

Function Prediction Using Neighborhood Patterns

Petko Bogdanov, Swaroop Jagadish, Ambuj Singh
Department of Computer Science, University of California, Santa Barbara, CA 93106

1 Introduction

Traditionally, function prediction is performed using sequence/structure homology and then manual verification in the wet lab. The first step, also called computational function prediction, directs the experimental design to a narrow set of possible annotations for unstudied proteins.

Recently, large interaction networks have been constructed using data from high-throughput techniques like Microarray co-expression analysis and Yeast2Hybrid experiments. These networks contain more information compared to sequence or structure alone. A molecular function is performed in the context of a biological process with multiple interacting agents. Hence, a natural next stage of computational function prediction includes the use of a protein’s interaction context within the network.

A recent survey [1], classifies most existing function prediction methods in two groups: *module assisted* and *direct methods*. Methods from the first group detect network modules and then perform a module-wide annotation enrichment [2]. They all differ in the manner they identify modules— some use graph clustering [3] while others use hierarchical clustering based on network distance [2]. Direct methods assume that similar functional annotations are neighbors in the network. The *Majority* method [4] predicts prevailing functions among the direct interactors of a target protein. This idea has later been extended to higher levels in the network [5]. *Indirect Neighbor* [6], utilizes direct and indirect functional associations, considering level 1 and level 2 associations. The *Functional Flow* method [7] simulates an annotation network flow from known to target proteins.

A common drawback of the methods in the two groups is their hypothesis that proteins with similar functions are always close in the network. Functional annotations in actual protein networks do not always corroborate this hypothesis. The direct methods are also limited to use information about neighbors up to a certain level. Thus, they cannot predict the functions of proteins surrounded by unannotated interaction partners.

2 Method

Proteins are presented as nodes in a protein interaction network and interactions between proteins as edges. The network is stochastic as the edges are weighted with the probability of interaction. Functional terms appear as attributes of each node.

Our technique’s main idea is that the functions need not be localized in the network and that functional annotations can be inferred based on patterns in the neighborhood. For example, proteins annotated with *GTPase Activity (GTPases)* perform the important role of biological switches. As they control diverse cellular processes, we expect that some *GTPases* would appear at different locations in the network as part of the related sub-networks. This behavior is confirmed by our analysis of the prediction accuracy for the *GTPases* in *C. elegans*. Hence the assumption that similar functions always cluster together in interaction networks does not always hold.

Figure 1 presents an overview of the key steps in our technique. To summarize a protein’s neighborhood we compute the steady state distribution of a *Random Walk with Restarts (RWR)* from the protein. A functional neighborhood profile is obtained from the neighborhood profile, reflecting the GO structure. Further, we define a novel distance metric between functional profiles based on the *Earth Mover’s Distance (EMD)* that incorporates the ontological relationships among the functions in the GO hierarchy. We state the function prediction as a classification problem and define three classifiers to solve it. First, we employ *k-Nearest-Neighbors (kNN)* classification to predict the function of a target protein. Noise in the input interaction data is handled by transforming the original multi-class classification problem into a set of two-class problems and consolidating their results. We employ a modified version of the kNN classifier that we call *Hierarchical kNN (HkNN)* as a second classifier. HkNN classifies a target protein in a top-down manner using training instances only from the most confident GO sub-hierarchy. Employing a third classifier, called *Highest Bin (HB)*, we capture functions that cluster in the network. The predictions of this classifier are the highest scoring classes in a protein’s profile. We combine kNN, HkNN and HB in an ensemble voting classification scheme *COMP*. COMP is robust to both functional classes that cluster on the network and ones that are at relatively large distances.

