Protein Fold Recognition Using Matrix Factorization Technique

Pooya Zakeri,[1,2] Jaak Simm[1,2], Adam Arany[1,2], Forough Amini[3], Mehdi Sadeghi[3,3] and Yves Moreau[1,2]

1) KU Leuven, 2)iMinds 3)Institute for Research in Fundamental Sciences (IPM)
4) National Institute of Genetic Engineering and Biotechnology,

Proteins/ Application poster, Application Fundamenta.

Most of protein fold predictor machines only cover less than 30 folds, which is far less than protein folds have been identified. Moreover, the typical approaches proposed for protein fold recognition often neglect the relationship between protein folds. These motivate us to formulate the protein fold recognition as a factorization of an incompletely filled binary protein-fold-matrix where the objective is to predict unknown values. Protein fold recognition database such as SCOP can be seen as an incomplete matrix (M×N) where each row is a protein and each column is a protein fold. Then, the SCOP matrix can be modeled by the two smaller matrices P and F, which, when multiplied, approximately reconstruct the SCOP matrix. We propose an extended version of the Bayesian probabilistic matrix factorization [1] with the added advantage of working with multiple protein features as side sources for completing protein-fold-matrix. Accordingly, two sequence-based protein features, including the predicted secondary structure and information extracted directly from position-specific scoring matrices, are incorporated into the proposed factorization model as side information.In order to validate our models in a more realistic task setting, we develop a prospective benchmark, based on the latest version of the SCOP database, which covers about 200 protein folds. The experimental results on our unbiased benchmark show that our proposed model can effectively improve the accuracy of the state-of-theart protein fold predictors such as GeoFold [2].

[1] doi: 10.1093/bioinformatics/btu118.

[2] doi: 10.1145/1390156.1390267.