ISLAMIC UNIVERSITY OF TECHNOLOGY

Finding Protein Complexes in Protein Interaction Networks: An Extension of Homogeneous Decomposition Method

By:

Solaiman Shawon (094410)

Abir Mahmud Emon (094434)

Supervised by:

Tareque Mohmud Chowdhury

Assistant Professor

Department of Computer Science and Engineering

A thesis submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Computer Science and Engineering

Academic Year: 2012-2013

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

A Subsidiary Organ of

The Organization of Islamic Co-operation (OIC)

Dhaka, Bangladesh.

October 23, 2013

Declaration of Authorship

We, Solaiman Shawon & Abir Mahmud Emon, declare that this thesis titled,'Finding Protein Complexes in Protein Interaction Networks: An Extension of Homogeneous Decomposition Method' and the work presented in it are our own. We confirm that:

- This work was done wholly while in candidature for a Bachelor degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.

Submitted By:

Solaiman Shawon (094410)

Abir Mahmud Emon (094434)

Finding Protein Complexes in Protein Interaction Networks: An Extension of Homogeneous Decomposition Method

Approved By:

Prof. Dr. M.A. Mottalib
Head of the Department,
Department of Computer Science and Engineering,
Islamic University of Technology.

Tareque Mohmud Chowdhury
Thesis Supervisor,
Assistant Professor,
Department of Computer Science and Engineering,
Islamic University of Technology.

ISLAMIC UNIVERSITY OF TECHNOLOGY

Abstract

CSE
Department of Computer Science and Engineering

Bachelor of Science in Computer Science and Engineering

Finding Protein Complexes in Protein Interaction Networks: An Extension of Homogeneous Decomposition Method

by
Solaiman Shawon (094410)
Abir Mahmud Emon (094434)

Detecting protein complexes from protein-protein interaction (PPI) network is becoming a difficult challenge in computational biology.In the post-genome era, databases of protein-protein interactions are growing too fast that many well established algorithms can not perform well in analyzing these.New forms of interaction datasets are increasingly being found where many of well known alorithms can not be applied.

Identifying protein complexes and the way they share components appears as an essential step in describing cellular biology on a molecular basis.Graph Theoretic approaches are very popular in predicting protein complexes as they give a good insight into protein interaction networks in terms of connectivity and neighborhood.A novel graph decomposition method named Homogeneous Decomposition is very well performing in predicting protein complexes.It not only shows the members of the complexes but also can describe the relations between them.But it can not go further in some kind of networks with different scenario.We intend to investigate these limitations and propose new descriptions refining the current ones to overcome the limitations.

Acknowledgements

Special thanks to:

Prof. Dr. M.A. Mottalib

Head of the Department

Department of Computer Science and Engineering

Special thanks to:

Tareque Mohmud Chowdhury

Thesis Supervisor

Assistant Professor

Department of Computer Science and Engineering

Contents

List of Figures

Dedicated to our parents...

Chapter 1

Introduction

1.1 Biological Networks

A biological network is any network that applies to biological systems. A network is any system with sub-units that are linked into a whole, such as species units linked into a whole food web. Biological networks provide a mathematical analysis of connections found in ecological, evolutionary, and physiological studies, such as neural networks.

1.1.1 Network Biology and Bioinformatics

Complex biological systems may be represented and analyzed as computable networks. For example, ecosystems can be modeled as networks of interacting species or a protein can be modeled as a network of amino acids. Breaking a protein down farther, amino acids can be represented as a network of connected atoms, such as carbon, nitrogen, and oxygen. Nodes and edges are the basic components of a network. Nodes represent units in the network, while edges represent the interactions between the units. Nodes can represent a wide-array of biological units, from individual organisms to individual neurons in the brain.Two important properties of a network are degree and betweenness.Degree (or connectivity) is the number of edges that connect a node, while betweenness is a measure of how central a node is in a network. Nodes with high betweenness essentially serve as bridges between different portions of the network (i.e. interactions must pass through this node to reach other portions of the network). In social networks, nodes with high degree or high betweenness may play important roles in the overall composition of a network.

Bioinformatics has increasingly shifted its focus from individual genes, proteins, and search algorithms to large-scale networks often denoted as -omes such as biome, interactome, genome and proteome. Such theoretical studies have revealed that biological networks share many features with other networks such as the Internet or social networks, e.g. their network topology.

1.1.2 Networks in Biology

1.1.2.1 Protein-protein interaction networks

Many protein-protein interactions (PPIs) in a cell form protein interaction networks (PINs) where proteins are nodes and their interactions are edges. PINs are the most intensely analyzed networks in biology. There are dozens of PPI detection methods to identify such interactions. The yeast two-hybrid system is a commonly used experimental technique for the study of binary interactions. Recent studies have indicated conservation of molecular networks through deep evolutionary time.Moreover, it has been discovered that proteins with high degree of connectedness are more likely to be essential for survival than proteins with lesser degrees.This suggests that the overall composition of the network (not simply interactions between protein pairs) is important for the overall functioning of an organism.

