IN SILICO DETERMINATION OF PROTEIN COMPLEXES IN YEAST

by
Hilal Candan Karabıyık
B.S. in Ch.E., Boğaziçi University, 2005

Submitted to the Institude for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University
2008

ACKNOWLEDGEMENTS

I offer my thanks to Prof. Betül Kırdar for her support and counsel throughout this thesis and encouragement she provided during my graduate study.

I would also like to thank to Prof. Zeynep İlsen Önsan and Assoc. Prof. Ebru Toksoy Öner for devoting their precious time to read and comment on my thesis.

I am sincerely grateful to my friend Yalçın helping me finding the way in the thesis and to my friends Esra, Saliha, Duygu, Dicle, Aylin, Güray, Nil, Sinem and Ayça for their help and encouragement during my thesis.

I would also like to thank to Hande Sengel, Fehiman Akmaz and Melek Orhon for their consideration given on the importance of education and my collaborators Arca, Aslı and Necla for their encouragement during my thesis.

Last, but certainly not the least, I would like to thank to my family for providing me with peace, support, advices and joy throughout my whole life. And most importantly, I'm grateful to Fatih Mehmet Güçlü for supporting me, for his ceaseless love and his never-ending encouragement; his contribution to my life is priceless.

ABSTRACT

IN SILICO DETERMINATION OF PROTEIN COMPLEXES IN YEAST

In this thesis, the protein complexes in an interaction network of 20,487 interactions between 4,944 proteins were determined by the integration of different biological datasets. We recruit different data sources that include co-expression measures (PCCA), interaction data (TI), and GO process (PSA), function (FSA) and localization similarity (LSA) information and CFinder algorithm to identify highly interactive modules in the interaction network. Due to the dense interactions inside a community, among 4944 proteins, CFinder could only clustered 1822 proteins inside one or more community. Among the 55 communities investigated, according to the cut-off value that was selected (0.64), 20 communities had FF value greater than 0.64 and all of the 20 communities were protein complexes or part of the complexes. 11 different complexes were found, 9 of them; nuclear exosome (NE), anaphase promoting complex (APC), transport protein particle (TRAPP), mRNA cleavage factor (mRNACF), RSC, spliceosomal uridine-rich small nuclear ribonucleoprotein (snRNP U1), mediator (M), ARP2/3, Transcriptional elongation factor (TEF) are transient complexes. Two of them, proteasome regulatory particle (PRP) and DNA directed RNA polymerase II complex (RNA 2) are permanent complexes. These 11 complexes contain overall 295 proteins. The total community frequency is 94.8%; 128 out of 135 proteins of k-clique communities are members of the complexes. % 100 of the proteins of the TRAPP complex is identified with a community frequency of 100 %.11 proteins of the APC complex was identified. 13 proteins of mRNACF were detected by the community k9.4 with a community frequency of %100. Transcriptional elongation factor consists in total of 19 proteins (SGD). We identified 7 proteins of transcriptional elongation factor in the community k7.1. Mediator complex consists of 20 proteins (SGD), and we identified 7 proteins of M with a community frequency of 100% in the community k7.6 PRP consist of 22 proteins (SGD).We identified 11 members of the PRP complex out of 12 proteins in community k10.4 and 14 members of PRP complex out of 16 proteins in column k9.6. Six components out of 8 proteins of k8.8 are the members of the RNA 2 complex.

ÖZET

MAYADAKİ PROTEİN KOMPLEKSLERİNİN *IN SILICO* SAPTANMASI

Bu tezde, 4.944 protein arasında 20.487 etkileşimin olduğu bir etkileşim ağyapının içerisindeki protein kompleksleri, değişik biyolojik verilerin birleştirilmesiyle bulundu. Genlerin ko-ekspresyon değerleri (PCCA), etkileşim verileri, GO proses (PSA), fonksiyon (FSA) ve hücre içindeki yer benzerlikleri (LSA) bilgileri ile etkileşim ağyapısı içerisindeki yüksek etkileşimli öbekleri ortaya çıkarmak amacıyla CFinder algoritmasını da içeren değişik veri kaynakları toplandı. Bir öbek içerisindeki yüksek etkileşimlerden ötürü 4.944 protein içerisinden CFinder yalnızca 1.822 proteini bir veya daha fazla öbeğin içerisine atabildi. İncelenen 55 öbek arasından, seçilen sınır değer dahilinde (0,64), 20 tane öbeğin FF değerinin 0,64'ün üzerinde ve bu 20 yumağın hepsinin de protein kompleksi veya protein kompleksinin bir parçası olduğu bulundu. 11 farklı kompleks ortaya çıkarıldı, 9 tanesi; nukleer egzozom (NE), anafaz promoting kompleks (APC), taşıyıcı protein parçacığı (TRAPP), mRNA ayırıcı faktörü (mRNACF), RSC, splisozom uridince zengin küçük nukleer ribonucleoprotein (snRNP U1), mediatör (M), ARP2/3 transkripsiyonal elongasyon factor (TEF), geçici komplekslerdir. İki tanesi, proteazom düzenleyici parçacık (PRP) ve DNA yönlendirici RNA polimeraz II kompleksi (RNA 2) , kalıcı komplekslerdir. Bu 11 protein kompleksleri toplamda 295 protein içerirler. Toplam öbek frekansı %94,8'dir; k-klik öbeklerinin 135 proteininin 125 tanesi protein komplekslerinin üyeleridir. TRAPP kompleksinin üyelerinin %100'ü , %100 öbek frekansıyla bulundu. APC'nin 11 üyesi bulundu. mRNACF'ün 13 proteini, k9.4 öbeği tarafından %100 öbek frekansıyla bulundu. APC'nin 11 üyesi ortaya çıkarıldı. Transkripsiyonel elongasyon faktörü (TEF) 19 protein içerir (SGD). TEF'in 7 proteini k7.1 öbeğiyle bulundu. Mediatör komleksi (M) 20 protein içermekte (SGD) ve M'nin 7 üyesi %100 öbek frekansıyla k7.6 tarafından ortaya çıkarıldı. 22 protein içeren PRP'nin 11 üyesi 12 protein içeren k10.4 öbeği tarafından, 14 üyesi 16 protein içeren k9.6'nın içinde bulundu. Sekiz protein içeren k8.8'in 6 elemanı da RNA 2'nin üyeleri olduğu bulundu.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS / ABBREVIATIONS

20-s P	20-s proteasome
APC	Anaphase promoting complex
CC4	Ctochrome c4
CLR	Ccytoplasmic large ribosome
COC	Cyto-chrome c oxidase
CS	Component similarity
CSA	Average component similarity
CSR	Cytoplasmic small ribosome
F0.F1	ATP synthase comlex
FF	Fitness function
FS	Functional similarity
FSA	Average functioanl similarity
M	Mediator
MLR	Mitochondrial large ribosome
mRNACF	mRNA cleavage factor
MSR	Mitochondrial small ribosome
NE	Nuclear exosome
PCC	Pearson correlation
PCCA	Average Pearson correlation
PR	Pre-replicative complex
PRP	Proteasome regulatory particle
PS	Process similarity
PSA	Average process similarity
R	Replication complex
RNA1	DNA directed RNA polymerase 1
RNA2	DNA directed RNA polymerase II complex
RNA3	DNA directed RNA polymerase 3
SAGA	Spt-Ada-Gcn5-Acetyltransferase

snRNP U1	Spliceosomal uridine-rich small nuclear ribonucleoprotein

1. INTRODUCTION

Protein complexes may well be the most relevant molecular units of cellular function. The activities of protein complexes have to be regulated both in time and space to integrate within the overall cell programs. The cell can be compared to a factory orchestrating individual assembly lines into integrated networks fulfilling particular and superimposed tasks. Recent proteome-wide studies provide insight into the properties of cellular protein complexes, their modular nature, their interaction with other complexes and the resulting preliminary organization chart of proteome (Gavin *et al.*, 2003).

Our understanding of protein complexes has mainly come from pioneering analysis of the molecular machines involved in transcription and translation. However, several recent studies demonstrate that protein complexes are ubiquitous and represent the molecular norm, rather than the exceptional functional units of proteomes. Complexes are composed of subunits that probably are the result of coordinated gene expression, concerted translation and assembly as well as transport, activity and degradation (Gavin *et al.*, 2003).

The composition of protein complexes is not easily inferred by studying pair-wise interactions on a large scale (Gavin *et al.*, 2003).Accordingly, recent large-scale analyses of protein complexes show rather poor overlap with data generated by two-hybrid screens (Ho *et al.*, 2002; Gavin *et al.*, 2002).

Understanding the molecular and functional circuitry among protein complexes in human cells, as well as the changes associated with time, disease states or cell type will be an invaluable tool for medicine. For example the majority of drug targets are proteins. Understanding of protein complex connectivity and modularity allows us to track the pathways and cellular functions affected by a given drug. It provides a molecular basis for the appreciation of drug secondary effects. In turn, this streamlines toxicology studies but may also provide suggestions for new medical uses of existing drugs (Gavin *et al.*, 2003).

A large body of information of the biological roles of genes has been accumulated and aggregated in the past decades of research, both from traditional experiments detailing the role of individual genes and proteins, and from newer experimental strategies that aim to characterize gene function on a genomic scale. It is clear that the goal of functional genomics can only be achieved by integrating information and data sources from the variety of these different experiments. Integration of different data is thus an important challenge for bioinformatics. The integration of different data sources often helps to uncover non-obvious relationships between genes, but there are also two further benefits. First, it is likely that whenever information from multiple independent sources agrees, it should be more valid and reliable. Secondly, by looking at the union of multiple sources, one can cover larger parts of the genome. This is obvious for integrating results from multiple single gene or protein experiments, but also necessary for many of the results from genome-wide experiments since they are often confined to certain (although sizable) subsets of the genome (Jansen *et al.*, 2002).

There are several benefits of combining experimental and computational data sources. Often, one may be able to uncover non-obvious and potentially significant relationships, such as those between expression and chromosomal positioning or subcellular localization (Jansen *et al.*, 2006). Moreover, the integration of multiple sources obviously increases the range of the genome that can be characterized. This benefit of increasing coverage is obvious for integrating many of the experiments for individual genes or proteins, but is also valid for the combination of multiple genomic-scale experiments.

There have been considerably fewer attempts to integrate *more than two* types of whole-genome data. One example was the combination of expression correlations, phylogenetic profiles and patterns of domain fusion to predict protein function (Marcotte *et al.*, 1999). In another study, a Bayesian framework was used to integrate expression, essentiality, and sequence motif data for the prediction of protein subcellular localizations (Gerstein *et al.*, 2000; Drawid *et al.*, 2000).

In this paper, we explore an example of such a data integration procedure. We focus on the prediction of protein complexes for *Saccharomyces cerevisiae*. For this, we recruit

different data sources that include expression profiles, interaction data, and GO process, function and localization information. Each of these data sources individually contains some weakly predictive information with respect to protein complexes, but this prediction can be improved by combining all of them.

This thesis comprises 4 chapters. In Chapter 2, some background aspects of the protein complexes are given. General description of protein complexes and a detailed literature survey to detect the protein complexes is reviewed.

Chapter 3 describes the materials and methods used in this thesis. Numerous biological data and several analysis techniques used explained briefly.

In chapter 4 results obtained from the determination of protein complexes are given and discussed.

Conclusions and recommendations for further studies are given in Chapter 5. This part is followed by appendices comprising supplementary information about the computer codes developed in the context of the thesis.

2. THEORETICAL BACKGROUND

2.1. Protein Complexes

Most cellular processes are carried out by multiprotein complexes. The identification and analysis of their components provides insight into how the ensemble of expressed proteins (proteome) is organized into functional units (Zotenko *et al.*, 2006). The complexity in biological systems arises not only from various individual protein molecules but also from their organization into systems with numerous interacting partners. In fact, most cellular processes are carried out by multi-protein complexes, groups of proteins that bind together to perform a specific task (Zotenko *et al.*, 2006).

2.1.1. Permanent and Transient Complexes

Some proteins form stable complexes, such as the ribosomal complex that consists of more than 50 proteins and three RNA molecules, while other proteins form transient associations and are part of several complexes at different stages of a cellular process (Zotenko *et al.*, 2006).

According to Nooren *et al.* (2003), in contrast to a permanent interaction that is usually very stable and thus only exists in its complexed form, a transient interaction in a transient complex, associates and dissociates *in vivo*. They also note that many protein protein interactions in a community do not fall into distinct types. Rather, a continuum exists between non-obligate and obligate interactions, and the stability of all complexes very much depends on the physiological conditions and environment. An interaction may be mainly transient *in vivo* but become permanent under certain cellular conditions. Folding data, as well as data on the dynamics of the assembly at different physiological conditions or environments, are often not available. However, the subcellular location of subunits and the function of the protein will often suggest the biologically relevant type of interaction; for example, interactions in intracellular signalling are expected to be transient, since their function requires a ready association and dissociation (Nooren *et al.*, 2003).

Another feature of the complexes is the essentiality of its subunits. Many of the identified protein complexes of *Saccharomyces cerevisiae* possess an invariant core, in which the biochemical role of each protein subunit is irreplaceable, and is seamlessly integrated into a higher-level function of the whole complex. In turn, the deletion phenotype of each core protein is determined by the role of the complex in the organism. If the given complex is essential for cell growth, the deletion of any core protein disrupts the complex's functional integrity, and subsequently renders the cell unviable. If, however, the cell is able to tolerate the loss of a complex's function, none of its specific core subunits are essential. The core is generally surrounded by several "halo" proteins that typically do not share a common deletion phenotype, functional classification, or cellular localization with the core subunits. This indicates that they likely represent temporal attachments, some acting as modifiers of the complex's function, whereas others are functionally unrelated proteins that spuriously attach to the surface of the core proteins (Von Mering *et al.*, 2002).

2.1.2. Regulation and Expression of Protein Complexes

In the work of Lichtenberg and his collabarators (2007), by constructing a cell cycle interaction network from protein-protein interaction data and information on gene expression during the cell cycle, they found that protein complexes were generally formed as a combination of static and dynamic proteins. Static proteins are constitutively expressed and dynamic proteins are periodically expressed. The dynamic proteins showed a clear tendency to be expressed right before the complex is known to become active and thus, by their change in abundance, control the assembly of the functional complex. This mechanism, which we termed *just-in-time assembly,* is illustrated in Figure 2.1. In this figure a single cell cycle-regulated subunit (shown in green) is sufficient to govern the activation of an entire complex, as proper assembly cannot take place until this final subunit is expressed. Phosphorylation often serves as an extra layer of regulation of the same subunits and may be required for assembly of the complex, for subcellular translocation into the proper compartment, and/or for targeted degradation and thus disassembly of the protein complex. When they studied protein complexes of known function across several organisms, they observed a clear tendency for transcription of genes encoding dynamic subunits to peak just before the complex is known to function. This link between the timing of transcription and complex activity strongly suggested them

that translational control plays only a minor role and supports a model in which protein synthesis is temporally controlled at the transcriptional level. This model is also supported by a recent genome-wide study of transcription and translation in fission yeast by Chen *et al.* (2007), and as well as by the many known examples where the temporal pattern of gene expression correlates closely with the changes in abundance of the protein (for example, cyclins and histones). Similarly, the lack of periodic fluctuation in the expression level of other complex subunits suggests that these proteins reside inside the cell at a constant level, either as individual components or as subcomplexes (Lichtenberg *et al.*, 2007).

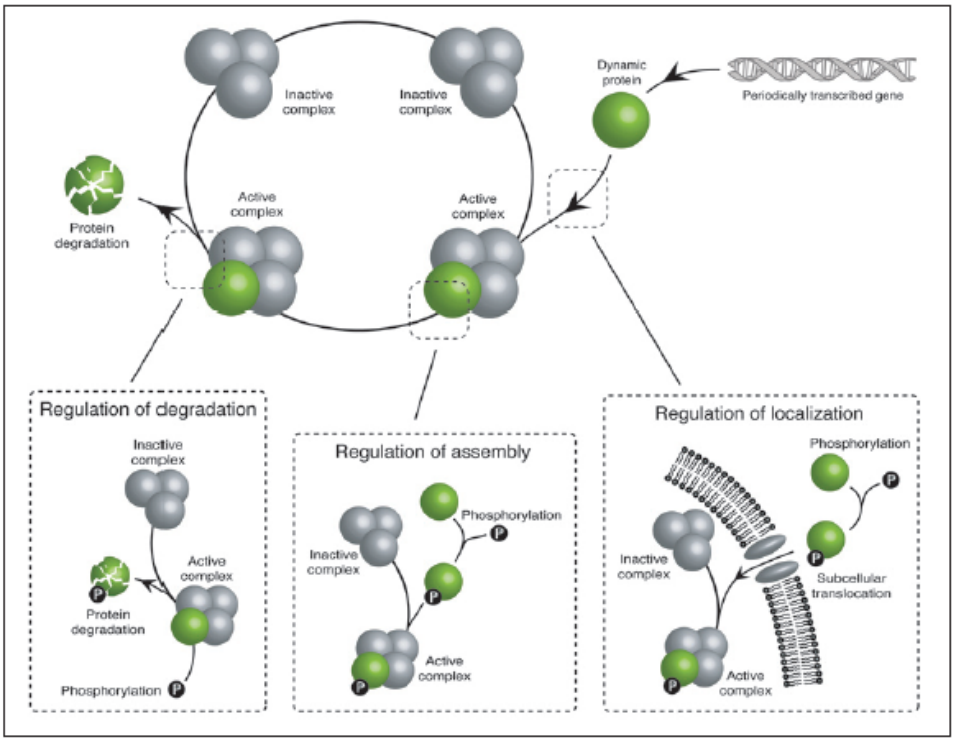

Figure 2.1. Just in time assembly of protein complexes (Lichtenberg *et al.*, 2007)

In another work (Jansen *et al.*, 2006), the yeast protein complexes were classified as either permanent or transient, permanent ones being maintained over a time course and they found that generally, permanent complexes, such as the ribosome and proteasome, have a particularly strong relationship with expression, while transient ones do not. However, they also found that that several transient complexes, such as the RNA polymerase II holoenzyme and the replication complex, can be subdivided into smaller permanent ones, which do have a strong relationship to gene expression.

Many protein complexes are established, particularly in the model organism *Saccharomyces cerevisiae*, yeast. The discovery of protein complexes is now performed genome wide; the elucidation of most protein complexes of the yeast is undergoing. In 2006 a group of scientists at Cellzome AG and the European Molecular Biology Laboratory in Heidelberg identified 494 protein complexes in *Saccharomyces cerevisiae* (Gavin *et al.*, 2006).

2.2. Methods for Finding Protein Complexes

Our knowledge regarding the identity of the building elements of specific complexes is limited and is based on selected biochemical approaches and genetic analyses. The only comprehensive protein interaction studies are based on *ex vivo* and *in vitro* systems, such as two-hybrid systems and protein chips, and need to be integrated with more-physiological approaches. Whenever it has been possible to retrieve and analyze particular cellular protein complexes under physiological conditions, the insight gained from the analysis has been fundamental for the biological understanding of their function, and has often taken the analysis well beyond the limits of genetic analysis. Prominent examples are the spliceosome, the cyclosome, the proteasome, the nuclear pore complex and the synaptosome. No systematic analysis of protein complexes from the same cell type using the same technique has yet been reported (Gavin *et al.*, 2002).

Complex cellular processes are modular and are accomplished by proteins in complex multi-protein assemblies. Often these multi-protein complexes act as highly efficient protein machines and perform activities related to complex biological phenomena, such as DNA replication, transcription, metabolism, and signal transduction. A variety of experimental and computational approaches have been employed to deduce the constituents of protein macromolecular complexes (Xiong *et al.*, 2005).

2.2.1. Experimental Methods

Data on protein complexes are collected from the study of individual systems, and more recently through high-throughput experiments, such as yeast two-hybrid (Y2H) and tandem affinity purification followed by mass spectrometry (TAP/MS) (Ho *et al.*, 2002).

The TAP/MS approach helps pinpoint proteins that interact with a tagged bait protein, either directly or indirectly, and are thus suited to identify multi-protein complexes. In fact, several research groups have systematically applied TAP/MS technology to study protein complexes involved in different signaling pathways.

In the work of Gavin *et al.* (2002) 1,739 genes are processed, including 1,143 human orthologues of relevance to human biology, and purified 589 protein assemblies. Bioinformatic analysis of these assemblies defined 232 distinct multiprotein complexes and proposed new cellular roles for 344 proteins, including 231 proteins with no previous functional annotation.

Experimental approaches such as the yeast two-hybrid genetic screen yield binary interaction data while more recent large scale methods combine tagged bait proteins and protein complex purification schemes with mass spectrometric measurements to identify protein complexes that contain three or more components (Xiong *et al.*, 2005).

2.2.2. Computational Methods

2.2.2.1. Graph Theoretical Methods. Modeling PPI networks with simple graphs has been used for many applications, one of which is the prediction of protein complexes within the PPI networks. Protein complexes generally correspond to dense subgraphs in the PPI network; thus, proteins in a given complex are highly interactive with each other. Previous approaches to graph-theoretic cluster prediction include simple clustering methods such as identification of *k*-cores (Przulj *et al.*, 2004; Bader *et al.*, 2004) super-paramagnetic clustering (Spirin *et al.*, 2003) and the highly connected subgraph approach Shamir (Hartwell *et al.*, 1999). The results suggest that true protein complexes exhibit certain graph-theoretic properties and functional homogeneity. Thus, using size, density and functional homogeneity as filtering criteria for network clusters is a reasonable approach to predict novel protein complexes. However, there are some problems with this approach. While protein complexes are usually expected to have high density in PPI networks, not all do. A related problem is the incompleteness of current PPI networks. The more complete and accurate our PPI and known protein complexes datasets are, the more accurately we can analyze the PPI networks. Further, the functional homogeneity, while accurate for the

most part, seems to be an incomplete, oversimplified model. Many known complexes show low functional homogeneity. Also, many proteins belong to multiple functional groups. In addition, many proteins are of unknown function. Even with such a simple filtering model and incomplete data, it can be managed to achieve very high matching rates between PPI network clusters and known protein complexes (Zotenko *et al.*, 2006).

