

Function Prediction Using Neighborhood Patterns

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

1 Introduction

Function prediction has traditionally been done using sequence/structure homology coupled with manual verification in the wet lab. The first step, referred to as computational function prediction, facilitates the functional annotation by directing the experimental design to a narrow set of possible annotations for unstudied proteins.

High-throughput techniques like Microarray co-expression analysis and Yeast2Hybrid experiments have allowed the construction of large interaction networks. There is more information in these networks compared to sequence or structure alone. Molecular functions are performed in the context of a biological process with multiple interacting agents. Hence, the next stage of computational function prediction is characterized by the use of a protein’s interaction context within the network to predict its functions.

According to a recent survey [1], most existing function prediction methods can be classified in two groups: *module assisted* and *direct methods*. Module assisted methods detect network modules and then perform a module-wide annotation enrichment [2]. The methods in this group 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 neighboring proteins in the network have similar functional annotations. The *Majority* method [4] predicts the three prevailing annotations among the direct interactors of a target protein. This idea has later been generalized to higher levels in the network [5]. Another approach, *Indirect Neighbor* [6], distinguishes between direct and indirect functional associations, considering level 1 and level 2 associations. The *Functional Flow* method [7] simulates a network flow of annotations from annotated proteins to target ones.

A common drawback of both the direct and module-assisted methods is their hypothesis that proteins with similar functions are always close in the network. As we show, functional annotations in actual protein networks do not corroborate this hypothesis. The direct methods are further limited to use information about neighbors up to a certain level. Thus, they are unable to predict the functions of proteins surrounded by unannotated interaction partners.

2 Method

A protein interaction network consists of nodes representing proteins, and edges representing interactions between proteins. It is a stochastic network as the edges are weighted with the probability of interaction. Each node is annotated with one or more functional terms.

The main idea behind our technique 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 the function *GTPase Activity*, called *GTPases*, perform the important role of biological switches. As they regulate diverse cellular processes, we expect that some *GTPases* would appear at different locations in the network as part of the process-related sub-networks. Our analysis of the prediction accuracy for the *GTPases* in *C. elegans* confirms this behavior. This is in contrast to the assumption made by direct and module assisted methods that similar functions always cluster together in interaction networks. Our technique is not conservative as it does not impose an exact topological match and uniqueness of the functional neighborhood patterns.

An overview of the key steps in our technique is presented in Figure 1. We summarize a protein’s neighborhood by computing the steady state distribution of a *Random Walk with Restarts (RWR)* from the protein. A protein’s neighborhood profile is then transformed into a functional neighborhood profile, reflecting the GO structure. Further, we devise a novel distance metric between neighborhood profiles based on the *Earth Mover’s Distance (EMD)*. This distance computation incorporates the ontological relationships among the functions in the GO hierarchy. We then pose the problem of function prediction as a classification problem and define three classifiers to solve it. We employ *k-Nearest-Neighbors (kNN)* classification to predict the function of a target protein. To address the problem of noise in the input interaction data, we transform the original multi-class classification problem into a set of two-class problems and consolidate their results. A second classifier that we employ is a modified version of the kNN classifier that we call *Hierarchical kNN (HkNN)*. HkNN exploits the hierarchical organization of the classes and classifies a target protein in a top-down manner, using training instances only from the most confident GO sub-hierarchy. To capture functions that cluster in the network, we devise a third classifier called *Highest Bin (HB)*. It annotates a target protein with the highest scoring classes in its neighborhood profile. We build an ensemble voting classification scheme *COMP* combining kNN, HkNN and HB. Its predictions are

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 introduce a novel self-contained approach of capturing

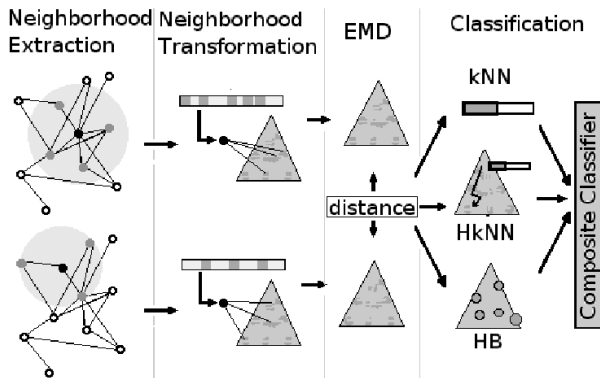


Figure 1: Key steps in our function prediction technique.

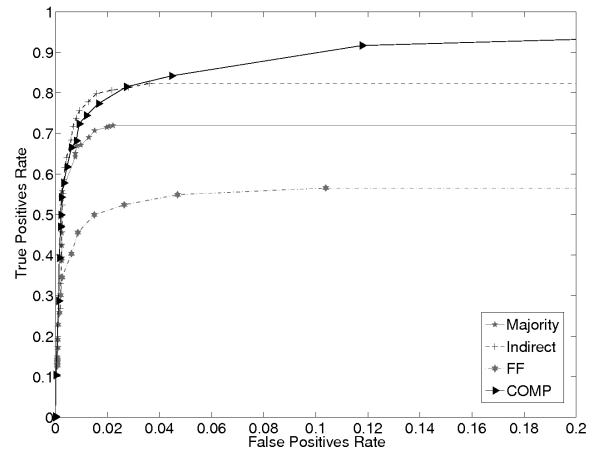


Figure 2: ROC Curves for the multi-class classification case in the FYI network.

the functional neighborhood of a target protein. (ii) We propose a novel distance function between functional neighborhood profiles that incorporates the hierarchical relationship among functional classes in GO. (iii) We define three classification algorithms based on kNN 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) We show that different functions have different network localization behaviors and devise a composite classification that is robust to this phenomenon. (v) Our technique improves the classification accuracy up to 13 % as compared to previous techniques.

3 Experimental Results

We compare our functional neighborhood technique to *Majority (MAJ)* [4], *Functional Flow (FF)* [7] and *Indirect Neighbors(IND)* [6] techniques. The abbreviation used for our Composite classifier is *COMP*.

The ROC curves for the multi-class classification scenario are presented in Figure 2 up to an FPR of 0.12. All existing methods assume that similar functions appear as neighbors and are limited to a network distance at which they infer functions. We observe a steep rise of the TPR for small FPR for all methods as all functionally similar co-localized proteins are correctly predicted. After that, the ROC curves of existing methods saturate. Our approach classifies additional informative proteins based on patterns in their functional neighborhoods. It achieves 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 approach for the problem of protein function prediction in a network setting. We devised a method that captures functional neighborhoods and defined a novel distance function based on EMD to compare them. We exploited the ontological relationships between GO annotations both in the distance computation and in the prediction process. We adapted existing machine learning algorithms to predict functions of target proteins. Our approach is self-contained and does not require manual tuning of parameters.

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

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 CM. 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.