1.1.2.2 Gene regulatory networks

The activity of genes is regulated by transcription factors, proteins that typically bind to DNA. Most transcription factors bind to multiple binding sites in a genome.As a result, all cells have complex gene regulatory networks. For instance, the human genome encodes on the order of 1,400 DNA-binding transcription factors that regulate the expression of more than 20,000 human genes.Technologies to study gene regulatory networks include ChIP-chip, ChIP-seq, CliP-seq, and others.

1.1.2.3 Metabolic networks

The chemical compounds of a living cell are connected by biochemical reactions which convert one compound into another. The reactions are catalyzed by enzymes. Thus, all compounds in a cell are parts of an intricate biochemical

network of reactions which is called metabolic network. It is possible to use network analyses to infer how selection acts on metabolic pathways.

1.1.2.4 Signaling networks

Signals are transduced within cells or in between cells and thus form complex signaling networks. For instance, in the MAPK/ERK pathway is transduced from the cell surface to the cell nucleus by a series of protein-protein interactions, phosphorylation reactions, and other events. Signaling networks typically integrate protein-protein interaction networks, gene regulatory networks, and metabolic networks.

1.2 Protein-Protein Interaction Networks

Proteins are involved in physical interactions with each other during two main processes: cellular signaling and complex assembly. In the first process, an extracellular signal or stimulus is transduced into the nucleus in order to initiate gene transcription. Signal transduction involves ordered sequence of biochemical reactions in which proteins activate other proteins usually by phosphorylation.The second process involving interactions is protein complex assembly, in which a set of proteins is assembled together to build a larger, more complex machine. Examples for such complexes are the RNA-polymerase and DNA-polymerase.

1.2.1 Interaction Discovery Methods

1.2.1.1 Yeast Two-Hybrid (Y2H)

The Yeast Two-Hybrid technique (Y2H) allows the detection of pair-wise protein interactions. Y2H exploits a modular property that is typical of many eukaryotic transcription factors. These factors can usually be decomposed into two distinct modules: one directly binds the DNA (BD, DNA-binding domain) and the other activates the transcription process (AD, transcriptional activating domain). The first component, BD, is able to bind the DNA even without the presence of the second component, AD. AD will activate transcription only if it is physically associated to a binding domain (BD). In the two-hybrid experiment, the two test

proteins (also referenced as the bait and the prey) are expressed as two fusion proteins (hybrids) with a DNA-binding domain (BD and the bait) and a transcriptional activating domain (AD and the prey respectively). Interaction between the bait and the prey proteins is identified via the expression of a reporter gene. A reporter gene is usually attached to another gene of interest, and holds a characteristic which can be easily identified in case it is activated. In the Y2H assay, LacZ is usually used as a reporter gene. Its expression causes blue color to appear and thus imply on an interaction.

1.2.1.2 Co-Immunoprecipitation (coIP)

The development of ultra-sensitive mass spectrometric techniques for protein identification has led to new interaction detection experimental procedures, other than Y2H. Immunoprecipitation is a technique of precipitating a protein antigen out of solution using an antibody that specifically binds to that particular protein. Protein complex immunoprecipitation (coIP) works by selecting an antibody that targets a known protein that is believed to be a member of a larger complex of proteins. By targeting this known member with an antibody it may become possible to pull the entire protein complex out of solution and thereby identify unknown members of the complex. Due to the difficulty in generating an antibody that specifically targets the known target proteins, the tagging method has been developed. Tagging proteins involves the fusion of tags either the 3 or the 5 ends of the protein of interest, and by that a single antibody (with this tag as its antigen) can be used for all target proteins.

1.2.2 Why are protein-protein interactions so important?

The binding of one signaling protein to another can have a number of consequences:

1. Such binding can serve to recruit a signaling protein to a location where it is activated and/or where it is needed to carry out its function.
2. The binding of one protein to another can induce conformational changes that affect activity or accessibility of additional binding domains, permitting additional protein interactions.

3. Imagine a cell in which, suddenly, the specific interactions between proteins would disappear. This unfortunate cell would become deaf and blind, paralytic and finally would disintegrate, because specific interactions are involved in almost any physiological process.

4. The study of protein-protein interactions has provided important insights into the functions of many of the known oncogenes, tumor suppressors, and DNA repair proteins.