Following the observation that protein interaction networks display a characteristic power-law like node degree distribution (Barabasi *et al*., 1999) a substantial body of research focused on statistical properties of protein interaction networks (Przulj *et al*., 2004). In 1999, Hartwell and his friends (Hartwell *et al.*, 1999) introduced a notion of a functional module, a group of cellular components and their interaction that can be attributed a specific biological function. The authors also suggested the modular organization of molecular interaction networks, where each functional module involves a small number of cellular components and is autonomous, i.e., its interaction with other modules is limited to a few cellular components. Subsequently, this assumption was used in several computational methods to identify protein complexes and functional modules in high-throughput protein interaction networks (Spirin *et al.*, 2003).

In the paper of Xiong *et al.* (2005), a hyperclique pattern discovery approach for extracting functional modules from protein complexes is presented. A hyperclique pattern is a type of association pattern containing proteins that are highly affiliated with each other. The analysis of hyperclique patterns shows that proteins within the same pattern tend to be present in the protein complexes together. Also, statistically significant annotations of protein in a pattern using the Gene Ontology suggest that proteins within the same hyperclique pattern more likely perform the same function and participate in the same biological process.

More interestingly, the 3-D structural view of proteins within a hyperclique pattern shows that these proteins physically interact with each other. For example in Figure 2.2, all of the proteins of the proteasome complexes are in contact. In addition they reveal that several hyperclique patterns corresponding to different functions can participate in the same protein complex as independent modules. Moreover a hyperclique pattern can be

involved in different complexes performing different biological functions (Xiong *et al.*, 2005).

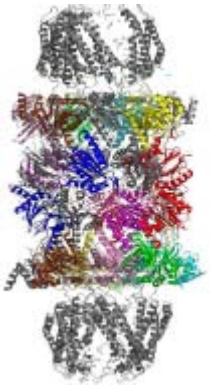

Figure 2.2. 3-D structural view of proteins within a proteasome complex (Xiong *et al.*, 2005)

The major problems of the general clustering techniques is that one node generally belongs to one group, however we know that most of the proteins involve in more than one community. To overcome this difficulty in 2005, Pella and his collaborators (Pella *et al.*, 2005) introduced CFinder a standalone application that reads a list of binary interactions, performs a search for dense subgraphs (groups), and unlike several currently used algorithms it allows for any node to belong to more than one group.

CFinder, is a fast program locating and visualizing overlapping, densely interconnected groups of nodes in undirected graphs, and allowing the user to easily navigate between the original graph and thereof these groups.

The search algorithm uses the Clique Percolation Method (CPM, see Derenyi *et al.*, 2005) to locate the k-clique percolation clusters of the network that may be interpreted as modules or communities. A k-clique is a complete subgraph on k nodes, and two k-cliques are said to be adjacent, if they share exactly k -1 nodes. A k-clique percolation cluster consists of all nodes that can be reached via chains of adjacent k-cliques from each other and the links in these cliques. Note that larger values of k correspond to a higher stringency during the identification of dense groups.

In gene (protein) association networks CFinder is used to predict the function(s) of a single protein and to discover novel modules. CFinder is also very efficient for locating the cliques of large sparse graphs.

One year later, Adamcsek *et al.* (2006) applied the full yeast interaction network to Cfinder and they showed that in protein association networks CFinder can be used to predict the function(s) of a single protein and to discover novel modules. CFinder is also very efficient for locating the cliques of large sparse graphs.

2.2.2.2. Integrative Method. Previous works on protein complexes generally concentrate on functional modules or clique patterns existing in the protein complexes. There are very few examples that focus on solely protein complexes. One of these works is the study of Jansen and his collaborators (Jansen *et al.*, 2002).

The integration of different data sources often helps to uncover non-obvious relationships between genes, but there are also two further benefits. First, it is likely that whenever information from multiple independent sources agrees, it should be more valid and reliable.

In the paper of Jansen *et al.*, (2002) authors focused on the prediction of membership in protein complexes for individual genes. For this, six different data sources that include expression profiles, interaction data, and essentiality and localization information are recruited. Each of these data sources individually contains some weakly predictive information with respect to protein complexes, but they show this prediction can be improved by combining all of them.

When multiple experiments cover the same genes, then there are other benefits from combining data. In general, the combination of different data sources should help to increase the reliability of the interpretation of experimental results. They used the MIPS complexes catalogue as the standard for known protein complexes. This study is preliminary but intended to show possible ways of combining new genome-wide datasets to ultimately determine all protein complexes. Similar ways of combining genome wide

datasets for predicting other kinds of biological information, such as biological functions or pathways could be possible as well.

In 2007, another integrative method is applied (Futschik *et al.*, 2007) to human protein network consisting more than 30.000 interactions. First of all modular structures are identified by applying CFinder algorithm, then GO annotation terms of the modules including process, localization and function annotations and co-expression of the proteins by using a large microarray data were investigated. Their analysis showed that many modules can be assigned to cellular processes. It also indicated that protein complexes and dynamic functional modules can be distinguished based on colocalization and co-expression, although there exists no rigorous threshold to distinguish them.

2.3. Systems Biology and *Omic* Data

In classical approach, the function of a gene or protein in a biological system was investigated as an individual property in the system in a separate time interval. However, recently an integrative analysis, which is called systems biology, evaluates the biological data from all levels of metabolism, from genome to metabolome. This omic information in several levels is used in order to view the studied organism as a whole (Ideker *et al.*, 2002) by investigating the behavior and the relationships of all of the components in a particular biological system while it is biological sense out of the data to identify new targets for metabolic engineering, mathematical models plays an important role, and systems biology is therefore associated with quantitative investigation of the biological system under study.

2.3.1. Transcriptome Analysis

Today DNA arrays are the common tool for transcriptome analysis. DNA arrays give information of which genes and to what extent they are expressed. Thus, changes in transcription from a given time or condition to another can be quantified and this may give indications of genes function and regulation. Genes that participate in the same cellular processes often share similar transcription profiles (Cho *et al.*, 1998; DeRisi *et al.*). By mining expression data, using software applying clustering algorithms that group together

genes with similar transcription profiles, genes with similar expression can be identified (Eisen *et al.*, 1998).

Information related to the regulation of genes can also be obtained from DNA arrays. Genes with similar transcription profiles might also be regulated by the same mechanisms, and therefore have similar regulatory elements in their promoters (Ideker *et al.*, 2001). Many methods have been described to identify regulatory elements in promoters (Hughes *et al.*, 2000)

The goal of functional genomics is to assign function to each gene within a genome. The ability to perform genome-wide transcription analysis was a major breakthrough in functional genomics. The sequencing of the *Saccharomyces cerevisiae* genome together with technical achievements enabled the first DNA arrays for genome-wide transcription analysis *Saccharomyces cerevisiae* in the late 1990's (DeRisi *et al.*, 1997).

There are now numerous examples of the use of transcriptome analysis in *Saccharomyces cerevisiae*, and transcription data is generated in staggering amounts. Also, all these data are available in literature via internet databases; such as MIPS (Munih Information Center for protein Sequences), CYGD (The Comprehensive Yeast Genome Database), SGD (*Saccharomyces* Genome Database).

2.3.2. Interactome Analysis

Analysis of protein-protein and protein DNA interactions is very important for unravelling signal transduction pathways and can also be used to generate functional annotations for proteins with unknown function. The yeast two hybrid system identifies pair of physically interacting proteins and recently it has been used for genome wide studies (Ito *et al.*, 2000).Two hybrid essays are easy to do and readily applicable to other eukaryotic systems, but not all proteins can be analyzed for their interactions by the essay, which also produces many false positive outcomes (Bro, 2003). However, future two-hybrid technologies might overcome some of the present day limitations and protein–protein interactions might also be analyzed with protein chips (Zhu *et al.*, 2001).

Analysis of protein-DNA interactions gives valuable information about transcription factors and is required to construct a complete map of transcriptional regulation in yeast. A methodology which can be applied genome wide, has recently been developed .This method , which combines DNA arrays with chromatin immunoprecipitation, identifies sequences that are directly bound by a specific transcription factor of interest *in vivo* and has already identified many new putative targets of some transcription factors (Iyer *et al.*, 2001). Interactions obtained from several experimental systems such as; affinity chromatography, affinity precipitation, biochemical essays, dosage lethality, purified complex, reconstituted complex and two hybrid system are published and maintained via several databases.

3. MATERIALS AND METHODS

3.1. Data

3.1.1. Yeast Protein-Protein Interaction

Data on the yeast protein interaction network were collected from the databases; The General Repository for Interaction Datasets (BioGRID (Breitkreutz *et al.*, 2003)), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Mering *et al.*, 2005) and DIP (Salwinski *et al.*, 2004).

STRING database consists of known and predicted protein protein interactions. It includes direct and indirect associations; they are derived from several sources, genomic context, co-expression patterns and literature knowledge.

DIP database includes experimentally determined interactions between proteins. It combines information from a variety of sources to create single, consistent set of protein-protein interactions. The data stored within the DIP database is a curated data both by manually by expert curators and also automatically using computational approaches.

The Biological General Repository for Interaction Datasets (BioGRID) is a curated biological database of protein-protein interactions created in 2003 (originally referred to as simply the General Repository for Interaction Datasets (GRID) (Breitkreutz *et al.*, 2003). It strives to provide a comprehensive resource of protein–protein interactions for all major species while attempting to remove redundancy to create a single mapping of protein interactions (Breitkreutz *et al.*, 2008).

Altogether, the interaction network used in the thesis contains over 20400 interactions between more than 4200 unique proteins for further analysis.

3.1.2. Yeast Transcriptional Interaction (Protein-DNA)

Yeast transcriptional integration data was obtained from genome-wide experimental studies (Lee *et al.*, 2002), and publicaly available databases, YEASTRACT (Teixeira *et al.*, 2006), TRANSFAC (Wingender *et al.*, 1995) and InCyte.

Yeast Search for Transcriptional Regulators and Consensus Tracking (YEASTRACT) is a curated repository of more than 27800 regulatory associations between transcription factors and target genes in *Saccharomyces cerevisiae*, based on more than 900 bibliographic references. It also includes 281 specific DNA binding sites for more than a hundred characterized TFs.

TRANSFAC is a database about eukaryotic transcription regulating DNA sequence elements and the Transcription factors binding to and acting through them.

InCyte is a drug discovery and development company and presents commercially not available database of protein-DNA interactions, it is kindly provided by Professor J.Nielsen, Chalmers University of Technology, Sweeden.

Altogether, over 14,900 regulatory interactions between 556 transcription factor and 3,913 target genes were obtained.

3.1.3. Protein Complexes

The list of test and training protein complexes selected from the CYGD catalogue (http://mips.gsf.de/proj/yeast/CYGD/db/) (Mewes *et al.*, 2002) for this analysis.

3.1.4. Gene Expression

Gene expression measurements were obtained from the Stanford Microarray Database (SMD) and included 14 different microarray data, across a wide variety of cellular states.

In the microarray data of Gasch *et al.* (2001), to characterize the role of the Mec1 pathway in modulating the cellular response to DNA damage, the expression patterns of the 6200 predicted yeast genes in response to MMS treatment and ionizing radiation measured in a total of 40 microarray hybridizations.

Microarray-based gene expression data of Galitski *et al.* (1999), ploidy dependent expression in isogenic *Saccharomyces cerevisiae* strains that varied in ploidy from haploid to tetraploid was investigated.

In the DNA microarray analysis of Robert *et al*. (2000) Signaling and Circuitry of Multiple MAPK Pathways with >97% known genes *Saccharomyces cerevisia* for the pheromone response is investigated. Two datasets of this work was included, one is the gene expression with respect to different α-factor concentration, second is with respect to time.

In the work of Yoshimoto *et al.* (2002) calcineurin signaling *in vivo* by exposing yeast cells to high extracellular levels of Ca^{2+} or Na^{+} was investigated and DNA microarrays was used to analyze the resulting changes in gene expression in the presence and absence of immunosuppressive drugs FK506.

In the work of DeRisi *et al.* (1997), DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift (diauxic shift) from fermentation to respiration.

In the work of Gasch *et al.* (2000), genomic expression patterns in the yeast *Saccharomyces cerevisiae* responding to diverse environmental transitions were explored. DNA microarrays were used to measure changes in transcript levels over time for almost every yeast gene, as cells responded to temperature shocks, hydrogen peroxide, the superoxide-generating drug menadione, the sulfhydryl-oxidizing agent diamide, the disulfide-reducing agent dithiothreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into stationary phase.

The work of Spellman *et al.* (1998) is a well known cell cycle data. DNA microarrays were used to analyze samples from yeast cultures synchronized by three independent methods: a factor arrest, elutriation, and arrest of a *cdc15* temperature-sensitive mutant.

Travers *et al.* (2000) determined the transcriptional scope of the unfolded protein response that regulates gene expression in response to stress in the endoplasmic reticulum using DNA microarrays.

Meneghini *et al.* (2003) and Mizughuchi *et al.*(2004) investigated microarray data to elucidate histone variant H2AZ that has an important role in the regulation of gene expression and the establishment of a buffer to the spread of silent heterochromatin and the role of histone variants such as H2AZ .

Ogowa *et al.* (2000) investigated phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* by genomic expression analysis.

Tai *et al.* (2005) compared the response of aerobic as well as anaerobic chemostat cultures of the yeast *Saccharomyces cerevisiae* to growth limitation by four different macronutrients (carbon, nitrogen, phosphorus, and sulfur).

3.1.5. GO Annotation Terms

Gene Ontology is a controlled vocabulary used to describe the biology of a gene product in any organism. There are 3 independent sets of ontologies which describe the molecular function of gene product, the biological process and the cellular component terms where the gene products locate. The ontology terms may have multiple parents and multiple relationships to their parents. In addition each term inherits all the relationships of its parent. Go annotations of yeast genes were obtained from the website of the consortium or from *Saccharomyces* Genome Database (Dolinski *et al.*, 2005).

3.2. Identification of Community

The communities in the protein-protein interaction network were identified by the detection of *k-cliques* which are the fully connected subgraph of *k* vertices. These *k*-cliques figure highly connected arrangements named as *k*-clique communities. These communities are the combination of all *k*-cliques that can be reached from each other through a series of adjacent *k*-cliques. The cliques sharing *k-1* nodes are defined as adjacent. Pella and co-authors (Pella *et al.*, 2005) previously developed a powerful tool CFinder based on clique percolation method for detecting overlapping *k*-cliques communities in networks. Clique percolation method first locates all *k*-cliques in a network and then identifies communities by carrying out standard component analysis of the clique-clique overlap. For our analysis, CFinder was implemented to detect highly connected communities in the yeast protein interaction network.

3.3. Identification of Fitness Function

3.3.1. PPI Measure

In the present study a MATLAB code was written to express interactions quantitatively in a group of communities obtained from the CFinder program. The proteins of a community were entered as the input. Assuming, n as the number of proteins in the community, a P matrix was formed as following; for a set of n proteins, if any two of the proteins, i and j in the community were interacting, the P(i, j) value was set to equal 1, if not P(i,j) value is set equal to 0.

Then, an average value PPI in the protein set was obtained by the Equation 3.1:

$$PPI(set)=\frac{\sum_{i\neq j} P(i,j)}{n^2-n} \tag{3.1}$$

In Figure 3.1, a community consisting of 5 proteins is given. As there are interactions between the proteins P1 and P2, P1 and P4; P(1,4) and P(1,2) is set equal to 1 and as there

are no interactions between the proteins P1 and P3, P2 and P5; P(1,3) and P(2,5) are set equal to 1 also. Finally a P matrix obtained. Note that, P matrix is a symmetric matrix.

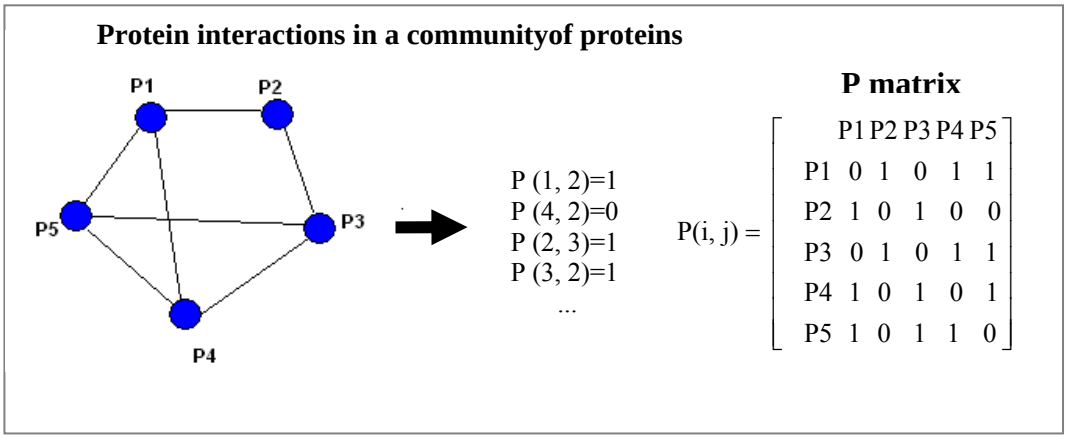

Figure 3.1. Representation of a community consisting of 5 interacting proteins and construction of P matrix

3.3.2. Gene Ontology Annotation Similarities (GO TERMS)

For the annotation of proteins, we utilized the Gene Ontology (GO) database supplying information about the assigned molecular function, biological process and cellular location. We assessed the significance whether the detected communities are enriched for proteins of certain functions, processes or locations by applying the following procedure written in MATLAB code.

Assuming n as the number of proteins in the community a PS matrix was formed as following PS (i, j) is set to 1 for each protein pair (i, j) that have same GO process annotation, if they have different process terms PS (i, j) is set to 0. An average process similarity measure (PSA) of the protein set is obtained by the Equation 3.2.

$$PSA = \frac{\sum_{I \neq J} PS(i, j)}{n^2 - n} \tag{3.2}$$

Assuming n as the number of proteins in the community a FS matrix was formed as following FS(i,j) is set to 1 for each protein pair that have same GO functional annotation,

if they have different function terms FS(i,j) is set to 0. An average function similarity measure (FSA) of the protein set is obtained by the Equation 3.3.

$$FSA = \frac{\sum_{I \neq J} FS(i, j)}{n^2 - n} \tag{3.3}$$

Assuming n as the number of proteins in the community a LS matrix was formed as following; LS(i,j) is set to 1 for each protein pair that have same GO component annotation, if they have different component terms LS(i,j) is set to 0. An average localization similarity measure (LSA) of the protein set is obtained by the equation 3.4.

$$LSA = \frac{\sum_{I \neq J} LS(i, j)}{n^2 - n} \tag{3.4}$$

3.3.3. Co- Expression

To assess co-expression of proteins, we utilized 14 different expression datasets. Each gene expression data was normalized with a mean of 0 and standard deviation of 1. For every expression data, the Pearson correlation coefficients (PCC (i, j)) were calculated for each protein pairs i and j (Equation 3.5). In this equation, x and y are 2 gene products. Assuming that expression profiles of both gene products are normally distributed, and have a profile length of n, the Pearson correlation coefficient (PCC(x, y)) was computed by MATLAB program according to the Equation 3.5.

$$PCC(x, y) = \frac{\sum_{i}^{n} (x_i - x)(y_i - y)}{\sqrt{\sum_{i}^{n} (x_i - x)^2 (y_i - y)^2}} \tag{3.5}$$

Pearson correlation coefficients of gene pairs were averaged according to the Equation 3.6. PCC_{datak} is the average Pearson correlation value of the pairs inside a community for one expression data. For each of 14 transcriptome data, average Pearson

coefficient (PCC_{datak}) for the pairs inside the community was calculated by using the equation 3.6.

$$PCC_{datak} = \frac{\sum_{i \neq j} PCC(i, j)}{n^2 - n} \quad (3.6)$$

14 aveage Pearson correlation values were averaged according to the Equation of 3.7.

$$PCCA = \frac{\sum_{k=1}^{14} PCC_{datak}}{14} \quad (3.7)$$

3.3.4. Co-Regulation

Assuming n as the number of proteins in the community a C (i, j) matrix was formed. If two proteins inside the community is regulated by the same transcription factor, C (i,j) is set to 1, if they are regulated with different transcription factors C (i, j) is set to 0. An average co-regulation measure (TI) of the community is obtained by the Equation 3.8.

$$TI = \frac{\sum_{I \neq J} C(i, j)}{n^2 - n} \quad (3.8)$$

3.3.5. Identification of Weights

Before applying C- Finder algorithm to interaction network, the PPI, PCCA, TI, PSA, CSA, FSA and finally fitness function (FF) values of some arbitrarily selected complexes were calculated and tested with other complexes.

