An Advance Technique in Graph Clustering For Protein To Protein Interaction Networks

K L V G K Murthy[1], Dr. R. J. Rama Sree[2]
[1, 2] Dept of CSE
[2]Professor&Head, Dept of Computer Science
[1]St. Marys Group of Institutions, Guntur, Research Scholar of Rayalaseema University, Kurnool.
[2]Rashtriya Sanskrit Vidya Peeth, Tirupathi, Research supervisor for Rayalaseema University, Kurnool.

Abstract- *Protein-protein interaction plays key role in predicting the protein function of target protein and drug ability of molecules. The majority of genes and proteins realize resulting phenotype functions as a set of interactions one of the most pressing issues of the post-genomic era is characteristic protein functions. Clustering Protein-Protein-Interaction networks may be a systems biological approach to the current problem. Traditional Graph clustering ways are crisp and allow only members of every node in at the most one cluster. However, most real-world networks contain overlapping clusters. Recently the requirement for scalable, correct and efficient overlapping graph clustering ways has been recognized and numerous soft (overlapping) graph clustering strategies are proposed. During this paper, AN efficient, novel, and quick overlapping clustering methodology is proposed based on purifying and filtering the coupling matrix (PFC). PFC is tested on PPI networks. The experimental results show that PFC methodology outperforms several existing ways by a number of orders of magnitude in terms of average statistical (hyper geometrical) confidence regarding biological enrichment of the identified clusters.*

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

I. INTRODUCTION

Protein–protein interactions (PPIs) are the physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by electrostatic forces including the hydrophobic effect. Many are physical contacts with molecular associations between chains that occur in a cell or in a living organism in a specific bimolecular context.[1]

Proteins rarely act alone as their functions tend to be regulated. Many molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their PPIs. These interactions make up the so-called interactomics of the organism, while aberrant PPIs are the basis of multiple aggregation-related diseases, such as Creutzfeldt–Jakob, Alzheimer's disease, and may lead to cancer..

PPIs have been studied from different perspectives: biochemistry, quantum chemistry, molecular dynamics, signal transduction, among others.[2] All this information enables the creation of large protein interaction networks – similar to metabolic or genetic/epigenetic networks – that empower the current knowledge on biochemical cascades and molecular etiology of disease, as well as the discovery of putative protein targets of therapeutic interest.

To describe the types of protein–protein interactions (PPIs) it is important to consider that proteins can interact in a "transient" way (to produce some specific effect in a short time) or to interact with other proteins in a "stable" way to build multi-protein complexes that are molecular machines within the living systems. A protein complex assembly can result in the formation of homo-oligomeric or hetero-oligomeric complexes. In addition to the conventional complexes, as enzyme-inhibitor and antibody-antigen, interactions can also be established between domain-domain and domain-peptide. Another important distinction to identify protein-protein interactions is the way they have been determined, since there are techniques that measure direct physical interactions between protein pairs, named “binary” methods, while there are other techniques that measure physical interactions among groups of proteins, without pair wise determination of protein partners, named “co-complex” methods.[1]

Homo-oligomers vs. hetero-oligomers

Homo-oligomers are macromolecular complexes constituted by only one type of protein subunit. Protein subunits assembly is guided by the establishment of non-covalent interactions in the quaternary structure of the protein. Disruption of homo-oligomers in order to return to the initial individual monomers often requires denaturation of the

complex.[5] Several enzymes, carrier proteins, scaffolding proteins, and transcriptional regulatory factors carry out their functions as homo-oligomers. Distinct protein subunits interact in hetero-oligomers, which are essential to control several cellular functions. The importance of the communication between heterologous proteins is even more evident during cell signaling events and such interactions are only possible due to structural domains within the proteins (as described below).

Stable interactions vs. transient interactions

Stable interactions involve proteins that interact for a long time, taking part of permanent complexes as subunits, in order to carry out structural or functional roles. These are usually the case of homo-oligomers (e.g. cytochrome c), and some hetero-oligomeric proteins, as the subunits of ATPase. On the other hand, a protein may interact briefly and in a reversible manner with other proteins in only certain cellular contexts – cell type, cell cycle stage, external factors, presence of other binding proteins, etc. – as it happens with most of the proteins involved in biochemical cascades. These are called transient interactions. For example, some G protein-coupled receptors only transiently bind to $G_{i/o}$ proteins when they are activated by extracellular ligands,[6] while some G_q-coupled receptors, such as muscarinic receptor M3, pre-couple with G_q proteins prior to the receptor-ligand binding.[7]Interactions between intrinsically disordered protein regions to globular protein domains (i.e. MoRFs) are transient interactions.[8]

Covalent vs. non-covalent

Main articles: Covalent bond and Non-covalent interactions

Covalent interactions are those with the strongest association and are formed by disulphide bonds or electron sharing. Although being rare, these interactions are determinant in some posttranslational modifications, as ubiquitination and SUMOylation. Non-covalent bonds are usually established during transient interactions by the combination of weaker bonds, such as hydrogen bonds, ionic interactions, Van der Waals forces, or hydrophobic bonds.[9]

Role of water

Water molecules play a significant role in the interactions between proteins.[10][11] The crystal structures of complexes, obtained at high resolution from different but homologous proteins, have shown that some interface water molecules are conserved between homologous complexes. The majority of the interface water molecules make hydrogen bonds with both partners of each complex. Some interface amino acid residues or atomic groups of one protein partner engage in both direct and water mediated interactions with the other protein partner. Doubly indirect interactions, mediated by two water molecules, are more numerous in the homologous complexes of low affinity.[12] Carefully conducted mutagenesis experiments, e.g. changing a tyrosine residue into a phenylalanine, have shown that water mediated interactions can contribute to the energy of interaction.[13] Thus, water molecules may facilitate the interactions and cross-recognitions between proteins.