5. Pharmacogenetic research has expanded to include the study of drug transporters, drug receptors, and drug targets.

1.3 Protein Complexes

A multiprotein complex (or protein complex) is a group of two or more associated polypeptide chains. If the different polypeptide chains contain different protein domain, the resulting multiprotein complex can have multiple catalytic functions. This is distinct from a multienzyme polypeptide, in which multiple catalytic domains are found in a single polypeptide chain.
Protein complexes are a form of quaternary structure. Proteins in a protein complex are linked by non-covalent proteinprotein interactions, and different protein complexes have different degrees of stability over time. These complexes are a cornerstone of many (if not most) biological processes and together they form various types of molecular machinery that perform a vast array of biological functions.
Many protein complexes are well understood, particularly in the model organism Saccharomyces cerevisiae (a strain of yeast). For this relatively simple organism, the study of protein complexes is now being performed genome wide and the elucidation of most protein complexes of the yeast is undergoing.

Chapter 2

Existing Methods

2.1 Individuation of Protein Complexes

A functional module is a group of cellular components, to which a specific biological function can be attributed. Consequently, molecular interaction networks can be organized in a set of modules of a small number of participants, which are low in interacting with other modules. A protein complex is a group of two or more associated proteins that interact by sharing the same biological goal. For example, the Breast Cancer Protein 1 (BRCA1) is known to participate inmultiple cellular processes by multiple protein complexes, such as in association with the BARD1 protein or with the Rad50, Mre11, Nbs1 proteins . Starting from a PPI network, complexes may be identified by searching for small and highly interconnected regions, called cliques. Predicted complexes can already be known, that is, their composition is known, or can denote a new protein complex. In this case, if the experiments confirm this relation, the algorithms can be used as predictors.

2.1.1 The Molecular Complex Detection Algorithm (MCODE)

The Molecular Complex Detection algorithm (MCODE), described in the early work of Bader takes in input an interaction network and tries to find complexes by building clusters. The rationale for MCODE is the separation of dense regions based on an ad hoc defined local density. MCODE has three main stages: (i) node weighting; (ii) complexes prediction; and (iii) post processing.
In its first stage, MCODE weights all vertices based on their local network density. The local area in which density is calculated is delimited by an ad hoc

defined subgraph structure called k-core. A k-core of a graph is the central most densely connected subgraph with minimal degree k. Thus, the core-clustering coefficient of a vertex v is the density of the highest k-core in the immediate neighborhood of v. Finally, the weight of a vertex is the product of the vertex core-clustering coefficient and the highest k-core level, kmax, of the immediate neighborhood of the vertex.
The resulting weighted graph is given as input to the second stage. Hence, the algorithm, starting from the highest weighted vertex, tries to span a region visiting vertices whose weight is above a certain threshold, called the vertex weight percentage (VWP). This stage stops when no more vertices can be added to the complex, and it is repeated by considering the next highest weighted network not already considered.
Finally, in the third stage, complexes are filtered when they do not contain at least a 2-core (i.e., a k-core with k = 2), that is, a graph of minimum degree equal to 2. The algorithm has two main options, fluff and haircut, that determine the characteristics of this phase.
The algorithm has two modes of execution: a direct mode in which the search starts from a given node and an undirect mode in which the seed is selected randomly.

2.1.2 The Markov Cluster Algorithm (MCL)

The Markov Cluster algorithm (MCL) finds clusters on a graph by simulating a stochastic flow and then analyzing its distribution. A network can be represented as a collection of paths sharing a starting point that guides a certain number of random walks. Observing the walks, we can see a particular behavior on the resulting flow: when a walker reaches a highly connected region, the walker will have small probability of getting out. In this way, considering the evolution of the flow, walks will reside in the regions with many edges, and walks linking highly connected regions will be more and more frequent.
So the MCL algorithm (i) simulates a collection of random walks within the network and (ii) iteratively weakens the flow where it is weak and increases the flow where it is strong (in the highly connected regions). This process will cause the apparition of a cluster structure, and will end when a set of regions with flow are separated by regions without flow.
This idea is implemented by building a stochastic matrix from the graph and then by simulating a flow with some algebraic operations. Formally, let us

consider a graph G and its adjacency matrix Mg, an associated Markov matrix is defined by normalizing all columns of Mg. Each value of this matrix represents the tendency of a node to be attracted by the other ones. Clearly, at the first step, each node is equally attracted by its neighbors. The evolution of the system, that is, of the flow, is computed by calculating the next power of this matrix. For any Markov matrix, the computation of successive powers causes a particular state in which each node is equally attracted from the others. However, the initially dense regions behave differently during the computation of the initial powers: nodes within dense regions are more attracted than the ones that are in the same region. The algorithm enhances this behavior with an operation, called 'inflation', that changes the matrix values in order to increase the probability of reaching a node in highly connected regions. The inflation operation, based on an inflation parameter greater than 1, influences the cluster structure; the greater the inflation parameter, the greater the number of clusters. Unlike the classical algorithms, the MCL does not suppose a defined cluster structure, that is, a fixed number of clusters.
Currently, MCL is implemented for Linux platforms and is freely available on the Internet.