Six permanent complex; 20-s Proteasome (20-s P), cytoplasmic large ribosome (CLR), cytoplasmic small ribosome (CSR), mitochondrial large ribosome (MLR), mitochondrial small ribosome (MSR), and DNA directed RNA polymerase 1 (RNA1) were used as training data to calculate the weights in the formula of fitness function which

is given by the Equation 3.9. The procedure for the identification of weights and the fitness function is given in Figure 3.2. First of all, the PCCA, TI, PPI, PSA, CSA, FSA measures were calculated as explained in the materials and methods for every protein complexes.

Fitness function calculation is given in Equation 3.9; it is a linear function containing all of the measures PCCA, TI, PPI, PSA, CSA and FSA. The weights were identified by using the Genetic algorithm framework (MATLAB Tool) to maximize sum of the FF values (Equation 3.10) of 6 permanent complexes (Figure 3.2). The objective function of the genetic algorithm is the maximize FFM value. The genetic algorithm is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and called recombination.

The basic idea here is, when the integrated data is implemented to an unknown community of protein, if it is a permanent complex, the result of the fitness function will be high, if it is not a permanent complex, for example a transient complex, or simply, a community of functional interacted protein, this number should be lower.

$$FF_{complex} = w1 * PCCA + w2 * PPI + w3 * TI + w4 * PSA + w5 * FSA + w6 * CSA \quad (3.9)$$

$$FFM = \sum_{i=1}^{6} FF_{complex} \quad (3.10)$$

Weights are given in the Table 3.1. The w1, w2, w3, w4, w5, and w6 are the weights of the co-expression (PCCA), co-regulation (TI), process (PSA), function (FSA) and component (CSA) similarity measures respectively.

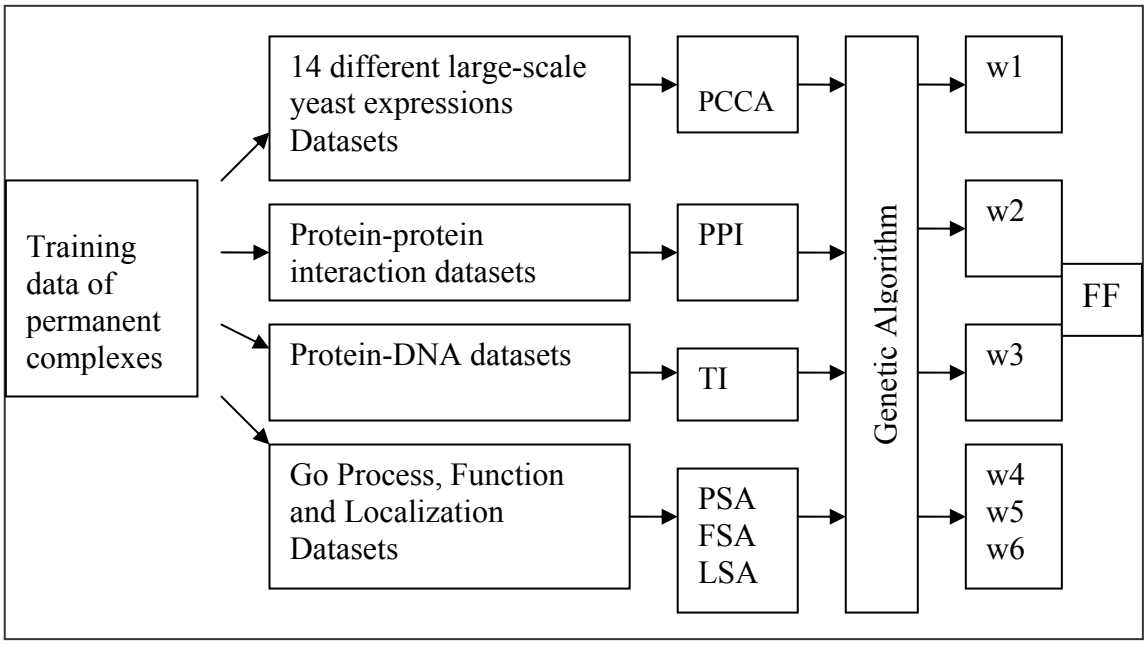

Figure 3.2. Procedure for the calculation of weights

Table 3.1. Weights calculated by Genetic Algorithm

w1	w2	w3	w4	w5	w6
PPI	**PCCA**	**TI**	**PSA**	**FSA**	**CSA**
0.2101	0.1992	0.0292	0.1941	0.1542	0.2131

As it can be seen from the Table 3.1, the weights are different, maximum weight value is the weight of the CSA and the minimum is the value of the weight of TI.

4. RESULTS AND DISCUSSION

In this thesis, we focused on the prediction of protein complexes from an interaction network using an algorithm that integrate several data sets. We recruited different data sources that include expression profiles, protein-protein interaction (PPI), transcriptional integration (TI), and GO annotation of proteins including function, process and localization criteria. Each of these data sources individually contains some weakly predictive information with respect to protein complexes, but this prediction was improved by calculating weights that maximize the solution of the linear fitness function measure (FF) with respect to the training data.

Firstly, the cut of value was identified according to the fitness functions of the test and training complexes as explained in the materials and methods. This cut off value was used to detect the protein complexes.

Secondly, the interaction network used in the thesis was applied to CFinder algorithm in order to identify highly interaction modules according to k-clique percolation method. Different measures mentioned above; protein-protein interaction (PPI), co-expression (PCCA), transcriptional integration (TI), GO annotation of proteins including function, process and localization criteria were calculated for these communities. Finally according to the cut-off value, some of the communities were assigned as protein complexes.

4.1. Determination of the Cut-off Value

Before applying CFinder algorithm to interaction network, FF values of some arbitrarily selected complexes were calculated and tested with other complexes.

The basic idea here is, when the integrated data is implemented to an unknown module of protein, if it is a protein complex, FF value will be high, if it is not a complex, a module of functionally interacted protein, and this number should be lower.

Fitness Function (FF) values were calculated using Equation 3.8 and weights given in Table 3.1 for the 9 well known Permanent complexes; 20-s Proteasome (20-s P), cytoplasmic large ribosome (CLR), cytoplasmic small ribosome (CSR), mitochondrial large ribosome (MLR), mitochondrial small ribosome (MSR), DNA-directed RNA polymerase 1 (RNA1), ATP synthase complex (F0.F1), cyto-chrome c oxidase (COC) and DNA-directed RNA Polymerase 3 (RNA3) and 6 well known transient complexes; anaphase promoting complex (APC), nuclear exosome (NE), Spt-Ada-Gcn5Acetyltransferase (SAGA), replication complex (R), pre-replicative complex (PR) and transcription factor-II (TFII) (Table 4.1).

Generally transient complexes have lower FF values, for example minimum FF value of permanent complexes is approximately equal to the maximum FF value of the transient complexes (0.75) and average FF value of the Permanent complexes is 0.81, however for the transient complexes this value is approximately 0.64.

After investigating the results in Table 4.1, we selected the cut off value of 0.64 which is the average FF value of the transient complexes. The reason to select this value can be explained as following; it is clear from the results that, due the some differences in the characteristics of permanent complexes, such as high co-expression value of its units etc, they have higher FF values; on the other hand, transient complexes have lower FF values.

If we had selected cut off value of 0.81, we would have probably not detected transient complexes from the interaction network. So, we expect that if there is a complex in an unknown network of community whether permanent or transient complexes, its FF value is probably greater than 0.64. However one should not forget that there might be some exceptional protein complexes that have lower FF value as in the case of replication (R) and cytochrome c-4 complexes (CC4).

Table 4.1. Fitness functions of well known protein complexes

Permanent Complexes	**Fitness Function (FF) Values**	**Transient Complexes**	**Fitness Function (FF) Values**
20s-P	0.8445	APC	0.7098
RNA 1	0.8086	CC4	0.6429
MSR	0.7787	PR	0.6220
MLR	0.7607	R	0.5265
CLR	0.7950	SAGA	0.5943
CSR	0.7692	TFII	0.7479
RNA3	0.8722		
F0.F1	0.7467		
COX	0.8787		
Average FF	**0.8060**	**Average FF**	**0.6406**

4.2. Identification of Communities by C-Finder Algorithm

The interaction network implemented contains 20,487 interactions between 4,944 unique proteins. The CFinder identifies highly connected structures of protein groups in these interactions. Therefore, the modules obtained from CFinder also contain both functional groups of protein that interacts at different time and place in a cell or some groups of proteins that form stable complexes, permanent complexes, such as the ribosomal complex that consists of more than 50 proteins and three RNA molecules and transient complexes that form transient associations and are part of several complexes at different stages of a cellular process (Zotenko *et al.*, 2006).

CFinder reads a list of binary interactions, compiles a search for dense subgraphs, and unlike several currently used algorithms, it allows for any node to belong to more than one group(Balazs *et al.*, 2006). Due to the dense interactions inside a community, among 4,944 proteins, CFinder could only clustered 1,822 proteins inside one or more community. Other proteins form binary interactions as in the case of proteins YPR201W and YLR171W or tree like interactions as in the case of protein YHR059W.

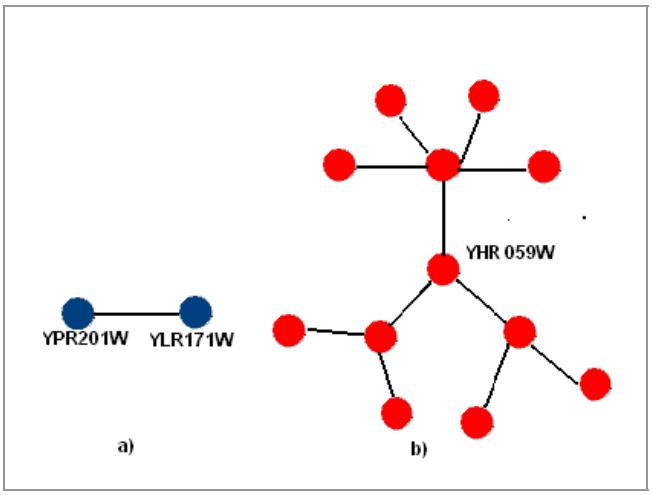

Figure 4.1. Interactions of proteins that are not member of any community (nodes are proteins and lines symbolize interaction between proteins)

Two communities of k-clique size of k=12 identified by CFinder when the interaction network is applied, were illustrated in 2 different ways in Figure 4.2. In Figure 4.2 (a) only the proteins that are overlapping between the 2 communities are given. In Figure 4.2 (b) proteins and interactions between and inside the communities are given; circles symbolize proteins and lines between the proteins shows the interactions between them. Community 1 is blue colored and is constituted of 16 proteins, purple circles are the 2nd community of the k-clique size of k=12 and red colored circles are the 8 overlapping proteins between the 2 communities.

CFinder also can give one community as a member of two different cliques, this is logical according to the definition of the community found by k-clique percolation method. In figure 4.3 nodes are the proteins and the interaction between the nodes is shown by edges. This is a special community called as 5-cliques, because it is a complete subgraph where every node has an interaction with the other 4 nodes. Additionally it is member of k-clique communities of 3, called k-clique community of size k=3, because it is union of all 3 –cliques that can be reached from each other through a series of adjacent 3-cliques, where adjacency means sharing $k-1$ nodes. As shown by the Figure, green and red triangles are the 3-cliques and they can be reached from each other by 2 nodes; YFL008 and YDL003.

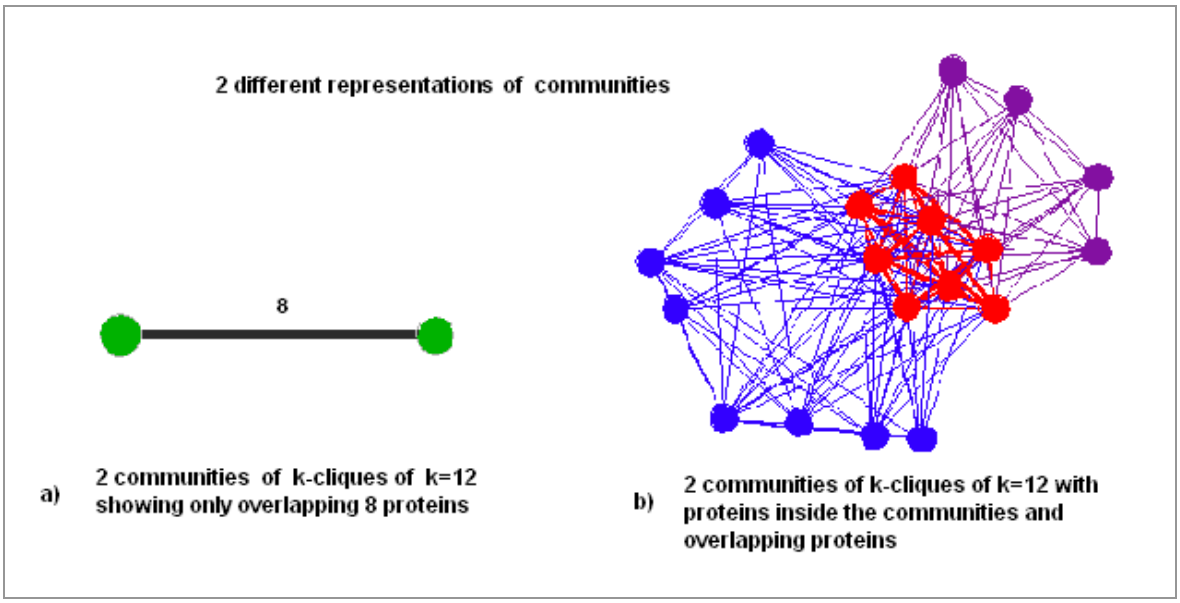

Figure 4.2. Communities of k-clique of k=12

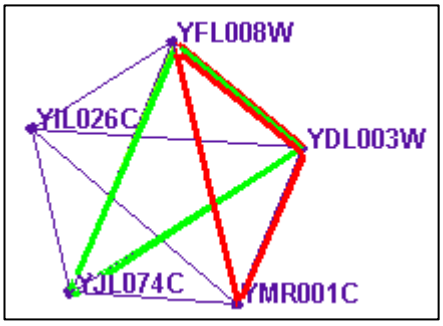

Figure 4.3. Community of k-clique of k =3 and 5-clique

CFinder algorithm is applied to assembled protein interaction network of the yeast to detect the modules of k-cliques. When the algorithm is applied, 427 k-clique communities of k ranging from 3 to 13 were found. Among 427 communities, 200 different communities had common overlapping proteins at least with one other community. In Figure 4.3, 200 overlapping communities are shown; green circles are constituted from densely interconnected proteins and the lines between the circles shows overlapping proteins between communities. There are some communities, not shown in the figure, have no overlapping proteins with other communities, and also, among 427 proteins, although they are members of different cliques, some communities are constituted exactly from the same proteins (Figure 4.3).

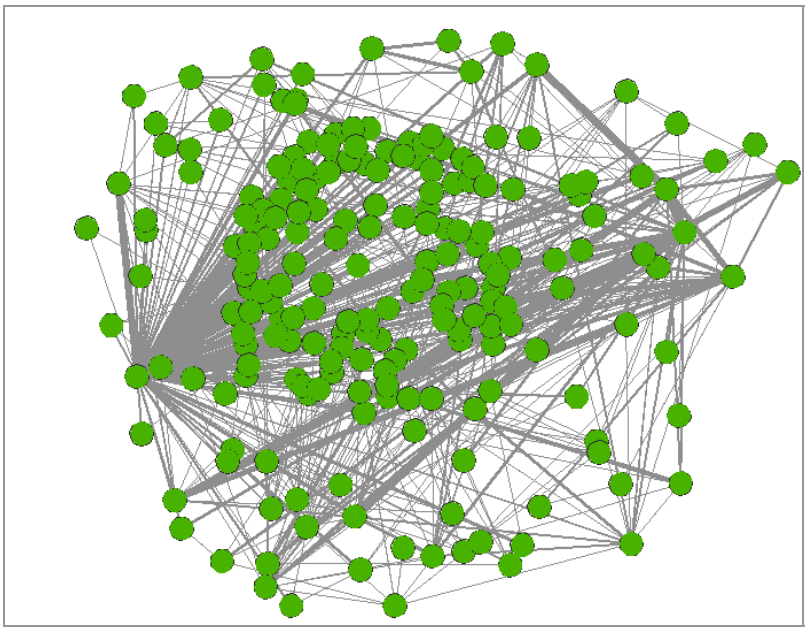

Figure 4.4. Overlapping 200 communities of CFinder (green nodes are the communities, lines show the overlaps between the communities)

In Figure 4.5, the distribution of the communities in different cliques is shown. Most of the communities are located on k-clique of k equals to 3 to 5, a clique of size of 3 includes 212 communities, the clique size of 4 includes 72 and a clique size of 5 contains 57 communities. There is an inverse relation between the k-clique size and the number of communities; the number of the communities that belong to a k-clique size decreases when the related clique size increases. For example 13-clique size contains only 2 communities and clique size of 12 contains only 3 communities.

In order to select highly interconnected communities that are not emerged by chance, we have created 10 random networks consisted of the same number of proteins and the same number of interactions as in the original interaction network, C-finder gave an average of 314 and 48 communities for k-clique of k=3 and 4 respectively (Figure 4.5). For k=5 only 4 communities found for 10 random networks run, and no community were found for k greater than 6. Therefore to increase the probability of finding complexes, in the rest of the work we investigated 7 to 13 cliques, in total 55 communities.

Among 55 communities, 23 of them have not any overlapping proteins with any other communities, 16 of them have overlapping proteins and the rest 16 are the same communities of some of those 39 communities (Figure 4.6).

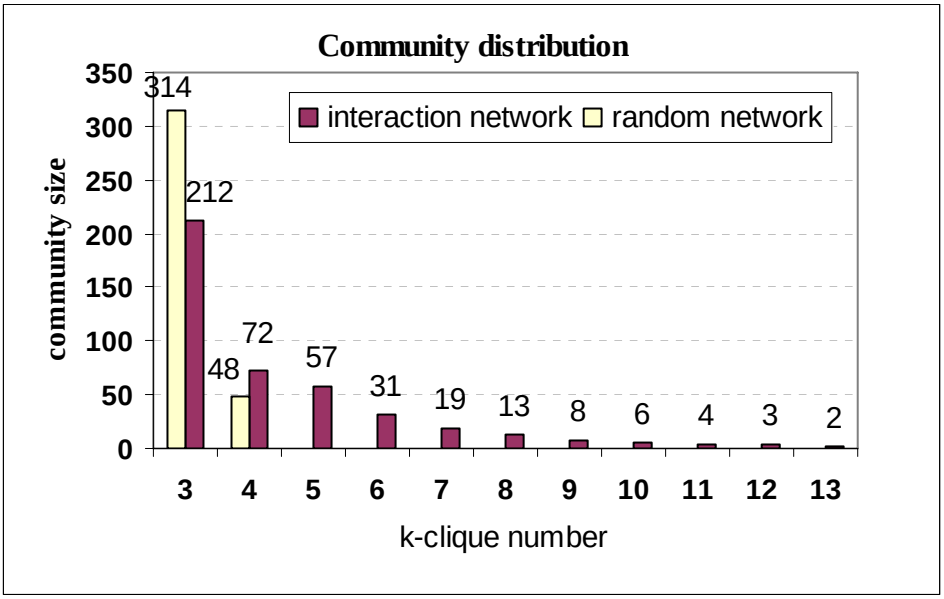

Figure 4.5. k- clique communities distribution

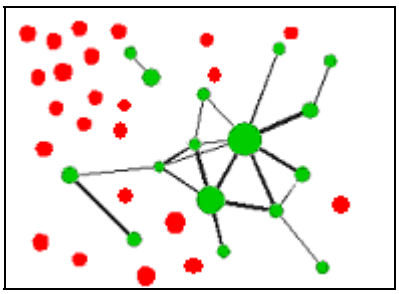

Figure 4.6. k-clique communities of k ranging 7 to 11 (nodes are the communities, the size of the nodes shows the number of proteins included in the corresponding modules. and the existence overlap between nodes are shown by lines, the width of the edges correlates with the number of linking proteins)

In Figure 4.6, 16 green nodes are the communities constituted of overlapping proteins, the lines shows the overlap between different communities. 23 red nodes on the other hand show non-overlapping communities. One node might have two or more community name. For example one of the red nodes in the Figure 4.6 represents the community of k10.0, k9.0, k8.0 and k7.3 which are constituted from the same proteins.

The number of communities that each protein participated was also investigated. We found that among the 4944 proteins interaction network as an input to CFinder, 1822 of them participates at least in one community and 3122 of them do not participate in any community. Most of the proteins participate in one or two communities, 895 proteins are only member of 1 community, 349 proteins are members of 2 communities. There are very few proteins that participate in more than 12 communities; only 2 proteins are members of 14 communities, 2 other proteins are the subunits of 17 communities and only one protein is member of 18 communities (Fig. 4.7).

To elucidate which kinds of proteins and what properties make those proteins participating in many communities, some examples of these proteins are given; YLR200W participates in 18 communities, YML094W and YNL153C are involved in 17 communities. All of those genes are subunits of the heterohexameric cochaperone prefoldin complex which binds specifically to cytosolic chaperonin and transfers target proteins to it (Vainberg *et al.*, 1998; Brew CT and Huffaker TC, 2002; Siegers *et al.*, 1999). Therefore, they have very dense interaction network and have interactions with many other proteins and biological pathways.

The proteins that participate in 14 communities are YGR086W which is primary component of eisosomes, large immobile cell cortex structures associated with endocytosis (Reinders *et al.*, 2007) and YLR085C which is an actin-related protein (ARP) that binds nucleosomes; a component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A . It is known that all of the nuclear ARPs that have been studied in detail are constituents of either ATP dependent nucleosome remodeling complexes or histone acetyltransferase complexes, both of which are involved in the modification of chromatin structure (Wu *et al.*, 2005). So, these proteins are involved in the regulation of transcription and other DNA transaction, therefore they have broad interactions with other proteins.

