A Soft Graph Clustering Method for Protein-Protein Interaction Network Analysis

Ying Liu
Division of Computer Science, Mathematics and Science
St. John's University
Queens, NY 11349
Email: liuy1 [AT] stjohns.edu

***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. In this paper, an efficient, novel, and fast overlapping clustering method is proposed based on purifying and filtering the coupling matrix (PFC). PFC is tested on PPI networks. The experimental results show that 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, Purifying and Filtering the Coupling Matrix, Overlapping Clusters, Functional Modules

I. 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.

II. 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.

III. 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 net works we used in the paper are yeast PPI networks (4873 proteins and 17200 interactions).

A. Coupling Matrix

Bibliographical coupling is an idea from text classification: If two documents (for example two scientific papers) share a significant number of cited references, they are likely to deal with similar topics. A coupling matrix in a network describes the number of shared neighbors (or paths of length two) for each node pair. For undirected graphs like PPI networks, this matrix is symmetric and can be easily obtained from the original adjacency matrix A by: $B = A * A$. Notably, for second degree neighbors, the entry in coupling matrix is nonzero, even if there is no edge between the nodes. The importance of second degree neighbors in PPI networks has been emphasized before in the literature. For example: [16] note that "A substantial number of proteins are observed to share functions with level-2 neighbors but not with level-1 neighbors."

B. Purification of the Coupling Matrix

Adjacency matrices of biological networks are in general very sparse. The coupling matrix described above is slightly denser. However, not all nonzero-values are equally valuable. In the purification step, we determine the number of nonzero values (in unweighted graphs like PPI-Networks, this corresponds to the row sum), the maximum entry and the minimum non-zero value for each line of the coupling matrix. Rows in which the minimum nonzero entry and the maximum value are relatively close are considered homogenous and left unchanged. For other rows, we delete nonzero entries that don't make a significant contribution to the row sum. The Purification Process is summarized below:

FOREACH row i of the Coupling Matrix B
IF $min(B(i,:)) < \lfloor max(B(i,:)) * \alpha) \rfloor$
THEN $B(i,:) = \lfloor B(i,:) ./ (Bavg\ (i) * \beta) \rfloor$
Where: "./" is the Matlab cell wise division operator, $\lfloor\ \rfloor$ is the basic floor operation and α and β are values less than and greater than 1 respectively.

This purification step is robust in regard to choice of values for its parameters. In particular in our experiment with a yeast PPI network, the results for $\alpha = 0.8$ and $\beta = 1.2$ did not differ from those for $\alpha = 0.7$ and $\beta = 1.3$.

C. Filtering of the purified coupling matrix

The set of nonzero entries in each line of the Purified Coupling matrix can be considered as a candidate cluster. For a network of n nodes, this generally means n candidate clusters. However, not all rows are equally interesting. The set of nonzero entries (the information content) of many rows is likely to be very similar to, or contained largely within the sets of nonzero entries of other rows. This means that many rows are likely to represent spurious or redundant clusters. In the filtering step, we address this problem and try to select the most relevant and interesting rows of the purified coupling matrix. The set of nonzero entries in each of the selected lines of the purified coupling matrix represent our final clusters. The filtering step of PFC is a flexible step. Two alternative filtering approaches are discussed below.

D. 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 [Spirin, V. 2003]. 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 *i*th node in the *i*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.

