Predicting domain-domain interactions using a parsimony approach

Katia S Guimaraes[1,2], Raja Jothi[1], Elena Zotenko[1,3], and Teresa M. Przytycka[1]

[1] National Center of Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894 USA
[2] Center of Informatics, Federal University of Pernambuco, Recife, PE 50732 Brazil
[3] Department of Computer Science, University of Maryland, College Park, MD 20742 USA

The knowledge of protein interactions provides crucial insights into their function within a cell. Proteins typically contain two or more domains, and a protein interaction usually involves binding between specific pairs of domains. Identifying such interacting domain pairs is an important step towards determining the protein-protein interaction network. Previous computational approaches geared towards predicting domain contacts detect only a small fraction of known interacting domain pairs, which suggests that the problem is largely unsolved.

One simple approach to solve this problem, the Association method, scores each domain pair by the ratio of the number of occurrences of a given pair in interacting proteins to the number of independent occurrences of that pair. Another previous method uses a expectation maximization algorithm (EM), which computes domain interaction probabilities that maximize the expectation of the observed protein-protein interaction network. More recently, a more insightful method, called Domain Pair Exclusion Analysis (DPEA) was proposed. DPEA relies on the analysis of the impact of the absence of each particular domain pair in the overall likelihood of the network.

In this work, we explore an alternative approach for domain-domain interaction prediction. We hypothesize that interactions between proteins evolved in a parsimonious way and that the set of correct domain-domain interactions is well approximated by the minimal set of domain interactions necessary to justify a given protein-protein interaction network. We refer to our approach as the ***Parsimonious Explanation*** (**PE**) method. We formulate PE as a linear programming optimization problem, where each potential domain-domain contact is a variable that can receive a value ranging between 0 and 1 (called ***LP-score***), and each edge of the protein-protein interaction network corresponds to one linear constraint. This formulation allows for a novel way of handling the noise (false positives) in the data. Namely, we construct our linear programming instance in a probabilistic fashion, in which the probability of including an LP constraint equals the probability with which the corresponding protein-protein interaction is assumed to be correct.

To control for possible overprediction of interactions between frequently occurring domain pairs, we assign a ***promiscuity and witnesses score*** (***pw-score***) to each domain-

domain potential contact. The pw-score, derived from two observations, measures the confidence in the prediction. First, domain-domain interactions supported by many witnesses (interacting pairs of single domain proteins that support it) are more likely to be correct than ones that have a few or no witnesses. Second, there are promiscuous domain-domain interactions that are scored high due to the frequency of their appearance and not to the specific topology of the protein-protein interaction network. Consistent with these observations, the pw-score rewards domain interactions that have many witnesses and penalizes promiscuous interactions.

We compare the PE method to previous ones based on the enrichment of domain pairs confirmed by PDB crystal structures among the topmost scoring pairs. A plot of accuracy versus sensitivity shows that PE outperforms the previous methods. We also compare the sets of experimentally confirmed domain pairs among the 3000 topmost scoring predictions of the PE method and the 3005 high-confidence pairs predicted by DPEA, shown to be the best among the currently available methods. We separate the predictions into easy and difficult ones. In the easy category are domain pairs supported by at least one ***witness*** (interacting pair of single-domain proteins). Interacting domain pairs that do not have such direct experimental evidence fall in the ***difficult*** category, as they are harder to detect for any method. The PE method recovers more experimentally confirmed interactions than DPEA in both classes. In particular, in the difficult class, it outperforms DPEA by an order of magnitude.

To further evaluate the PE method, we also estimated sensitivity and accuracy measures using an experiment aimed at identifying the mediating domain pair(s) among all potential domain contacts in interacting proteins pairs that contain at least one PDB crystal structure confirmed domain pair. Every potential domain pair occurring in such set of 1,780 protein pairs that was not in the gold standard set was considered a gold standard negative. The average estimated measure of accuracy and sensitivity for the PE method were 75.3% and 76.9%, respectively, while for the DPEA method were 42.5% and 36.9%, respectively.

The results indicate that the parsimony principle provides a correct approach for detecting domain-domain interactions from the topology of the protein-protein interaction network.

An article with the full version of this work will appear in ***Genome Biology,*** published by BioMed Central.

References

[1] Sprinzak E and Margalit H (2001). Correlated sequence-signatures as markers of protein-protein interaction. ***Journal of Molecular Biology***, 311 (4): 681-692.
[2] Deng M, Mehta S, Sun F, and Chen T (2002). Inferring domain-domain interactions from protein-protein interactions. ***Genome Research***, 12:1540-1548.
[3] Riley R, Lee C, Sabatti C, and Eisenberg D (2005). Inferring protein domain interactions from databases of interacting proteins. ***Genome Biology*** 6(10):R89.