Therefore, we can conclude from the above in formations that CFinder gives highly interactive groups of proteins, but as most of the experiments in the literature are based on the transcription and signaling mechanisms, there is an enriched data on the interaction of

the proteins involved in these types of processes. Therefore we expect the complexes related on the transcription or signaling processes enriched in the communities.

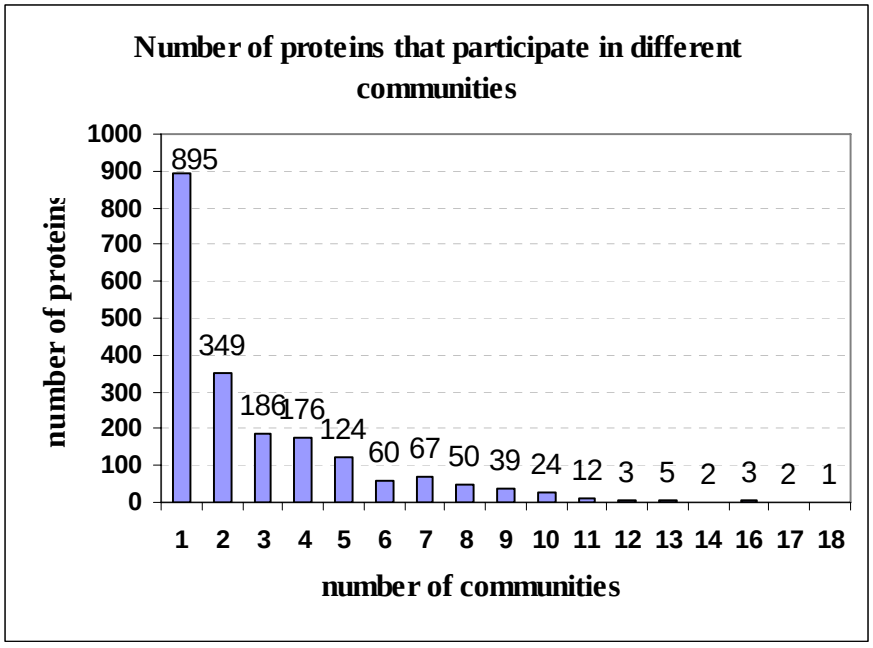

Figure 4.7. Number of proteins that participate in different communities

4.3. Identification of Protein Complexes

In this work, taking into account the presence of a highly statistically significant modular structure of the yeast interaction network, the k-clique communities, k size ranging from 7 to 13, including a total of 55 communities were investigated. Fitness function values of the 55 communities were calculated and the results are given in Figure 4.5.). Twenty communities which have FF values greater than the cut-off value (>0.64) were selected (Figure 4.8).

In order to address whether the related communities obtained from CFinder algorithm includes transient or permanent complexes, the GO component terms associated with the members of each community (cut-off> 0.64) were determined using GO term finder (SGD). Interestingly among the 20 communities, all of them were protein complexes or part of the complexes (Table 4.2). So cut-off value is very effective to select communities enriched with complexes.

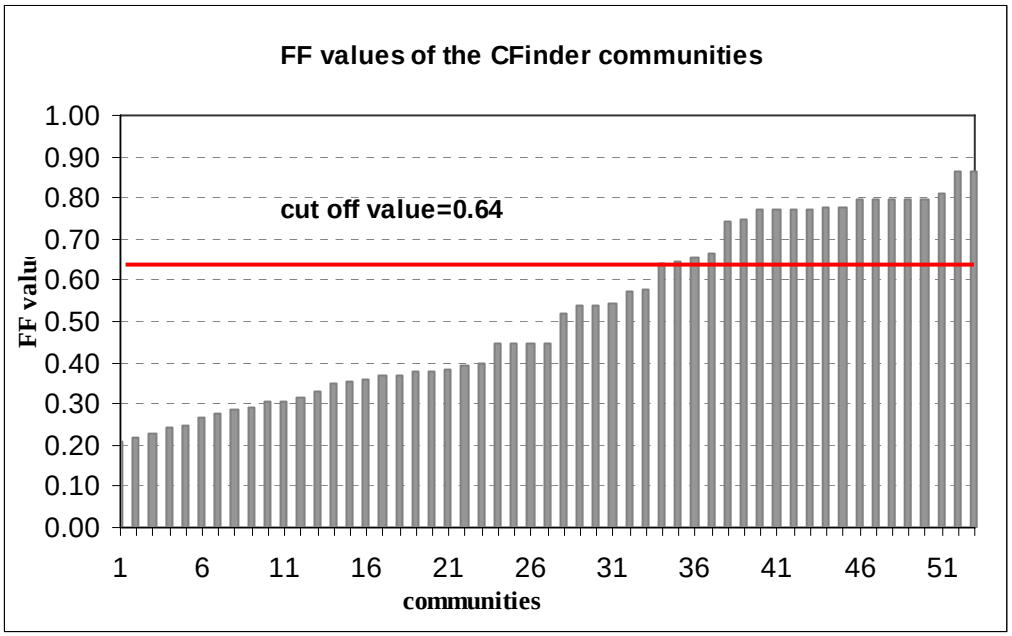

Figure 4.8. FF values of CFinder communities

GO terms associated with members of 20 communities indicated the presence of 11 different complexes. 9 of them ; nuclear exosome (NE), anaphase promoting complex (APC), transport protein particle (TRAPP), mRNA cleavage factor (mRNACF), RSC, spliceosomal uridine-rich small nuclear ribonucleoprotein (snRNP U1), mediator (M), ARP2/3, transcriptional elongation factor (TEF) are transient complexes.

Two of them, proteasome regulatory particle (PRP) and DNA directed RNA polymerase II complex (RNA 2) are permanent complexes. Community frequency is the ratio of the number of the genes annotated as member of the complex to the total number of genes in the community.

These 11 complexes were reported to contain overall 295 proteins. The total community frequency is a very high number which is 94.8%; 128 out of 135 proteins of k-clique communities are members of the complexes (Table 4.2). Some of the communities have got the same proteins; therefore they have the same GO localization annotation terms as in the case of k9.6 and k7.17 that are annotated as proteasome regulatory particle.

Table 4.2. Communities that have cut off values>0.64 and annotated protein complexes

k-clique communities[1]	**GO component term**	**Community frequency[2]**	**Number of genes in the complex**
k10.0,k9.0,k8.0,k7.3	TRAPP	10 / 10 genes, 100.0%	10
k 7.0	ARP 2/3	7 / 7 genes, 100.0%	8
K7.8	NE	11 / 11 genes, 100.0%	13
K9.3	NE	9 / 9 genes, 100.0%	13
k11.0, k10.1, k9.2, k8.2, k7.5	APC	11 / 11 genes, 100.0%	16
k9.4	mRNACF	13 / 13 genes, 100.0%	20
k7.4	RSC	11 / 12 genes, 91.7%	17
k7.9	SnRNPU1	11 / 12 genes, 91.7%	19
k7.1	TEF	7 / 7 genes, 100.0%	19
k7.6	M	7 / 7 genes, 100.0%	20
k9.6	PRP	14 / 16 genes, 87.5%	22
K10.4	PRP	11 / 12 genes, 91.7%	22
k7.7	POLII	6/ 8 genes, 75.0%	12
TOTAL: 20 community	**11 complexes**	**128/135 genes, 94.8%**	**295**

1 k.n1.n2 : n1 is the k- clique number , n2 is the community number.

2 community frequency of greater than 50 % is included in the table.

4.3.1. Transport Protein Particle (TRAPP)

TRAPP is a multisubunit vesicle tethering factor composed of 10 subunits as involved in endoplasmic reticulum-to-Golgi trafficking (Kim *et al.*, 2006).

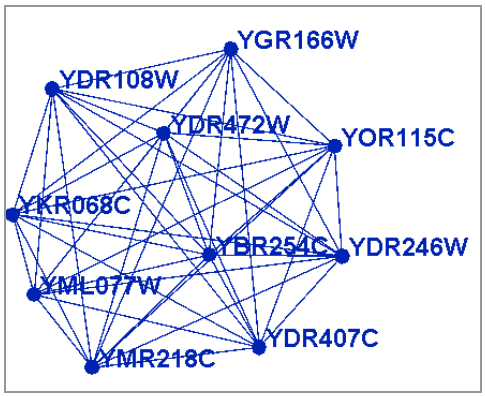

Figure 4.9. The members of the communities k10.0, k9.0, k8.0, k7.3 and TRAPP complex (nodes are the proteins and the lines show the interactions inside the community)

All of the communities k10.0, k9.0, k8.0 and k7.3 consist of the same 10 proteins (Table 4.2), so they have the same GO annotation term associated with the transport protein particle (TRAPP). Ten known members of the complex could be identified by the FF value of related communities, so all of the proteins of the TRAPP complex is identified with a community frequency of 100 %. Identified community is shown in Figure 4.9, it is a complete subgraph of 10 proteins, where every protein has an interaction with the other 9 proteins in the community. The community is also member of other k-cliques communities, this is logical according to the definition of k-clique community. Every complete subgraph is member of the k value smaller than the nodes of the community. For example this community is a complete subgraph of 10 proteins and it is a member of k=9 because it is also a union of all 9–cliques that can be reached from each other through a series of adjacent 9-cliques, where adjacent means k-1, 8 nodes, the same is true for k=8 and k=7.

4.3.2. Arp 2/3

Arp 2/3 complex is an actin-binding protein which is important for assembling the actin. It has been proposed to influence cell shape and motility *in vivo*. The complex consists of 8 subunits (Machesky et al, 1999).

The community k7.0 is a complete subgraph of 7 proteins (Figure 4.10) which are members of the Arp 2/3 complex. There is one other protein, YLR429W, which was not detected in k7.0 but is also annotated to GO component term Arp2/3 complex with category *colocolize with (SGD)*. "*Colocalizes with" term* appears to annotate gene products transiently or peripherally associated with an organelle or complex to the relevant cellular component term; in cases where the resolution of an assay is not accurate enough to say that the gene product is a bona fide component member (www.yeastgenome.org).

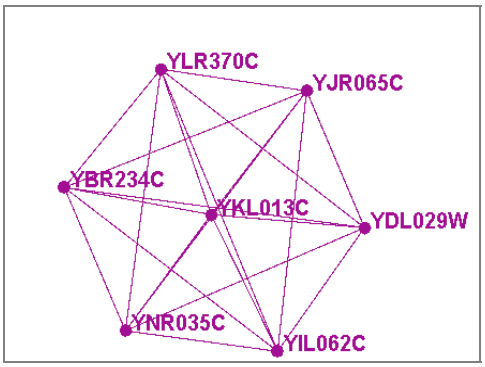

Figure 4.10. Members of the community k.7.0

According to Humphries and his collaborators (Humphries *et al.*, 2002), YLR429W encodes a yeast coronin (Crn1) that physically associates with the Arp2/3 complex and inhibits WA- and Abp1-activated actin nucleation *in vitro*, although there is not enough evidence that the YLR429W is a member of the complex. In the interaction network used in the thesis, YLR429W has not any interactions with the members of the Arp 2/3. Therefore it is not included in the community 7.0. Experimental evidence for the interactions of the YLR429W are necessary to identify whether it is member of the complex or not, but now, may be due the false negative interactions in the interaction network, YLR429W seems not to be a member of the complex.

4.3.3. Nuclear Exosome (NE)

The nuclear exosome (NE) complex is involved in multiple RNA processing and degradation pathways. How exosome is recruited to particular RNA substrates and then chooses between RNA processing and degradation modes is unknown (Vasiljeva *et al.*, 2006). NE consists of 12 subunits (SGD). One additional protein YNL251C is also referred with nuclear exosome term but within the category of *colocalizes with*. As explained before, this category is used in cases where the resolution of an assay is not accurate enough to say that the gene product is a bona fide component member (www.yeastgenome.org). According to the work of Vasiljeva *et al.* (2006), RNA binding protein YNL251C complexed with its partners Nab3, Sen1, and cap binding complex, physically interacts with the nuclear form of exosome, stimulates the RNA degradation activity of the exosome in vitro but there is not enough evidence that it is directly member of the Nuclear exosome (NE) complex.

In Table 4.3, members of the communities and the NE complex with the interaction number of the proteins within the communities are given. Nine members of the NE complex were identified in the community k9.3 with a community frequency of 100 %. It is a complete subgraph where every protein has 8 interactions. The bold members YOL142W and YNL251C, in the NE column are non-identified proteins by the communities. YOL142W, YNL251C are members of the NE but are not identified in the communities. YNL251C has not any interactions with the proteins in the complex, this might be due to the false negatives in the interaction network, or as mentioned before, because there is not enough evidence to say that the this gene product is a bona fide component member of NE it might not be the member of the complex also. On the other hand YOL142W have 5 interactions with the members of the NE. Therefore with other members of the community k7.8 it might be in a k-clique community of k less or equals to 6.

Eleven proteins of the NE complex is identified by the community k7.8 (Table 4.3), in addition to the 9 proteins of the community k9.3, in k7.8, 2 more proteins of NE, YGR158C and YHR081W are also identified.

YGR158C and YHR081W could be included in 7-clique community, because they have 6 and 7 interactions with the members of the community and k7.8 is a union of communities of 7–cliques where the members should have at least 6 interactions with the members of the community.

Table 4.3. Proteins of NE and the communities k9.3, k7.8 and k6.13 with interaction number (Int no) of some proteins (non-identified proteins are written in bold)

NE	Int no	k9.3	Int No	k7.8	Int no
YCR035C		YCR035C	8	YCR035C	9
YDL111C		YDL111C	8	YDL111C	9
YDR280W		YDR280W	8	YDR280W	10
YGR095C		YGR095C	8	YGR095C	9
YGR158C		YGR195W	8	YGR158C	6
YGR195W		YHR069C	8	YGR195W	10
YHR069C		YNL232W	8	YHR069C	9
YHR081W		YOL021C	8	YHR081W	7
YNL232W		YOR001W	8	YNL232W	10
YOL021C				YOL021C	10
YOL142W	**5**			YOR001W	9
YOR001W					
YNL251C*	**0**				

**Annotated with the term colocolize with*

4.3.4. Anaphase Promoting (APC)

Anaphase Promoting (APC) is another complex identified by CFinder. The APC is a multisubunit E3 ubiquitin ligase that targets cell-cycle-related proteins for degradation by the 26 S proteasome (SGD). The APC contains 16 subunits and is regulated by the binding of co-activator proteins and by phosphorylation. It is not known why the APC contains 16 subunits when many other ubiquitin ligases are small single-subunit enzymes.

In table the list of the proteins of APC and the communities associated with the GO component term of APC is given. 5 communities, k11.0, k10.1, k9.2, k8.2 and k7.5 are

annotated as APC component. They all contain the same 11 proteins of the APC complex. Non-identified 5 proteins of the APC are written in bold in Table 4.4.

The community k11.0 is a complete subgraph of 11 proteins, where every protein has 10 interactions with other proteins in the community (Table 4.4). According to the definition of the k-clique community, it is also member of other k-clique communities such as k10.1, k9.2, k8.2, k7.5.

Table 4.4. Proteins of APC and communities, k11.0, k10.1, k9.2, k8.2 and k7.5 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

APC	**Int no**	**k11.0/ k10.1/k9.2/k8.2 /k7.5**	**Int no**
YBL084C		YBL084C	10
YDL008W		YDL008W	10
YDR118W		YDR118W	10
YDR260C	**1**	YFR036W	10
YFR036W		YGL240W	10
YGL003C	**0**	YHR166C	10
YGL116W	**0**	YKL022C	10
YGL240W		YLR102C	10
YGR225W	**-**	YLR127C	10
YHR166C		YNL172W	10
YIR025W	**0**	YOR249C	10
YKL022C			
YLR102C			
YLR127C			
YNL172W			
YOR249C			

Undetected proteins of the APC complex are YGR225W, YGL116W, YGL003C, YIL025C and YDR260C. Among them YGR225W was not included in the interaction network. YGL003C, YGL116W and YIL025C have not any interaction with other members of the APC and YDR260C has only one interaction in the interaction network (Table 4.4). In the literature 16 proteins are annotated as members of APC from the direct assays, so they have physical interactions with other members of APC, the interactions

between some of proteins of the complex are not listed in our interaction network, as there are some false negatives in the interaction network.

4.3.5. mRNA Cleavage Factor (mRNACF)

The posttranscriptional maturation of eukaryotic mRNA 3' ends is an essential step in gene expression. This maturation occurs in two steps that are tightly coupled *in vivo*, but can be experimentally uncoupled *in vitro*. First, the nascent transcript is cleaved at a specific site downstream of the translational stop codon, and then a polyadenylate tail is added. This processing is affected by a multisubunit complex (mRNACF), cleavage factor that provides the nuclease activity and also confers specificity to a template-independent poly(A) polymerase . The complex consist of 4 subunits; 2 cleavage factors, CF I and CF II recognize the processing signals of the RNA and perform the endonucleolytic cleavage, whereas CF I, polyadenylation factor I, and the single-polypeptide PAP are required for the polyadenylation step (Gros and Moore., 2001). In total, mRNACF consists of 20 subunits (Table 4.5).

13 proteins of mRNACF were detected by the community k9.4 with a community frequency of 100 % (Table 4.2). However, 7 proteins written in bold in Table 4.5; YDR228C YGL044C, YKL018W, YNL222W, YOL123W, YOR179C, YOR250C were not in these communities due to the low interactions between the community members. In the interaction network, YNL222W, YOL123W, YOR179C, YOR250C have 3 or 4 interactions (Table 4.5) and YGL044C, YKL018W have not any interactions with the members of the mRNACF. As some proteins in the interaction network have no interactions, this might be due to the false negatives in the interaction data used. Because there are several experiments proving the physical interactions of those proteins with the members of mRNACF (Gross and Moore, 2001).On the other hand as YDR228W has an interaction number of 7 (Table 4.5), it might be included k –clique communities of k smaller than 7.

Table 4.5. Proteins of mRNACF and community k9.4 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

mRNACF	Int no	k9.4	Int no
YAL043C		YAL043C	12
YDR195W		YDR195W	12
YDR228C	**7**	YDR301W	12
YDR301W		YER133W	9
YER133W		YGR156W	9
YGL044C	**0**	YJR093C	11
YGR156W		YKL059C	12
YJR093C		YKR002W	12
YKL018W	**0**	YLR115W	12
YKL059C		YLR277C	10
YKR002W		YMR061W	9
YLR115W		YNL317W	12
YLR277C		YPR107C	11
YMR061W			
YNL222W	**3**		
YNL317W			
YOL123W	**3**		
YOR179C	**4**		
YOR250C	**3**		
YPR107C			

4.3.6. RSC

RSC complex is a complex identified by the community k7.5. RSC is a 15-subunit complex with the capacity to remodel the structure of chromatin. It exhibits a DNA-dependent ATPase activity stimulated by both free and nucleosomal DNA and a capacity to perturb nucleosome structures. It is essential for mitotic growth (Cairns *et al.*, 1996).

Community k7.5 contains 11 proteins that are annotated by RSC complex out of 12 proteins (Table 4.6). The proteins written in bold in the column of RSC (Table5.6) are unidentified proteins by the community. YCR020W-B, YGR056W, YGR275W, YBL006C are not included in the community because only YGR275W have interactions with the 4

proteins of the community k7.5 and others have not any interactions with the proteins of the RSC complex (Table 4.6).The protein YBR245C written in bold in the column of the community k7.5 is not member of the RSC complex, actually it is the member of the imitation-switch (ISWI) class of ATP-dependent chromatin remodeling complexes; ATPase that forms a complex to regulate transcription elongation, and a complex to repress transcription initiation. There are interactions between YBR245C and 6 subunits of k7.5 and 8 units of RSC. Furthermore, additional proteins interact, thereby linking the promoter nucleosome targeted ISW1 a complex to RSC function (Van Vugt *et al.*, 2007). So it might be a potential key protein linking two functional groups together (Futschik *et al.*, 2007), or might be member of the RSC complex.

Table 4.6. Proteins of RSC and community k7.4 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

RSC	Int no	k7.4	Int no
YCR020W-B	**0**	**YBR245C**	**6**
YCR052W		YCR052W	7
YDR303C		YDR303C	10
YFR037C		YFR037C	10
YGR056W	**0**	YIL126W	7
YGR275W	**4**	YKR008W	7
YIL126W		YLR033W	10
YKR008W		YLR357W	10
YLR033W		YML127W	7
YLR321C		YMR033W	10
YLR357W		YMR091C	10
YML127W		YPR034W	6
YMR033W			
YPR034W			
YBL006C	**0**		

4.3.7. Spliceosomal Uridine-Rich Small Nuclear Ribonucleoprotein (snRNP U1)

Spliceosomal uridine-rich small nuclear ribonucleoprotein **(**snRNP U1) is the first to bind the splicing complex to form the commitment complex during spliceosome assembly.

The branch-point bridging protein binds to a snRNP of U1 and the spliceosome cycle proceeds to the A complex. Six members of the snRNP U1 and six other proteins form commitment complex. That is the initiation step for the whole spliceosome assembly, and is crucial (Seraphin and Rosbash, 1989).

