Integrating Biological Insights with Topological Characteristics for Improved Complex Prediction from Protein Interaction Networks

Sriganesh Maniganahalli Srihari
(MSc., NTU Singapore)
(B.Tech. (Hons.), NIT Calicut, India)

A THESIS SUBMITTED FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE
2012

To Swami Brahmananda, for the life that made this happen

Acknowledgements

This thesis edifies an unremitting debt I owe to my advisor Professor Hon Wai Leong. I am incredibly grateful for his mentorship, training, support, and most importantly friendship. From him, I learnt the hallmark of a good researcher is to be not afraid to venture out of the "borders" created by others and to approach scientific questions from an alternative prespective. The most I enjoyed while working with him were the research discussions where coarse ideas were refined and polished into interesting pieces of research work to eventually become part of this thesis. I particularly liked two qualities in his approach towards evaluating research. First, analyzing at every step of the methodology pipeline instead of merely the final output ("open up the 'black box'", he would say). Second, adopting the right "yardstick" where required - analyzing some aspects at the nanoscale while others from a bird's eye view. His high regard for excellence has had a lasting impact on my outlook on research, by inspiring me to pursue and achieve wider and more impactful goals through long and relentless effort instead of merely settling for smaller mediocre goals, and by teaching me the *art* of patience during this pursuit. His influence has also been on my writing, both as a product and as a process, to explain the most complicated of scientific concepts in the simplest possible manner, yet maintaining its preciseness as well as conciseness. His belief in maintaining a healthy and active relationship among all members of his research group by involving a mix of technical talks and informal discussions over tea not only exposed me to new and exciting subjects beyond my research, but also helped to kill some of the monotonicity and loneliness of PhD days. His friendship and support, especially during my trying times, will be a valuable source of resilience and inspiration for years to come. In fact I will try my best to imbibe and retain some of his qualities when I embark upon guiding my students someday in the future.

The influence of Professor Limsoon Wong, who readily agreed to be part of my thesis committee, has been serendipitously complementary. Himself being an expert in the field (Bioinformatics), his suggestions and timely comments helped me see the bigger picture and applicability of my research, and significantly influenced the path taken in this thesis. I am extremely grateful as well as impressed by how he always allocated time (almost instantly) whenever I requested for a discussion. I thank Professors Limsoon Wong and Wing-Kin Sung for their time, effort and commitment as members of my thesis committee. I look forward to even closer collaborations with them in the future.

My special thanks to former and present members of the Computational Biology Lab: Dr. Kang Ning for taking active interest in my work, Nan Ye, Hufeng Zhou and Dr. Francis Ng for all the enthusiastic discussions, Melvin Zhang and Dr. Ket Fah Chong for their constant suggestions and feedback. My thanks also to my friends at NUS, especially the 'tea gang': Sucheendra Palaniappan, Sudipta Chattopadhyay, Manoranjan Mohanty, Dr. Dhaval Patel, Harish Katti, Ashwin Nanjappa and Abhinav Dubey for good times in both work and play. My thanks also to NUS and the School of Computing in particular for providing me the environment and assistance to pursue my research.

My special thanks to Prof. Srinivasan Parthasarathy (the Ohio-State University) for his valuable guidance during all the collaborative works we did together. Harkening back to my undergraduate days (at NIT Calicut), I am especially indebted to Dr. K. Muralikrishnan, Dr. V. K. Govindan, Mr. Abdul Nazeer and Ms. N. Saleena for inspiring us towards higher academic pursuits. Great teachers seldom know that they become secret inspirations for their students for many years to come. Finally, thanks to my family, father, mother, sister Dr. Sulakshana and wife Preeti for their constant love and affection, and Preeti for putting up with me during those uninteresting days when the only thing on my mind was work.

Sriganesh M. Srihari

Christmas Day, 2011

Singapore

Summary

Most biological processes within the cell are carried out by proteins that physically interact to form stoichiometrically stable *complexes*. Even in the relatively simple model organism *Saccharomyces cerevisiae* (budding yeast), these complexes are comprised of many subunits that work in a coherent fashion. These complexes interact with individual proteins or other complexes to form functional modules and pathways that drive the cellular machinery. Therefore, a faithful reconstruction of the entire set of complexes (the 'complexosome') from the physical interactions among proteins (the 'interactome') is essential to not only understand complex formations, but also the higher level cellular organization.

This thesis is about devising and developing computational methods for accurate reconstruction of complexes from the interactome of eukaryotes, particularly yeast. The methods developed in this thesis integrate biological knowledge from auxiliary sources (like biological ontologies, literature on experimental findings, etc.) with the rich topological properties of the network of protein interactions (for short, PPI network) for accurate reconstruction of complexes. However, complex reconstruction is a very challenging problem, mainly due to the 'imperfectness' of data: scarcity of credible interaction data (current estimates put the coverage even in the well-studied organism yeast to only $\sim$70%), presence of high levels of noise (between 15% and 50% false positive interactions), and incompleteness of auxiliary sources.

To counter these challenges, this thesis addresses the problem in progressive stages. In the first stage, it proposes a refinement over a general density-based graph clustering method called Markov Clustering (MCL) by incorporating "core-attachment" structure (inspired from findings by Gavin and colleagues, 2006) to reconstruct complexes from the yeast PPI network. This improved method (called

MCL-CAw) refines the raw MCL clusters by selecting only the "core" and "attachment" proteins into complexes, thereby "trimming" the raw clusters. This refinement capitalizes on reliability scores assigned to the interactions. Consequently, MCL-CAw reconstructs significantly higher number of 'gold standard' complexes (∼30% higher) and with better accuracies compared to plain MCL. Comparisons with several 'state-of-the-art' methods show that MCL-CAw performs better or at least comparable to these methods across a variety of reliability scoring schemes.

In spite of this promising improvement, being primarily based on density, MCL-CAw fails to recover many complexes that are "sparse" (and not "dense") in the PPI network, mainly due to the lack to sufficient credible PPI data. In the second stage, the thesis presents a novel method (called SPARC) to selectively employ functional interactions (which are conceptual and not necessarily physical) to non-randomly 'fill topological gaps' in the PPI network, to enable the detection of sparse complexes. Essentially, SPARC employs functional interactions to enhance the "incomplete" clusters derived by MCL-CAw from sparse regions of the network. SPARC achieves this through a novel Component-Edge (CE) score that evaluates the topological characteristics of clusters so that they are carefully enhanced to reconstruct real complexes with high accuracies. Through this enhancement, MCL-CAw and other existing methods are capable of reconstructing many sparse complexes that were missed previously (an overall improvement of ∼47%).

As an extension to these methods, in the third stage, the thesis incorporates temporal information to study the dynamic assembly and disassembly of complexes. By incorporating the yeast cell cycle phases in which proteins in cell-cycle complexes show peak expression, the thesis reveals an interesting biological design principle driving complex formation: a potential relationship between 'staticness' of proteins (constitutive expression across all phases) and their "reusability" across temporal complexes.

This thesis contributes towards the ultimate goal of deciphering the eukaryotic cellular machinery by developing computational methods to identify a substantial complement of complexes from the yeast interactome and by revealing interesting insights into complex formations. Therefore, this thesis is a valuable contribution in the areas of computational molecular and systems biology.

Publications and Softwares

Publications

A major portion of this thesis has been published in the following:

- Srihari, S., Ng, H.K., Ning, K., Leong, H.W.: **Detecting hubs and quasi cliques in scale-free networks**. *International Conference on Pattern Recognition (ICPR)* 2008, **1(7)**:1–4.
- Srihari, S., Ning, K., Leong, H.W.: **Refining Markov Clustering for complex detection by incorporating core-attachment structure**. *International Conference on Genome Informatics (GIW)* 2009, **23(1)**:159–168.
- Srihari, S., Leong, H.W.: **Extending the MCL-CA algorithm for complex detection from weighted PPI networks**. *Asia Pacific Bioinformatics Conference (APBC)* 2010, Poster.
- Srihari, S., Ning, K., Leong, H.W.: **MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure**. *BMC Bioinformatics* 2010, **11(504)**.
- Ning, K., Ng, H.K., Srihari, S., Leong, H.W.: **Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology**. *BMC Bioinformatics* 2010, **11(505)**.
- Srihari, S., Leong, H.W.: **"Reusuability" of 'static' protein complex components during the yeast cell cycle**. *International Conference on Bioinformatics (InCoB)* 2011, Poster 220.
- Srihari, S., Leong, H.W.: **Employing functional interactions for the characterization and detection of sparse complexes from yeast PPI networks**. *Asia Pacific Bioinformatics Conference (APBC)* 2012, To appear.

Softwares

The following softwares along with the relevant datasets are available for free:

- **MCL-CAw**: A download-and-install implementation of the MCL-CAw algorithm for complex detection.
- **SPARC:** A download-and-install implementation of the SPARC algorithm for sparse complex detection.

Downloadable from:
`http://www.comp.nus.edu.sg/~srigsri/Web/Complex_Prediction.html`

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Unfortunately, the proteome is much more complicated than the genome.

The Scientific American, April 2002

- *Carol Ezzel [1]*

Bruce Alberts in a survey [2] (1998) termed large *assemblies* of proteins as protein *machines* of the cell. This was precisely because, like machines invented by humans, these protein assemblies comprise of highly specialized parts, and perform functions of the cell in a highly coherent manner. It is not hard to see why protein machines are advantageous to the cell than individual proteins working in an uncoordinated manner. Compare, for example, the speed and elegance of the machine that simultaneously replicates both strands of the DNA double helix with what could be achieved if each of the individual components (DNA polymerase, DNA helicase, DNA primase, sliding clamp) acted in an uncoordinated manner [2, 3].

But the devil is in the details. Though they might seem like individual parts assembled to perform arbitrary functions, these machines can be overly specific and enormously complicated. For example, consider the spliceosome. Composed of 5 small nuclear RNAs (snRNAs or "snurps") and more than 50 proteins, this machine is thought to catalyze an ordered sequence of more than 10 RNA rearrangements as it removes an intron from an RNA transcript [2]. In fact the discovery of this intron splicing process won Phillip A. Sharp and Richard J. Roberts the 1993 Nobel Prize in Physiology and Medicine[1].

[1] `http://nobelprize.org/nobel_prizes/medicine/laureates/1993/illpres/index.html`

When one examines these protein assemblies, now known to be in the order of hundreds even in the simplest of eukaryotic cells, and the kind of cellular activities they are involved in, one is reminded of the baffling paintings in an art exhibit composed of an intricate interplay of form, color, light and shade. But perhaps this is because we do not fully understand what the cell needs to accomplish with each of its protein assemblies just like how an amateur art appreciator does not fully understand the deeper expressions the artist is trying to convey through each of her strokes.

Given this intricacy and ubiquity of protein assemblies, a serious attempt towards identification, classification and comparative analysis of all such assemblies is essential not only to understand them in more depth, but also to decipher the higher level organization of the cell.

To proceed on such a vast exploration, the quest is to first crack the proteome - a concept so novel that the word *proteome* did not even exist a decade ago. The proteome is the entire library of proteins expressed in an organism [6]. With the dawn of the 21st century and the introduction of "high-throughput" techniques in molecular biology, cataloging this library of proteins has become feasible. Though the cataloging of information about human proteins has still a long way to go, notable progress has been done for simpler organisms like *Escherichia coli* (bacteria) and *Saccharomyces cerevisiae* (yeast), which can give us enlightening insights into the cellular machinery. After all, considering the 3.8 billion years of the history of evolution, we humans appearing 200,000 years ago are mere increments, and therefore what is fundamentally true of these smaller organisms should be fundamentally true of us. As the late French geneticist Jacques Monod put it, only half in jest, 'Anything that is true of *E. coli* must be true of elephants, except more so' [6]. Naturally, the same must be true of humans!

Just like how organizing our home libraries can involve a lot of time and effort, and school libraries even more so, where books need to be carefully chosen, categorized, ordered and arranged so that they can be of effective use, the categorizing and organizing of the large-scale data churned out from these high-throughput techniques can also involve significant time and effort so that we make the *right* sense out of them. Once this task is reasonably done, this data can be effectively and

efficiently mined and analysed to decipher new insights into cellular mechanisms. Towards this end, the major research questions being pursued are: "How to organize and store the large quantities of data?", "How to interpret and categorize or classify this data?", "How to differentiate between useful and erroneous (noisy) data?", "How to analyze this data and interpret the findings to fill the gaps in our present knowledge?", etc. The task of answering these questions certainly calls for enormous computational analyses (by computer scientists) that can effectively complement experimental techniques (by molecular biologists).

1.1 Research scope

One of the important areas where large-scale data has been employed is to identify and map the entire complement of protein assemblies from organisms. Depending on the functional, spatial and temporal context, protein assemblies can be categorized broadly into a number of types, and one way to do so is [4],

1. *Complexes:* These are stoichiometrically stable structures formed by physical interactions among proteins at specific time and space, and are responsible for distinct functions within the cell. Complexes can be both permanent (example, proteasomes) or transient (example, a kinase and its substrate).

2. *Functional modules:* These are typically formed when two or more complexes interact with each other or individual proteins in a 'time-dependent' manner to perform a particular function and dissociate after that (for example, the complexes and proteins forming the DNA replication machinery).

