Toward a Metric Learning Model for Protein Fold Recognition Using a Novel Feature Extraction Technique Based on the Mixture of Evolutionary and Secondary Structural Information

Presenting Author: Pooya Zakeri

1)KU Leuven. 2)iMinds

Co-Author(s):
Forough Amini,
Institute for Research in Fundamental Sciences (IPM)
Mehdi Sadeghi,
1) National Institute of Genetic Engineering and Biotechnology, 2)Institute for Research in Fundamental Sciences (IPM)
Yves Moreau,
1) KU Leuven, 2)iMinds

Abstract

It has been demonstrated that integrating the evolutionary and secondary structural information could be useful in explicating the relationship between primary and tertiary structure in proteins. In this work, we propose a protein feature extraction technique based on the blending of evolutionary and secondary structural information. A protein feature vector is created by summing up each column of the same predicted secondary structure elements in the position-specific scoring matrix and dividing by the length of the protein domain.
Then, the kernel-based kNN is employed to address the prediction of protein folds. Nevertheless, the performance of kNN-based fold predictors depends crucially on the metric used to measure distances between protein sequences. To overcome this limitation, we develop for the first time a protein fold predictor based on the large margin nearest neighbor algorithm, which learns a Mahalanobis distance metric for the kNN algorithm in a supervised fashion to improve its performance.
In order to evaluate our models in a more realistic task setting, we develop a time-stamped benchmark based on the SCOP database. Our models are trained based on the proteins of known fold discovered before a certain time using protein features released prior to that time. Then, we assess our model on the prediction of proteins of known fold reported afterwards. The experimental results on our prospective benchmark which covers about two-hundreds folds demonstrate that our developed model based on our proposed feature extraction technique can effectively improve the accuracy of the state-of-the-art protein fold predictors, such as TAXFOLD and GeoFold.