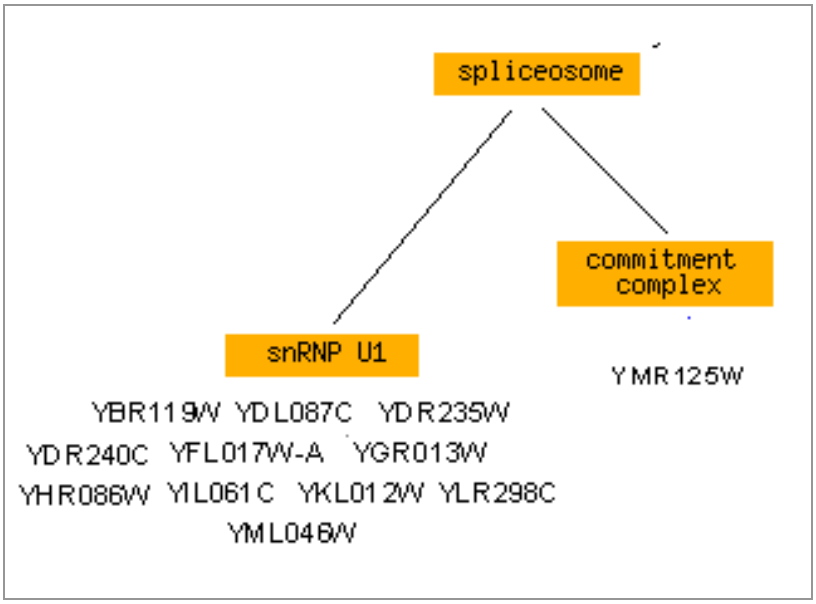

Figure 4.11. Distribution of the proteins of the community k7.9 over GO component terms

SnRNPU1 consist of 19 proteins (SGD). In Figure 4.10, the GO component terms of the 12 members of the community k7.9 is given.11 out of 12 proteins of k7.9 are members of snRNP U1 (Table 4.2). Only one member of it, written in bold in Table 4.7, YMR125W, is not included in the snRNP U1 complex. Actually it is a member of the commitment complex that is crucial for the start point of the splicesome assembly. In total 12 proteins; six members of snRNP U1, YMR125W and 5 other proteins form commitment complex. YMR125W might be a potential key protein for the formation of commitment complex from the snRNPU1 proteins as it has high interactions with the members of snRNP U1, or it might be a member of the snRNP U1 also.

7 proteins written in bold in Table 4.7 are undetected proteins of the snRNP U1 by the community k7.9. The interaction number of the unidentified proteins is low, YGR074W and YPR182W have 4 interactions; YER029C and YLR275W have 3 interactions YLR147C has 2 interactions, YOR159C has only one interaction with the

members of the k7.9. Therefore they are not included in the k-clique community of k7.9 where the proteins have 7 or more interactions (Table 4.7). As snR19 is not included in the interaction network, it will not be detected in any community.

Table 4.7. Proteins of snRNP U1 and community k7.9 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

snRNP U1	**Int no**	**k7.9**	**Int no**
YBR119W		YBR119W	
YHR086W		YDL087C	9
YML046W		YDR235W	7
YKL012W		YDR240C	10
YDR235W		YFL017W-A	9
YER029C	**3**	YGR013W	9
YGR074W	**4**	YHR086W	10
YLR275W	**3**	YIL061C	11
YLR147C	**2**	YKL012W	10
YOR159C	**0**	YLR298C	11
YFL017W-A		YML046W	7
YPR182W	**4**	YMR125W	9
YIL061C			
snR19	**-**		
YDR240C			
YGR013W			
YLR298C			

4.3.8. Transcriptional Elongation Factor (TEF)

The transcription elongation factor TEF is a component of RNA polymerase II preinitiation complexes. Transcription of protein-coding genes by RNA polymerase II is a dynamic process that begins with the formation of a preinitiation complex at the promoter and proceeds through initiation, elongation, termination, and, finally, reinitiation.

During the transition from initiation to elongation, promoter-specific contacts between RNA 2 and the preinitiation complex are disrupted as the polymerase begins messenger RNA synthesis and traverses into the ORF. The efficiency of elongation by RNA 2 is regulated by a number of Transcriptional elongation factors (Kim *et al.*, 2007).

Elongation factors bind to enhance the rate of transcription. Elongation occurs at approximately 60 nucleotides /second.

Table 4.8. Proteins of TEF and the community k7.1 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

TEF	Int no	k7.1	Int no
YAL021C	**0**	YBR279W	6
YBR279W		YGL207W	6
YDR138W	**0**	YGL244W	6
YER164W		YLR418C	6
YGL207W		YML069W	6
YGL244W		YOL145C	6
YGR063C	**-**	YOR123C	6
YGR116W	**0**		
YHR167W	**1**		
YLR418C			
YML010W	**0**		
YML062C	**0**		
YML069W			
YNL139C	**0**		
YNL230C	**0**		
YOL145C			
YOR123C			
YPL046C	**0**		
YPR133C	**0**		

TEF actually consist of several complexes as Paf1, FACT, DSIF, nucleoplasmic THO, ELL-EAF and transcription elongation factor b. The FACT complex is an abundant complex that has been shown to reorganize the structure of the nucleosome. In this way, the FACT complex may play a role in DNA replication and other processes that traverse the chromatin, as well as in transcription elongation. Paf1 is a multiprotein complex that associates with RNA polymerase II and general RNA polymerase II transcription factor complexes and may be involved in both transcriptional initiation and elongation (Krogan et al., 2002). DSIF is a heterodimeric protein complex which is expressed in eukaryotes from yeast to man, it is an inhibitory elongation factor that promotes RNA polymerase II

transcriptional pausing, but can also stimulate transcriptional elongation under certain conditions. ELL-EAF is a heterodimeric protein complex that acts as an RNA polymerase II elongation factor as the transcription elongation factor b. Not all of the physical interactions between these complexes have been unraveled (Krogan *et al.*, 2002).

Transcriptional elongation factor consists in total of 19 proteins (SGD). We identified 7 proteins of transcriptional elongation factor in the community k7.1. When further investigated, 5 proteins of the k7.1 are member of the Paf1 complex that consists of 7 proteins and 2 members of the k7.1 are member of the FACT complex that consists of 2 proteins. According to S. Squazzo and his friends (2002) Paf1 complex have physical interactions with the FACT complex, this information validate our results. Other 11 proteins which are written in bold in Table 4.11, are not identified in the community k7.9. Among them, YGR063C is not in the interaction network, and beside from YHR167W which has a 1 interaction, other 9 proteins have not any known interactions with any of the members of the community of the k7.1. This is why they are not included in the community k7.1 which is a complete subgraph where every member has interactions with other 6 members of the community (Table 4.11). The reason for those proteins not included in the community might be due to the false negatives in the interaction network.

4.3.9. Mediator (M)

Mediator (M) is a multiprotein complex that functions as a transcriptional coactivator. The mediator complex is required for the successful transcription of nearly all class II gene promoters in yeast. M functions as a coactivator and binds to the C-terminal domain of RNA polymerase II holoenzyme, acting as a bridge between this enzyme and transcription factors (Biddick and Young, 2005).

Mediator complex consists of 20 proteins (SGD), and we identified 7 proteins of M with a community frequency of 100% in the community k7.6 (Table 4.2, Table 4.9). 13 non identified proteins are written in bold in Table 4.9. These 13 members could not be identified in the community k7.6, because they have less than 5 interactions with the members of the community k7.6 (Table 4.9). They might be in the k-clique communities where k number is less or equal than 6.

Table 4.9. Proteins of M and the community k7.6 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

M	Int no	K7.6	Int no
YBL093C	**5**	YBR193C	6
YBR193C		YBR253W	6
YBR253W		YDL005C	6
YDL005C		YER022W	6
YDR308C	**3**	YGL025C	6
YER022W		YOL135C	6
YGL025C		YOR174W	6
YGL127C	**4**		
YGR104C	**5**		
YHR041C	**5**		
YHR058C	**4**		
YLR071C	**3**		
YMR112	**3**		
YNL236W	**4**		
YNR010W	**4**		
YOL051	**5**		
YOL135C			
YOR174W			
YPL129W	**3**		
YPR070W	**3**		
YPR168W	**2**		

4.3.10. Proteasome Regulatory Particle (PRP)

The ubiquitin-proteasome pathway regulates a wide variety of biological processes (Hershko and Ciechanover, 1998). The proteasome degrades proteins conjugated to ubiquitin and thus plays a central role in this pathway (Finley, 2002). The two major subcomplexes of the proteasome are the proteolytic core particle and regulatory particle (PRP). The PRP is thought to bind ubiquitin chains (Lam *et al.*, 2002), unfold the attached substrate protein, and translocate the substrate into the core proteasome.

Table 4.10. Proteins of PRP and the communities, k9.6 and k10.4 with interaction number (Int No) of some proteins (non-identified proteins are written in bold)

RP	Int no	k10.4	Int no	k9.6	Int no
YDL007W		YDL097C	10	YBR217W	8
YDL097C		YDL147W	10	YDL007W	11
YDL147W		YDR394W	11	YDL097C	15
YDR394W		YER021W	11	YDL147W	13
YDR427W		YFR004W	11	YDR394W	14
YER021W		YFR010W	9	YDR427W	10
YFR004W		YFR052W	9	YER021W	13
YFR010W		YGL004C	9	YFR004W	13
YFR052W		YHR027C	11	YFR010W	11
YGL048C		YHR200W	11	YFR052W	12
YHR027C		YKL145W	10	YGL004C	11
YHR200W		YOR261C	11	YGL048C	11
YKL145W				YHR027C	14
YOR261C				YHR200W	15
YPR108W	**11**			YKL145W	15
YDL020C	**7**			YOR261C	13
YDR363W-A	**3**				
YGR232W	**5**				
YIL075C	**5**				
YLR421C	**4**				
YOR117W	**7**				
YOR259C	**7**				

PRP consist of 22 proteins (SGD).We identified 11 members of the PRP complex out of 12 proteins in community k10.4 and 14 members of RP complex out of 16 proteins in column k9.6 (Table 4.10). Unidentified 7 proteins are written bold in Table 4.13 at the column of PRP. Among the unidentified proteins although YPR108W have 11 interactions with the members of k9.6 which is a union of all 9-cliques that can be reached from each other through a series of adjacent 8-cliques, it is not a member of community k9.6 because mathematically YPR108W does not a member of such 9 clique. It is forms an adjacency of 7 cliques. Other unidentified proteins generally have less than 8 interactions, therefore not

included in the communities of k10.4 at which members have minimum of 10 interactions and k9.6 at which members have minimum 8 interactions.

One protein of k10.4 (YGL004C) and 2 proteins of k9.6 (YBR217W, YGL004C) written in bold in Table 4.13 are not member of the PRP complex. Actually YGL004C is an uncharacterized protein and its molecular function is unknown but according to Huh et al (2003). YGL004C is a putative non-ATPase subunit of the 19S regulatory particle of the 26S proteasome, which also adjusts our results. There are also 5 proteins of PRP with molecular functions unknown which are YGR232W, YER021W, YDL147W, YOR261C and YDR363W-A. Although we do not know the function of YGL004C, physical interactions between the members of PRP show that it might be a member of the PRP.

4.3.11. RNA Polymerase II (RNA2)

RNA polymerase II core complex (RNA 2) is an enzyme that catalyzes the transcription of DNA to synthesize precursors of mRNA and most snRNA and microRNA. It consists of 12 subunits.

Table 4.11. Proteins of the RNA 2 and the community k8 with interaction number (Int no) of proteins (non-identified proteins are written in bold)

RNA 2	Int no	K8.8	Int no
YBR154C		YBR154C	7
YDL140C		YDL140C	7
YDR404C		YDR404C	7
YGL070C		YGL070C	7
YHR143W-A	**-**	**YGR005C**	**7**
YIL021W		**YGR186W**	**7**
YJL140W		YIL021W	7
YOL005C	**2**	YOR151C	7
YOR151C			
YOR210W	**3**		
YOR224C	**6**		
YPR187W	**5**		

In Table 4.11 members of the RNA 2 and community k8.8 is given.6 components out of 8 proteins of k8.8 are the members of the RNA 2 complex. 2 members written in bold in the column of k8.8 are not members of RNA 2, YGR005C and YGR186W, are not subunits of the core complex RNA 2, but are the components of the TFIIF that helps to speed up the polymerization process. TFIIF is a part of RNA polymerase, a holoenzyme that is constituted of 72 subunits. Therefore although they are not members of RNA2 YGR005C and YGR186W have very high interactions with other members of the k8.8. 6 of the components of the RNA 2 written in bold in Table 4.11, are not identified by the community k8.8. Among the unidentified proteins, YHR143W-A is not in the interaction network used. YOL005C and YOR210W have 2 and 3 interactions with the members of k8.8. And YOR224C and YPR187W have 5 and 6 interactions with the members of k8.8. Those two proteins may be involved with other members of k8.8 in k-clique communities of k less than 7.

4.3.12. Communities that Have Cut-off Values <0.64

Cut off value determined from FF values of the test data, provided us to detect the communities which are complexes. We referred communities as complexes if they contain member of complexes with a community frequency of greater than 50 %. Among the 55 communities, all of the 20 communities that have FF value greater than the cut off value had proteins which were member of some complexes. On the other hand, among the 35 communities which had lower FF values than the cut off value, only 5 communities had members which were also proteins of some complexes. According to this information we can say that FF value calculated from the PCCA, FSA, PSA, CSA and TI measures is very successful to detect the complexes. Although here we only investigated 55 communities, it will be beneficial firstly to easily filter communities from their FF values in a network of communities with size greater than few hundreds, and then investigate these communities which contain probably members of the complexes.

5 communities that have lower FF values than the cut off value were also annotated as complexes; RNA cleavage factor (mRNACF), proteasome regulatory particle (PRP) and prefolding complex (PC).

We had already detected PRP and mRNACF complexes from the communities that had FF values greater than the cut off value. But those complexes had FF value very near to the cut-off value, and they had community frequencies greater than the values in Table 4.12. For example the communities k9.6 and k10.4 were annotated as PRP with community frequency of 87.5% and 91.7%, had FF values of approximately 0.64 and 0.65. The community k9.4 which was annotated as mRNACF had a community frequency of 100% and had a FF value of 0.67. But now as the community frequency decreases, which means that they contain one or more proteins that are not the member of the related complexes they had lower FF values than the cut off value; the communities, k8.5 and k7.11 were annotated as mRNACF with a community frequency of 92.9% and had a FF value of ~0.54, k8.8 and k7. 17 were annotated as PRP with a community frequency of 78.9% and 81% and they had FF values of 0.54 and 0.57 respectively. And finally k9.7 was annotated as PC with a community frequency of 50% and had a FF value of 0.35.

Table 4.12. Communities that have a cut-off value <0.64 and annotated protein complexes

k-clique communities[1]	**GO component term**	**Cluster frequency[2]**	**Number of genes in the complex**
K8.5-K7.11	mRNACF	13 / 14 genes, 92.9%	20
K8.8	PRP	15 / 19 genes, 78.9%	22
K7.17	PRP	17/21 genes 81%	22
k9.7	PC	5 / 10 genes, 50.0%	6
Total: 5 communities	**3 complexes**	**50/64 genes 78%**	

Some of the complexes might have normally lower FF values for example very near to the cut-off value, when one or more proteins which are not members of the complexes are included in these communities, they got lower FF values than the cut-off value, therefore we can not detect these communities.

4.4. k-clique Communities with k =6

In this thesis we calculated FF values of the k-clique communities of k ranging from 7 to 13. When we run the program with 10 random network, only the communities of k-

clique of 3 and 4 and neglected number of communities in the k=5 were formed. As highly connected communitities of k equals to 6 and 5 are not also emerged by chance these cliques might probably contain communities which are members of the complexes.

In order to identify whether k-clique communities of k=6 consist of complexes, we directly investigated the GO component terms of the members of these 31 communities of k-clique communities of k=6. We found that 24 of these communities were annotated as 18 different complexes (Table 4.17). We have already identified 10 out of 18 of these complexes written in bold in Table 4.17, in k-clique communities of k ranging from 7 to 11. However in k-cliques of 6 we have identified 2 additional members of nuclear exosome (NE) complex, 4 additional members of mediator complex (M) and 1 additional member of the mRNA cleavage factor complexes. So, as the k-clique number decreases more members of the complexes can be identified in the communities. On the other hand, the contrary situation is also true: more proteins that are not member of the complexes also involved in the communities of smaller k. In total those 18 complexes have 447 members. All of the communities have 208 proteins annotated as complexes out of 248 proteins, so they have a community frequency of 81%. 40 of these proteins are not member of the complexes. The average community frequency was approximately %94.8 for the k-cliques communities of k ranging 7 to 13 (Table 4.2). As the clique size decreases the community frequency also decreases because as k number decreases, the number of interaction in k-cliques decreases, and as the interaction network is rich in proteins that have lower protein-protein interactions, more proteins involved in these communities although they are not member of the complexes. We expect for k size smaller than 5 will be more enriched with proteins that are not members of the complexes.

The minimum value of the community frequencies was 75% for the communities that have FF values greater the cut-off value for k clique communities k>6 (Table 4.2). However k-clique communities of k=6 were rich in communities that have lower frequencies. For example there are 6 communities that have community frequencies lower. By the usage of fitness function (FF) for k-clique communities of smaller k, we expect to eliminate the communities which lower community frequencies or in other words which are not rich in members of complexes. It is valuable to calculate FF values of the k-clique communities of k=6 in order to check this filtration criteria.

Table 4.13. k-clique communities of k=6 with related complexes (SGD) (complexes written in bold were already identified by k-clique communities of k >6)

k-clique* communities	GO term	Community frequency	Number of genes in the complex
6.2	**Arp2/3**	7 /7 genes, 100.0%	7
6.3	transcription elongation factor complex	7 /7 genes, 100.0%	19
6.4	DNA-directed RNA polymerase III complex	10 /10 genes, 100.0%	17
6.5	**PRP**	11/12 genes, 91.7%	46
6.6	transcription factor TFIID complex	7 /13 genes, 53.8%	15
6.7	SLIK (SAGA-like) complex	6 /6 genes, 100.0%	16
6.8	DNA-directed RNA polymerase I complex	10 /10 genes, 100.0%	14
6.1	**exosome (NE)**	13 /13 genes, 100.0%	14
6.11	**DNA-directed RNA polymerase II, core complex (RNA2)**	7 /9 genes, 77.8%	12
6.12	retromer complex | AmiGO	3 / 6 genes, 50.0%	5
6.13	**TRAPP complex**	10/10 genes, 100.0%	10
6.14	**RSC complex**	11 /12 genes, 91.7%	17
6.15	eukaryotic translation initiation factor 2B complex	5 /9 genes, 55.6%	5
6.16	**APC**	11 /11 genes, 100.0%	16
6.17	**M**	11/12 genes, 91.7%	20
6.18	small subunit processome	9 /9 genes, 100.0%	48
6.19	**snRNP U1**	10 /12 genes, 83.3%	19
6.2	**snRNP U1**	9 /11 genes, 81.8%	19
6.21	**mRNA cleavage factor complex**	14 /15 genes, 93.3%	20
6.22	protein kinase CK2 complex	4 / 8 genes, 50.0%	4
6.23	snRNP U6	7 / 12 genes, 58.3%	9
6.24	proteasome accessory complex	16 /22 genes, 72.7%	22
6.26	transcription factor TFIID complex	5 /6 genes, 83.3%	15
6.27	90S preribosome	5 /6 genes, 83.3%	58
	Total 18 complexes	**208 /248 genes, 81%**	**447**

* k-clique communities with the frequency of % 50 and higher are listed.

4.5. Reliability of the Method

In order to check the reliability of the method the contribution of different data types used in the calculation of the FF was investigated according to distinct complexes grouped in Table 4.14. Six well known permanent complexes; 20-s proteasome (20-s P), cytoplasmic large ribosome (CLR), cytoplasmic small ribosome (CSR), mitochondrial large ribosome (MLR), mitochondrial small ribosome (MSR), DNA-directed RNA polymerase 1 (RNA1) used in the weight calculation were notated as GROUP 1, 9 well known transient complexes anaphase promoting complex (APC), nuclear exosome (NE), Spt-Ada-Gcn5Acetyltransferase (SAGA), replication complex (R), pre-replicative complex (PR) and transcription factor-II (TFII) were notated as GROUP 2. Nine transient complexes; nuclear exosome (NE), anaphase promoting complex (APC), transport protein particle (TRAPP), mRNA cleavage factor (mRNACF), RSC, spliceosomal uridine-rich small nuclear ribonucleoprotein (snRNP U1), mediator complex (M), ARP2/3, transcriptional elongation factor (TEF) detected from the k- clique communities of k ranging 7-13 were notated as GROUP 3 and the k-clique communities of k ranging from 7 to 13 and that are not complexes were notated as GROUP 4.

Table 4.14. Groups of complexes.

Name of the complexes	**GROUPS**
6 well known permanent complexes	GROUP 1
9 well known transient complexes	GROUP 2
11 transient complexes detected	GROUP 3
Communities that are not complexes	GROUP 4

4.5.1. CSA Results of the Communities and Complexes

If the protein complex localizes in a particular subcellular compartment, then all its subunits should be present in the same compartment as well (Jansen et al, 2000). If all of the members of the k-clique community are annotated in the same cellular compartment, the resulting component similarity (CSA) is 1. Therefore when the localization of the components of the complexes (whether transient or permanent) investigated, they should have a CSA value nearly 1. In some cases one or more uncharacterized proteins might be

in the communities, these proteins might have different cellular compartments then the members of the communities and these proteins might decrease the CSA value.