3. *Signaling pathways:* These comprise of ordered succession of 'time-dependent' interactions among proteins, but does not require all components to co-localize in time and space (for example, the MAPK pathway controlling mating response).

In summary, there are distinct types of assemblies and we can derive a variety of criteria to categorize them; many of these criteria can overlap, and any one criteria in isolation will fail to encompass all types of assemblies [4, 5]. But, among all the types defined above, complexes are the most clearly defined assemblies. They can be considered the fundamental functional units formed by physical interactions

among proteins in time and space. Here, the focus is primarily on the detection and analysis of complexes, however, occassionally in the presence of 'timing information' we attempt to understand functional modules as well.

Large-scale experimental identification of complexes can be done by *in vitro* "pull down" of cohesively interacting groups of proteins. Very broadly, this procedure comprises of a 'bait' protein introduced into a solution of cell lysate, and purified together with its physically binding 'preys'. The individual component proteins in this complex can then be identified by Mass Spectrometry analysis. However, the exhaustiveness of this procedure depends on the baits used. There is no way to identify all possible complexes unless all possible baits are tried. Further, a chosen bait may not physically interact with all components in its complex, and hence multiple baits need to be tried to identify the complete complex. Additionally, a protein might be involved in more than one distinct complexes, which means each protein has to be verified for both as a bait and as a prey, and that too in multiple purifications. In these 'combinatorial trials' there can also occur "errors" due to *in vitro* experimental conditions, which can either result in contaminants within the complexes or washing out of weakly associated proteins. Of course, there is a monetary cost factor also involved in performing these experiments.

One way to overcome these difficulties is to use the "pull-down" complexes to first infer the *physical interactions* among the constituent proteins. This is done either as interactions between the bait and its preys in a complex (like the "spokes" of a wheel), or as interactions among all proteins in a complex (like a "matrix"), or a suitable combination of both. If a significant number of such physical interactions can be inferred and catalogued, distinct groups of proteins forming complexes can be isolated from them: proteins within a complex form many interactions with each other than with proteins not in the complex. Quite naturally, such an procedure cannot be done manually, and therefore calls for specialized computational techniques that can decipher the complexes from the set of interactions.

The scope of this thesis is to design and develop effective computational techniques for identifying protein complexes from physical interactions catalogued from such high-throughput experiments.

1.2 Research methodology

In computational analysis, protein interactions from an organism are typically assembled in the form of a network with the proteins as nodes and the interactions among them as edges, commonly called *protein-protein interaction network* or *PPI network*. Such a network provides a 'global picture' of the entire set of interactions. This network is rich in topological properties that can give vital evidences or insights into cellular organization. For example, it was found that the degree distribution of proteins in the network is not random, but instead roughly follows a power law indicating the presence of a few high-degree proteins (called "hubs") which when disrupted can cause the network to breakdown (this is commonly referred to as the "scale-free" property) [7, 8]. Similarly, the 'betweenness centrality' for a protein is the total number of shortest paths in the network that pass through that protein, and corresponds to the topological 'centrality' of the protein [9]. These "hubs" and 'central' proteins in the network likely correspond to essential or lethal proteins within the cell [10, 11].

In this thesis, we design and develop computational methods for identifying protein complexes from PPI networks (see Figure 1.1). Typically, the approaches proposed for identifying complexes from PPI networks fall within the purview of the following steps:

1. Constructing the PPI network from the individual physical interactions;
2. Identifying candidate complexes from the network; and
3. Evaluating the identified complexes against *bona fide* complexes, and validating the novel complexes.

Although promising, complex identification from PPI networks still requires careful attention in handling errors and noise and reconstructing complexes with high accuracies. The specific techniques and algorithms developed in this thesis are motivated by the following desirable properties for the results in this thesis:

1. Detecting possibly all complexes and with high accuracies;
2. Effective countering of noise observed in experimental datasets; and

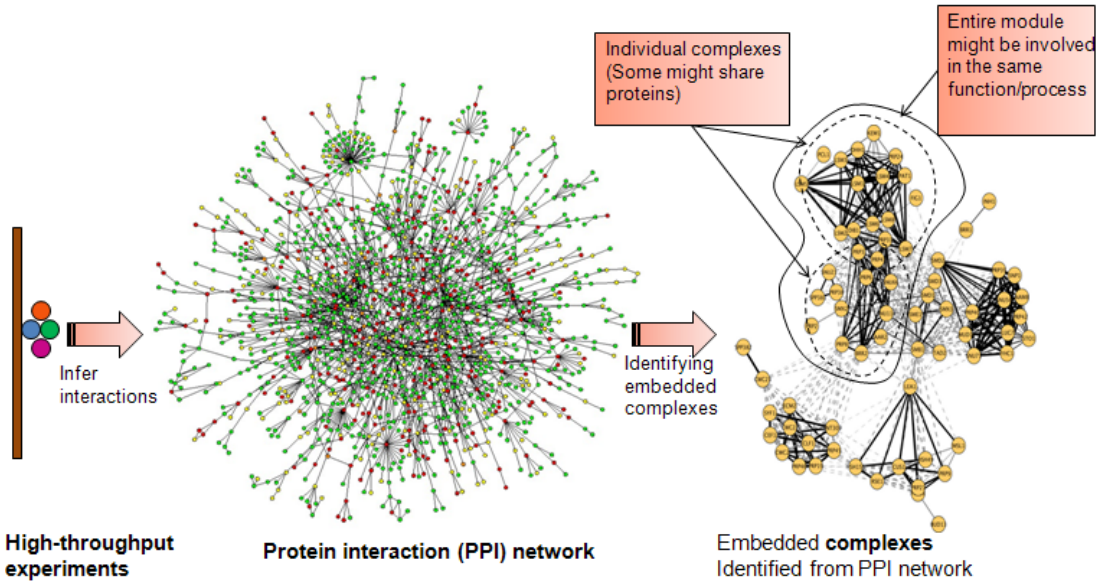


Figure 1.1: Research objective: Reconstructing protein complexes from the network of protein interactions.

3. In-depth analyses of detected complexes to gain deeper and possibly novel insights into biological phenomena.

To achieve the aforesaid desired results, we devise novel methods to integrate a variety of known biological information and insights with the rich topological properties of the PPI network. This auxiliary biological knowledge can be in the form of organizational, structural, functional or evolutionary information gathered about proteins, interactions and complexes from experimental and other studies, and catalogued in literature and databases. The broad methodology followed is to "encode" this auxiliary biological knowledge as topological structures in the PPI network. By implementing this methodology, we capitalize on both the biological knowledge as well as the topological properties of the PPI network for detecting complexes.

1.3 Contributions of the thesis

This thesis contributes several new principles and procedures of inquiry into the computational analysis of PPI networks in general, and complex detection in particular. The main constributions are listed below:

1. **A 'foresightful' survey and taxonomy of existing computational methods:**

   From the time high-throughput experimental techniques were first introduced for inferring protein interactions (by Uetz et al. in 2000 [12] and Ito et al. in 2001 [13]), computational techniques began parallely gaining popularity to analyse the large amounts of data being continuously catalogued (one of the first attempts in computational complex prediction was by Bader and Hogue in 2003 [14]). It is almost a decade now, and newer and more reliable experimental techniques have been introduced that have in turn inspired many new computational methods making use of these improved datasets. While surveys and comparative assessments have periodically come out on these computational methods, an extensive taxonomy that gives us a "sense of time" when the methods were developed and relates them to experimental improvements, has not been presented till date.

   In this thesis (Chapter 3), we present a comprehensive taxonomy of computational methods (we identify close to 20 methods) developed for complex detection over the years. We present this taxonomy as two snapshots - a chronology-based "bin-and-stack" and an algorithmic methodology-based 'tree'. This taxonomy condenses the history of complex detection, and has a capability, what we believe, to show directions for future research in this area.

2. **An improved complex detection method using core-attachment insights:**

   In 2006, Gavin and colleagues [15], for the first time, studied the organizational structure within yeast complexes on a genome-wide scale. Their findings revealed an inherent modularity among proteins within complexes, organized as two distinct sets - "cores" and "attachments". This revelation inspired several computational methods to reconstruct complexes, ours being one of the earliest, by identifying "core" and "attachment" proteins from their topological properties within the PPI network.

   In Chapter 4 of this thesis, we present this new method to reconstruct yeast complexes. Our method provides two levels of "controls" to be stringent or

lenient while identifying the "core" and "attachment" complex proteins from "dense" regions. This helps us to "trim" our predictions instead of considering whole "dense" regions as complexes. The initial "dense" regions are identified using a popular but general graph clustering method called Markov Clustering (MCL) [16], and therefore we consider our method (called MCL-CAw) as a 'customization' of MCL to detect complexes by incorporating "Core-Attachment" structure. We demonstrate that MCL-CAw reconstructs on average ∼30% higher number of complexes than MCL.

A reliability *weight* or score is typically assigned to interactions in the PPI network to account for the biological variability and technical limitations of experimental conditions. The 'w' in MCL-CAw refers to the ability of our method to capitalize on such weights, and therefore handle noise in biological datasets. We demonstrate through extensive analysis that such scoring aids to significantly improve complex prediction, and that MCL-CAw shows consistent performance across a variety of scoring schemes.

A significant portion of these results were published first as a preliminary version in the proceedings of the 20^{th} International Conference on Genome Informatics (GIW) 2009 [17], and later as a substantially extented version in BMC Bioinformatics (2010) [18].

3. **A quantitative definition to the notion of complex "derivability":**

In this thesis (Chapter 5), we test the credibility of the key assumption underlying all existing computational methods that complexes form "dense" regions within the PPI network. We define the notion of complex "derivability", that is, whether a complex is derivable or not from a given PPI network, and if yes to what extent. We present a measure (called the Component-Edge or CE score) to quantitatively capture this notion effectively. We show that this measure strongly correlates with the actual complex derivation capability of computational methods, and use it to demonstrate that overly relying on the 'denseness' assumption in the wake of insufficient PPI data can cause "sparse" complexes to be missed.

A significant portion of these results were published in the International Jour-

nal of Bioinformatics Research and Applications (2012) [19], invited from the 10^{th} Asia Pacific Bioinformatics Conference (APBC) 2012.

4. **A novel improvement to detect "sparse" complexes by employing functional interactions:**

   Our experiments reveal that many complexes are "sparse" (and not "dense") in the PPI network, rendering methods that over rely on the 'denseness' assumption of complexes ineffective in detecting these "sparse" complexes. In Chapter 5, we characterize these "sparse" complexes using our proposed CE score. Going further, we present a novel method called SPARC which employs functional interactions to elevate some of the "sparse" complexes to "dense", enabling existing methods to detect these complexes satisfactorily. Functional interactions are logical associations inferred from a variety of biological information to "encode" affinity beyond just physical interactivity. This is, to our knowledge, the first such work that combines functional with physical interactions to detect complexes, particularly the "sparse" ones. Our experiments show that SPARC aids existing methods to reconstruct on average ∼47% higher number of complexes.

   A significant portion of these results were published in the International Journal of Bioinformatics Research and Applications (2012) [19], invited from the 10^{th} Asia Pacific Bioinformatics Conference (APBC) 2012.

5. **Novel biological insights deciphered from detected complexes:**

   Finally, to demonstrate the impact of the developed computational methods, in Chapter 6 we employ the detected complexes to understand some of the phenomena driving complex formations in yeast. We incorporate auxiliary biological information in the form of protein essentiality and the yeast cell-cycle phase in which the proteins are transcribed to reveal two interesting insights: (i) Essential proteins have a higher tendency to function in groups, many of which are complexes; (ii) The relatively higher enrichment of 'staticness' (constitutive expression) in proteins shared among 'time-based' complexes, hinting towards the biological design principle of temporal "reusability" of 'static' proteins for temporal complex formations.

Some portions of these results were published in BMC Bioinformatics (2010) [18] and as a poster in the 10^{th} International Conference on Bioinformatics (InCoB) 2011 [20].

1.4 Organization of the thesis

Chapter 2 presents background on protein interaction networks required for understanding the details of this thesis. The chapter provides concise information on some of the experimental and computational techniques used to infer the interactions, and the limitations and challenges in these techniques, particularly those leading to inherent noise in experimental datasets. **Chapter 3** surveys existing computational methods developed for reconstructing complexes from protein interaction networks. It dwelves into their merits and demerits, and identifies challenges and limitations to motivate the subsequent chapters. **Chapter 4** proposes a new computational method (MCL-CAw) for reconstructing complexes. **Chapter 5** identifies some of the overlooked loopholes in MCL-CAw, and proposes an improvement (called SPARC) to address these loopholes. **Chapter 6** analyses the reconstructed complexes to gain deeper and novel biological insights into complex organization, and thereby provides a fitting sign off to the methods developed in this thesis. **Chapter 7** draws the final curtain by summarizing the main contributions of the thesis, discussing the significance of the results, identifying some of the limitations, and thereby recommending directions for future research.

CHAPTER 2

Techniques for inferring protein interactions

All mass is interaction.