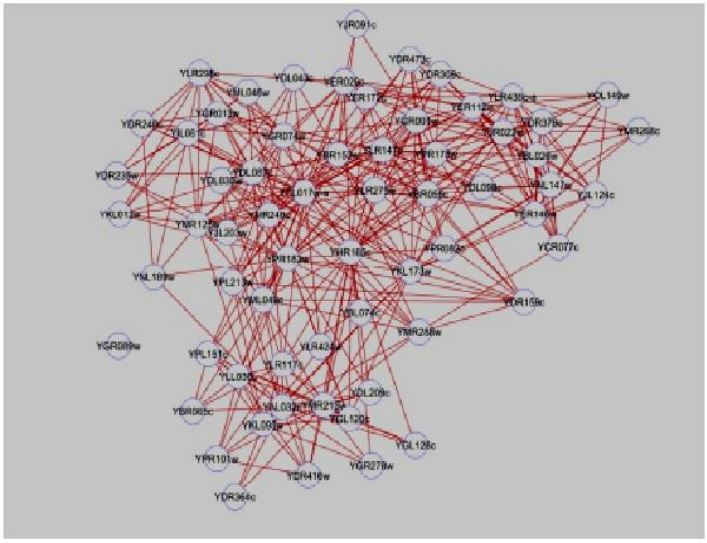

Fig. 1. This Figure shows the community for the row labeled "YKL173w" in the purified coupling matrix of yeat PPI network. It is one of the clustered selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01.

E. 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 Binary version of the Purified Coupling Matrix B
Calculate Overlap Matrix $O = B * B$
Normalize $O(i, j)$ by Size of Module j
Calculate Corroboration Matrix $C = \lfloor O ./ \alpha \rfloor$
Where: $0.5 < \alpha \leq 1$; and "./" is the Matlab cellwise division.
Calculate Common Corroborator Matrix C Com$= C * C'$
Rank the rows of Ccom by the sum of their entries
Interpret Ccom as description of a directed Confirmation graph between clusters, where the direction of confirmation is from lower ranked to higher ranked rows.
Select clusters whose in-degree in the confirmation graph is higher than a threshold and whose out degree is 0.

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.

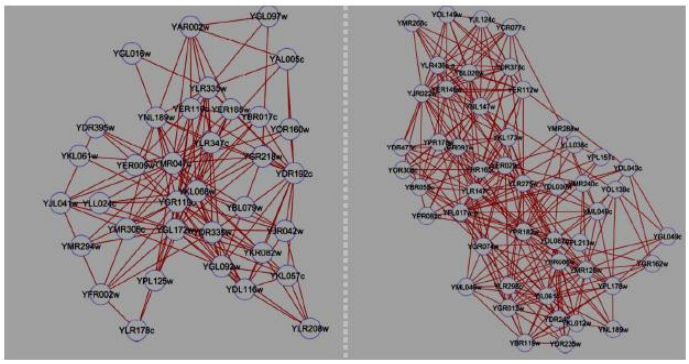

Fig. 2. Two of the clusters selected by PFC2. The left Figure shows the selected community for the row labeled "YDR335w" in the purified coupling matrix. Out of the 35 proteins in this community, 29 belong to MIPS Funcat 20.09.01(nuclear transport). The right Figure shows the selected community for the 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).

IV. Discussion Experimental Results and Discussions

The results for two versions of PFC (PFC1: using the local criteria filter and PFC2: using the corroboration based filter) 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] as available from its author's webpage, 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 **Error! Reference source not found.**.

Table 1. Comparison of results from different methods.

Method	Cluster Count	Average Cluster Size	Average Enrichment	Network Coverage	Diversity
Betweenness based	20	302.70	-15.11	0.58	19/20
Chinese Whisper	38	23.45	-12.11	0.17	32/38
C Finder	68	14.50	-15.70	0.19	48/68
PFC 2	40	25.4	-19.40	0.17	36/40
PFC 1	183	44.76	-19.35	0.31	55/183

. To ensure comparability of results from different methods, for each pair of methods it was determined which functional categories are covered by clusters from both methods. For example, Clusters obtained by Chinese Whisper fell into 32 different MIPS-Funcat categories and those found by C-Finder fell into 48 different categories, 18 of which were also among the 32 categories associated with results from Chinese Whisper. We then compared for each method pair and each common category the best enrichment results.

Table 2. Pair-wise Overlaps between Top Rated Functional Categories of all Methods

Method	**BW**	**CW**	**C-**	**PFC1**	**PFC2**
BW	19	6	12	12	11
CW	6	32	18	18	16
C-Finder	12	18	48	25	28
PFC1	12	18	25	55	22
PFC2	11	16	28	22	36

Table 3 In-depth comparison of performances of PFC1 and C-Finder by MIPS category. Lines where PFC outperforms C-Finder are in bold.