2.1.3 Complex Prediction via Clustering

The work of Altaf-Ul-Amin presents another approach for finding complexes that is also based on the clustering of an interaction network. The rationale for the algorithm is the building of a cluster as a dense region embedded into a sparse region. The algorithm is organized logically in five major steps: (i) initialization; (ii) termination check; (iii) selection of a starting node; (iv) cluster growth; and (v) output.
In the first step the algorithm takes as input an undirected graph and initializes its main variables: cluster density, cluster property, and cluster ID. The algorithm calculates the minimum value of density for each generated cluster, that is, the ratio of the number of edges present in the cluster and the maximum possible number of edges in the cluster. The cluster property cpn,k of any node n, with respect to any cluster k of density dk and size _Nk_, is the ratio between (i) the total number of edges between the node n and each of the nodes of the cluster and (ii) the product between the density and the size of the cluster dk. The cluster identifier (ID) k is initialized to 1.

In the second step the algorithm verifies the termination conditions, and if the graph has no edges, the algorithm will end.
Conversely, if the termination check fails, the algorithm enters the third step, namely selection of starting node, selecting a node as a starting point to build a new cluster.
Hence, in the fourth step, namely, cluster growth, the algorithm adds nodes to the cluster chosen from the neighbors of the starting node. Neighbors are labeled in priority in order to guide the cluster formation.
Finally, when a cluster is generated, it is removed from the graph and the cluster ID k is incremented.
The algorithm is polynomial, and its complexity in the worst case is O(N3), where N is the number of nodes. This complexity is due to the cost of sorting clusters.

2.1.4 Complex Prediction via Restricted Neighborhod Search Clustering (RNSC)

The algorithm tries to find complexes by clustering protein interaction networks and filtering the generated clusters. In this scenario, clustering the network corresponds to its decomposition into different subsets of nodes inducing dense subgraphs. The algorithm, after an initial random clustering, uses the Restricted Neighborhood Search (RNSC), a cost-based local search algorithm based on the tabu heuristic.
The RNSC assigns a cost to each partition of nodes, and hence searches this space. The algorithm uses two cost functions, an integer-valued cost function (known as nave function) and a real-valued function (called a scaled function). The first function acts as a fast preprocessor and the second one assigns a more sophisticated cost on the initial network clustering, which can be random or userdefined.
Finally, the generated clusters are filtered. Small clusters are discarded for two reasons: (i) small known complexes generally have low density in known PPI networks; and (ii) the overlapping of a small predicted complex and a true complex has more probability of occuring by chance.
Hence, predicted clusters are matched to a known biological complex if the predicted cluster is entirely contained in the complex, or if it overlaps a large proportion of the complex.

2.1.5 Complex Identification through Chordal Graphs

The methods presented so far reside on the interpretation of a complex as a particular region in the graph, and this region is defined on the basis of an ad hoc density measure. These regions can be interpreted as functional modules of the graph, that is, a set of components interacting for targeting a specific goal. The translation of such a concept in a graph property is the rationale for different approaches.
These groups are represented by a particular clique (both maximal or defined ad hoc) on the graph. The proposed method is based on a possible decomposition for a chordal graph called a clique tree representation. Although not every PPI network is represented by a chordal graph, the authors developed a framework that generalizes this representation. The developed algorithm builds a so-called 'Tree of Complexes' whose nodes are complexes.

2.2 Limitations of these algorithms

All of the above mentioned methods for predicting protein complexes have the following common limitations:

1. They are based on the idea of finding dense subgraphs however, they differ in the definition of a dense subgraph and the procedure to cluster the nodes into dense subgraphs. In order to achieve a breakthrough, we need a deeper understanding of the proteins within these complexes.

2. They are unable to detect complexes which contain few proteins or few interactions.

3. They have not incorporated biological knowledge such as structural or evolutionarily relationships between proteins within the complexes.

4. The datasets applied to these problems have been derived from highthroughput experiments, which, in the case of PPI, are known to have both high false-positive and high false-negative rates.

Almost all of these limitations can be solved by applying Homogeneous Decomposition Method.

Chapter 3

Modular and Homogeneous Decomposition

3.1 Modular Decomposition

Under the modular decomposition model, a module M is a graph inside the given graph G so that all the vertices in M have exactly the same neighbors outside M (let us call that the neighborhood property). The aim of the modular decomposition is to decompose a graph into non-trivial modules (at least two), and then to iterate the decomposition process on the resulting modules until all modules are made of one vertex (such modules are the leaves of the decomposition tree). Thus, the children of each node in the modular decomposition tree are its modules, whether they are internal vertices or leaves. Figure 4b shows one decomposition of G in Figure 1a into modules. Module β may be furtherly decomposed into modules with vertex sets 4,6 and 9, while all the other modules have only trivial decompositions into 1-vertex modules (i.e. leaves).