The contributions of our work are as follows:(i)We devise a novel self-contained method of capturing the functional neighborhood of a target protein. (ii) We propose a novel hierarchy-aware distance function between

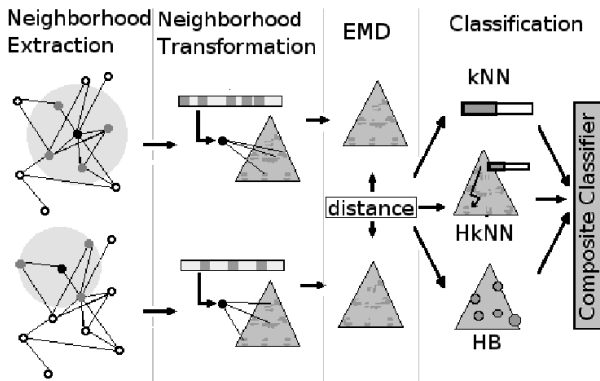


Figure 1: Key steps in our function prediction technique.

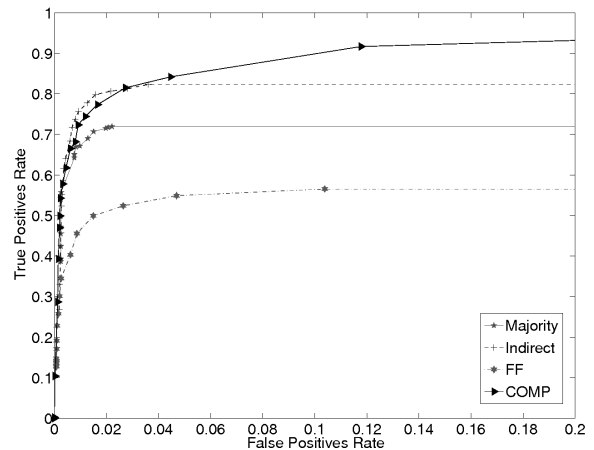


Figure 2: ROC Curves for the multi-class classification case in the FYI network.

functional neighborhood profiles (iii) We employ three classification approaches that take into account the hierarchical organization of the annotations and the inherent noise in the interaction data to predict the function of a target protein. (iv) Our analysis shows that functions have different network localization behaviors and we devise a composite classifier that is robust to this phenomenon. (v) We improve the classification accuracy up to 13% as compared to previous techniques.

3 Experimental Results

We use *Majority (MAJ)* [4], *Functional Flow (FF)* [7] and *Indirect Neighbors(IND)* [6] techniques for comparison with our functional neighborhood technique *COMP*.

We compare the ROC curves of all methods for the multi-class classification scenario (Figure 2). The existing methods are limited to a network distance at which they infer functions as they assume that similar functions appear as neighbors. For small False Positive Rates (FPR) all functionally similar co-localized proteins are correctly predicted, hence a steep rise of the True Positive Rate (TPR) is observed for all methods. The ROC curves of existing methods then saturate. *COMP* classifies correctly additional informative proteins based on patterns in their functional neighborhoods. It reaches a $TPR = 0.92$ for $FPR = 0.12$ and a $TPR = 0.99$ for $FPR = 0.53$ (not shown).

4 Conclusion

We proposed a novel method for the problem of protein function prediction in a network setting. Our approach captures protein functional neighborhoods and includes a novel distance function based on EMD to compare them. The ontological relationships between GO annotations are reflected both in the distance computation and in the prediction process. We employed existing machine learning algorithms to predict functions of target proteins.

We performed leave-one-out validation experiments in which our technique outperforms existing techniques by as much as 0.13 in AUC value. We analysed three protein interaction networks and revealed different function localization trends, all of which our technique was able to capture and classify.

5 Acknowledgments

This work is supported in part by National Science Foundation (0612327) and the Institute for Collaborative Biotechnologies (DAAD19-03-D-0004.)

References

- [1] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 2007.
- [2] K. Maciag, S.J. Altschuler, M.D. Slack, N.J. Krogan, A. Emili, J.F. Greenblatt, T. Maniatis, and L.F. Wu. Systems-level analyses identify extensive coupling among gene expression machines. *Molecular Systems Biology*, 2006.
- [3] R. Dunn, F. Dudbridge, and C.M. Sanderson. The use of edge-betweenness clustering to investigate the biological function in protein interaction networks. *BMC Bioinformatics*, 2005.
- [4] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature*, 2000.
- [5] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 2001.
- [6] H. Chua, W. Sung, and L. Wong. Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 2006.
- [7] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21:i302-i310.