- Richard Feynman

statement titled "Principles" (c. 1950), as quoted in [21]

Proteins interact with each other in a highly specific manner, and protein interactions play a key role in many cellular processes. In order to get a global picture of these interactions, especially for system level studies, these interactions are typically assembled in the form of a protein interaction network (PPI network). Over the past decade or so, several high-throughput studies have been developed for screening interactions on a genome-wide scale resulting in the cataloging of vast amounts of interaction data from several organisms, in turn leading to larger and more complete PPI networks that can be systematically studied and analyzed to extend our knowledge about cellular processes. But, in order to study and analyse PPI networks, we need to first understand the major promises and limitations of these high-throughput techniques, and the approaches used to verify, validate and complement the diverse experimental data produced from these techniques, which is the subject of this chapter. A reader familiar with the domain may skip this chapter and refer back to relevant sections if required.

2.1 High-throughput experimental techniques for inferring interactions

Protein interactions can be analyzed by different genetic, biochemical and biophysical high-throughput techniques, some of which are listed in Table 2.1 and diagrammatically shown in Figure 2.1. Some techniques such as yeast two-hybrid (Y2H) [12,13,22] and protein-fragment complement assay (PCA) [23] enable identification of binary physical interactions between proteins, while other techniques like affinity purification (AP) [24] enable "pull down" of whole complexes from which the binary interactions can be inferred, and still others like synthetic lethality [25] enable detection of functional (indirect) associations among proteins apart from physical (direct) interactions.

Technique	Living cell assay	Interaction type
Yeast two-hybrid [12, 13, 22]	In vivo	Physical binary
Protein-fragment complement assay [23]	In vivo	Physical binary
Affinity purification-MS [24]	In vitro	Physical complex
Synthetic lethality [25]	In vitro	Functional association

Table 2.1: Some high-throughput experimental techniques for screening protein interactions.

2.1.1 Yeast two-hybrid

Yeast two-hybrid or Y2H is an *in vivo* technique based on the fact that many eukaryotic transcription activators have at least two distinct domains, one that directs binding to a promoter DNA sequence (BD) and other that activates transcription (AD). It was demonstrated that splitting BD and AD inactivates transcription, but the transcription can be restored if a DNA-binding domain is physically associated with an activating domain [26]. Accordingly, a protein of interest is fused to BD. This chimeric protein is cloned in an expression plasmid, which is then transfected into a yeast cell. A similar procedure creates a chimeric sequence of another protein fused to AD. If the two proteins physically interact, the reporter gene is activated. Numerous variants of Y2H have been developed for detecting interactions in higher eukaryotic cells like mammalian cells.

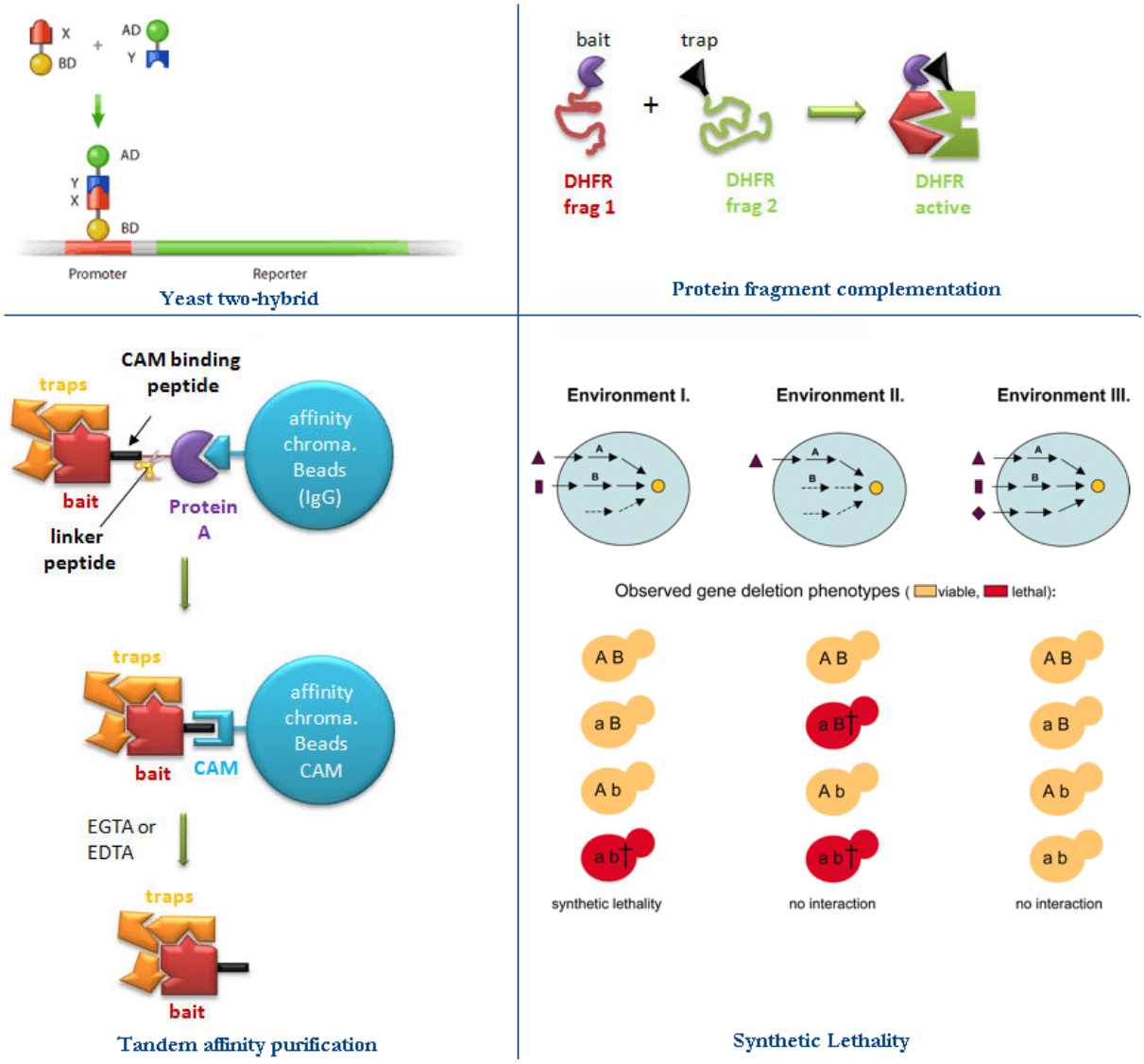


Figure 2.1: Some of the high-throughput experimental techniques developed for screening protein interactions: yeast two-hybrid, tandem affinity purification, protein fragment complementation and synthetic lethality.

One of the first genome-wide Y2H screens from yeast by Uetz et al. [12] and Ito et al. [13] inferred 692 and 841 putative interactions, respectively. The overlap between the two screens was only about 20%. Investigations into the small overlap revealed several limitations in the Y2H technique: bias towards nonspecific interactions and bias against membrane proteins, proteins initiating transcription by themselves cannot be targeted in Y2H experiments, and the use of sequence chimeras can affect the structure of target protein [26].

2.1.2 Affinity purification followed by mass spectrometry

Complementing the *in vivo* Y2H technique are the *in vitro* Affinity Purification followed by Mass Spectrometery (AP-MS) techniques for high-throughput screening of interactions. These comprise of two steps - affinity purification and mass spectrometery. The most common technique uses the tandem affinity purification (TAP) tag. In the TAP approach, the protein of interest (bait) is TAP-tagged and purified from a cell lysate together with its binding partners (preys) after washing out the contaminants. The components of each such purified complex are screened by gel electrophoresis, and identified by MS.

The first two large TAP-MS screens of yeast by two seperate groups, Gavin et al. (2002, 2006) [15, 27] and Krogan et al. (2006) [28], showed 7592 and 7123 protein interactions identified with high confidence, respectively. Subsequently, several other groups improved on these AP-MS techniques to identify significantly many more interactions (for a survey, see [26]).

Comparing with the Y2H technique, AP-MS can report whole complexes and can therefore report on higher-order interactions beyond binary. However, Y2H has the advantage of being an *in vivo* technique and of detecting transient interactions.

2.1.3 Protein-fragment complementation assay

Protein-fragment complementation assay or PCA is another *in vivo* technique based on the principle of splitting a protein into two fragments, each of which cannot function alone [23]. These fragments are fused to potentially interacting protein partners, and if complementation upon interaction leads to restored function, then the interaction between the partners in inferred.

Although PCA is similar to Y2H, it requires the reconstitution of a separate (third) protein to detect the interaction between two partners. But, PCAs have advantage over Y2H because they can be employed to identify interactions between membrane proteins, and also between membrane and membrane associated proteins [26].

2.1.4 Synthetic lethality

Synthetic lethality is a genetic interaction method which produces mutations or deletions of two separate genes which are viable alone but cause lethality when combined together in a cell under certain conditions [25]. Since these mutations are lethal, they cannot be isolated directly and should be synthetically constructed. Synthetic interaction can point to possible physical interaction between two gene products, their participation in a single pathway, or a similar function (functional associations) [25, 26].

2.2 Constructing PPI networks from interaction datasets

The pairwise (binary) physical interactions inferred among proteins using different experimental techniques are assembled into a PPI network with the proteins as nodes and the interactions among them as edges in the network. However, some techniques like TAP-MS offer only whole complexes comprising of preys showing high affinities to baits instead of pairwise binary interactions. To infer the binary interactions from TAP-MS complexes, their topologies are represented as collections of hypothetical pairwise interactions, for which there are two kinds of models: "spoke" and "matrix" [15, 28–31].

The spoke model assumes that the protein bait interacts directly with each of the prey proteins, like spokes of a wheel. The spoke model is useful to reduce complexity of data visualization, but necessarily misses out on several prey-prey interactions that may be true. The matrix model assumes that all proteins within a complex have pairwise interactions with each other. The matrix model contains all possible true interactions, but necessarily has a large number of false interactions as well. The empirical evaluations [29, 32, 33] of pull-down data from Gavin et al. (2006) [15] showed about 19.8% true interactions and 39% false interactions in

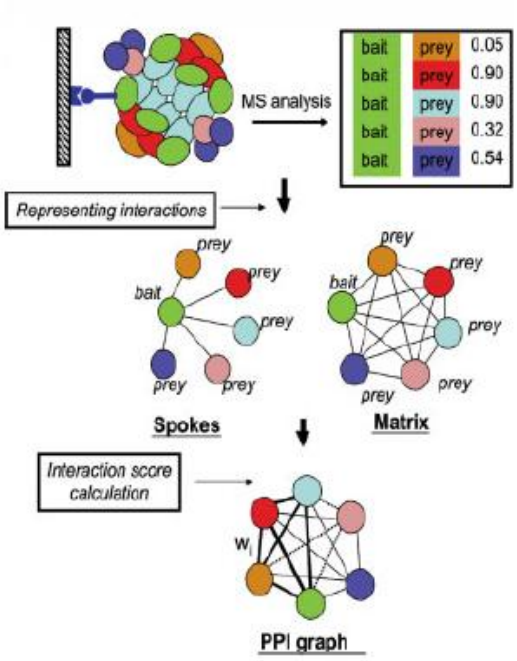


Figure 2.2: Deriving scored PPI network from TAP/MS purifications [31]: The "pulled-down" complexes from TAP/MS experiments are assembled as 'spoke' and 'matrix' models to infer the interactions among the constituent proteins.

the spoke model, and 68.8% true interactions and 308.7% false interactions in the matrix model. Therefore, typically a balance is struck between the two models that covers most of the true interactions without accepting in too many false interactions. Several groups including Gavin et al. [15] have used such a combination of spoke and matrix models. The complete picture for the network construction is summarized in Figure 2.2.

2.3 Gaining confidence in high-throughput datasets

Although high-throughput techniques have been successful in large-scale screening of protein interactions, several recent analyses and reviews [32–35] have highlighted the prevalence of spurious interactions in high-throughput data. Consequently, a crucial challenge in adopting such data is separating the subset of credible interactions from the background noise.

2.3.1 False positives and true negatives in interaction datasets

The spurious interactions (false positives) in high-throughput screens may arise from technical limitations in the underlying experimental techniques. The Y2H system, in spite of being *in vivo*, does not consider the localization, time and cell context in different cell types while testing for binding partners. On the other hand, *in vitro* "pull downs" are carried out using cell lysates in an environment where every protein is present in the same "uncompartmentalized soup". Therefore, even though two proteins interact, it is not certain that they will interact under real conditions. Opportunities are high for proteins to interact promiscuously with partners that they never normally come across in an intact cell and for 'sticky' molecules to function as bridges between two other proteins [35]. Recent analysis [26] have shown that only 30-50% of high-throughput interactions are biologically relevant.

In addition to spurious interactions, another challenge is to be able to cover the whole complement of interactions (the 'interactome'). The comparisons [26, 32–34] between datasets from different techniques have shown striking lack of correlation, each technique producing a unique distribution of interactions suggesting that the techniques have specific strengths and weaknesses. A major drawback of most techniques is that many interactions may depend on certain post-translational modifications such as disulfide bridge formation, glycosylation and phosphorylation, which may not occur properly in the adopted system. Many of these techniques also show bias towards abundant proteins and against certain kind of proteins like membrane proteins. For example, AP-MS techniques predict relatively few interactions for proteins involved in transport and sensing (transmembrane proteins), while Y2H being targeted in the nucleus fail to cover extracellular proteins [26].

2.3.2 Estimating the reliabilities of interactions

The integration of high-throughput datasets from multiple experimental sources can certainly help in enriching true interactions and covering a sizeable fraction of the interactome. However, the prevalence of spurious interactions continues to remain a challenge, which magnifies further upon integration of datasets. In order to

separate credible interactions from background noise, the reliabilities of individual interactions are estimated so that less reliable interactions can be selectively filtered.