MIPS-Funcat	**Best Cluster found by PFC 1 in this category**	**Best Cluster found by C-Finder for this category**
01.03.16.01	-8.8964	-18.8542
10.01	**-15.6145**	**-5.0963**
10.01.05.01	-7.2082	-7.6082
10.01.09.05	**-12.2059**	**-8.0998**
10.03.04	**-10.7738**	**-10.2305**
11.02.01	**-31.0417**	**-17.072**
11.02.03.01	**-33.0918**	**-14.9017**
11.02.03.04	**-39.7314**	**-15.5374**
11.04.01	-47.6756	-74.1431
11.04.03	**-62.9351**	**-60.2463**
11.04.03.01	**-91.0033**	**-30.6045**
11.04.03.05	-28.9036	-30.4952
12.04.01	**-28.4907**	**-18.4372**
14.01	-7.8672	-8.8271
14.07.04	**-25.0625**	**-9.955**
14.13	-35.6546	-41.4309
14.13.01.01	**-40.2921**	**-12.8711**
16.19.01	-7.6759	-12.3029
20	-10.4059	-13.7993
20.09	-5.6343	-7.8097
20.09.01	**-51.1643**	**-39.0378**
20.09.07	**-23.2574**	**-22.8797**
20.09.07.03	**-32.8137**	**-19.041**
34.01.01.03	**-18.8907**	**-18.7371**
42.04.03	**-32.7875**	**-26.508**

V. 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. Figures 3-6 are the clustering results of two versions of PFC (PFC1: using the local criteria filter and PFC2: using the corroboration based filter). The results demonstrate 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.

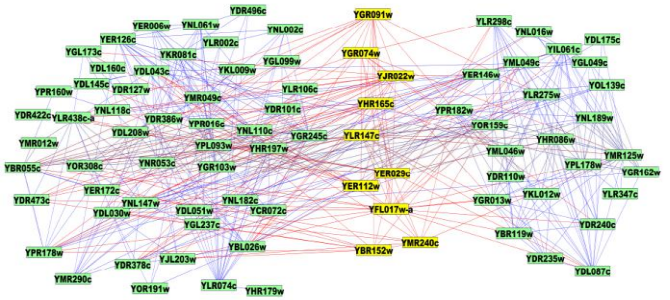

Fig. 3. PFC1 result #1: There is a relatively large overlap (yellow nodes). All 10 overlap nodes are involved in "nuclear mRNA splicing, via spliceosome-A". The same is true for ca.25% (12 out of 45) of the green nodes to the left and 68% (17 out of 25) of the green nodes to the right of the overlap. Furthermore, two of the overlap nodes are also involved in spliceosome assembly the total number of such nodes in the entire network is 19.

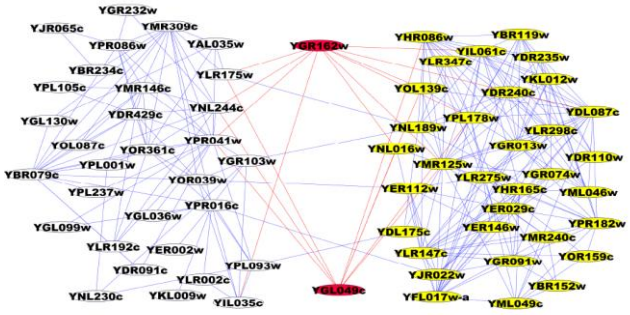

Fig. 4. PFC1 result #2: The dominant function for the left module is translation initiation (10 out of 31) for the right module, it is nuclear mRNA splicing (27 out of 33); both overlap nodes are involved in translation initiation and Protein-RNA complex assembly.

