from NetSurfP [5] are used. We design a feature vector (FV) created by summing up each column of the same SSE in the PSSM and dividing by the length of the protein domain, followed by a normalization that scales each score to the range of [0,1]. In the similar way, two additional FVs are created by considering the transition and distribution of SSE [3] in each column of PSSM. To this end, three FVs are merged together.

K-nearest neighbor (kNN) and support vector machine algorithms are among the most used machine learning methods in the prediction of protein folds [3, 4]. In particular, several modified versions of the kNN algorithm have been proposed to address the PFR problem. They predict the fold of a new protein with the fold obtained from the majority vote for the k closet proteins of known structure. Closeness is calculated with a pre-defined metric such as the Euclidian distance. As a result, the performance of kNN-based fold predictors depends strongly on the metric used to measure distances between protein sequences. LMNN is an approach that learns a Mahalanobis distance metric for the kNN algorithm in a supervised fashion to enhance its performance. In addition to LMNN, the kernel-based kNN is employed to address the prediction of protein folds.

Results & Conclusions

Furthermore, it is necessary to validate protein fold predictors in a more realistic task setting that their respective performance fairly reflect their ability to predict the fold of a protein with unknown fold. Most of protein fold predictor machines have been evaluated through cross validation experiment. This often leads to an overestimate of the real performance. Moreover, they only cover less than 50 folds, which is far less than protein folds have been identified. However, it has been estimated that the large majority of protein domains belong to less than 1000 protein folds.

In order to overcome these issues we develop an unbiased and prospective benchmark based on the SCOP database. It is at the heart of our assessment strategy to: (1) train our model based on the proteins of known fold discovered before a certain time (version-1.75, 2009) using protein features released prior to that time, and(2) evaluate our model on the prediction of proteins of known fold reported afterwards (version-2.05, 2015). We construct two benchmarks named fold100 and fold200 that cover about 100 and 200 folds respectively.

The experimental results on our prospective benchmark show that our developed model based on our proposed feature extraction methods can effectively improve the accuracy of the state-of-the-art protein fold predictor. While both the Geometric kernel fusion framework [1] (developed by integration of evolutionary-based and secondary structure-based kernels[1]), and the kernel-based kNN designed based on our proposed feature extraction method present promising results for predicting protein fold on the fold100 and fold200 datasets, the proposed model offers a significant improvement for predicting protein fold on the benchmarks.

References

1. Weinberger, K.Q.; Saul L.K. JMLR 10:207–244 (2009).
2. Murzin A.G. et al. J Mol Biol 247:536–540 (1995).
3. Zakeri P. et al. Bioinformatics 30(13):1850-1857 (2014).
4. Yang J.Y.; Chen X. Proteins 79:2053–2064 (2011).
5. YPetersen BMC Struct Biol 9:51 (2009).