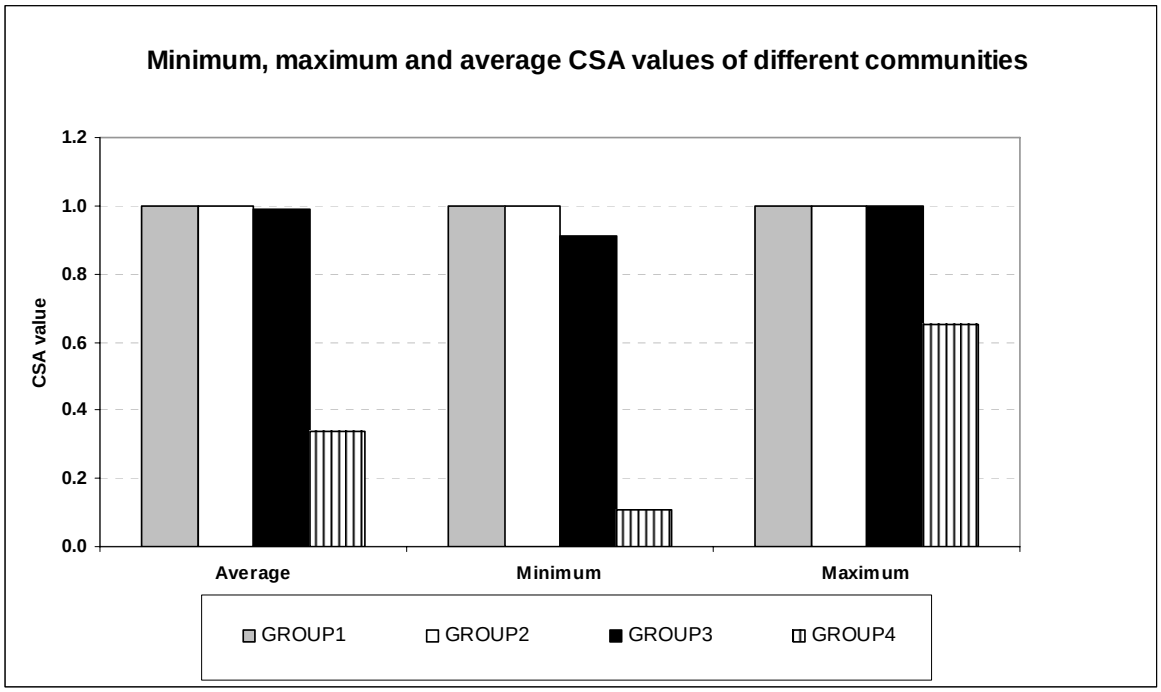

Figure 4.12. Average, minimum and maximum CSA values of the complexes and communities

According to the results (Figure 4.12), the average of the CSA values of the well known permanent complexes (GROUP 1) and transient complexes (GROUP 2) were calculated to be equal to 1. The average CSA value for the predicted transient complexes by their FF values (GROUP 3) was approximately 0.99. It is not equal to 1 because the CSA value of the community k7.9 associated with the snRNPU1 complex is equal to 0.91, as the community k 7.9 includes a protein from another localization which is a protein of commitement complex. As a result we can say that the members of the permanent and transient complexes locate in the same subcellular location as mentioned in the literature and therefore their CSA value is equal to 1. However the inclusion one or more protein into the communities that are not member of the protein complex, or any misinformation about the localization of the proteins may cause a decrease in CSA value. For the GROUP 4, both the maximum and average CSA values of the communities are very low when compared with the CSA value of the complexes (Figure 4.12). Therefore

CSA measure is an efficient parameter to differentiate a complex from a community that is not a complex.

4.5.2. FSA Results of the Communities and Complexes

The function of any protein complex depends on the function of its subunits for example if the complex has a particular biochemical function, then this most likely also provides a functional definition for its subunits (Jansen *et al.*, 2002). The permanent complexes have higher FSA values than the transient ones because the members of the transient complexes might have different functional properties as they are participated in more than one function. For example some of proteins of the SAGA complex have different functional properties as transcriptional regulator activity, transcriptional binding and TATA binding protein binding activity. An average FSA value of approximately 1 was calculated for the permanent complexes (GROUP 1) as expected (Figure 4.13).

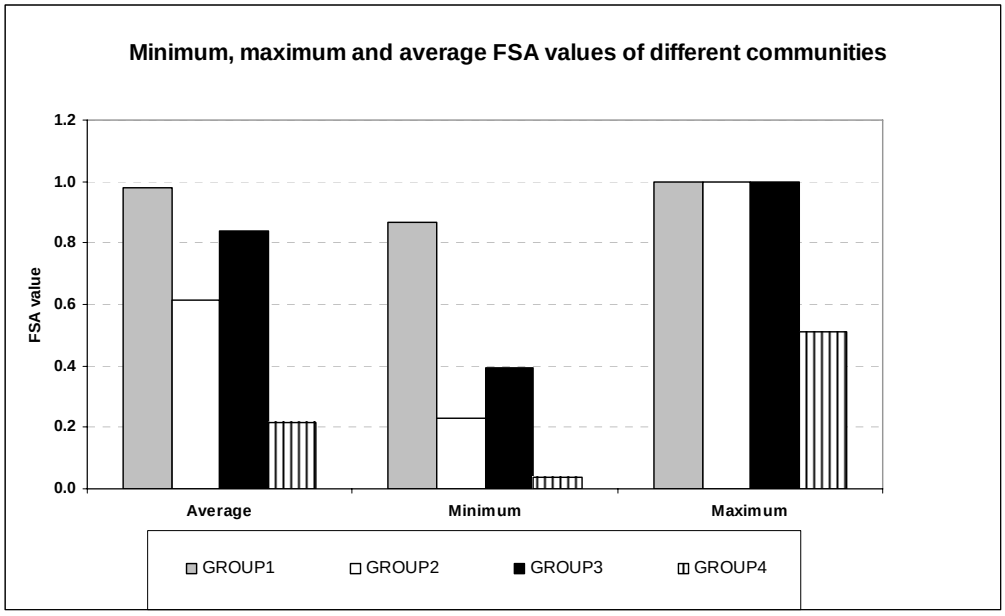

Figure 4.13. Average, minimum and maximum FSA values of the complexes and communities

Well known transient complexes in GROUP 2 and predicted transient complexes identified by FF values in GROUP 3 were found to have average FSA values of 0.60

and 0.83 which are lower than the average FSA value of well known permanent complexes in GROUP 1 (Figure 4.13).

Additionally FSA values of the communities that are not complexes (GROUP 4) have lower FSA value (0.2), this shows that although they have high interactions they are not concerned in the same biological function. So, FSA measure is an efficient parameter to differentiate a complex from a community at which members are highly interacted but that is not a complex.

4.5.3. PSA Results of the Communities and Complexes

As in the case of localization, functional annotation, well known permanent complexes (GROUP 1) well known transient complexes (GROUP2) and predicted transient complexes by their FF value (GROUP 3) were found to have average PSA values of approximately equal to 1 (Figure 4.14). This is because as the proteins in the protein complexes are expected to have a role in the same metabolic process, they have the same process term (Jansen *et al.*, 2002).

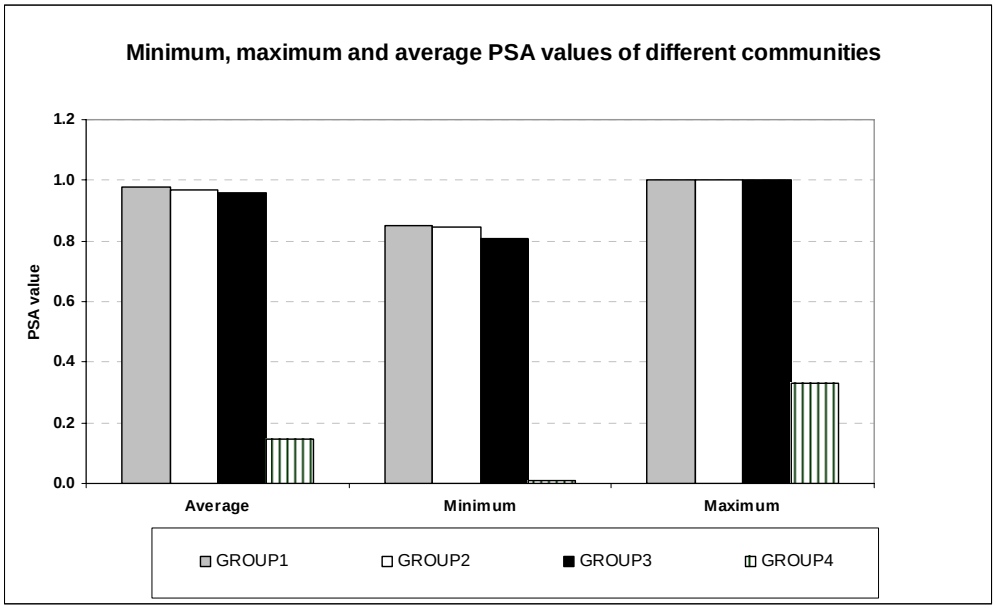

Figure 4.14. Average, minimum and maximum PSA values of the complexes and communities

The communities that are not complexes in GROUP 4 were found to have very low PSA values which are equal to, approximately 0.18. This result shows that although they have high interactions they are not concerned in the same biological process. Therefore, PSA measure is an efficient parameter to differentiate a complex from a community that is not a complex.

When the data of the average process and functional annotation similarity measures (PSA, FSA) of the complexes and communities which are not complexes are plotted (Figure 4.15), it is observed that communities which are not complexes (GROUP 4) were located near to the origin, transient complexes (GROUP 2&3) were in the middle area as some of them have lower FSA and PSA values and the permanent complexes (GROUP 1) have at the upper corner, as they have high FSA and PSA values. Some of the complexes may have lower PSA, FSA values because they may compromise some uncharacterized proteins. Additionally transient complexes might have lower FSA and PSA values because due to the transient associations of the proteins they may be involved more than one biological process and functions, however by using PSA, FSA measures one can separate a protein complex from a community that have lower FSA and PSA values (Figure 4.14).

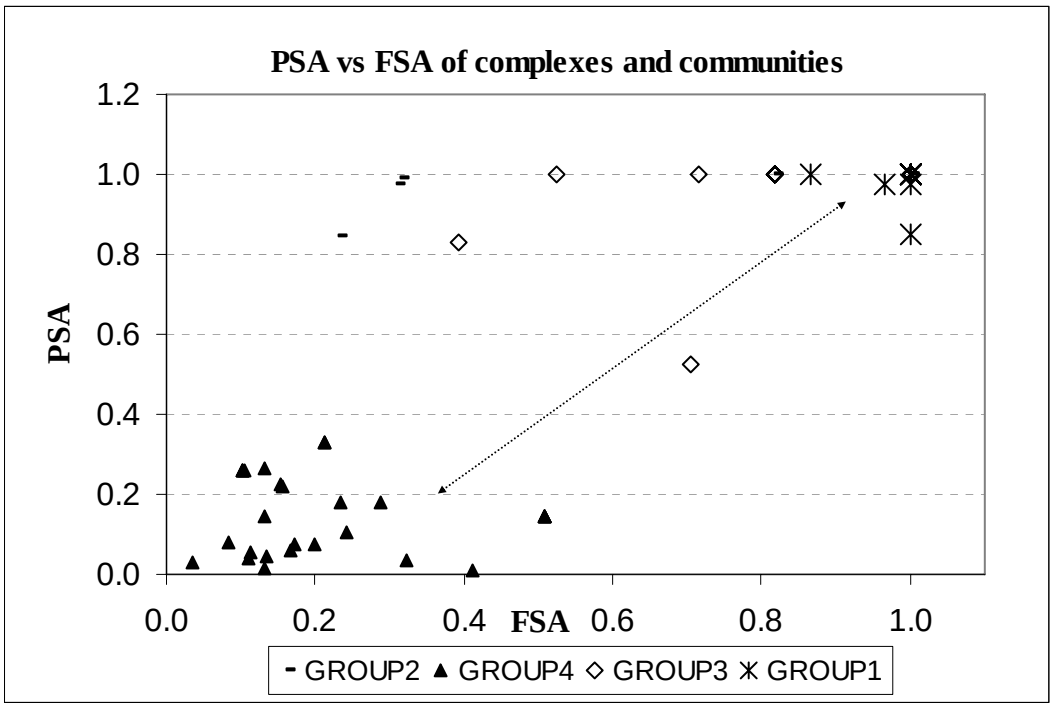

Figure 4.15. FSA vs. PSA values of the complexes and communities

4.5.4. Correlation of Expression Profiles (PCCA)

In Figure 4.16 the expression correlation data for permanent complexes in GROUP 1 and transient complexes in GROUP 2 are plotted by using the expression data from the studies by Gasch *et al.* (2001) and by Galitski *et al.* (1999). In the Gasch *et al.* (2001) mutations in components of the ATR/Mec1 pathways are introduced into the experiments in order to create DNA damage, which results in hypersensitivity to DNA-damaging agents in yeast. Thus, cells have evolved complex surveillance mechanisms that monitor genomic integrity during normal cell-cycle progression and in response to DNA damage, and they orchestrate a multifaceted response to DNA damage to ensure accurate transmission of genetic information (Gasch *et al.*, 2001). On the other hand in the Galitski *et al.* (1997) microarray-based gene expression data is analyzed to identify genes showing ploidy dependent expression in isogenic *Saccharomyces cerevisiae* strains that varied in ploidy from haploid to tetraploid. These genes are induced or repressed in proportion to the number of chromosome sets, regardless of the mating type.

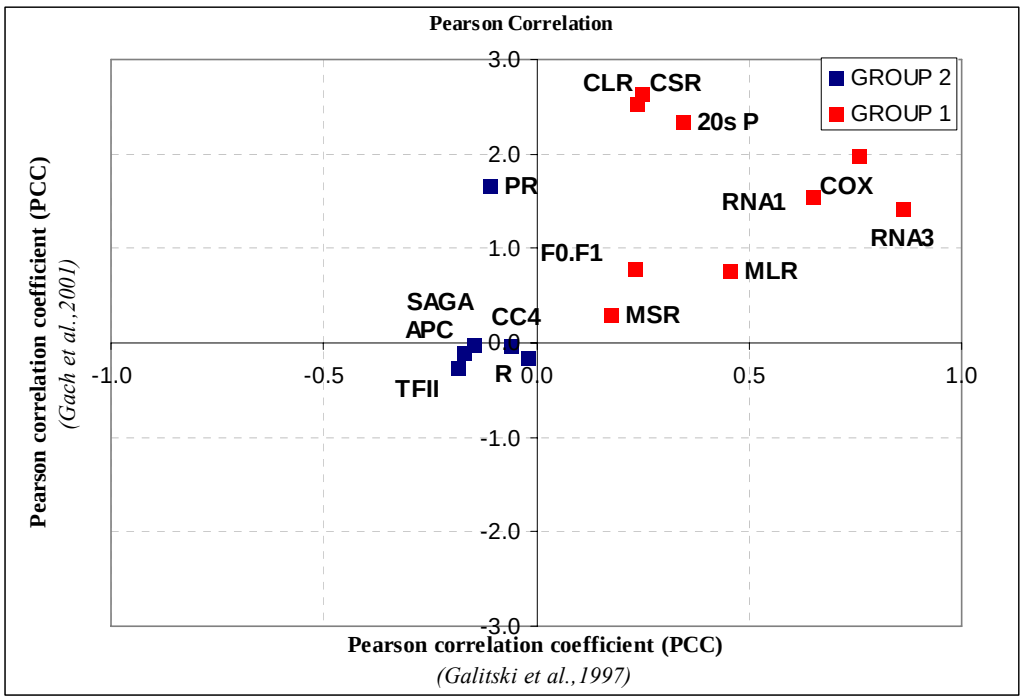

Figure 4.16. Expression correlation of test and training complexes for Galitski *et al.* (1997) and Gasch *et al.* (2001) data

Accordingly in the Figure 4.16, permanent complexes; cytoplasmic ribosomes (CLR and CSR), RNA1, RNA3, 20s Proteasome (20 s P) , F0.F1 and mitochondrial complexes (MLR, MSR) in GROUP 1 displayed higher average expression correlation (PCCA) when compared to the complexes SAGA, TFII, APC,R,CC4 complexes in GROUP 2. The fact that PR complex have higher co-expression value for the Gasch *et al.* (2001), microarray data, some of the components of PR may act as permanent complexes. Or higher co-expression of PR members than other transient complexes may be caused by experimental noise in the microarray data.

When expression correlation values of other 4 microarray data are plotted (Figure 4.17, Figure 4.18), it was observed that subunits of the same permanent protein complex show significant co-expression, in terms of similarities of expression profiles; the genes encoding especially the member proteins of cytoplasmic ribosomes, RNA1 and RNA3 and 20s Proteasome, COX showed high expression correlations.

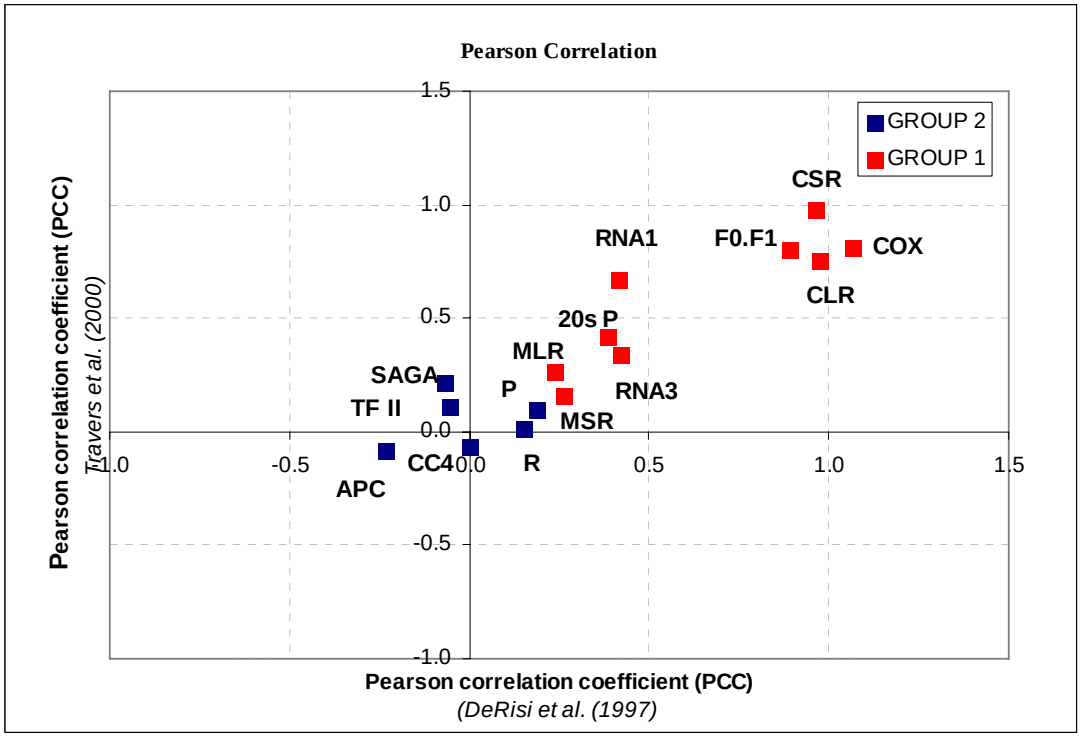

Figure 4.17. Expression correlation of test and training complexes for DeRisi *et al.*(1997) and Travers *et al.* (2000) data

The genes encoding the members of transient complexes as SAGA, anaphase promoting complexes, TFII, replication and prereplication (PR) complexes were found to have lower Pearson correlation coefficients which are approximately zero. However, the genes encoding members of mitochondrial ribosomal complexes generally were found to have lower expression correlation value.

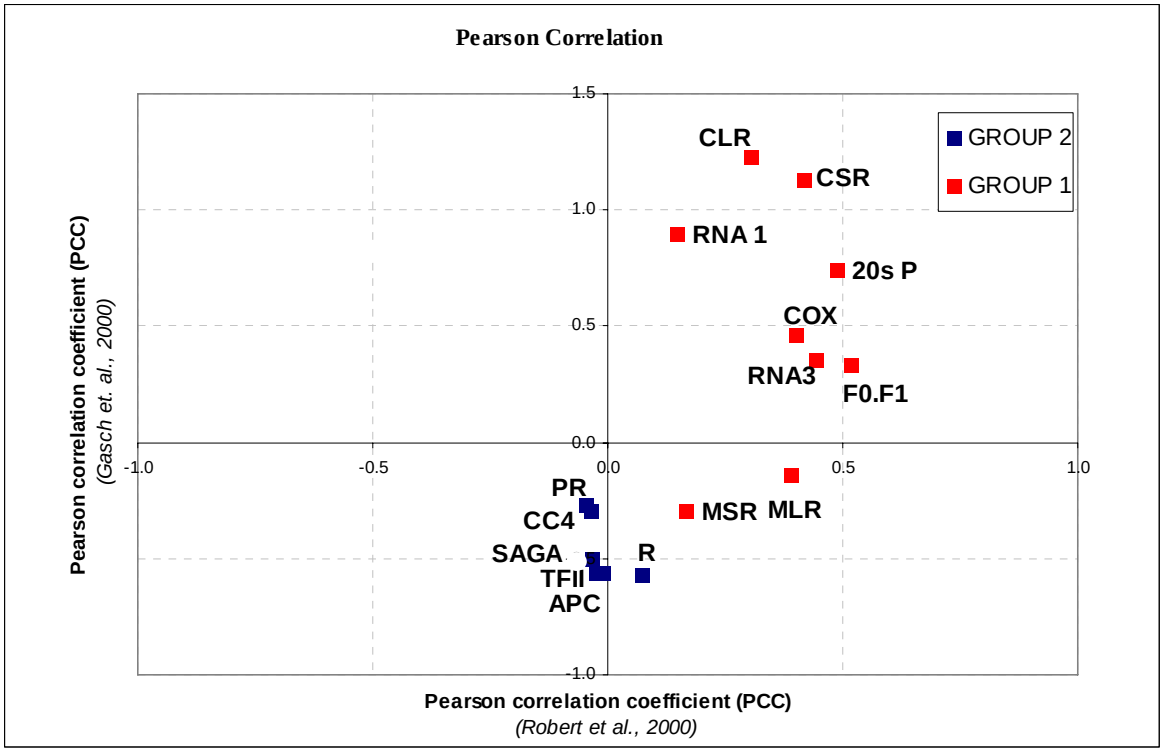

Figure 4.18. Expression correlation of permanent and transient complexes by using the microarray data from the studies of Gasch *et al.* (2000) and Robert *et al.* (2000)

Although permanent complexes were calculated to have high correlation values, in some microarray data they displayed a lower correlation values in some experiments which might be due to the type of the experiment. Therefore, 14 microarray data were taken into account to decrease this type of variations. The average Pearson correlation coefficients of 14 microarray data were calculated and presented for the test; training and the predicted complexes are given in Figure 4.19.