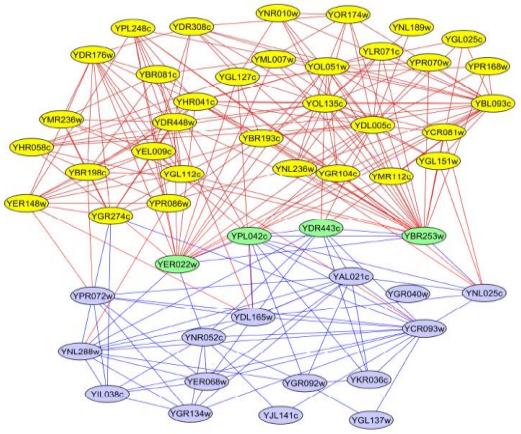

Fig. 5. PFC2 result #1: The main functions of the top and bottom clusters are identical: on both sides, over 80% of the nodes are involved in "transcription from RNA polymerase II promoter" and this is also the main function of all of the overlap nodes. However, the bottom part also contains a specialized module for poly tail shortening: all 7 node in the entire network that are involved in poly tail shortening are gathered here

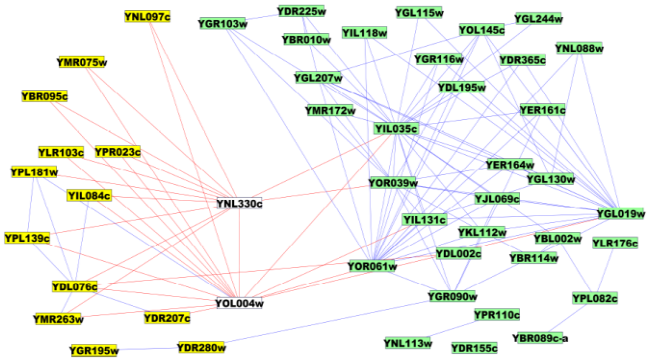

Fig. 6. PFC2 result #2.10 out of 13 yellow nodes are involved in histone deacytlation(left), 21 out of 33 green nodes are involved in transcription, DNA dependent (right); both white nodes are involved in both functions.

VI. CONCLUSIONS

This paper introduced PFC, a new clustering concept based on purification and filtering of a coupling (common neighbor) matrix. It discussed two very different filtering methods resulting in two flavors of PFC. PFC consists of only a few matrix multiplications and manipulations and is therefore very efficient. Both flavors of PFC seem to outperform 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 he overlap nodes relates to the functions of the respective clusters.

REFERENCES

[1] Altschul, SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, no. 17: 3389.

[2] Thompson, JD, DG Higgins, and TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22, no. 22: 4673-4680.

[3] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. *Science* 278, no. 5338: 631.

[4] Hartuv, E., R. Shamir. 2000. A clustering algorithm based on graph connectivity. *Information processing letters* 76, no. 4-6: 175-181.

[5] King, A. D., N. Przulj, and I. Jurisica. 2004. Protein complex prediction via cost-based clustering. *Bioinformatics (Oxford, England)* 20, no. 17 (Nov 22) : 3013-3020. PMID: 15180928; bth351 [pii].

[6] Spirin, V., L. A. Mirny. 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences* 100, no. 21: 12123-12128.

[7] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, no. 7043 (Jun 9) : 814-818. PMID: 15944704; nature03607 [pii].

[8] Derenyi, I., et al. 2005. Clique percolation in random networks. *Physical Review Letters* 94, no. 16: 160202.

[9] Adamcsek, B., G. et al. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, no. 8: 1021-1023.

[10] Biemann, C. 2006. Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the HLT-NAACL-06 workshop on textgraphs-06, new york, USA*

[11] Van Dongen, S. 2000. A cluster algorithm for graphs. *Report-Information systems* , no. 10: 1-40.

[12] Pinney, J. W., D. R. Westhead. 2006. Betweenness-based decomposition methods for social and biological networks. In *Interdisciplinary statistics and bioinformatics.* Edited by S. Barber, P. D. Baxter, K. V. Mardia and R. E. Walls. Leeds University Press.

[13] Gregory, S. 2007. An algorithm to find overlapping community structure in networks. *Lecture Notes in Computer Science* 4702: 91.

[14] Girvan, M., M. E. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, no. 12 (Jun 11) : 7821-7826. PMID: 12060727; 99/12/7821 [pii].

[15] MIPS. *The functional catalogue (FunCat).* 2007. Internet on-line. Available from <http://mips.gsf.de/projects/funcat>. [10/2/2008, 2008].

[16] Chua, H. N., W. K. Sung, and L. Wong. 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics (Oxford, England)* 22, no. 13 (Jul 1) : 1623-1630. PMID: 16632496; btl145 [pii].