Reliability scoring schemes offer a score (weight) to each interaction in the PPI network, which typically encodes the reliability (confidence) of the physical interaction between the protein pair. The score accounts for the biological variability and technical limitations in the experiments. For example, Gavin et al. [15] combined the spoke and matrix models using a 'socio-affinity' scheme which quantized the log-ratio of the number of times two proteins were observed together as a bait and a prey, or a prey and a prey, relative to what would be expected from their frequency in the dataset. On the other hand, Krogan et al. (2006) [28] used machine learning techniques (Bayesian networks and C4.5-decision trees) trained using diverse evidences to define the confidence scores between proteins in their spoke modeled PPI dataset.

Subsequent to these two scoring schemes, several other schemes [29,36,38–41,43, 45–47] have been developed to score PPI networks (see a survey, see [42]). Collins et al. [36] developed a Purification Enrichment (PE) scoring system to generate the 'Consolidated network' from the matrix modeled relationships of the Gavin et al. and Krogan et al. datasets. Collins et al. used a Bayes classifier to generate the PE scores in the Consolidated network by incorporating training data from hand-curated co-complexed protein pairs, Gene Ontology (GO) [37] annotations, mRNA expression patterns, and cellular co-localization and co-expression profiles. This new network was shown to be of high quality - comparable to that of PPIs derived from small-scale experiments stored at the Munich Information Center for Protein Sequences (MIPS). Hart et al. [38] generated a Probabilistic Integrated Co-complex (PICO) network by integrating matrix modeled relationships of the Gavin et al., Krogan et al. and Ho et al. datasets using a measure similar to socio-affinity scores. Zhang et al. [29] used Dice coefficient (DC) to assign affinities to protein pairs, and evaluated their affinity measure against socio-affinity and PE measures. They concluded that DC and PE offered the best representation for protein affinity among the three schemes. Chua et al. [39] and Liu et al. [40] developed network topology-based scoring systems called Functional Similarity Weight (FS Weight) and Iterative-Czekanowski-Dice (Iterative-CD), respectively, to assign

reliability scores to the interactions in networks. Friedel et al. [41] developed a bootstrapped scoring system based on random sampling to score TAP-MS interactions from Gavin et al. and Krogan et al. Kuchaiev et al. [43] embedded PPI networks into Euclidean spaces and modeled them as geometric random graphs to de-noise the networks based on geometric distances (the same group showed earlier that geometric random graphs are the best models for PPI networks [44]). Voevodski et al. [45] used PageRank, a random walk-based method used in context-sensitive web search, to define the affinities between proteins within PPI networks. More recently, Jain et al. [46] (2010) developed Topological Clustering Similarity Scheme (TCSS) that used the knowledge captured in Gene Ontology [37] to assess the reliabilities of interactions. Breitkreutz et al. [47] (2010) developed the Significance Analysis of Interactome (SAINT) scoring to detect non-specifically binding proteins based on peptide counts, an additional type of experimental data generated using a peptide identification phase in their screens. SAINT employs a mixture of Poisson distributions to heuristically compute posterior probabilities of specific interactions based on the peptide counts.

We classified these scoring schemes into three broad categories (Table 2.2): (i) Sampling or counting-based, (ii) Evidence-based, and (iii) Solely topology-based.

Sampling or counting	Evidence based	Solely topology
Dice coefficient [29]	Bayesian networks [28]	FS Weight [39]
Socio-affinity [15]	Purification enrichment [36]	Iterative CD [40]
Hart sampling [38]	Gene Ontology-based [46]	Geometric embedding [43]
Bootstrap sampling [41]	SAINT [47]	PageRank affinity [45]

Table 2.2: Broad classification of affinity scoring schemes for reliability estimation of protein interactions.

2.4 Computational techniques for inferring interactions

Although high-throughput techniques produce large amounts of data, the covered fraction of the interactomes from many organisms are far from complete. The low interaction coverage and the need for verification of high-throughput data calls for the development of computational techniques to predict protein interactions. However, these techniques can have two kinds of limitations: (i) many of these techniques use experimental data to infer new interactions leading to an inherent bias in their

predictions; (ii) many of these techniques do not predict physical interactions directly but rather infer the functional associations between potentially interacting proteins. Despite these limitations, computational techniques have proved an effective complement to experimental techniques for analyzing interactions. These techniques can be useful for choosing potential targets for experimental screening or for independently validating experimental data [26].

Protein physical or functional interactions are predicted computationally using various kinds of genome inference methods that use genomic or proteomic context to infer interactions. We discuss a few of them here.

Genes with closely related functions encoding potentially interacting proteins are often transcribed as a single unit, an *operon*, in bacteria and are co-regulated in eukaryotes. Different methods have been developed to predict operons in bacterial genomes based on intergenic distances [48]. Analysis of gene order conservation within three bacterial and archaeal genomes found that 63%-75% of co-regulated genes interact physically [49]. Similar results were found for eukaryotes like yeast and worm [50].

The phylogenetic profile method is based on the hypothesis that functionally linked and potentially interacting nonhomologous proteins co-evolve and have orthologs in the same subset of fully sequenced organisms. Indeed, components of complexes and pathways should be present simultaneously in order to perform their functions [26]. A phylogenetic profile is constructed for each protein, as a vector of N elements, where N is the number of genomes. The presence or absence of a given protein in a given genome is indicated as '1' or '0' at each position of a profile. Proteins or their profiles can then be clustered using a bit-distance measure, and those proteins from the same cluster are considered functionally related.

The Rosetta Stone approach infers protein interactions from protein sequences in different genomes. It is based on the observation that some interacting proteins or domains have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein [51]. Gene fusion apparently occurs to optimize co-expression of genes encoding for interacting proteins. In *Escherichia coli*, the Rosetta Stone method found 6,809 potentially interacting pairs of nonhomologous proteins; both proteins from each pair had significant sequence similarity to a single

protein from some other genome. Analysis of pairs found by this approach revealed that for more than half of the pairs both members were functionally related [51].

2.5 Protein interaction databases

As a result of the large variety of experimental and computational methods developed for detecting and characterizing protein interactions, several databases have been set up to catalogue, study and analyze these interactions. Some publicly available databases and their Web sources are listed in Table 2.3.

The Database of Interacting Proteins (DIP) [52] contains experimentally determined (Y2H and TAP-MS) protein interactions and includes a core subset of interactions that have passed a quality assessment (for example, literature-based verification).

The Biomolecular Interaction Network Database (BIND) [53], now called Biomolecular Object Network Database (BOND), includes high-throughput experimental protein interactions, and also protein-small molecule interactions and protein-nucleic acid interactions.

The BioGrid [54] is a database of protein and genetic interactions gathered from several high-throughput experiments, while STRING [55] is a database of physical (direct) and functional (indirect) interactions gathered from several experimental as well as computational techniques.

The MIPS Comprehensive Yeast Genome Database (CYGD) [56] and the MIPS Mammalian Protein-Protein Interaction Database (MPPI) [57] is a comprehensive catalogue of yeast and mammalian protein interactions and hand-curated complexes, while the Human Protein Reference Database (HPRD) [58] stores interactions specific to human.

Apart from these, the Interlogous Interaction Database (I2D) [59] and Predictome [60] database integrate interactions from multiple sources, and also interactions between orthologous proteins inferred across species (“interlogs”).

Database	Interaction type	URL/FTP
DIP [52]	P	http://dip.doe-mbi.ucla.edu
BIND [53]	P,M	http://bind.ca
BioGrid [54]	P,G,M	http://thebiogrid.org
STRING [55]	E,M,F	http://string-db.org/
CYGD [56]	P,M	http://mips.helmholtz-muenchen.de/genre/proj/yeast/
MPPI [57]	P,M	http://mips.helmholtz-muenchen.de/proj/ppi/
HPRD [58]	E,M	http://www.hprd.org/
I2D [59]	E,M,F	http://ophid.utoronto.ca/
Predictome [60]	E,M,F	http://predictome.bu.edu/

Table 2.3: Protein interaction databases and their Web sources. The interaction types are: high-throughput experimental-protein (P), high-throughput experimental-genetic (G), manual (M) and functional/predicted (F).

2.6 Outlook

In this chapter, we summarized some of the experimental and computational techniques developed to infer interactions among proteins, and the strengths and limitations of these techniques. Yeast is the most widely mapped eukaryote with more than two million experimentally and computationally inferred interactions catalogued in public databases. However, a significant fraction of these interactions is spurious and unvalidated making the credibility of these datasets difficult to be accurately estimated. Considering only the subset of multi-validated interactions, recent estimates put the covered fraction of the yeast interactome between 60% and 70% [26, 32–34], leaving significant room for new interactions to be still discovered and validated. For other organisms like mammals, this gap is even more appalling.

However, with the recent advancements in experimental and computational techniques for inferring, verifying and analyzing protein interactions, faster progress is being done to catalogue credible interactions from several organisms. As constantly new data is being generated, the analyses and surveys on these datasets are also being constantly updated to give us a sense as to where we currently stand. The picture is not as bleak as it seems.

As new data is being generated, newer ways are being devised to study and analyze this data to decipher unknown cellular principles. The detection and analysis of *protein complexes* from this large-scale interaction data is one such focused study that has emerged, and significant progress has been done over the last few years, as we shall see in the next chapter.

CHAPTER 3

Methods for complex detection from protein interaction networks

Dante can be understood only within the context of Italian thought, and Faust would be unthinkable if divorced from its German background; but both are part of our common cultural heritage.

Nobel Lecture, 29^{th} June 1927

- *Gustav Stresemann*

The complexes of proteins working together to achieve modular biological functions through a series of physical interactions constitute the fundamental (functional) units within the cell. From a biological perspective, this modularity is a result division of labor and of evolution to provide robustness against mutation and chemical attack [4]. From a topological perspective, this modularity is a result of proteins within complexes being densely connected to each other than to the rest of the PPI network [29].

Typically, the process of identifying complexes from high-throughput interaction data involves the following steps: (i) Integrating high-throughput datasets from multiples sources and assessing the reliabilities of interactions; (ii) Constructing the PPI network; (iii) Identifying the modular subnetworks from the network to generate a candidate list of complexes; (iv) Evaluating the identified complexes against *bona fide* complexes, and validating and assigning roles to the novel complexes.

3.1 Review of existing methods for complex detection

In this section, we review, classify, compare and evaluate some representative works done till date on the computational prediction of protein complexes from PPI networks. We begin describing these methods by first mentioning some definitions and terminologies widely adopted across these works.

3.1.1 Definitions and terminologies

A PPI network is modeled as an undirected graph $G = (V, E)$, where V is the set of proteins and $E = \{(u, v) : u, v \in V\}$ is the set of interactions among protein pairs. For any protein $v \in V$, $N(v)$ is the set of direct neighbors of v, while $deg(v) = |N(v)|$ is the degree of v. The interaction density of G is defined as $density(G) = \frac{2.|E|}{|V|.(|V|-1)}$. This is a real number between 0 and 1, and typically quantifies the "richness of interactions" within G: 0 for a network without any interactions and 1 for a fully connected network. The clustering coefficient $CC(v)$ measures the "cliquishness" of the neighborhood of v: $CC(v) = \frac{2.|E(v)|}{|N(v)|.(|N(v)|-1)}$, where $E(v)$ is the set of edges in the neighborhood of v. If the interactions of the network are reliability scored (weighted), these definitions can be extended to their corresponding weighted versions: $deg_w(v) = \sum_{u \in N(v)} w(u, v)$, $density_w(G) = \frac{\sum_{e \in E} w(e)}{|V|.(|V|-1)}$, and $CC_w(v) = \frac{\sum_{e \in E(v)} w(e)}{|N(v)|.(|N(v)|-1)}$, where $w : E \rightarrow \mathcal{R}$ is a scoring function on the interactions in E. There are several interesting variants proposed for weighted clustering coefficient CC_w; for a survey see [61].

3.1.2 Taxonomy of existing methods

Although at a very generic level most existing methods make the key assumption that complexes are embedded among densely-interacting groups of proteins within PPI networks, these methods vary considerably either in the algorithmic methodologies or the kind of biological insights employed to detect complexes. Accordingly, we classified some popular complex detection methods into two broad categories (a soft classification): (i) methods based solely on graph clustering; (ii) methods based

on graph clustering and some additional biological insights. These biological insights may be in the form of functional, structural, organizational or evolutionary information known about complexes or their constituent proteins from experimental or other biological studies.

We present this classification in two snapshots. The first snapshot, shown in Figure 3.1, gives a *chronology-based* "bin-and-stack" classification, while the second snapshot, shown in Figure 3.2 gives a *methodology-based* "tree" classification of the methods.

In the chronology-based classification, we *binned* methods based on the years in which they were developed, and stacked them based on the kind of biological insights used. It is interesting to note from this classification that, over the years, as researchers tried to improve the basic graph clustering ideas, they also incorporated a variety of biological information into their methods. *Note* that we will keep returning back to this "bin-and-stack" chronology-based classification in subsequent chapters of this thesis, and adding new "data points" and/or "layers" to it.