When a graph is broken up into modules, one must be able to build the graph again using only its modules and a logical rule (otherwise the decomposition looses information). The logical rule is stored in the node of the tree corresponding to the graph as a character with values 0, 1 or P as follows (see Fig. 2a for illustration):

- 0 means that G is the union of all its modules, without any edge joining vertices from different modules. In this case, G is a 0-graph (or 0-module)

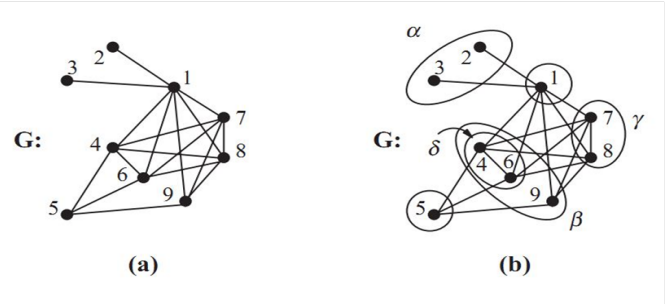

FIGURE 3.1: (a) Example graph G and (b) its decompositions into modules α, 1, β, 5 and γ.

and its corresponding node is a 0-node. Modules α and β in Figure 1b are 0-modules.

- 1 means that G is the join of all its modules, obtained by adding an edge between every pair of vertices from two different modules. In this case, G is a 1-graph (or 1-module) and its corresponding node is a 1-node. Module γ of G and module δ with vertex set 4,6 of β are 1-modules.

- P means that G is obtained from its modules by performing, between any pair of modules, either a union or a join, according to the characteristic graph C(G), whose vertices correspond to the modules, and whose edges correspond to the join operations. In this case, G is a P-graph (or P-module) and its corresponding node is a P-node. The whole graph G in Figure 1a is a P-module with modules α,1, β,5, γ (Fig. 1b), which have to be combined together according to the characteristic graph C(G) in Figure 2b to build G again. Notice here that the characteristic graph of G is obtained by shrinking in G each module into a single vertex.

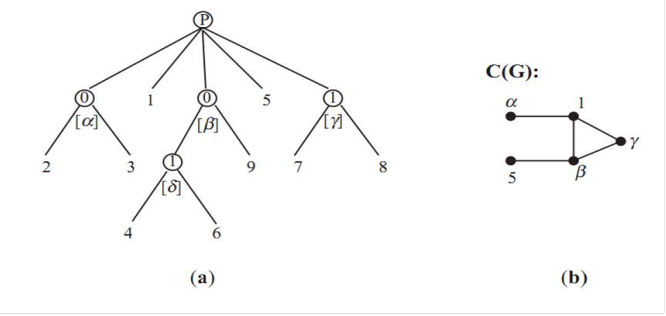

FIGURE 3.2: The modular decomposition tree of the example graph G (a) and the characteristic graph C(G) associated to the root (b), which is a P-node.

3.2 Homogeneous Decomposition

The main drawback of the modular decomposition is its unability to further decompose the characteristic graphs associated to the P-modules. The homogeneous decomposition partially solves this problem by identifying P-modules with a specific structure, for which a further decomposition is proposed. It is not meant to replace the decomposition into modules but to refine it, once the modules have been computed and the characteristic graph has been built. The homogeneous decomposition offers, therefore, an obvious qualitative improvement to the modular decomposition, which is described here in an intuitive manner and is illustrated in examples.
A graph G furtherly decomposable by an homogeneous decomposition is called a W-graph (or W-module) which has the wheel structure depicted in Figure 3a. Such a module has a characteristic graph (Fig. 3b) made of a hub (which is a clique) and of a set of single vertices around the hub that have neighbors in the hub but are not joined to each other. The h-modules of a W-module are the graphs (which are also cliques) made of a single vertex and all its neighbors in the hub. Note that h-modules do not have the neighborhood property as other modules do. These notions are illustrated on the example graph G in Figure 4a and 4b.

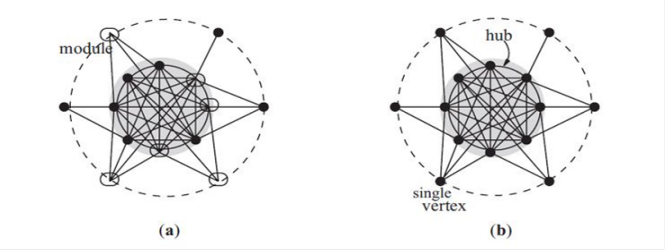

FIGURE 3.3: (a) Wheel structure of a graph. (b) Characteristic graph with emphasized hub and single vertices.The dotted circle simply highlights the wheel structure, it does not indicate edges.

