A Graph-based Cluster Ensemble Method to Detect Protein Functional Modules From Multiple Information Sources

Yuan Zhang∗, Liang Ge∗, Nan Du+, Kebin Jia∗, Aidong Zhang+
∗Beijing University of Technology
Beijing, 100124, China
zhangyuan@emails.bjut.edu.cn
kebinj@bjut.edu.cn

+State University of New York at Buffalo
Buffalo, 14260, U.S.A.
liangge, nandu, azhang@buffalo.edu

ABSTRACT
Many works have been done to identify functional modules in Protein-Protein Interaction (PPI) networks but the results are far from satisfaction. One main reason is that the PPI data generated from high-throughput experiments is noisy and incomplete. Solving the problem goes beyond what a single data source can provide and thus requires the integration of multiple information sources. To address this problem, we hereby propose a graph-based cluster ensemble method which integrates gene ontology (GO) and gene expression data with PPI networks. Experimental results show that our method is superior to the baseline methods and demonstrate the benefits of integrating multiple biological information sources and diverse clustering methods.

1. INTRODUCTION

Predicting protein functional modules among the interacted proteins is a significant problem for understanding the biological processes at the cellular level. PPI networks are the most important sources of information for functional module detection. It’s a common practice to adopt clustering methods to find the protein functional modules [2, 8]. However, we observe that utilizing single information source or criterion to find protein functional modules has suffered the following three limitations. Firstly, it is well known that a PPI network contains a lot of noise and errors: the rate of false positive links is sometimes up to 50% [11]. Also, the gene expression data is reported to be noisy and incomplete [6]. Given a data source with so much noise and errors, it is expected that the clustering results are not satisfactory. Secondly, the clusters produced by these methods which are based on single information source may lack biological meaning. The third limitation is caused by the clustering methods. Most module detection methods are case dependent, and their results typically depend on the specific initial conditions, seeds selection, or parameter setting. Furthermore, different clustering methods tend to get different and unstable cluster partitions based on their own objective functions.

The challenge of protein module detection changes to how to effectively integrate multiple biological data sources. Here we propose a graph based cluster ensemble method to approach the problem. Extensive experimental results show that our method is superior to the baseline methods. The evaluation results have also shown that our method has greatly improved the prediction of protein functional modules than just using individual clustering methods. Also, by adding additional information, the full graph based cluster ensemble method has achieved more accurate functional modules than solely using the bipartite graph.

2. GRAPH BASED CLUSTER ENSEMBLE

2.1 Base Similarity Graphs and Base Clusters
We adopt two different protein correlation measurements that integrate Gene Ontology information and gene expression data with the PPI network. The Pearson correlation coefficient value (normalized to the range of 0 to 1) is used to calculate the co-expression correlation. GO-driven Similarity is calculated based on the information content of GO terms and Lin’s model [7] is implemented. The co-expression correlation and GO-driven similarity are denoted as CoExp and Sim respectively.

Three different clustering methods, including hierarchical agglomerative clustering, K-means, and Markov clustering (MCL), are performed on those two correlation matrices to obtain the base cluster partitions. Each of the three methods will achieve its own objective function individually and produce diverse clustering results for our cluster ensemble model.

2.2 Bipartite Graph Model
As illustrated in Figure 1, we construct a bipartite graph based on the affiliation of proteins and clusters, denoted as BGENS. Suppose we are given a set of (T) clustering partitions G = G1, G2, ..., GT of the original data. Each clustering partition, Gt, t ∈ {1, 2, ..., T}, consists of a set of clusters from a certain clustering method with one binary matrix. BGENS constructs the graph GB(G, E), where V = VG \cup VN, VG represents all the cluster nodes from the base clustering partitions, and VN is the N protein nodes in the data. E is the pairwise connectivity relation between the protein nodes and the clusters, where if the protein node i belongs to cluster node j, E(i, j) = E(j, i) = 1 and 0 otherwise. Here we get the adjacency matrix Aβ of the bipartite graph.
where $P_s$ is the binary matrix describing protein interaction and $G_s$ is the binary matrix of cluster similarity. We calculate $P_s$ based on a joint measurement of Co-expression correlation and GO-driven similarity:

$$P_s(p_i, p_j) = \begin{cases} 1 & \text{if } \text{InSim}(p_i, p_j) > \beta, \\ 0 & \text{otherwise}, \end{cases}$$  

where $\text{InSim}$ is the joint similarity measurement:

$$\text{InSim}(p_i, p_j) = (1 - \alpha) \times \text{CoExp}(p_i, p_j) \times \text{Ppi}(p_i, p_j) + \alpha \times \text{Sim}(p_i, p_j) \times \text{Ppi}(p_i, p_j),$$  

where $\text{Ppi}$ is PPI network adjacency matrix and $\alpha$ is a real positive number between 0 and 1 to specify how much the GO-driven similarity measurement contributes to the joint matrix.

