Prediction of Protein-Protein Interactions Using Protein Signature Profiling

Mahmood A. Mahdavi and Yen-Han Lin*

Department of Chemical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.

Protein domains are conserved and functionally independent structures that play an important role in interactions among related proteins. Domain-domain interactions have been recently used to predict protein-protein interactions (PPI). In general, the interaction probability of a pair of domains is scored using a trained scoring function. Satisfying a threshold, the protein pairs carrying those domains are regarded as "interacting". In this study, the signature contents of proteins were utilized to predict PPI pairs in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens.* Similarity between protein signature patterns was scored and PPI predictions were drawn based on the binary similarity scoring function. Results show that the true positive rate of prediction by the proposed approach is approximately 32% higher than that using the maximum likelihood estimation method when compared with a test set, resulting in 22% increase in the area under the receiver operating characteristic (ROC) curve. When proteins containing one or two signatures were removed, the sensitivity of the predicted PPI pairs increased significantly. The predicted PPI pairs are on average 11 times more likely to interact than the random selection at a confidence level of 0.95, and on average 4 times better than those predicted by either phylogenetic profiling or gene expression profiling.

Key words: protein-protein interaction, protein signature, ROC curve

Introduction

Protein-protein interaction (PPI) is the key element of any biological process in a living cell. Proteins interact through their functional subunits (*1*). Protein domains, active sites, and motifs (collectively called signatures) are sub-sequence functional and conserved patterns that are essential to the functioning of individual cells and are the interfaces used in interactions at the protein level (*2*). With the completion of genome sequences of many organisms, genome-wide characterization of protein signatures is now practical. Although proteins are specified by unique amino acid sequences, the signature content of a protein sequence is crucial to determining interactions in which the particular protein is involved.

Protein signature (domain) information has been used to predict PPI. Naively, when two proteins are known to interact, their homologs in other organisms are assumed to interact based on comparative analysis (*3*). Domain contents of the interacting partners are utilized as input to predict more accurate interactions in another organism (*4*). Intermolecular or intramolecular interactions among protein families that share one or more domains are implemented to infer interactions among proteins (*5*). Domain-domain relationships are used to predict interactions at the protein level. In the association method (*6*), interacting domains are learned from a dataset of experimentally determined interacting proteins, where one protein contains one domain and its interacting partner contains the other domain. The probabilistic model of maximum likelihood estimation (MLE) outperforms the association method through taking the experimental errors into account. Following a recursive calculation procedure, in MLE method probabilities for domain-domain interactions are predicted based on the observation of interactions between their corresponding proteins. Then the prediction is extended to the protein level, assuming that two proteins interact if and only if at least one pair of domains from the two proteins interact (*7*). Potentially interacting domain (PID) pairs are extracted from an ex-

***Corresponding author.**
E-mail: yenhan.lin@usask.ca