Refinement of Pairwise Potentials via Logistic Regression to Score Protein-Protein Interactions

Kiyoto Tanemura[1], Jun Pei[1], and Kenneth Merz[1]

[1]Michigan State University

May 5, 2020

Abstract

Protein-protein interactions (PPIs) are ubiquitous and functionally of great importance in biological systems. Hence, the accurate prediction of PPIs by protein-protein docking and scoring tools is highly desirable in order to characterize their structure and biological function. Ab initio docking protocols are divided into the sampling of docking poses to produce at least one near-native structure, then to evaluate the vast candidate structures by scoring. Concurrent development in both sampling and scoring is crucial for the deployment of protein-protein docking software. In the present work, we apply a machine learning model on pairwise potentials to refine the task of protein quaternary structure native structure detection among decoys. A decoy set was featurized using the Knowledge and Empirical Combined Scoring Algorithm 2 (KECSA2) pairwise potential. The highly unbalanced decoy set was then balanced using a comparison concept between native and decoy structures. The resultant comparison descriptors were used to train a logistic regression (LR) classifier. The LR model yielded the optimal performance for native detection among decoys compared to conventional scoring functions, while exhibiting lesser performance for the detection of low root mean square deviation (RMSD) decoy structures. Its deployment on an independent benchmark set confirms that the scoring function performs competitively relative to other scoring functions. All data and scripts used are available at: https://github.com/TanemuraKiyoto/PPI-native-detection-via-LR .

Hosted file

submission1_refinement_pairwise_potentials.pdf available at https://authorea.com/users/300318/articles/429959-refinement-of-pairwise-potentials-via-logistic-regression-to-score-protein-protein-interactions