In our experiments, we choose $\alpha = 0.7$ according to the experimental optima. $G_s$ in Equation 2 is the binary matrix of Jaccard similarity of clusters. In our experiments, we set $\beta = 0.8$.

$$G_s(g_i, g_j) = \begin{cases} 1 & \text{if } \frac{|g_i \cap g_j|}{|g_i \cup g_j|} > \beta, \\ 0 & \text{otherwise}. \end{cases}$$  

### 2.4 Graph Partition

With the graph we constructed above, the cluster ensemble problem is reduced into a graph partition problem. We use the spectral clustering method which is a well-studied and efficient graph partition algorithm. The goal of the spectral clustering is to partition the graph into $K$ parts with the objective of minimizing the cut (the sum of the weights of those edges connecting different parts in the graph) [9]. Spectral clustering normalizes the similarity matrix, the $A^F$ in our case, with the diagonal matrix $D(i, i) = \sum_j A^F(i, j)$ which is actually the degree matrix of $A^F$. And we get the normalized Laplacian matrix $L = D^{-1}A^F$. Then it transforms the normalized matrix into $K$-dimensional space, whose coordinates are the $K$ largest eigenvectors and the new matrix is then clustered via K-means clustering.

### 3. EXPERIMENT AND RESULTS

#### 3.1 Data sets

Our PPI data set is from Gavin [5] which contains 2551 proteins annotated with their GO function terms and 21413 interactions, which are retrieved by mass spectrometry of tandem affinity purification data (TAP). We downloaded GO-slim file from http://www.geneontology.org/. We used the January 14, 2012 version of the GO slim mapping file and chose the Biological Process (BP) hierarchy to calculate the GO-driven similarity of proteins. The gene expression data is from Brem [3] in which each gene is described by 131 expression values. We evaluate our results with the CYC2008 benchmark dataset [10], a comprehensive catalogue of manually curated 408 heteromeric protein complexes in S. cerevisiae reliably backed by small-scale experiments from the literature.

#### 3.2 Comparison With Baseline Methods

We compared the performance of our methods with the previous methods: the Flownet [4], MCODE [2], CFinder [1], the spectral clustering method [9], RWR [8], and MCL. We implemented these baseline methods on the TAP data.
with optimal parameters set and the comparison results are presented in Figure 2.

Based on the comparison, our method outperforms other approaches by getting higher F-measure and Recall rates, which demonstrates that our method has efficiently found more matched modules from the PPI network than others. In detail, we have the following observations: 1) FGENS gets higher F-measure and Recall than the methods MCODE, CFinder and RRW which just use the PPI network topology and almost achieves the Precision of CFinder. Furthermore, FGENS obtains better results in all three evaluation criteria than the Flownet method which combines GO similarity information. All these results demonstrate that the integration of multiple information sources greatly benefits the functional module detection. 2) FGENS gets better performance than the other cluster ensemble methods, i.e., INENS and CLENS. This indicates that our graph based cluster ensemble method is more effective than other ensemble methods and gets more consensus and meaningful modules. 3) By adding the additional information of proteins and clusters to the bipartite graph, our graph model captures more precise belonging relationships than BGENS according to the higher rates of the three evaluation criteria.

### 3.3 Analysis of Ensemble Power

To evaluate the superior performance of the graph based cluster ensemble method, we compared the ensemble result with the results of the single clustering partitions in detail. The results for single clustering partitions are denoted as GoKm, CoexKm (for K-means), GOHie, CoexHier (for Hierarchical clustering) and GOMCL, CoexMCL (for MCL) in Table 1. Note that FGENS gets higher Precision, Recall and F-measure than all the single clustering partitions, proving that the integration of multiple sources has great power to predict protein functional modules in PPI networks. We extracted all the matched modules from the clustering methods and counted the number of proteins they covered and the modules they detected. The results are shown in Table 2. FGENS retrieves a larger number of matched modules and covers more proteins than all these base clustering methods. Also, the results of FGENS are slightly better than BGENS which has omitted the original protein interactions and the clusters’ relationships.

### 4. CONCLUSIONS

This paper has presented a graph-based cluster ensemble method integrating multiple data sources to identify protein functional modules from PPI networks. We applied our method on the TAP data and evaluated the results with the hand-curated complexes from CYC2008 benchmark dataset.

From the experimental results we see that our cluster ensemble framework is an efficient and effective method to integrate different biological information sources for identifying more stable and balanced functional modules. The diversity of multiple biological information sources and the clustering methods have enforced our results rather than pulled it down.

### 5. ACKNOWLEDGMENTS

The research work is supported by National Natural Science Foundation of China under Grant No. 30970780 and Ph.D. Programs Foundation of Ministry of Education of China under Grant No. 2009110310005.

### 6. REFERENCES


