An Efficient Algorithm for Protein-Protein Interaction Network Analysis to Discover Overlapping Functional Modules

Ying Liu

1Department of Computer Science, Mathematics and Science, College of Professional Studies, St. John’s University, Queens, NY 11439

Abstract - One of the most pressing problems of the post genomic era is identifying protein functions. Clustering Protein-Protein-Interaction networks is a systems biological approach to this problem. Traditional Graph Clustering Methods are crisp, and allow only membership of each node in at most one cluster. However, most real world networks contain overlapping clusters. Recently the need for scalable, accurate and efficient overlapping graph clustering methods has been recognized and various soft (overlapping) graph clustering methods have been proposed. Previously we proposed an efficient, novel, and fast overlapping clustering method is proposed based on purifying and filtering the coupling matrix (PFC). In this paper, we further improved PFC. The experimental results show that the improved PFC method outperforms many existing methods by a few orders of magnitude in terms of average statistical (hypergeometrical) confidence regarding biological enrichment of the identified clusters.

Keywords: Protein-Protein Interaction networks; Graph Clustering; Overlapping functional modules; Coupling Matrix; Systems biology

1 Introduction

Homology based approaches have been the traditional bioinformatics approach to the problem of protein function identification. Variations of tools like BLAST [1] and Clustal [2] and concepts like COGs (Clusters of orthologous Groups) [3] have been applied to infer the function of a protein or the encoding gene from the known a closely related gene or protein in a closely related species. Although very useful, this approach has some serious limitations. For many proteins, no characterized homologs exist. Furthermore, form does not always determine function, and the closest hit returned by heuristic oriented sequence alignment tools is not always the closest relative or the best functional counterpart. Phenomena like Horizontal Gene Transfer complicate matters additionally. Last but not least, most biological Functions are achieved by collaboration of many different proteins and a proteins function is often context sensitive, depending on presence or absence of certain interaction partners.

A Systems Biology Approach to the problem aims at identifying functional modules (groups of closely cooperating and physically interacting cellular components that achieve a common biological function) or protein complexes by identifying network communities (groups of densely connected nodes in PPI networks). This involves clustering of PPI-networks as a main step. Once communities are detected, a hypergeometrical p-value is computed for each cluster and each biological function to evaluate the biological relevance of the clusters. Research on network clustering has focused for the most part on crisp clustering. However, many real world functional modules overlap. The present paper introduces a new simple soft clustering method for which the biological enrichment of the identified clusters seem to have in average somewhat better confidence values than current soft clustering methods.

2 Previous Work

Examples for crisp clustering methods include HCS [4], RNSC [5] and SPC [6]. More recently, soft or overlapping network clustering methods have evolved. The importance of soft clustering methods was first discussed in [7], the same group of authors also developed one of the first soft clustering algorithms for soft clustering, Clique Percolation Method or CPM [8]. An implementation of CPM, called CFinder [9] is available online. The CPM approach is basically based on the “defective cliques” idea and has received some much deserved attention. Another soft clustering tool is Chinese Whisper [10] with origins in Natural Language Processing. According to its author, CW can be seen as a special case of the Random Walks based method Markov-Chain-Clustering (MCL) [11] with an aggressive pruning strategy.

Recently, some authors [12, 13] have proposed and implemented betweenness based [14] Clustering (NG) method, which makes NG’s divisive hierarchical approach capable of identifying overlapping clusters. NG’s method finds communities by edge removal. The modifications involve node
removal or node splitting. The decisions about which edges to remove and which nodes to split are based on iterated all pair shortest path calculations.

In this paper, we present a new approach, called PFC, which is based on the notion of Coupling matrix (or common neighbors). In the rest of the paper, we first describe PFC and compare its results with the best results achieved by the aforementioned soft approaches. The second part of this work aims to illustrate the biological relevance of soft methods by giving several examples of how the biological functions of overlap nodes relate to biological functions of respective clusters.

3 PFC Method

The method introduced here is based on the purification and filtering of coupling matrix, PFC. PFC is a soft graph clustering method that involves only a few matrix multiplications/manipulation. Our experimental results show that it outperforms the above mentioned methods in terms of the p-values for MIPS functional enrichment [15] of the identified clusters. The PPI networks we used in the paper are yeast PPI networks (4873 proteins and 17200 interactions).

Liu and Foroushani [16] proposed a PFC filtering by simple, local criteria. In this paper, we propose a new PFC approach, filtering by corroboration.

3.1 Filtering by Simple, Local Criteria

The first Filtering approach is motivated by assumptions about the nature of the data and size of the target clusters. PPI data are for the most part results of high throughput experiments like yeast two hybrid and are known to contain many false positive and many false negative entries. For certain, more thoroughly studied parts of the network, additional data might be available from small scale, more accurate experiments. In PFC, the emphasis lies on common second degree neighbors and this can magnify the effects of noise. Under the assumption that Nodes with low degree belong in general to the less thoroughly examined parts of the network, it is conceivable that the current data for the graph around these low nodes contains many missing links. Missing links in these areas can have dramatic effects on the constellation of second degree neighbors. This means the Coupling data for low degree nodes is particularly unreliable. On the other hand, many extremely well connected nodes are known to be central hubs that in general help to connect many nodes of very different functionality with each other, hence, their second degree neighbors compromise huge sets that are less likely to be all functionally related. Additionally, it has been shown that most functional modules are meso-scale [6]. There are also some fundamental physical constrains on the size and shape of a protein complex that make very large modules unlikely. Taking these considerations into account, a filter is easily constructed by the following rules:

- Discard all clusters (rows of purified coupling matrix) where the labeling node (the _th node in the _th row) has a particularly low (<14) or particularly high (>30) degree. Discard all clusters where the module size is too small (<35) or particularly large (>65).