Protein-protein interactions (PPIs) handle a wide range of biological processes, including cell-to-cell interactions and metabolic and developmental control [1]. Protein-protein interaction is becoming one of the major objectives of system biology. Noncovalent contacts between the residue side chains are the basis for protein folding, protein assembly, and PPI [2]. These contacts induce a variety of interactions and associations among the proteins. Based on their contrasting structural and functional characteristics, PPIs can be classified in several ways [3]. On the basis of their interaction surface, they may be homo- or heterooligomeric; as judged by their stability, they may be obligate or nonobligate; as measured by their persistence, they may be transient or permanent [4]. A given PPI may be a combination of these three specific pairs. The transient interactions would form signaling pathways while permanent interactions will form a stable protein complex.

Homology based approaches have been the traditional bioinformatics approach to the problem of protein function identification. Variations of tools like BLAST [1] and Clustal and concepts like COGs (Clusters of orthologous Groups) 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, hyper geometrical 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).

3.1 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 obs erved to share functions with level-2 neighbors but not with level-1 neighbors."

3.2 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 $\beta = 1.3$.

3.3 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 _ nodes, this generally means _ 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.

3.4 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.

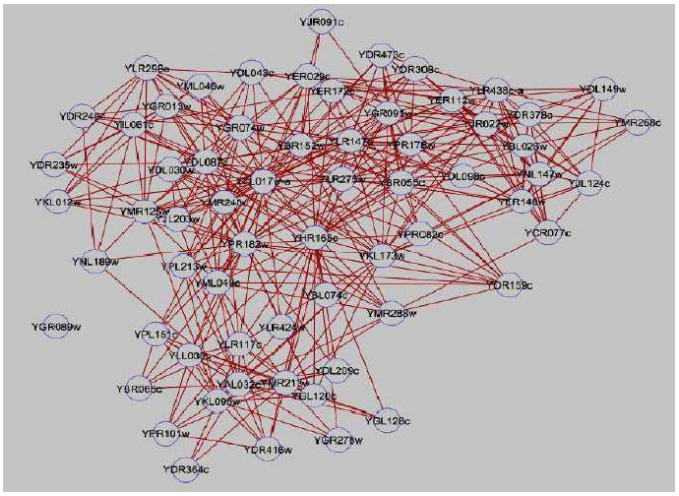

Figure 1 This Figure shows the community for the row labeled "YKL173w" in the purified coupling matrix o f yeat PPI network. It is one of the clustered selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01.

IV. 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 hyper geometric 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 is a 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

Method	Cluster Count	Average Cluster Size	Average Enrichment	Network Coverage	Diversity
Between's 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	183	44.76	-19.35	0.31	55/183

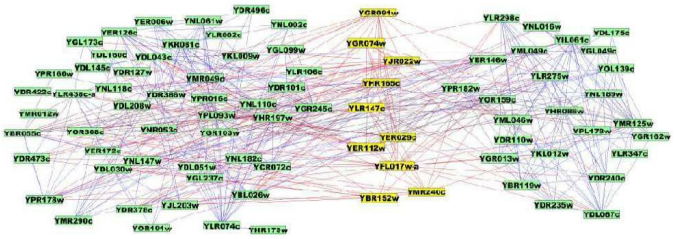

Figure 2. Result #1: There is a relatively large overlap (yellow nodes). All 10 overlap nodes are involved in "nucle ar mRNA splicing, via splice some-A". The same is true for c a.25% (12 out of 45) of the green nodes to the left and 68% (17 out ofOf the green nodes to the right of the overlap. Furthermore, two of the overlap nodes are also involved in splice some assembly the total number of such nodes in the entire network is 19.

V. 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.

REFERENCES

[1] Liu, Ying & Foroushani, Amir. (2011). An Efficient Soft Graph Clustering Method for PPI Networks based on Purifying and Filtering the Coupling Matrix. .

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

[3] Thompson, JD, DG Higgins, and TJ Gibson. "CLUST ALW: 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, 1994

[4] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. "A genomic perspective on protein families". Science 2 78, no. 5338: 631, 1997.

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

[6] King, A. D., N. Przulj, and I. Jurisica. "Prote in complex prediction via cost-based clustering". Bioinformati cs 20,: 3013-3020, 2004.

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

[8] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". Nature 435, no. 70 43 (Jun 9): 814-818, 2005.

[9] Derenyi, I., et al. "Clique percolation in rand om networks". Physical Review Letters 94, no. 16: 1602 02, 2005.

[10] Adamcsek, B., G. et al. "CFinder: locating cliq ues and overlapping modules in biological networks". Bioinf ormatics 22, no. 8: 1021-1023, 2006.

[11] Biemann, C. "Chinese whispers-an efficient gra ph clustering algorithm and its application to natural language processing problems". In Proceedings of the HLT-NAA CL-06 workshop on textgraphs-06, new york, USA, 2006.

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

[13] Pinney, J. W., D. R. Westhead. "Betweenness-ba sed 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, 2000.

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

[15] Girvan, M., M. E. Newman. "Community structure in social and biological networks". PNAS 99: 7821-7826 , 2002.

[16] Chua, H. N. et al. "Exploiting indirect neighb ours and topological weight to predict protein function from

protein-protein interactions". Bioinformatics 22: 1623-1630 , 2006.

[17] MIPS. The functional catalogue (FunCat). 2007. <http://mips.gsf.de/projects/funcat>.