In the methodology-based classification, we distributed the methods to different branches of a *classification tree* based on the kind of computational strategy used. At the first level from the root, we grouped these methods into those based solely on graph clustering, and those employing additional biological insights. At subsequent levels, we further divided these methods based on the kind of algorithmic strategies used, into: (i) methods employing merging or growing of clusters; (ii) methods employing repeated partitioning of networks; and (iii) methods employing network alignment. The methods employing merging or growing clusters go "bottom-up", that is, typically start with small "seeds" (for example, triangles or cliques), and repeatedly add or remove proteins or merge clusters based on some similarity measures to arrive at the final set of complexes. On the other hand, the methods based on network partitioning go "top-down", that is, repeatedly partition or break the network into multiple subnetworks based on certain divisive criteria. The methods based on network alignment use multiple networks (typically from different species) to arrive at isomorphic regions that likely correspond to complexes, the inituition being that proteins belonging to real complexes should generally be conserved through the evolution process to act as an integrated functional unit [29].

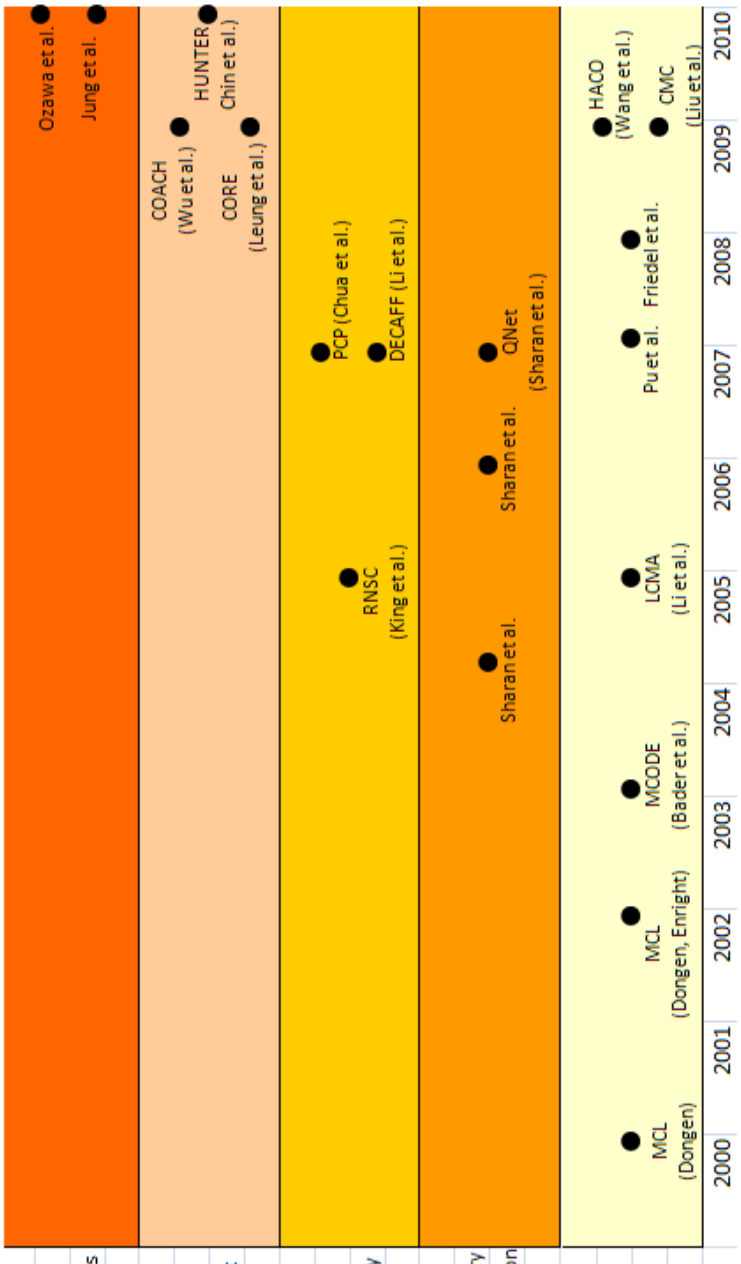


Figure 3.1: The "Bin-and-Stack" classification: Chronological binning of complex detection methods based on biological insights used. It is interesting to note that over the years, as researchers have tried to improve the basic graph clustering ideas, they have also incorporated newer biological information into their methods.

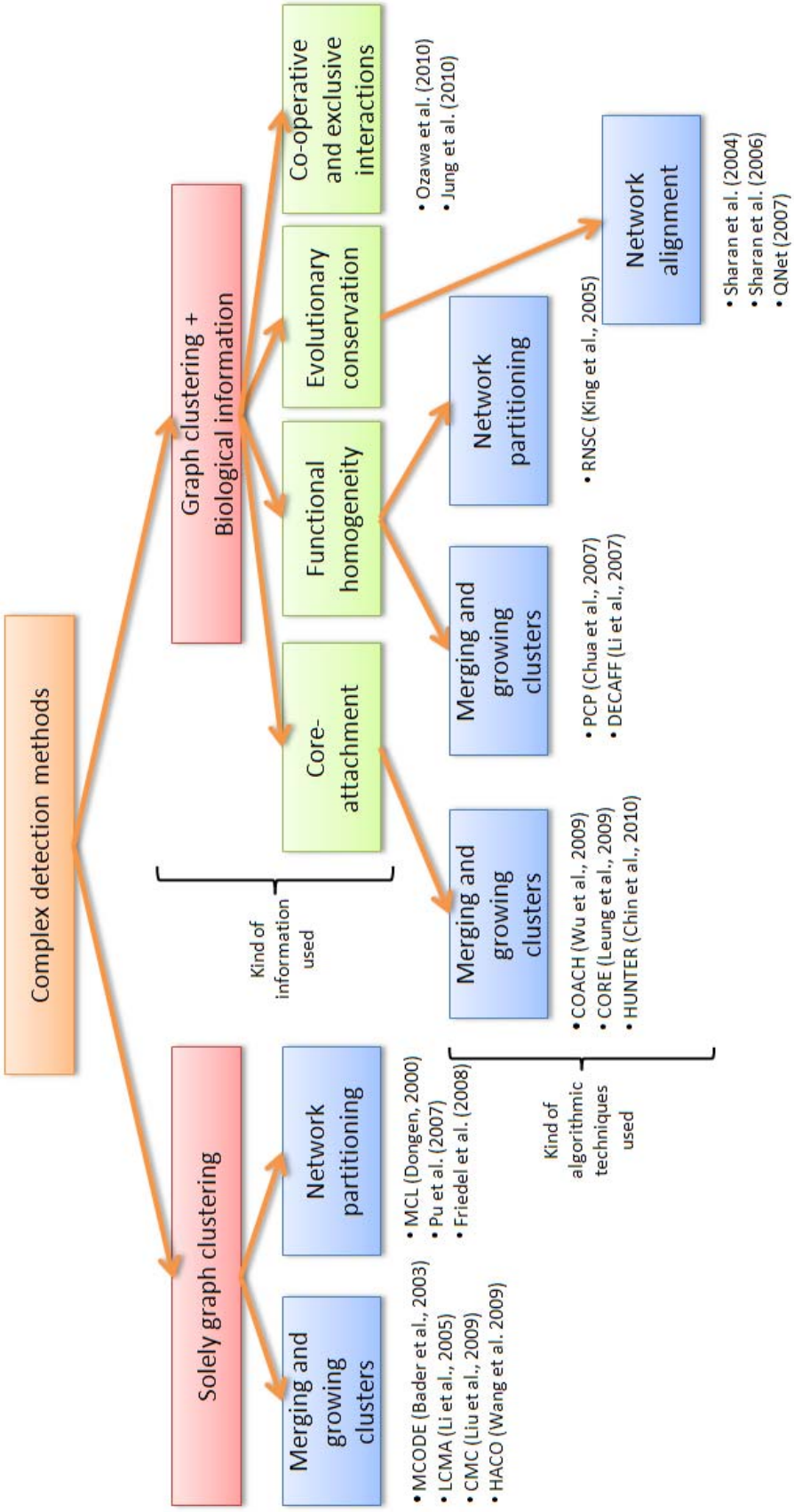


Figure 3.2: The 'Tree' classification: Classification of existing methods for complex detection based on the algorithmic methodologies used. Primarily three methodologies are adopted: merging and growing clusters, network partitioning and network alignment.

3.1.3 Methods based solely on graph clustering

Most methods that cluster the PPI network into multiple dense subnetworks make use of solely the topology of the network.

Molecular COmplex DEtection (MCODE)

MCODE, proposed by Bader and Hogue (2003) [14], is one of the first computational methods (and therefore, seminal) developed for complex detection from PPI networks. The MCODE algorithm operates in mainly in two stages, vertex weighting and complex prediction, and an optional third stage for post-processing.

In the first stage, each vertex v in the network $G = (V, E)$ is weighted based on its neighborbood density. Instead of directly using clustering coefficient, MCODE uses core-clustering coefficient which measures the density of the highest k-core in the neighborhood of v. This amplifies the weighting of densely connected regions in G. In the second stage, the vertex v with the highest weight is used to seed a complex. MCODE then recursively moves outwards from the seed vertex, including vertices into the complex whose weight is a given percentage (vertex weight parameter - VWP) away from the seed vertex. A vertex once added to a complex is not checked subsequently. The process stops when there are no more vertices to be added to the complex, and is repeated using the next unseeded vertex. At the end of this process multiple non-overlapping complexes are generated. The optional third stage performs a post-processing on the complexes generated from the second stage. Complexes without 2-cores are filtered out, and new vertices in the neighborhood with weights higher than a given 'fluff' parameter are added to existing complexes. The resultant complexes are scored and ranked based on their densities. The time complexity of the algorithm is $O(|V|.|E|.h^3)$, where h is the vertex size of the average vertex neighbourhood in the network G.

Markov CLustering (MCL)

The Markov Clustering (MCL) algorithm, proposed by Stijn van Dongen (2000) [16], is a general graph clustering algorithm that simulates random walks (called *flow*) to extract out relatively dense regions within networks. In biological applications, it was first applied to cluster protein families and ortholog groups [62] before it proved

to be effective in detecting complexes from protein interaction networks [31, 41, 63].

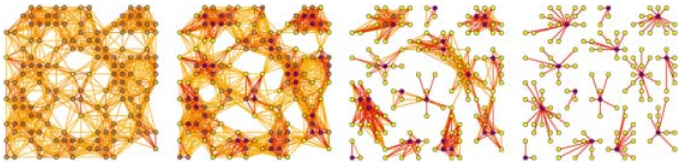


Figure 3.3: How MCL works [16]: Repeated expansion and inflation in MCL separates the network into multiple non-overlapping regions.

MCL manipulates the adjacency matrix of networks with two operators called *expansion* and *inflation* to control the random walks (flow). Expansion models the spreading out of the flow, while inflation models the contraction of the flow, making it thicker in dense regions and thinner in sparse regions. These parameters boost the probabilities of intra-cluster walks and demote those of inter-cluster walks. Mathematically, expansion coincides with normal matrix multiplication, while inflation is a Hadamard power followed by a diagonal scaling (see the pseudocode in Algorithm 1). Therefore, MCL is highly efficient and scalable. The iterative expansion and inflation separates the network into multiple non-overlapping regions, depicted in Figure 3.3 (one can view an animated example from `http://www.micans.org/mcl/`).

Algorithm 1 Markov Clustering (Graph G)

```
Add loops to G;
Inflation I to some value;
Set M1 to be a matrix of random walks on G;

while (change) do
   M2 := M1*M1; /* Expansion */
   M1 := Inflate(M2, I) /* Inflation */
   change := difference (M1, M2);
end while
Clusters := Compontents of M1;
```

Clustering based on merging Maximal Cliques (CMC)

CMC was proposed by Liu et al. (2009) [64] to detect complexes from PPI networks based on repeated merging of maximal cliques. Some earlier algorithms like CFinder [65] and Local Clique Merging Algorithm (LCMA) [66] also adopted clique merging to find dense neighborhoods, but the distinct advantage of CMC over these algorithms is its ability to work on weighted networks and to find relatively low density regions (in subsequent improved versions of CMC).

CMC begins by enumerating all maximal cliques in the PPI network using the Cliques algorithm proposed by Tomita et al. [67]. Although enumerating all maximal cliques is NP-hard, this does not pose a problem in PPI networks because these networks are usually sparse. CMC then assigns a score to each clique C based on its weighted density, which considers the reliabilities (weights) of the interactions within the clique:

$$Score(C) = \frac{\sum_{u,v \in C} w(u,v)}{|C|.(|C|-1)}. \tag{3.1}$$

CMC then ranks these cliques in decreasing order of their scores and iteratively merges or removes highly overlapping cliques based on their inter-connectivity scores. The inter-connectivity score of two cliques C_i and C_j is based on the non-overlapping regions of the two cliques and is defined as:

$$Inter\_score(C_i, C_j) = \sqrt{\frac{\sum_{u \in (C_i - C_j)} \sum_{v \in C_j} w(u,v)}{|C_i - C_j|.|C_j|} . \frac{\sum_{u \in (C_j - C_i)} \sum_{v \in C_i} w(u,v)}{|C_j - C_i|.|C_i|}} \tag{3.2}$$

CMC determines whether two cliques C_i and C_j sufficiently overlap: $|C_i \cap C_j|/|C_j| \geq overlap\_thresh$. If so, C_j is either removed or merged with C_i based on the inter_score: if the $inter\_score(C_i, C_j) \geq merge\_thresh$, then C_i and C_j are merged, else C_j is removed. Finally, all the resultant merged clusters are output as the predicted complexes.