The homogeneous decomposition introduces two new logical rules in the decomposition tree (Fig. 4c), described by characters Wand H used to label internal nodes:

- W means that the graph G is obtained from its modules (stored as children) and its h-modules (stored as a specific child which is an H-node) by recovering a wheel structure. This happens when G is a W-graph (or W-module) and in this case its corresponding node is called a W-node.

Graph G in Figure 1a is a W-module whose corresponding W-node is shown in Figure 4c.

- H means that the internal node stores the h-modules of its father, which is necessarily a W-module, in the following form: each h-module is stored in a child labeled by a vertex set whose first element is the single vertex identifying the h-module and the other elements are its neighbors in the hub. In this case, the node is called an H-node. Although this is not necessary, in our figures and so as to simplify explanations, the vertex set of the hub labels the H-node. The h-modules of the characteristic graph C(G) of the example graph G, identified in Figure 4b, are stored in the H-node in Figure 4c.

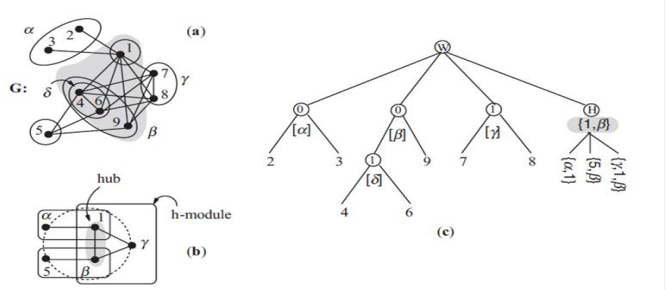

FIGURE 3.4: (a) Wheel structure of G. (b) Characteristic graph C(G) with hub (in gray) and its three h-modules. (c) Homogeneous tree of G.

3.3 Limitations of Homogeneous Decomposition

Some drawbacks that weve found regarding homogeneous decomposition are discussed as follows:

3.3.1 Multiple Hub

In some extended scenarios of protein interaction networks there may be multiple potential hubs.Homogeneous Decomposition can not describe this scenario.An illustration is shown below:

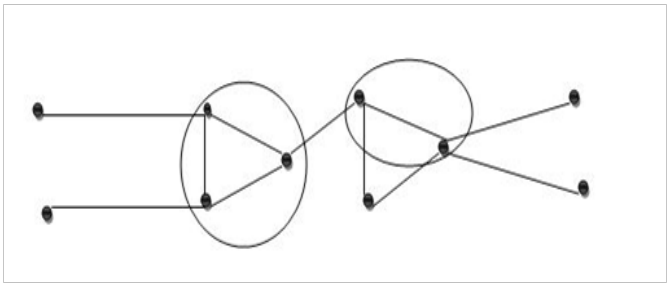

FIGURE 3.5: A theoretical character graph C(G) with two potential hubs(encircled).

3.3.2 Wheel Structure

In some scenarios, there are nodes connected to the attachment nodes but not to the hub nodes.Homogeneous Decomposition can not describe this scenario. An illustration is given below:

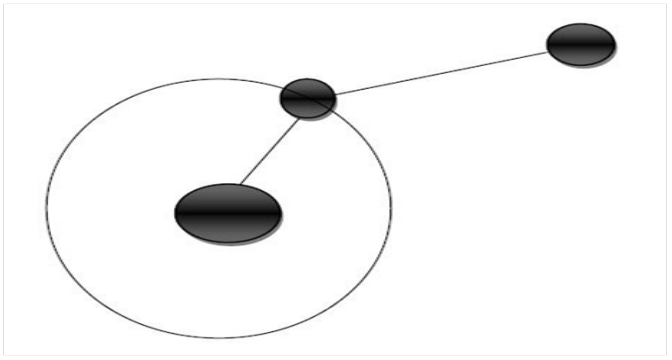

FIGURE 3.6: A theoretical wheel structure where extra node is connected to an attachment node.

3.3.3 Sub-complexes

Homogeneous Decomposition can identify subcomplexes in a protein coplex and interactions between them. But it can not relate the properties of sub-complexes(functionally and spatially more homogeneous, enriched with essential proteins etc.) by its interaction description.

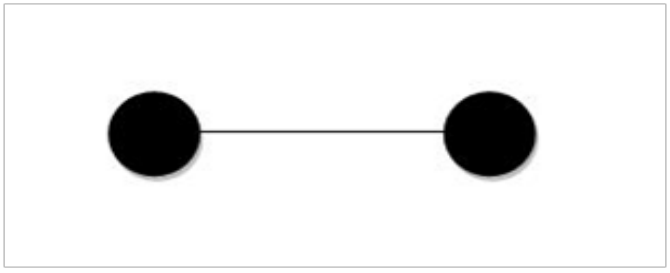

FIGURE 3.7: Character graph after all modular decomposition in protein complexes.