The selected minimum and maximum values for degree of labeling nodes and module size are heuristically motivated. The intervals can be easily changed to obtain or discard more clusters, but the enrichment results for these intervals seem reasonably good. The peak log value for the enrichment of selected clusters is at -91.00 and the average lies at -18.99.

Using this filter, by clustering yeast PPI networks, PFC yields 151 clusters from 52 different Functional categories. Figure 1 gives an example.

3.2 Filtering by Corroboration

Filtering by local criteria gives impressing results but it does not guarantee that a few of the remaining clusters do not overlap in majority of their elements. Although PFC is an overlapping clustering algorithm, very large overlaps between clusters are bound to indicate presence of redundant clusters. At the same time, repeated concurrence of large groups of proteins in different rows does reinforce the hypothesis that these groups are indeed closely related, and that the corresponding rows represent a high quality cluster. These observations can be used to construct an alternative filter that removes both low quality and redundant clusters from the coupling matrix. The main idea is that a line A is corroborated...
by a Line B if the majority of nonzero elements in A are also nonzero in B. The following summarizes this filter:

Given the sparse nature of the involved matrices, this Corroboration based filter can be implemented very efficiently in Matlab. It discards by design redundant clusters (out-degree>0 in the confirmation graph indicates that there is a similar cluster with a higher rank) and retains only high quality clusters (clusters with a high in-degree in the confirmation graph have been confirmed by presence of many other clusters with similar structure). The ranking by row sum helps consolidate and summarize relevant parts of smaller clusters into larger ones. Figure 2 gives two examples of clusters selected by this approach on Yeast-PPI network.

Figure 1: Two of the clusters selected by PFC2. The left Figure shows the selected community for the row labeled “YDR335w” in the purified coupling matrix. Out of the 35 proteins in this community, 29 belong to MIPS Funcat 20.09.01 (nuclear transport). The right Figure shows the selected community for the row labeled “YKL173w” in the purified coupling matrix. It is one of the clustered selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01 (Splicing).

4 Experimental Results and Discussions

The results of the PFC are compared with results obtained by other soft clustering methods. A PPI network of yeast with 4873 Nodes and 17200 edges is used as the test data set. The other methods are an in-house implementation of Pinney and Westhead’s Betweenness Based proposal [12], Chinese Whisper [10], CPM as implemented in C-Finder [9]. Whenever other methods needed additional input parameters, we tried to choose parameters that gave the best values. The results from different methods are summarized in Table 1.

4.1 Biological Functions of Overlap Nodes

The hypergeometric evaluation of individual clusters is the main pillar in assessing the quality of crisp clustering methods. For soft clustering methods, further interesting questions arise that deal with relationships between clusters. A possible conceptual disadvantage, production of widely overlapping, redundant clusters was addressed in previous sections. Figure 2 and Figure 3 are clustering results of the PFC. The result demonstrates an important advantage of soft methods against crisp ones: They show how soft clustering can adequately mirror the fact that many proteins have context dependent functions, and how in some cases overlap nodes can act as functional bridges between different modules.

Table 1 Comparison of results from different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Cluster Count</th>
<th>Average Cluster Size</th>
<th>Average Enrichment</th>
<th>Network Coverage</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness based</td>
<td>20</td>
<td>302.70</td>
<td>-15.11</td>
<td>0.58</td>
<td>19/20</td>
</tr>
<tr>
<td>Chinese Whisper</td>
<td>38</td>
<td>23.45</td>
<td>-12.11</td>
<td>0.17</td>
<td>32/38</td>
</tr>
<tr>
<td>C-Finder</td>
<td>68</td>
<td>14.50</td>
<td>-15.70</td>
<td>0.19</td>
<td>48/68</td>
</tr>
<tr>
<td>PFC1</td>
<td>183</td>
<td>44.76</td>
<td>-19.35</td>
<td>0.31</td>
<td>55/183</td>
</tr>
<tr>
<td>PFC2</td>
<td>40</td>
<td>25.4</td>
<td>-19.40</td>
<td>0.17</td>
<td>36/40</td>
</tr>
</tbody>
</table>
result #1: The dominant function for the left module is translation initiation (10 out of 31) for the right module, it is nuclear mRNA splicing (27 out of 33); both overlap nodes are involved in translation initiation and Protein-RNA complex assembly.

result #2. There is a relatively large overlap (yellow nodes). All 10 overlap nodes are involved in “nuclear mRNA splicing, via spliceosome-A”. The same is true for ca.25% (12 out of 45) of the green nodes to the left and 68% (17 out of 25) of the green nodes to the right of the overlap. Furthermore, two of the overlap nodes are also involved in spliceosome assembly the total number of such nodes in the entire network is 19.

5 Conclusions

This paper introduced PFC, a new clustering concept based on purification and filtering of a coupling (common neighbor) matrix. It discussed a very different filtering method. PFC consists of only a few matrix multiplications and manipulations and is therefore very efficient. The PFC outperforms current soft clustering methods on PPI networks by a few orders of magnitude in terms of average statistical confidence on biological enrichment of the identified clusters. The paper illustrated the importance of soft clustering methods in systems biology by giving a few concrete examples of how the biological function of the overlap nodes relates to the functions of the respective clusters.

6 References