Some other methods based on graph clustering

Apart from these discussed methods, three other methods worth mentioning here are LCMA (2005) [66], PCP (2007) [68] and HACO (2009) [69]. The LCMA algorithm first locates cliques within local neighborhoods using vertex degrees and then merges them based on overlaps to produce complexes. Protein Complex Prediction (PCP) uses FS Weight scoring to remove unreliable interactions and add indirect interactions, and then merges cliques to produce the final list of complexes. On the other hand, HACO uses hierarchical agglomerative clustering to produce the intial set of (non-overlapping) clusters. Proteins are then assigned to multiple clusters based on their interactions to produce the final list of overlapping clusters.

A few other recently proposed (2010 - 2011) methods include those by Zhang et al. [70], Ma et al. [71], Wang et al. [72] and Chin et al. [73]. These use the property of "bridgeness" of cross-edges among clusters along with the internal connectivities to detect complexes.

3.1.4 Methods incorporating core-attachment structure

Gavin and colleagues (2006) [15] performed large-scale analysis of yeast complexes and found that the proteins with complexes were divided into two distinct groups, "cores" and "attachments". The cores formed central functional units of complexes, while the attachment proteins aided these cores in performing their functions. Several computational methods were proposed to reconstruct complexes from PPI newtorks by capitalizing on this structural organization.

Wu Min et al. (2009) [75] proposed the COACH method which reconstructs complexes in two stages - it identifies dense core regions, and subsequently includes proteins as attachments to these cores. Figure 3.4 summarizes how COACH identifies core and attachment proteins to build complexes.

Leung et al. (2009) [76] proposed the CORE method to identify protein cores within the PPI network. They defined the probability of two proteins p_1 and p_2 (of degrees d_1 and d_2, respectively) to belong to the same core using two main factors: whether the two proteins interact or not and the number of common neighbors m between them. The probability that p_1 and p_2 have $\geq i$ interactions and $\geq m$ common neighbors is calculated under the null hypothesis that d_1 edges connecting p_1 and d_2 edges connecting p_2 are randomly assigned in the PPI network according to a uniform distribution. This probability is used to arrive at a p-value for whether p_1 and p_2 belong to the same core. Subsequently, CORE merges sets of core proteins of sizes two, three, etc. until further increase in size is not possible, to produce the final set of cores. CORE then scores and ranks the predicted cores based on the number of internal and external interactions in them. The attachments are added to these cores in a manner similar to COACH to produce the final set of complexes.

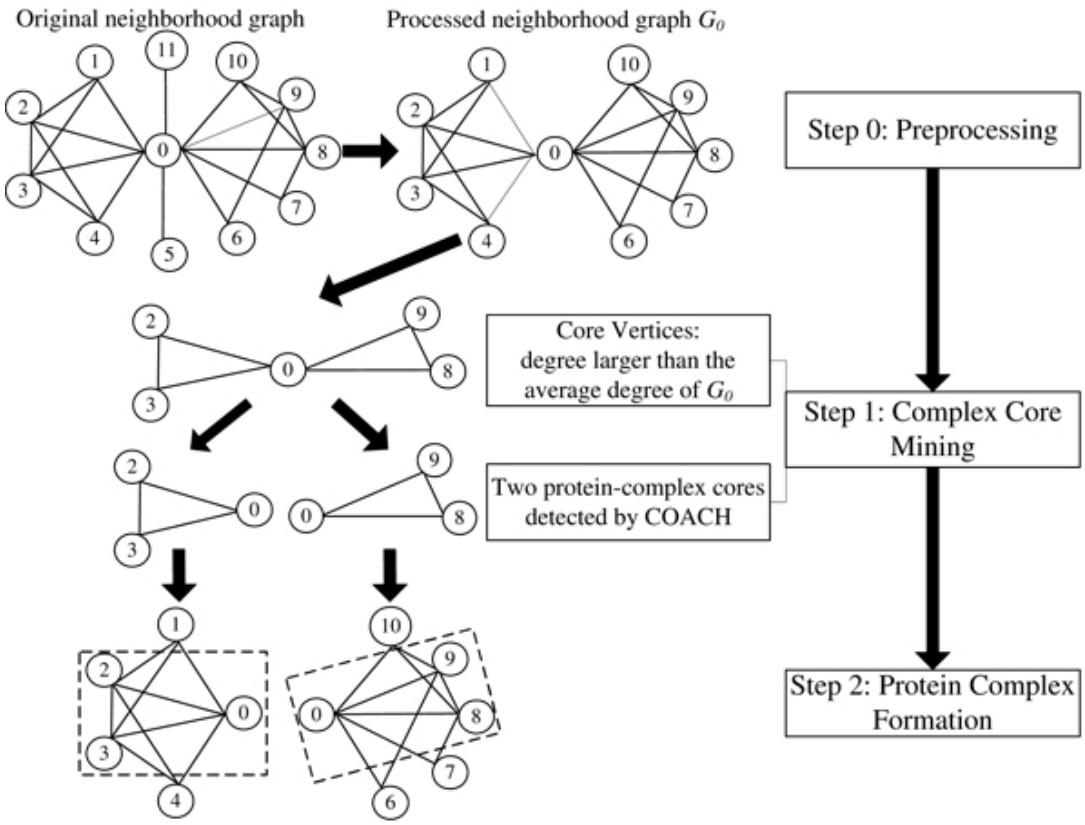


Figure 3.4: The identification of core and attachment proteins in COACH [75]: The cores are first identified based on vertex degrees in the neighborhood graphs. Attachment proteins are then appended to these cores to build the final complexes.

3.1.5 Methods incorporating functional information

Proteins within complexes are generally enriched with same or similar functions [15, 29]. If the functional information for proteins from an organism are available, then this information can be combined with topological information from PPI networks for the reconstruction of complexes from the organism. One possible way to incorporate functional information is to score the interactions based on the functional similarity between the interacting pairs of proteins. Alternately, functional annotations (for example, from Gene Ontology [37]) can be used to aid decisions where including or excluding a protein into complexes purely based on topological information might be difficult.

Restricted Neighborhood Search Clustering (RNSC)

King et al. (2004) [77] proposed the RNSC algorithm that combines topological and Gene Ontology information to detect complexes. The algorithm operates in two steps - it begins by clustering the PPI network and then filters the clusters based on cluster properties and functional homogeneity.

The network $G = (V, E)$ is first randomly partitioned into multiple subnetworks, which is essentially a partitioning of the node set V. The algorithm then iteratively moves nodes from one cluster to another in a randomized fashion till an integer-valued cost function is optimized. A common problem among such clustering algorithms is the tendency to settle in poor local minima. To avoid this, the RNSC algorithm adopts diversification moves, which shuffle the clustering by occasionally dispersing the contents of a cluster at random. Once the clustering process is completed, clusters of small sizes or densities (the lower bound on cluster sizes and densities are experimentally determined) are discarded. Finally, a p-value is calculated using functional annotations (from GO) for each cluster that measures the functional homogeneity of the clusters. All clusters above a certain p-value are discarded to produce the final list of predicted complexes. Based on experiments, King et al. recommend cluster density cut-off of 0.70 and p-value cut-off of 10^{-3}.

Dense neighborhood Extraction using Connectivity and conFidence Features (DECAFF)

Li et al. (2007) [78] proposed the DECAFF algorithm which essentially is an extention of the LCMA algorithm [66] proposed earlier by the same group. DECAFF identifies dense subgraphs in a neighborhood graph using a hub-removal algorithm. Local cliques are identified in these dense subgraphs and merged based on overlaps to produce clusters. Each cluster is assigned a functional reliability score, which is the average of the reliabilities of the edges within the cluster. All clusters with low reliabilities are discarded to produce the final set of predicted complexes.

The PCP algorithm [68] described earlier can also be categorized into this set of methods because PCP uses a weighting scheme based on functional similarity (though the similarity is inferred from topology) to assign reliability scores to interactions, and then uses a clique merging strategy to detect complexes.

3.1.6 Methods incorporating evolutionary information

The increasing availability of PPI data from multiple species like yeast, fly, worm and some mammals has made it feasible to use insights from cross-species analysis for detection of (conserved) complexes. The assumption is that proteins belonging to real complexes should generally be conserved through the evolution process to act as an integrated functional unit [29].

Sharan et al. proposed methods (2005-2007) [79, 80] for detection of conserved complexes across species based on the evolution of PPI networks. In these methods, an orthology network (network alignment graph) is constructed from the PPI networks of different species, which essentially represents the orthologous proteins and their conserved interactions across the species. For a protein pair $\{u_1, v_1\}$ in network G_1 of species S_1 and (u_2, v_2) in G_2 of species S_2, the orthology network G_{12} contains the pair $\{u, v\}$ if u_1 is orthologous to u_2, and v_1 is orthologous to v_2. The edge (u, v) is weighted by the sequence similarities between the pairs $\{u_1, v_1\}$, and $\{u_2, v_2\}$. Any subgraph in G_{12} is therefore a conserved subnetwork of G_1 and G_2. Such candidate subgraphs are then evaluated for parts of conserved complexes. Based on this idea, a tool *QNet* [81] was developed which returns conserved complexes from different species when queried using known complexes from yeast.

3.1.7 Methods based on co-operative and exclusive interactions

The overlapping binding interfaces in a protein may prevent multiple interactions involving these interfaces from occurring simultaneously [82]. In other words, the set of interactions in which a protein participates may be either co-operative or mutually exclusive. The information about the co-occurrence or exclusiveness of interactions can therefore be useful for predicting complexes with higher accuracy. This information can be gathered from the interacting domains of protein pairs or the three-dimensional structures of the interacting surfaces.

Ozawa et al. (2010) [83] proposed a refinement method over MCODE and MCL to filter predicted complexes based on exclusive and co-operative interactions. They used domain-domain interactions to identify conflicting pairs of protein interactions in order to include or exclude proteins within candidate complexes. Based on their results, the accuracies of predicted complexes from MCODE and MCL improved by two-fold.

On the other hand, Jung et al. (2010) [84] used structural interface data to construct a simultaneous PPI network (SPIN) containing only co-operative interactions and excluding competition from mutually exclusive interactions. MCODE and LCMA algorithms tested on this SPIN displayed a sizeable improvement in correctly predicted complexes.

Even though incorporating information about co-operative and exclusive interactions shows promising improvement in complex detection algorithms, there are still several practical problems related to this approach. Gathering more data on conflicting interactions, especially based on three-dimensional structures of interfaces, needs to be addressed before this approach can be more easily adopted.

3.1.8 Incorporating other possible kinds of information

In a recent foresightful survey by Przytycka et al. [85], the application of network *dynamics* (temporal information) into current computational analysis is discussed at good lengths, especially with respect to detection of complexes and pathways from protein interaction networks. The authors suggest that if sufficient information

about the 'timing activities' of proteins can be obtained, the dynamical nature of the underlying organizational principles in interaction networks can be better understood. This shift from static to dynamic network analysis is vital to understanding several cellular processes, some of which may have been wrongly understood due to ignoring dynamic information.

3.1.9 Comparative assessment of existing methods

Considering the wide variety of proposed methods for complex detection, one can gauge the seriousness in the research effort towards computational identification and categorization of complexes. Several surveys and experiments [86–88] have focused on the comparative analyses of these proposed methods for complex detection. Each new work on complex detection also comes with detailed comparative analyses of the new method with some earlier methods. However, due to the differences in PPI and benchmark datasets, evaluation criteria, thresholds and parameters used, and the subset of methods considered for these comparative assessments, different works arrive at different conclusions about the performance of methods. Here, we present a summary of some widely accepted surveys dealing with comparative assessments of complex detection methods.

One of the first comprehensive assessments of algorithms was performed by Brohee and van Helden (2006) [86]. They performed a detailed empirical comparison between MCODE [14], MCL [16], RNSC [77] and Super-paramagnetic Clustering (SPC) [89]. These algorithms were tested on PPI datasets from high-throughput experiments, and the resultant complexes were evaluated against benchmark complexes from MIPS [90]. Additionally, the PPI datasets were introduced with artificial noise (random edge addition and deletion) to test the robustness of these algorithms. They concluded that MCL and RNSC outperformed MCODE and SPC in terms of precision (the proportion of correctly predicted complexes) and recall (the proportion of correctly derived benchmarks). RNSC was robust to variation in its input parameter settings, while the performance of the other three varied widely for parameter changes. MCL was remarkably robust even upon introducing 80%-100% random noise. Overall, the experiments confirmed the general superiority of MCL over the other three algorithms.

Vlasblom et al. (2009) [87] compared MCL with another clustering algorithm, Affinity Propagation (AP) [91] on unweighted as well as weighted PPI networks. The initial unweighted network was built from a set of 408 hand-curated complexes from Wodak lab [92] followed by random addition and removal of edges to mimic real PPI networks. The weighted network was obtained from the Collins et al.'s work [36], generated from Gavin and Krogan datasets [15,28]. They concluded that MCL performed considerably better than AP in terms of accuracy and separation of predicted clusters, and robustness to random noise. In particular, MCL was able to achieve about 90% accuracy and 80% separation compared to only 70% accuracy and 50% separation of AP on unweighted PPI networks with introduced random noise. MCL was able to discover benchmark complexes even at high (40%) noise levels.