Chapter 4

Our Proposal

4.1 Multiple Hub

Our conceptual proposal is All potential hubs will form a single hub together where each of the potential hubs will be connected to atleast one of other hubs. That means there will be atleast one path available from between every potential hubs. An illustration is given below:

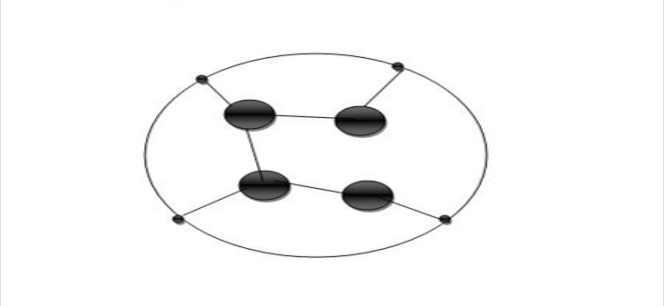

FIGURE 4.1: A theoretical representation of our first proposal.

4.2 Wheel Structure

Nodes connected to the hub node through any attachment node will be in the same complexes where the attachment nodes are members of that complex.An illustration is given below:

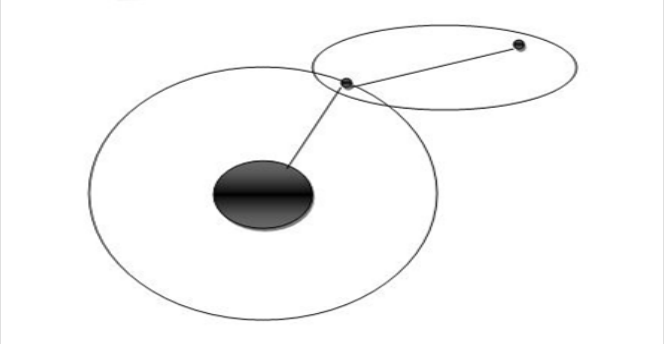

FIGURE 4.2: A theoretical representation of our second proposal.

4.3 Sub-complexes

Homogeneous Decomposition can identify subcomplexes in a protein complex and interactions between them.But it can not relate the biological counterpart of sub-complexes by its interaction description.The biological significance of protein subcomplexes recently investigated experimentally are:

- SCs are functionally and spatially more homogeneous than complete protein complexes.
- The abundance of subcomplex proteins is less variable than the abundance of other proteins.
- SCs are enriched with essential and synthetic lethal proteins.
- Mutations in SC-proteins have higher fitness effects than mutations in other proteins.

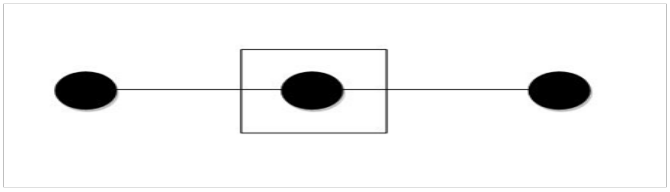

FIGURE 4.3: A theoretical representation of our third proposal.

Among these the first property is exploited to propose functions for so far unknown proteins of S.cerevisiae.
We want to investigate within protein complexes to relate the biological functions of sub-complexes with the properties of their corresponding interactions between them in the homogeneous tree.

4.4 Applications of our extensions

4.4.1 A Theoretical Network

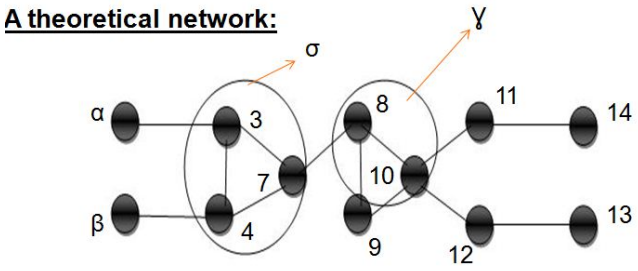

FIGURE 4.4: Character graph C(G) of our theoretical protein interaction network.

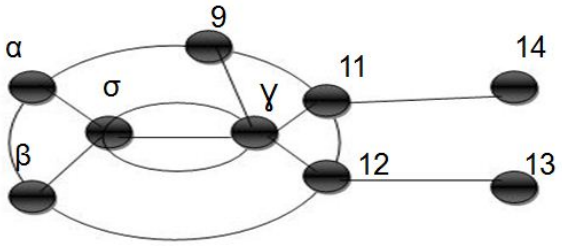

FIGURE 4.5: Wheel structure of the above characteristic graph.