The distribution of the average Pearson correlation values (PCCA) of the complexes was investigated and assumed that PCCAs may be classified into 3 groups. The 1st group consists of permanent complexes that have high PCCA values. This group includes 6 well

known complexes of the GROUP 1 namely, 20-sP, CSR, CLR, F0.F1, RNA3, COX and PRP which is a permanent complex identified in the thesis by its FF values, was also located in this region. Two transient complexes and Arp 2/3 of GROUP 3 that have high PCCA values were also in this region. 2nd group consists of generally permanent complexes; that have lower PPCA values than the complexes in the 1st group; MLR, MSR and RNA 1 and RNA and one transient complex TEF of the GROUP 3. 3rd group however generally consists of well known transient complexes (GROUP 2) and transient complexes predicted by their FF values (GROUP 3) that have lower PCCA values. Although not shown in Figure 4.19, the communities that are not complexes (GROUP 4) have also lower PCCA values as transient complexes.

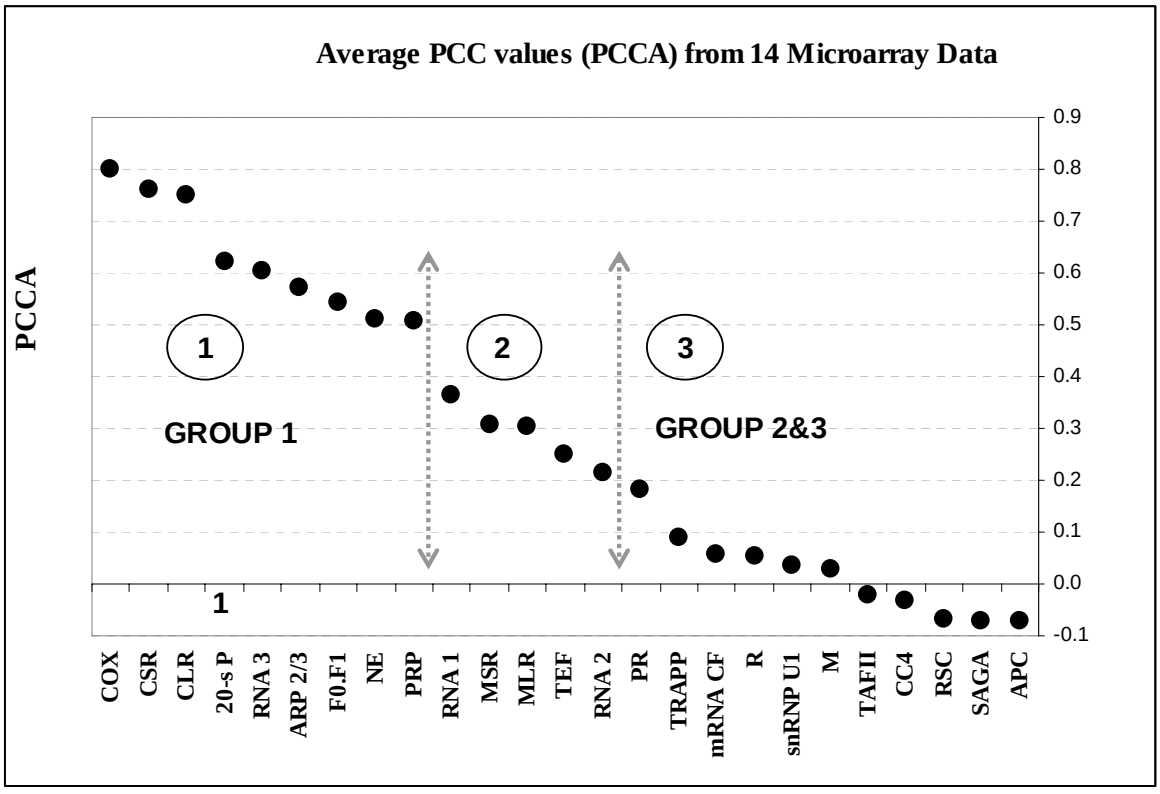

Figure 4.19. Average Pearson correlation coefficient (PCCA) of complexes

The Figure 4.20 represents average, minimum and maximum PCCA values of the well known permanent complexes (GROUP 1), well known transient complexes (GROUP 2), predicted transient complexes by their FF values (GROUP 3) and the communities that are not complexes (GROUP 4). Permanent complexes were identified to have higher PCCA value with an average of 0.55. However transient complexes in GROUP 2 and 3

were calculated to have an average PCCA values between 0 and 0.25 approximately. The PCCA value of the communities that are not complexes (GROUP 4) was calculated to be equal to 0.2. This result indicated that the members of the transient complexes are not co-expressed in a cell, they act as group of random proteins as the communities of GROUP 4.

Consequently according to these results, we can conclude that permanent complexes are generally highly co-expressed, however some permanent complexes as mitochondrial ribosomes and RNA2 consist of proteins that are not highly co-expressed. Some of the transient complexes are well co expressed in the cell as in the case of nuclear exosome (NE) and Arp 2/3. However, on the contrary to permanent complexes, transient complexes generally consist of proteins that are not well co-expressed in the cell; actually they tend to have a co-expression level as random groups of proteins. Those results are good agreement by the previous studies (Jansen *et al.*, 2002). Therefore it can be concluded that co-expression may be used valuable measure to differentiate a permanent complex from a transient complex or a community that is not a complex.

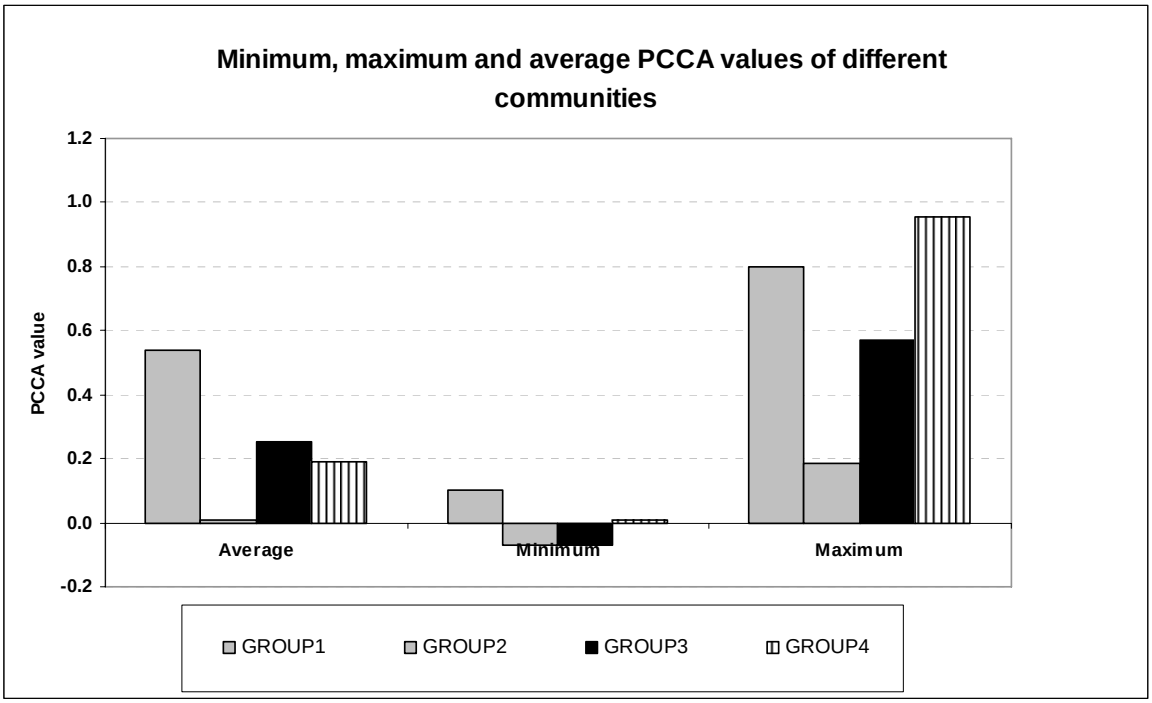

Figure 4.20. Average, minimum and maximum PCCA values of the complexes and communities

4.5.5. Protein-Protein Interaction (PPI)

The average, minimum and maximum PPI values of the all of the complexes in GROUP 1, 2, 3 and 4 were calculated and presented in Figure 4.21.

CFinder gives communities with high PPI values (>0.6). However both the test and validation complexes used at the start point were less connected complexes (PPI<0.6) than the resulting communities obtained by the CFinder because not all of the subunits in a protein complex are in structural closeness and thus do not physically interact with one another (Futschik *et al.*, 2007). They generally form a coherent structural unit as a whole with common properties.

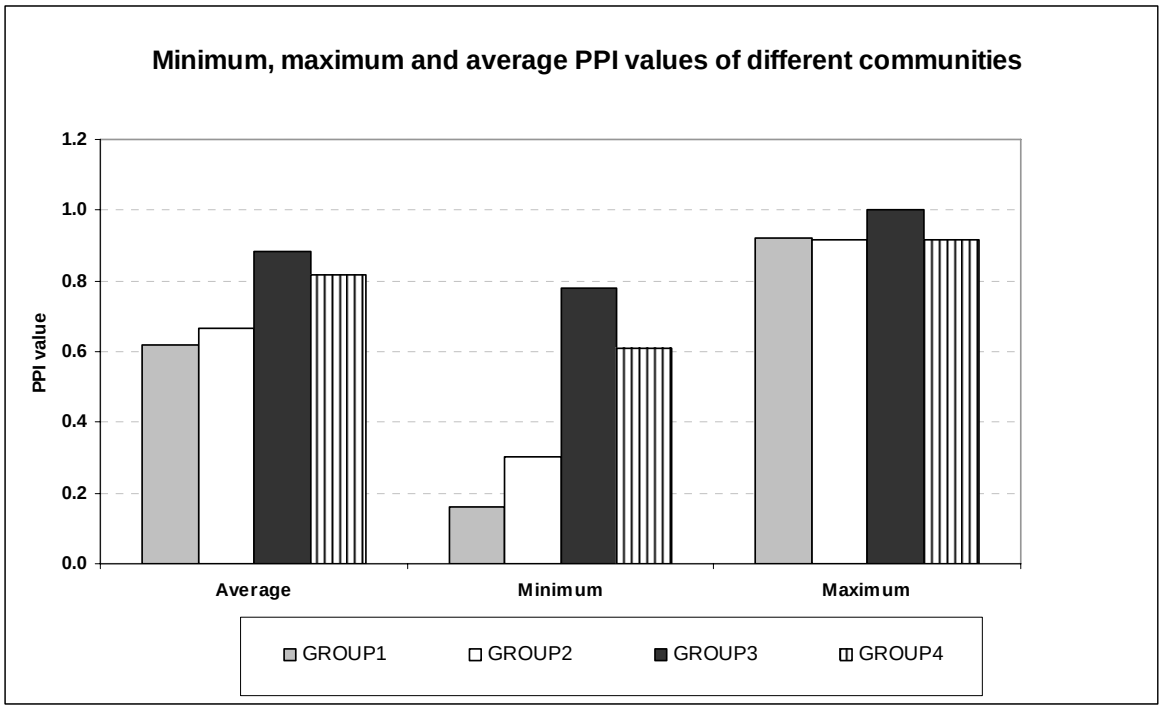

Figure 4.21. Average, minimum and maximum PPI values of the complexes and communities

Therefore it should be better not to include this measure in the calculation of FF value because approximately all of the communities have high average PPI values; we could not differentiate a complex from the community by their PPI measures.

4.5.6. Transcriptional Regulation (TI)

In order to map coregulated protein complexes, an algorithm was developed to find regulation degree of the dense protein clusters in the PPI network coregulated by one or more TFs. A high TI number is an indicator of a co-regulation of the proteins in the module by comparatively lower number of TFs. Such protein modules were termed co-regulated protein clusters. In the Figure 4.22, TI values of the complexes in different groups and the communities that are not complexes are given. The permanent complexes were identified to have high TI values so they tend to be co-regulated. The regulation of the transient complexes was found to be similar to that of the random communities that are not assigned complexes. The low co-regulation of the members of the transient complexes is due to the just in time assembly of some of its subunits (Figure 2.1). Generally very few members of the transient complexes are sufficient to govern the activation of an entire complex (Lichtenberg *et al.*, 2007).

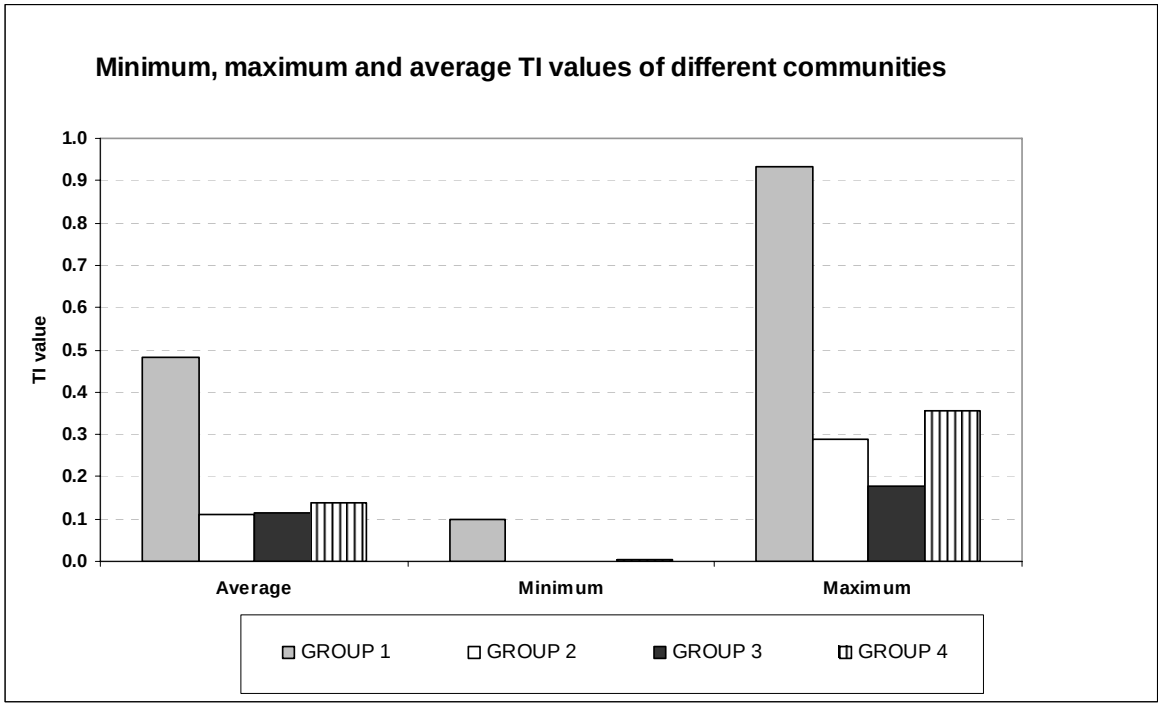

Figure 4.22. Average, minimum and maximum TI values of the complexes and communities

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

In this thesis the protein complexes in an interaction network consisting of 20487 interactions between 4944 proteins were determined by an integration of different biological datasets. We recruit different data sources that include co-expression measures (PCCA), interaction data (TI), and GO process (PSA), function (FSA) and localization similarity (LSA) information. Each of these data sources individually contains some weakly predictive information with respect to protein complexes, but this prediction is improved by combining all of them by the fitness function (FF). The weights in the formula of FF values were determined according to the measures of 6 well known permanent complexes namely 20-s Proteasome (20-s P), cytoplasmic large ribosome (CLR), cytoplasmic small ribosome (CSR), mitochondrial large ribosome (MLR), mitochondrial small ribosome (MSR), DNA-directed RNA Polymerase 1 (RNA1). The weights calculated by genetic algorithm were 0.2101, 0.1992, 0.0292, 0.1941, 0.1542, and 0.2131 for the co-expression (PCCA), co-regulation (TI), process (PSA), function (FSA) and component (CSA) similarity measures respectively. The average FF value, 0.64, was set as the cut-off value by using the measures of 6 well known transient complexes; anaphase promoting complex (APC), nuclear exosome (NE), Spt-Ada-Gcn5Acetyltransferase (SAGA), replication complex (R), Pre-replicative complex (PR) and transcription factor-II (TFII)

When the CFinder algorithm is applied to the interaction network, due to the dense interactions inside a community, among 4944 proteins, CFinder could only clustered 1822 proteins inside one or more community; 427 k-clique communities of k ranging from 3 to 13 were found. Although they were members of different cliques, some communities are constituted exactly from the same proteins. According to the results of the applied 10 random networks, C-finder gave an average of 314 and 48 communities for k-clique of k=3 and 4 respectively. For k=5 only 4 communities found for 10 random networks runs, and no community were found for k greater than 6. To increase the probability of finding complexes, we investigated 7 to 13 cliques, in total 55 communities were investigated.

According to the cut-off value that was selected (0.64), 20 communities had FF value greater than 0.64. All of the 20 communities were protein complexes or part of the complexes; 11 different complexes were found, 9 of them; nuclear exosome (NE), anaphase promoting complex (APC), transport protein particle (TRAPP), mRNA cleavage factor (mRNACF), RSC, spliceosomal uridine-rich small nuclear ribonucleoprotein (snRNP U1), mediator (M), ARP2/3, transcriptional elongation factor (TEF) are transient complexes. 2 of them, proteasome regulatory particle (PRP) and DNA directed RNA polymerase II complex (RNA 2) are permanent complexes. These 11 complexes contain overall 295 proteins. The total community frequency is 94.8%; 128 out of 135 proteins of k-clique communities are members of the complexes. 100 % of the proteins of the transport protein particle (TRAPP) was identified with a community frequency of 100 %. Seven proteins of the anaphase promoting complex (APC) complex consisting of 8 proteins were identified with a community frequency of 100 %. 11 members of the nuclear exosome (NE) were identified in the community k7.8 with a community frequency of 100 %. 11members of snRNP U1 that consist of 19 members is identified with a community frequency of 91.7%. 7 proteins of the Arp 2/3 complex that consist of 8 proteins were identified with a community frequency of 100 %. 11 of RSC complex that consist of 19 proteins were identified with a community frequency of 91.7%. 13 proteins of mRNACF that consist of 20 proteins were detected with a community frequency of 100 %. Transcriptional elongation factor (TEF) consists in total of 19 proteins (SGD). We identified 7 proteins of transcriptional elongation factor that consist of 19 proteins with a community frequency of 100 %. Mediator complex (M) consists of 20 proteins (SGD), and we identified 7 proteins of M with a community frequency of 100%. PRP consist of 22 proteins (SGD). We identified 11 members of the PRP complex with a community frequency of 91.7 % and 14 members of PRP complex with a community frequency of 87.5 %. 6 proteins of the RNA 2 is identified with a community frequency of 75%.

Among the 35 communities which had lower FF values than the cut off value, only 5 communities had members which were also proteins of some complexes; RNA cleavage factor (mRNACF), proteasome regulatory particle (PRP) and prefolding complex (PC).

To identify whether k-clique communities of k=6 consist of complexes, we directly investigated the GO component terms without calculating FF values of the members of the

31 communities of k-clique communities of k=6. We found that 24 of these communities were annotated as 18 different complexes.10 of these complexes were already identified by the k-clique communities of k>6. As the k-clique number decreases more members of the complexes were identified in the communities, 2 additional members of nuclear exosome (NE) complex, 4 additional members of mediator complex (M) and 1 additional member of the mRNA cleavage factor complexes were identified in k clique communities of k=6. On the other hand the contrary situation is also true: more proteins that are not member of the complexes also involved in the communities of smaller k; in total those 18 complexes have 447 members and all of the communities have 208 proteins annotated as complexes out of 248 proteins, so they have a community frequency of 81%. 40 of these proteins are not member of the complexes.

The above complexes were generally related with transaction of DNA as most of the experiments in the literature are based on the transcription and signaling mechanisms, there is an enriched data on the interaction of the proteins involved in these types of processes. Therefore the complexes related to the transcription or signaling processes were enriched in the communities.