More recently (2010), Li et al. [88] performed a detailed comparative evaluation of several algorithms: MCODE [14], MCL [16], CORE [76], COACH [75], RNSC [77] and DECAFF [78]. These algorithms were tested on PPI datasets from DIP [52] and Krogan et al. [28]. The DIP network consisted of 17203 interactions among 4930 proteins, while the Krogan dataset consisted of 14077 interactions among 3581 proteins. They used a total of 428 benchmark complexes from MIPS [90], Aloy et al. [93] and SGD [94]. A cluster P from a method was considered a correct match to a benchmark complex B using the Bader score $|V_P \cap V_B|^2/(|V_P|.|V_B|) \geq 0.20$, where V_P denotes the number of proteins in P, and V_B denotes the number of proteins in B. Based on this criteria, the precision (the proportion of correctly matched clusters), recall (the proportion of benchmark complexes matched) and F1-measure (the harmonic mean of precision and recall) values were calculated. The comparisons between precision, recall and F1-measures of these algorithms is shown in Figure 3.5 (adapted from [88]). The methods are arranged in chronological order, and it is interesting to note that over the years, the F1-measures have improved. Li et al. concluded that MCL, RNSC, CORE, COACH and DECAFF attained the best recall values. MCODE was able to achieve the highest precision, but it produced very few clusters resulting in very low recall.

Plugging into the "bin-and-stack" classification: We benchmarked these methods based on their performance on two more recent datasets: the raw (unscored)

and scored (using Purification Enrichment [36]) networks comprising of data from Gavin et al. [15] and Krogan et al. [28], as shown in Figure 3.6. For each method, we show the values *before / after* scoring. This figure clearly demonstrates that incorporating biological information together with affinity scoring significantly boosts performance. Therefore, our taxonomy has the potential to reveal interesting insights based on the trend of methods.

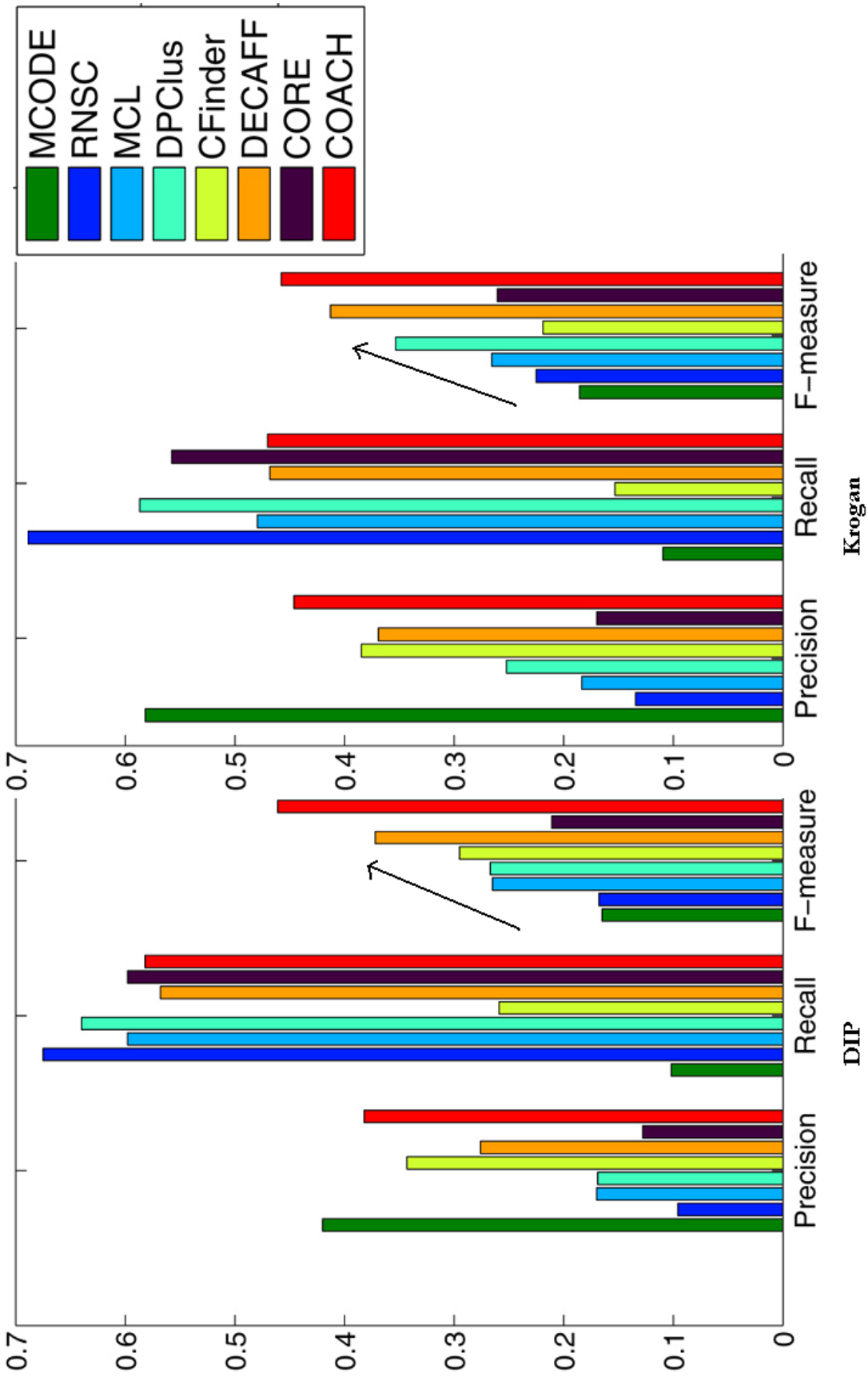


Figure 3.5: Comparative performance of complex detection methods in terms of precision, recall and F-measure on DIP and Krogan datasets (adapted from [88]). The methods are arranged in chronological order, and it is interesting to note that over the years, the F1-measures have improved.

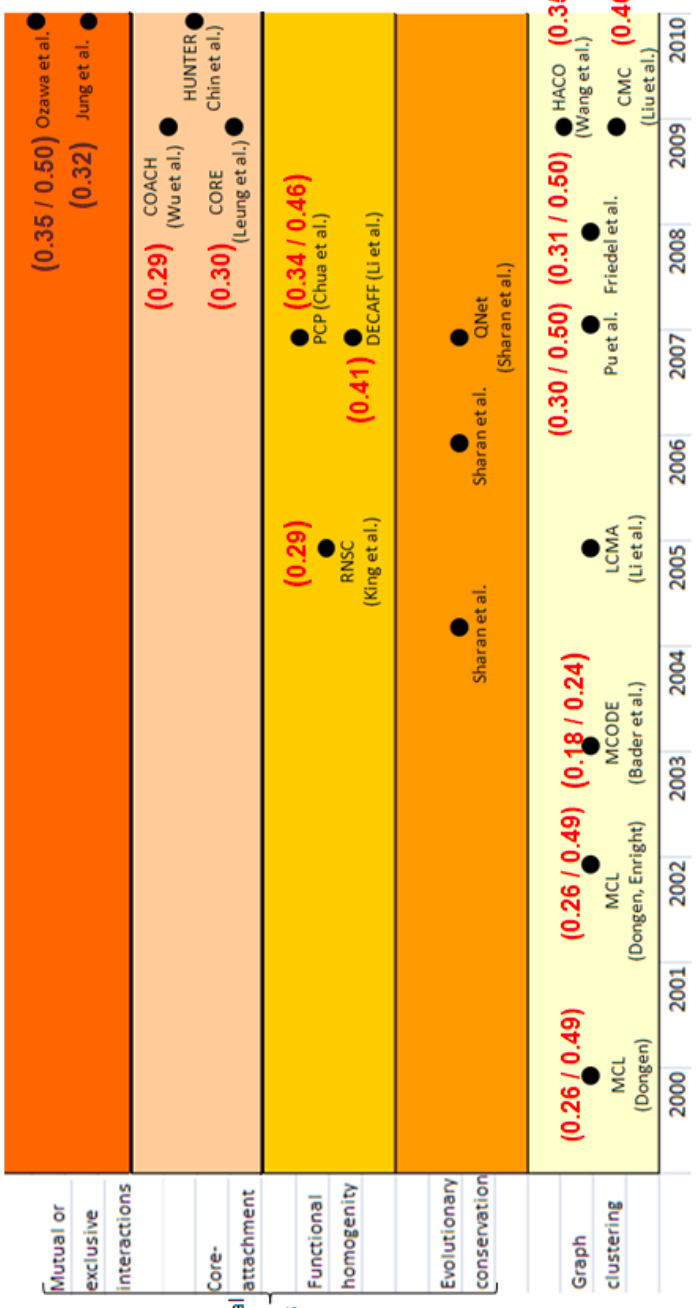


Figure 3.6: "Plugging-in" F1-measure values of existing methods into our "Bin-and-Stack" classification. The two values for each method mean (before / after) affinity scoring of interactions. This figure clearly demonstrates that incorporating biological information together with affinity scoring significantly boosts performance. Therefore, our taxonomy has the potential to reveal interesting insights based on the trend of methods.

3.2 Challenges and lessons from current practice

The review of experimental techniques for inferring protein interactions (Chapter 2), and computational methods for complex detection reveal several challenges facing identification of complexes from high-throughput interaction datasets. We have broadly classified these challenges into two categories: (i) challenges originating from real biological datasets; (ii) challenges originating from existing computational techniques.

Challenges from real biological datasets

Even though over the last few years, several independent high-throughput experiments (see Chapter 2 for a survey) have helped to catalogue enormous amount of protein interactions from organisms such as yeast, these individual datasets are the best available, they show surprising lack of correlation with each other, and some bias towards high abundance proteins and against proteins from certain cellular compartments (like cell wall and plasma membrane) [26, 32–34]. Also, each dataset still contains a substantial number of false positives (noise) that can compromise the utility of these datasets for more focused studies like complex reconstruction [36]. In order to reduce the impact of such discrepancies, a number of data integration and affinity scoring schemes have been devised. These affinity scores encode the reliabilities (confidence) of physical interactions between pairs of proteins. Therefore, the challenge now is to detect meaningful as well as novel complexes from protein interaction (PPI) networks derived by combining multiple high-throughput datasets and by making use of these affinity scoring schemes.

Challenges from existing complex detection methods

Even though there have been numerous methods developed for complex detection, all them suffer from low recall, which is mainly due to the lack of sufficient credible interactions and the presence of noise (spurious interactions) in the datasets.

From the study of existing methods we notice that every method, in one way or another, relies on the assumption that complexes are embedded among dense regions of the network. However, the overall recall of the methods is not very

impressive, indicating that relying too much on this assumption in the wake of insufficient credible interaction data causes these methods to miss many complexes that are of low densities in the network.

In addition to this, noise in datasets can also be a limiting factor. But, this noise can be countered to a certain extent by capitalizing on scoring schemes that assign reliability scores to the interactions (Chapter 2). However, currently there are very few methods that capitalize on these scores, and even if they do, these methods do not perform uniformly across all scoring schemes and are tied to one or two schemes.

Lessons learnt

We list the "take-home" lessons from this chapter that can help to improve complex detection:

1. Combining interaction datasets from multiple sources improves interaction coverage: increases the true positives and reduces the false negatives [36];

2. Adopting reliability scores for interactions is useful to remove many false positive interactions [36];

3. Incorporating biological information along with topology of PPI networks improves performance (Figure 3.6);

4. The assumption that complexes form "dense" regions in PPI networks is not entirely valid in the wake of insufficient credible data.

Keeping in mind these lessons, we proceed to the next chapter where we develop a new computational method to detect complexes from protein interaction networks by utilizing core-attachment modularity and capitalizing on reliability scores assigned to interactions.

CHAPTER 4

Refining Markov Clustering for complex detection by incorporating core-attachment structure

You know my method. It is founded upon the observations of trifles.

The Boscombe Valley Mystery, 1892
The Adventures of Sherlock Holmes
- Sir Arthur Conan Doyle

Our approach to reconstruct complexes from protein interaction networks is inspired from the findings by Gavin et al. (2006) [15] on the "core-attachment" modularity structure found in yeast and other eukaryotic complexes. The intuition behind our approach is that if yeast complexes indeed possess this "core-attachment" structure, most of the dense regions within PPI networks that correspond to real complexes should adhere to such a structure. Therefore, if we consciously search for such embedded structures among these dense regions, we should be able to accurately extract out complexes rather than considering whole of the dense regions as complexes as is done in most methods. This helps to reduce the number of incorrectly included (loosely-connected) proteins within predicted complexes, and thereby help to reconstruct complexes with better accuracies.

For finding the initial set of dense regions within PPI networks, we use the MCL clustering algorithm [16,62,63]. We then identify the "core" and "attachment" sets of proteins from the MCL clusters. This gives us two levels of "controls" to be stringent or lenient while identifying the complex proteins within dense regions. We name our algorithm as MCL-CAw, where the 'w' describes the ability of the algorithm to work on weighted (scored) PPI networks.