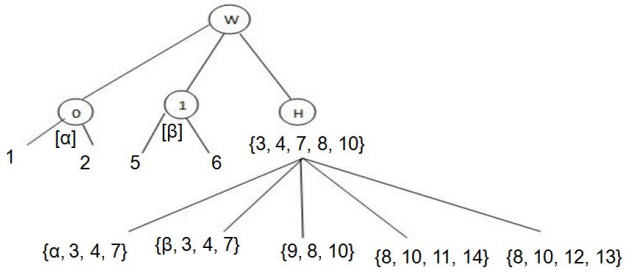

FIGURE 4.6: Homogeneous tree of previous characteristic graph by applying our extensions.

4.4.2 A Protein Complex

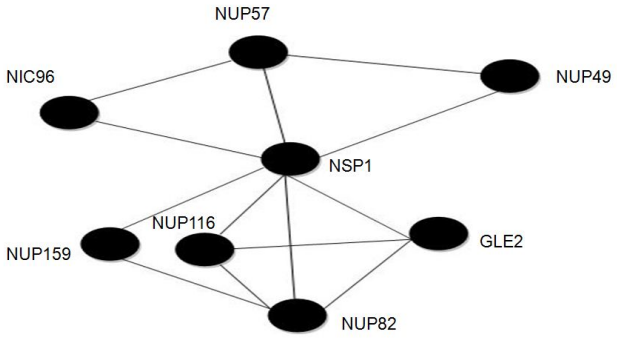

FIGURE 4.7: A protein complex extracted from PIN of Yeast by experimental methods.

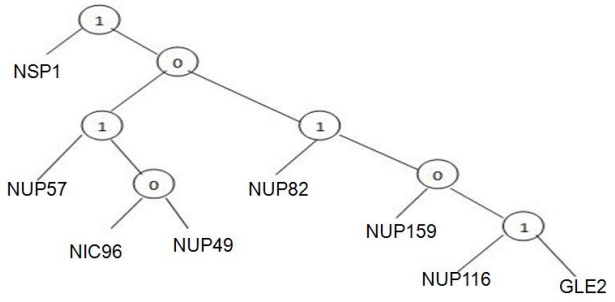

FIGURE 4.8: Identical tree of both Modular and Homogeneous Decomposition.

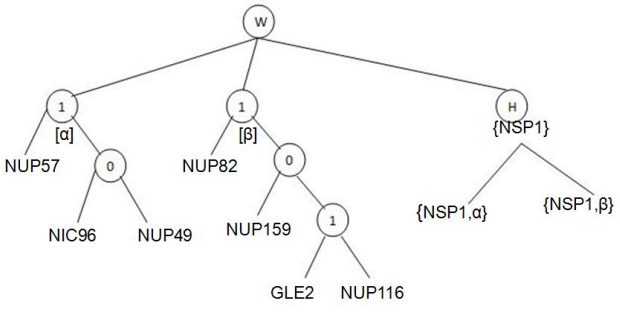

FIGURE 4.9: Homogeneous tree of previous complex by applying our extensions.

Chapter 5

Future Work

We could not implement our proposed extensions due to shortage of time and resources. In near future, we wish to implement these proposals in Perl and check the performance with protein interaction databases and compare it with the performance of Homogeneous Decomposition. Also we wish to further dig into the interpretation of biological relations between protein sub-complexes through our extensions.

Chapter 6

References

1. Gagneur,J. et al. (2004) **Modular decomposition of protein-protein interaction networks.** *Genome Biol.*, 5, R57.

2. Del Mondo,G.; Eveillard,D.; Rusu,I. et al. (2009) **Homogeneous decomposition of protein interaction networks:refining the description of intra-modular interactions.** *Bioinformatics*, 25, 926932.

3. Cannataro,M; H. Guzzi,P; Veltri,P. et al. (2010) **Protein-to-Protein Interactions: Technologies, Databases, and Algorithms.** *ACM Computing Surveys*, 43, 1.

4. Ma,X; Gao,L. et al. (2012) **Discovering protein complexes in protein interaction networks via exploring the weak ties effect.** *BMC Systems Biology*, 6:S6.

5. Spirin,V. and Mirny,L.A. (2003) **Protein complexes and functional modules in molecular networks.** *Proc. Natl Acad. Sci. USA*, 100, 12123 12128.

6. Hollunder,J.; Beyer, A.; Wilhelm,T. et al. (2007) **Protein Subcomplexes- Molecular Machines With Highly Specialized Functions.** *IEEE Transactions on Nanobioscience*, 6, 1.

7. Barabsi,A.; Gulbahce,N.; Loscalzo, J. et al. (2011) **Network medicine: a network-based approach to human disease.** *NatureRev*, v-12.

8. Bader, G. and Hogue, C. et al. (2003) **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics*, 4, 1, 2.