Due to the false negatives and lack of information in the interaction network some proteins were not included in the communities that were annotated as complexes of k ranging 7 to 13 or not included in the interaction network. The function of YGL004C is not known in the literature but it is a putative member of the PRP complex, physical interactions between the members of PRP and YGL004C showed that it might be a member of the PRP. YBR245C have an interactions with 8 units of RSC, it might be a member of the RSC complex.

In order to check the reliability of the method, the contribution of different data types to the calculation of the FF was investigated according to 4 distinct groups of complexes; 6 well known permanent complexes, 9 well known transient complexes, 11 transient complexes detected by their FF values, and the communities that are not complexes. It is found that the average of the CSA, FSA, PSA values of the well known permanent complexes and transient complexes were higher compared to the communities that are not complexes. Some complexes might have lower FSA and PSA values. Some of the complexes may have lower PSA, FSA values because they may compromise some

uncharacterized proteins. Additionally transient complexes might have lower FSA and PSA values because due to the transient associations of the proteins so they may be involved more than one biological process and functions. The permanent complexes have higher FSA values than the transient ones because the members of the transient complexes might have different functional properties as they are participated in more than one function.

It was also observed that subunits of the same permanent protein complex show significant co-expression, in terms of similarities of expression profiles; the genes encoding especially the member proteins of cytoplasmic ribosomes, RNA1 and RNA3 and 20s Proteasome, COX showed high expression correlations, whereas the genes encoding the members of transient complexes as SAGA, anaphase promoting complexes , TFII, replication and prereplication (PR) complexes were found to have lower Pearson correlation coefficients which are approximately zero. However, the genes encoding members of mitochondrial ribosomal complexes generally were found to have lower expression correlation value. According to the PCCA values, we can conclude that permanent complexes are generally highly co-expressed; however some permanent complexes as mitochondrial ribosomes and RNA2 consist of proteins that are not highly co-expressed. Some of the transient complexes are well co expressed in the cell as in the case of nuclear exosome (NE) and Arp 2/3. However, on the contrary to permanent complexes, transient complexes generally consist of proteins that are not well co-expressed in the cell; actually they tend to have a co-expression level as random groups of proteins. Those results are good agreement by the previous studies (Jansen *et al.*, 2002). Therefore it can be concluded that co-expression may be used valuable measure to differentiate a permanent complex from a transient complex or a community that is not a complex.

The permanent complexes were also identified to have high TI values so they tend to be co-regulated. The regulation of the transient complexes was found to be similar to that of the random communities that are not assigned complexes. The low co-regulation of the members of the transient complexes is due to the just in time assembly of some of its subunits (Figure 2.1). Generally very few members of the transient complexes are sufficient to govern the activation of an entire complex (Lichtenberg *et al.*, 2007).

CFinder gives communities with high PPI values (>0.6), however both the permanent and transient complexes used at the start point are less connected complexes (PPI<0.6) than the resulting communities obtained from the CFinder because not all of the subunits in a protein complex are in structural closeness and thus do not physically interact with one another (Futschik *et al.*, 2007).

5.2. Recommendations

In the thesis, the FF values of 7 to 13 cliques were investigated. By the usage of fitness function (FF) for k-clique communities of smaller k, we expect to eliminate the communities which have lower community frequencies. For further work, it is valuable to calculate FF values of the k-clique communities of k=6 in order to check this filtration criteria.

Here we should take into account that the applied CFinder algorithm for enlightening of modular structures is restrictive, since it necessitates fully connected cliques; among the 4944 proteins 1822 of them participates at least in one community and 3122 of them do not participate in any community. However not all of the complexes are highly interacted, to identify less densely linked modules, different methods may be favorable. The main difference between permanent and transient complexes is co-expression of their units and co-regulation of the members, in the thesis although permanent complexes have higher FF values than the transient complexes, there are no thresholds to differentiate a permanent complexes than the transient ones. Therefore for further work, in addition to the FF values, filtering criteria taking co-expression levels of complexes into account can be used to differentiate the two types of complexes.

A related problem with the method is the incompleteness of current PPI networks. The more complete and accurate our PPI and known protein complexes datasets are, the more accurately we can analyze the PPI network. Some of the proteins were not in the interaction network due to the false negatives in the interaction network. Further, when using GO annotation terms in the calculation of FSA, the functional homogeneity, while accurate for the most part, seems to be an incomplete, oversimplified model. Many known complexes show low functional homogeneity (Zotenko *et al.*, 2006).

Some of the proteins have high interactions with the members of the complexes, therefore they might be members of the related complexes; however more experimental results as protein interaction data from the two yeast hybrid or tandem affinity purification followed by mass spectrometry, are necessary to check the membership of the potential members of the complexes.

APPENDIX A: MATLAB PROGRAMS

A.1. Calculation OF PCCA

```
clc
clear all
%input: gene expression profiles of genes in a complex (D matrix)
%method: using normalized profiles
%output: average correlation in a complex

loadD  %D is the expressiın profiles of genes
z=zeros(1,size(D,2))
for i=1:size(D,1)
   z=[z; zscore(D(i,:))]
end
z2=z(2:size(z,1),:); % normalized profiles (mean=0; std=1)
PCC=corrcoef(z2'); %Correlation matrix
PCC2=abs(PCC);
MEANCOMP=(sum(sum(PCC2))-size(PCC2,1))/(size(PCC2,1)^2-size(PCC2,1))
```

A.2. Calculation of TI, PPI, PSA, FSA, LSA

```
clear all;
CLD=[ ]  % enter the matrix inside the paranthesis
tic
for i=1:size(CLD,1)
for j=i:size(CLD,1)
if CLD(i,2)==CLD(j,2);
CLM(CLD(i,1),CLD(j,1))=1;
CLM(CLD(j,1),CLD(i,1))=1;
```

```
end
end
end
toc
n=size(CLM,2);
A=[]
for i=1:n-1;
if sum(CLM(:,i))==0;
i ;
A=[A,i];
end
end
size(A,2);
sum(sum(CLM))
averageCLM=(sum(sum(CLM))-(size(CLM,1)-size(A,2)))/((size(CLM,1)-
size(A,2))^2-(size(CLM,1)-size(A,2)))
```

REFERENCES

Adamcsek B., G. Palla, I.J. Farkas, I. Derényi and T. Vicsek, 2006, "Cfinder: Locating Cliques and Overlapping Modules in Biological Networks", *Bioinformatics*, Vol. 22, pp.1021-1023.

Barabasi A. and R. Albert, 1999, "Emergence of Scaling in Random Networks", *Science*, Vol.86, pp. 509–512.

Biddick R and E.T. Young, 2005, "Yeast Mediator And its Role in Transcriptional Regulation", *C. R. Biol.* Vol. 328, pp. 773-782.

Bouwmeester T., A. Bauch, H. Ruffner, P. Angrand , G. Bergamini , K. Croughton , C. Cruciat , D. Eberhard , J.Gagneur and S. Ghidelli, 2004, "A Physical and Functional Map of the Human TNF-Alpha/NF-Kappab Signal Transduction Pathway ", *Nature Cell Biology*, Vol. 6, pp. 97–105.

Breitkreutz B.J., C. Stark and M. Tyers, 2003, "The GRID: The General Repository for Interaction Datasets", *Genome Biology*, Vol.4, pp.86-87.

Breitkreutz B.J., C. Stark, T. Reguly, L. Boucher, A. Breitkreutz , M. Livstone , R. Oughtred, D. Lackner, J. Bähler, V. Wood, K. Dolinski and M. Tyers, 2008, "The BioGRID Interaction Database", *Nucleic Acids Res*., Vol. 36, pp. 637-640.

Brew C.T. and T.C. Huffaker, 2002, "The Yeast Ubiquitin Protease, Ubp3p, Promotes Protein Stability", *Genetics,* Vol. 162, pp. 1079-1089.

Bro C. and J. Nielsen, 2004, "Impact Of 'Ome' Analyses on Inverse Metabolic Engineering", *Metab. Eng.,* Vol. 6, pp. 204-211.

Cairns B.R., Y. Lorch, Y. Li, M. Zhang, L. Lacomis , H. Erdjument-Bromage, P. Tempst, J. Du , B. Laurent and R.D. Kornberg, 1996, "RSC, an Essential, Abundant Chromatin-Remodeling Complex" *Cell*, Vol. 87, pp. 1249–1260.

Chen D., R.M. Wilkinson, S. Watt , C.J. Penkett, W.M. Toone, N. Jones and J. Bähler, 2007, "Multiple Pathways Differentially Regulate Global Oxidative Stress Responses in Fission Yeast", M*ol Biol Cell*, Vol. 19, pp. 308–317.

Cho R.J., M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka , M. Wolfsberg, A. E. Gabrielian, D. Landsman, D.J. Lockhart and R.W. Davis, 1998, " A genome wide Transcriptional Analysis of the Mitotic Cell Cycle", *Mol. Cell,* Vol. 2, pp. 65-73

DeRisi J.L., V.R. Iyer and P.O. Brown, 1997, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale" *Science,* Vol. 278, pp. 680-686.

Drawid A. and M. Gerstein, 2000, "A Bayesian System Integrating Expression Data with Sequence Patterns for Localizing Proteins: Comprehensive Application to the Yeast Genome" *J. Mol. Biol*., Vol. 301, pp. 1059−1075.

Drawid A, R., Jansen, and M. Gerstein, 2000, "Genome-Wide Analysis Relating Expression Level with Protein Subcellular Localization", *Trends Genet.*, Vol. 16, pp. 426−430.

Eisen M.B., P.T. Spellman, P.O. Brown, D. Botstein, 1998, "Cluster Analysis and Display of Genome-wide Expression Patterns", *PNAS USA*, Vol. 95, pp. 14863-14868.

Finley D., 2002, "Ubiquitin Chained and Crosslinked", *Nat. Cell Biol*, Vol.4, pp. 121–123.

Futschik M.E., G. Chaurasial, A. Tschaut, J. Russ, M.M. Babu and H. Herzel1, 2006, "Functional and Transcriptional Coherency of Modules in the Human Protein Interaction Network", *Cell*, Vol. 127, pp. 817-830.

Galitski T., A.J. Saldanha, C.A. Styles, E.S. Lander and G.R. Fink, 1999, "Ploidy Regulation of Gene Expression", *Science*, Vol. 285, pp. 251-254.

Gavin A.C., M. Bosche , R. Krause , P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon and C.M. Cruciat, 2002, "Functional Organisation of the Yeast Proteome by Systematic Analyses of Protein Complexes", *Nature*, Vol. 415, pp. 141-147.

e

Gavin A.C. and G. Superti-Furga, 2003, "Protein Complexes and Proteome Organization from Yeast to Human", *Current Opinion in Chemical Biology*, Vol. 7, pp. 21-27.

Gasch A.P., M. Huang, S. Metzner, D. Botstein, S.J. Elledge and P.O. Brown, 2000, "Genomic Expression Responses to DNA-Damaging Agents and The Regulatory Role of the Yeast ATR Homolog Mec1p", *Mol Biol Cell*, Vol. 12, pp. 2987-3003.

Gasch A.P., P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein and P.O. Brown, 2000, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes", *Mol Biol Cell*, Vol. 11, pp. 4241-4257.

Glickman M.H., D.M. Rubin, O. Coux, I. Wefes, G. Pfeifer, Z. Cjeka, W. Baumeister, V.A. Fried and D. Finley, 1998, "A Subcomplex of the Proteasome Regulatory Particle Required For Ubiquitin-Conjugate Degradation and Related to the Cop9-Signalosome and eif3", *Cell*, Vol. 94, pp. 615-623.

Hartwell L.H., J.J. Hopfield, S. Leibler and A.W. Murray, 1999, "From Molecular to Modular Cell Biology", *Nature*, Vol. 402, pp. 47-52.

Hershko, A. and A. Ciechanover, 1998, "The Ubiquitin System", *Annu. Rev. Biochem*., Vol. 67, pp. 425–479.

Ho Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore , S.L. Adams, A. Millar, P. Taylor, K. Bennet and K. Boutilier, 2002, "Systematic Identification of Protein Complexes in *Saccharomyces Cerevisae* by Mass Spectrometry", *Nature*, Vol. 415, pp. 180-183.

Horsey E.W., J. Jakovljevic, T.D. Miles, P. Harnpicharnchai and J.L. Woolford , 2004, “Role of the Yeast Rrp1 Protein in the Dynamics of Pre-ribosome Maturation”, *RNA*, Vol. 10, pp. 813-827.

Humphries C.L., H.I. Balcer, J.L. D'Agostino, B. Winsor, D.G. Drubin, G. Barnes , B.J. Andrews and B.L. Goode, 2002, “Direct regulation of Arp2/3 complex activity and function by the actin binding protein coronin”, *Cell Biology*, Vol. 159, pp. 993-1004.

Ideker T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler and J.K. Eng, 2001, “Integrated Genomic and Proteomic Analysis of a Systematically Perturbed Metabolic Network”, *Science*, Vol. 292, pp. 929-934.

Ito T., T Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, 2001, “A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome”, *Proceedings of the National Academy of Sciences USA*, Vol. 98, pp. 4569–4574.

Jansen R., D. Greenbaum and M. Gerstein, 2006, “Relating Whole-Genome Expression Data with Protein-Protein Interactions”, *Genome Res.,* Vol. 12, pp. 37-46.

Jansen R., N. Lan, J. Qian and M. Gerstein, 2002, “Integration of genomic datasets to predict protein complexes in yeast”, *Journal of Structural and Functional Genomics,* Vol.13, pp. 71–81.

Kim Y.G., S. Raunser, C. Munger, J.Wagner, Y.L. Song, M. Cygler, T. Walz, B.H. Oh and M. Sacher, 2006, “The Architecture of the Multisubunit TRAPP I Complex Suggests a Model for Vesicle Tethering”, *Cell*, Vol. 127, pp. 817 – 830.

Kobor M.S., S. Venkatasubrahmanyam, M.D. Meneghini, J.W, Gin, J.L. Jennings, A.J. Link, H.D. Madhani and J. Rine, 2004, “A Protein Complex Containing the Conserved Swi2/Snf2-Related ATPase Swr1p Deposits Histone Variant H2A.Z into Euchromatin”, *PLoS Biol*., Vol. 2, pp.131-132.

Krogan N.J., M. Kim, S.H. Ahn, G. Zhong, M.S. Kobor, G. Cagney, A. Emili, A. Shilatifard, S. Buratowski and J.F. Greenblatt, 2002, "RNA Polymerase II Elongation Factors of *Saccharomyces cerevisiae*: a Targeted Proteomics Approach", *Mol Cell Biol*., Vol. 22, pp.6979-6992.

Lam Y.A., T.G. Lawson, M. Velayutham, J.L. Zweier and C.M. Pickart,2002, "A Proteasomal ATPase Subunit Recognizes the Polyubiquitin Degradation Signal", *Nature* , Vol. 416, pp.763–767.

Lichtenberg U. and T.S. Jensen, 2007, "Evolution of Cell Cycle Control Same Molecular Machines, Different Regulation", *Cell Cycle*, Vol. 6, pp. 1819-1825.

Marcotte E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates and D. Eisenberg, 1999, "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences", *Nature* , Vol. 402, pp. 83−86.

Meneghini M.D., M. Wu, H.D. Madhani, 2003, "Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent *Heterochromatin*", *Cell*, Vol. 112, pp. 725-736.

Mizuguchi G., X. Shen, J. Landry, W.H. Wu, S. Sen and C. Wu, 2004, "ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex", *Science*, Vol. 3003, pp. 343-348

Nooren M.A. and M. Thornton , 2003, " Diversity of protein–protein interactions", *The EMBO Journal*, Vol. 22, pp. 3486–3492.

Ogawa N., J. DeRisi and P.O. Brown, 2000, "New Components of a System for Phosphate Accumulation and Polyphosphate Metabolism in *Saccharomyces cerevisiae* Revealed by Genomic Expression Analysis", *Mol Biol Cell*, Vol. 11, pp. 4309-4321.

Przulj N., D. Corneil D and I. Jurisica, 2004, "Modeling Interactome: Scale-Free or Geometric?", *Bioinformatics*, Vol. 20, pp.3508–3515.

Reinders J., K. Wagner , R.P. Zahedi , D. Stojanovski, B. Eyrich, M. van der Laan , P. Rehling , A. Sickmann, N. Pfanner and C. Meisinger, 2007, "Profiling Phosphoproteins Of Yeast Mitochondria Reveals a Role of Phosphorylation in Assembly of the ATP Synthase", *Mol Cell Proteomics*, Vol. 6, pp. 1896-1906.

Roberts C.J., B. Nelson, M.J. Marton, R. Stoughton, M.R. Meyer, H.A Bennett, Y.D. He, H. Dai , W.L. Walker,T.R. Hughes, M. Tyers, C. Boone and S.H. Friend, 2000, "Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles", *Science*, Vol. 287, pp. 873-880.

Seraphin B. and M. Rosbash, 1989, "Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing", *Cell, Vol.* 59, pp. 349–358.

Siegers K., T. Waldmann , M.R. Leroux , K. Grein , A. Shevchenko , E. Schiebel and F.U. Hart , 1999, "Compartmentation Of Protein Folding *in Vivo*: Sequestration of Non-Native Polypeptide by the Chaperonin-Gimc System", *EMBO J.*, Vol. 18, pp. 75-84.

Spellman P.T., G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher, 1998, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization", *Mol. Biology of the Cell*, Vol. 9, pp. 3273-3297.

Spirin V. and L. Mirny, 2003, "Protein Complexes and Functional Modules in Molecular Networks", *Proceedings of the National Academy of Sciences USA*, Vol. 100, pp. 12123–12128.

Squazzo**,** S.L., P.J. Costa, D.L. Lindstrom, K.E. Kumer, R. Simic, J.L. Jennings, A.J. Link, K.M. Arndt and G.A Hartzog, 2002, "The Paf1 Complex Physically and Functionally

Associates with Transcription Elongation Factors *in Vivo*", *EMBO J., Vol.* 21**,** pp. 1764–1774.

Stefan Gross S. and C. Moore, 2001, "Five Subunits are Required for Reconstitution of the Cleavage and Polyadenylation Activities of *Saccharomyces cerevisiae* Cleavage Factor I", *PNAS*, Vol. 98, pp. 6080-6085.

Tai L.S., V.M. Boer, P. Daran-Lapujade, M.C. Walsh, J.H. de Winde, J.M. Daran and J.T. Prokt, 2004, "Two Dimensional Transcriptome Analysis in Chemostat Cultures: Combinatorial Effects of Oxygen Availability and Macronutrient Limitation in *Saccharomyces cerevisiae*", *J.Biol. Chem*, Vol. 280, pp. 437-447.

Travers K.J., C.K. Patil, L. Wodicka, D.J. Lockhart, J.S. Weissman and P. Walter, 2000 , "Functional and Genomic Analyses Reveal an Essential Coordination Between the Unfolded Protein Response and Er-Associated Degradation" , *Cell*, Vol. 101, pp. 249-258.

Uetz P., L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan and P. Pochart, 2000, "A Comprehensive Analysis of Protein-Protein Interactions in *Saccharomyces cerevisiae*", *Nature*, Vol. 403, pp. 623–627.

Vainberg I.E., S.A. Lewis , H. Rommelaere , C. Ampe , J. Vandekerckhove , H.L. Klein and N.J. Cowan , 1998, "Prefoldin, a Chaperone That Delivers Unfolded Proteins to Cytosolic Chaperonin", *Cell*, Vol. 93, pp. 863-873.

Van Vugt J.F.A., R. Michael, C. Campsteijn, C and Logie, 2007, "The Ins and Outs of ATP-Dependent Chromatin Remodeling in Budding Yeast: Biophysical and Proteomic Perspectives", *Biochim Biophys Acta*., Vol. 1769, pp. 153-171.

Vasiljeva L. and S. Buratowski, 2006, "Nrd1 Interacts with the Nuclear Exosome for 3′ Processing of RNA Polymerase II Transcripts", *Molecular Cell*, Vol. 21, pp. 239-248

Wu W.H., S. Alami, E. Luk, C.H. Wu , S. Sen , G. Mizuguchi , D.Wei and C. Wu, 2005, "Swc2 Is a Widely Conserved H2AZ-Binding Module Essential for ATP-Dependent Histone Exchange", Nat *Struct Mol Biol* , Vol. 12, pp.1064-1071.

Xiong H., X. He, C. Ding, Y. Zhang, V. Kumar and S.R. Holbrook, 2005, "Identification of Functional Modules in Protein Complexes via Hyperclique Pattern", *Pacific Symposium on Biocomputing*, Vol. 10, pp. 221-232.

Yoshimoto H., K. Saltsman, A.P. Gasch, H.X. Li, N. Ogawa, D. Botstein, P.O. Brown and M.S. Cyert, 2002, "Genome-Wide Analysis of Gene Expression Regulated by the Calcineurin/Crz1p Signaling Pathway in *Saccharomyces cerevisiae*", *J. Biol Chem,* Vol. 277, pp. 31079-31088.

Zhou H. and F. Winston, 2001, "NRG1 Is Required for Glucose Repression of the SUC2 and GAL Genes of *Saccharomyces cerevisiae*", BMC *Genet*., Vol. 1, pp. 2-5.

Zotenko E., K.S Guimaraes, R. Jothi and T.M. Przytycka, 2006, "Decomposition of Overlapping Protein Complexes: A Graph Theoretical Method for Analyzing Static and Dynamic Protein Associations", *Algorithms for Molecular Biology*, Vol 1, pp.1-7.