We chose MCL because it is simple, scalable, robust to noise and performs reasonably well for general graph clustering compared to most other clustering algorithms like k-means, super-paramagnetic clustering (SPC) and affinity propagation (AP) (see Chapter 3 or [86, 87]). Secondly, MCL is a well-studied algorithm both for general graph clustering as well as complex detection [16, 62, 63, 86–88]. Its advantages and limitations are well-known under different scenarios. In addition to these, we also identified some limitations of MCL specific to complex detection, which further motivate our approach:

1. It is well-known that a protein may be recruited by more than one complex for performing functions [15, 31, 69]. However, MCL produces only non-overlapping complexes, arbitrarily assigning shared proteins to only one of them.

2. Our experiments revealed that MCL produces many noisy clusters that either do not match real complexes or reduce the accuracies of correctly predicted complexes. For example, when we ran MCL on PPI datasets from Gavin et al. [15] and Krogan et al. [28], the average Jaccard accuracies of predicted clusters when matched to the Wodak lab [92] benchmark was only 0.472 and 0.448, respectively (Table 4.1). Upon evaluation of these predicted clusters, we found that MCL had included several additional (noisy) proteins that reduced the accuracies of these clusters.

	# Clusters			Avg Jaccard		
PPI Dataset	Predicted	Matched	Missed	Predicted	Matched	Missed
Gavin 2006	232	53	179	0.472	0.694	0.282
Krogan 2006	632	81	551	0.448	0.627	0.173

Table 4.1: Low accuracies of predicted clusters of MCL from Gavin and Krogan datasets (criteria for a match: Jaccard score ≥ 0.50).

4.1 Gavin's "Core-attachment" model of yeast complexes

Even though likely to be expected from eukaryotic complexes and already hypothesized in some earlier works [74], the experiments by Gavin and colleagues (in 2006) [15] formed the first large-scale assessment that revealed distinct core-attachment organization of proteins within yeast complexes. Gavin et al. used a TAP-MS technique [24] to "pull down" complexes from *Saccharomyces cerevisae* (budding yeast). Binary interactions were inferred from these TAP-MS complexes using a combination of "spoke" and "matrix" models, and scored using a 'socio-affinity' index. Clustering these interactions using a matrix-based iterative approach generated 491 distinct complexes that matched the hand-curated complexes from MIPS with 83% coverage and 78% accuracy. Careful analyses of these complexes revealed distinct modularity structure vital to the performing of biological functions. Based on this, Gavin et al. proposed their model of yeast complexes. Complexes are composed of two distinct groups of proteins - "cores" and "attachments". The cores range from 1 - 23 proteins in size (average 3.1 ± 2.5) and form the main functional parts within complexes, while the attachments aid these cores in performing their functions. Among these attachments are tightly-coupled subsets of proteins called "modules" that always function in cohesion.

A note on interpreting the Gavin model for complex prediction: The Gavin "Core-attachment" model has been interpreted in different ways in computational works [75, 76, 96] to predict complexes, though the model *per se* is general enough to include all interpretations. Works like [75, 96] allow the same set of core

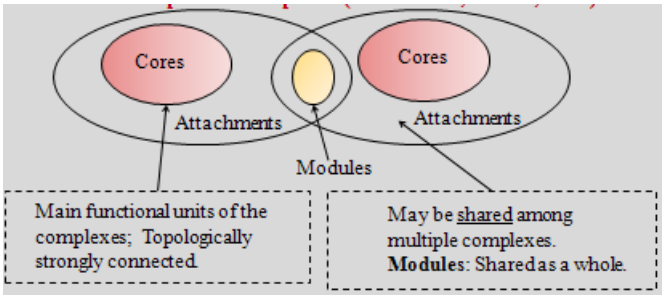


Figure 4.1: A pictorial representation of our interpretation of Gavin et al.'s "core-attachment" model [15] of yeast complexes.

proteins to participate in multiple complexes: these cores interact with different attachments to form different complexes. On the other hand, works like [76] consider the cores to be unique to complexes. In our interpretation of the Gavin model, we inherently put a restriction on the uniqueness of cores, but allow the attachments to be shared among multiple complexes (depicted in Figure 4.1). Even though the repetition of the same set of cores across complexes is possible, this is mainly seen in complex *isoforms*. These complex isoforms comprise of complexes that have almost the same protein compositions and therefore perform very similar functions (for example, the RNA polymerase complexes I, II and III) and in cohesion. Since there are very few such cases of complex isoforms relative to the total number of complexes with distinct sets of proteins (and functions) in eukaryotes, we do not consider the sharing of whole cores among complexes as a strong property in our interpretation of the Gavin model. Instead we allow the attachment proteins to be shared among complexes, which is essential to capture the sharing of proteins among "non-isoformic" complexes (for example, the sharing of Yor076c among the Exosome and Ski complexes), which is more important to understand the "cross-talk" among functional categories. However, a negative effect of not allowing core-sharing is that we might bundle together complex isoforms into a single cluster during computational predictions. But, this is not a serious problem because usually the bundled complexes are very similar in their functionalities and are better studied as a whole (in fact Gavin et al. also combine together the complex isoforms in their study; see Figures 1a and 1b in [15]). Nevertheless, in the next chapter we do propose a way to segregrate the individual complex isoforms from the cluster wherever such a study warrants.

4.2 The MCL-CAw algorithm

The MCL-CAw algorithm broadly consists of two phases. In the first phase, we partition the PPI network into multiple dense clusters using MCL. Following this (in the second phase), we post-process (refine) these clusters by incorporating core attachment structure to produce the final complexes. This procedure can be divided into the following steps:

1. Clustering the PPI network using MCL

2. Categorizing proteins as cores within clusters
3. Filtering away noisy clusters
4. Recruiting proteins as attachments into clusters
5. Extracting out complexes from clusters
6. Ranking the predicted complexes

Our PPI network is represented as $G = (V, E)$, where V is the set of proteins, and E is the set of interactions between these proteins. For each edge $(p, q) \in E$, there is a confidence score (weight) $0 \leq w(p, q) \leq 1$ encoding the affinity between the proteins p and q. These affinity scores depend on the scoring system used.

Clustering the PPI network using MCL

The first step of our algorithm is to partition (cluster) the PPI network using MCL [16], which simulates random walks (called a flow) to identify relatively dense regions in the network. The inflation coefficient parameter I in MCL is used to regulate the granularity of the clusters - higher the value more finer are the generated clusters (how to choose I in practice is discussed in the "Results" section). On PPI networks, MCL has a tendency to produce large clusters (sizes ≥ 25) which house several smaller complexes. If such large clusters are produced, we iteratively recluster them (hierarchical clustering) using a higher inflation value.

After MCL-based clustering, we obtain a collection of k disjoint (non-overlapping) clusters $\{C_i : C_i = (V_i, E_i), 1 \leq i \leq k\}$, where $V_i \subseteq V$ and $E_i \subseteq E$.

Categorizing proteins as cores within clusters

Microarray analysis by Gavin et al. [15] of their predicted complex components showed that a large percentage of pairs of proteins within cores were co-expressed at the same time during cell cycle and sporulation, consistent with the view that cores represent main functional units within complexes. Three-dimensional structural and yeast two-hybrid analysis showed that the core components were most likely to be in direct physical contact with each other. To reflect these findings in our algorithm, we expect:

- Every complex we predict to comprise of a non-empty set of core proteins; and
- The proteins within these cores to display relatively high degree of physical interactivity among themselves that with other proteins.

We categorize a protein $p \in V_i$ to be a *core* protein in cluster $C_i = (V_i, E_i)$, given by $p \in Core(C_i)$, if:

- The *weighted in-connectivity of p with respect to* C_i is at least the *average weighted in-connectivity of* C_i, given by: $d_{in}(p, C_i) \geq d_{avg}(C_i)$; and
- The weighted in-connectivity of p with respect to C_i is greater than the *weighted out-connectivity of p with respect to* C_i, given by: $d_{in}(p, C_i) > d_{out}(p, C_i)$.

The weighted in-connectivity $d_{in}(p, C_i)$ of p with respect to C_i is the total weight of interactions p has with proteins within C_i. Similarly, the weighted out-connectivity $d_{out}(p, C_i)$ of p with respect to C_i is the total weight of interactions p has with proteins outside C_i. These are given by $d_{in}(p, C_i) = \sum\{w(p, q) : q \in V_i\}$ and $d_{out}(p, C_i) = \sum\{w(p, q) : q \notin V_i\}$, respectively. The average weighted in-connectivity $d_{avg}(C_i)$ of cluster C_i is therefore the average of the weighted in-connectivities of all proteins within C_i, given by $d_{avg}(C_i) = \frac{1}{|C_i|} . \sum_{q \in V_i} d_{in}(q, C_i)$.

Filtering noisy clusters

Consistent with the assumption that every complex comprises of a set of core proteins, we consider a cluster as noisy if it does not contain a core of at least two proteins as per our above criteria. We discard all such noisy clusters.

Recruiting proteins as attachments into clusters

Microarray analysis by Gavin et al. [15] of their predicted complex components showed that attachment proteins were closely associated with core proteins within complexes and yet showed a greater degree of heterogeneity in expression levels, supporting the notion that attachments might represent non-stoichiometric components. Also, attachment proteins were seen shared between two or more complexes,

consistent with the view that the same protein may participate in multiple complexes [31, 69]. On the other hand, the application of MCL to PPI networks yields clusters that do not share proteins (that is, non-overlapping clusters). Mapping these clusters back to the PPI network shows that proteins having similar connectivities to multiple clusters are assigned arbitrarily to only one of the clusters. These proteins might as well be assigned to multiple clusters. To reflect these findings in our algorithm, we expect the attachment proteins to be those proteins within complexes that are:

- Non-core proteins;
- Closely interacting with the core proteins; and
- May be shared across multiple complexes.

We consider the following criteria to assign a non-core protein p belonging to a cluster C_j (called donor cluster) as an *attachment* in an acceptor cluster C_i (the donor and acceptor clusters may be the same), that is, $p \in Attach(C_i)$:

- Protein p has sufficiently strong interactions with the core proteins $Core(C_i)$ of the cluster C_i;
- The stronger the interactions among the core proteins, the stronger have to be the interactions of p with the core proteins;
- For large core sets, strong interactions are required to only some of the core proteins or, alternatively, weaker interactions to most of them.

Combining these criteria, we assign non-core p as an attachment in the acceptor cluster C_i, that is $p \in Attach(C_i)$, if:

$$I(p, Core(C_i)) \geq \alpha . I(Core(C_i)) . \Big(\frac{|Core(C_i)|}{2} \Big)^{-\gamma}, \tag{4.1}$$

where $I(p, Core(C_i))$ is the total weight of interactions of p with $Core(C_i)$, given by $I(p, Core(C_i)) = \sum \{w(p,q) : q \in Core(C_i)\}$, while $I(Core(C_i))$ is the total weight of interactions among the core proteins of C_i, given by $I(Core(C_i)) = \frac{1}{2} . \sum \{w(q,r) : q, r \in Core(C_i)\}$. The power function is normalized to yield 1 for core sets of size 2. The parameters α and γ are used to control the effects of $I(Core(C_i))$ and core size $|Core(C_i)|$. For a simple illustration, let $\alpha = 0.5$ and

$\gamma = 1$, and consider all interactions to be of equal weight 1. Therefore, p is attached to a core set of four proteins, if the total weight of its interactions with the core proteins is at least 3, which is possible if p is connected to at least three core proteins (how to choose values for α and γ in practice is discussed in the "Results" section). This step also ensures that non-core proteins having sufficiently strong interactions with multiple core sets are recruited as attachments to all those core sets.

Extracting out complexes from clusters

For each cluster we group together its constituent core and attachment proteins to define a unique complex. We expect all the remaining proteins within the cluster to have weaker associations with this resultant complex, and therefore categorize them as noisy proteins. Additionally, since these resulting complexes include attachment proteins that potentially may be recruited by multiple complexes, our predicted complexes adhere to the protein-sharing phenomenon observed in real complexes [15, 31, 69]. We discard all complexes of size less than 4 many of which may be false positives because it is difficult to predict small real complexes purely based on topological information (also noted in [64, 76]).

For each cluster C_i, we define a unique complex $Complex(C_i)$ as:

$$Complex(C_i) = \{Core(C_i) \cup Attach(C_i)\}. \tag{4.2}$$

Each interaction (p, q) within this complex carries the affinity score (weight) $w(p, q)$ observed in the PPI network.

Ranking the predicted complexes

As a final step, we output our predicted complexes in a reasonably meaningful order of biological significance. For this, we rank our predicted complexes in decreasing order of their weighted densities. The *weighted density* of a predicted complex C'_i is given by [64]:

$$Weighted\ density\ \ WD(C'_i) = \frac{\sum_{p,q \in C'_i} w(p, q)}{|C'_i|.(|C'_i| - 1)}. \tag{4.3}$$

The *unweighted density* of a predicted complex is defined in a similar way by setting

the weights of all constituent interactions to 1. This blindly favors very small complexes, or complexes with proteins having large number of interactions without considering the reliabilities of those interactions. On the other hand, the weighted density considers the reliabilities of such interactions. If two complexes have the same unweighted density, the complex with higher weighted density is ranked higher.

4.3 Experimental results

4.3.1 Preparation of experimental data

We gathered high-confidence Gavin and Krogan-Core interactions for yeast deposited in the public database BioGrid [54] (`http://thebiogrid.org/`) (version as of July 2009). These were assembled from bait-prey and prey-prey relationships (a combination of the 'matrix' and 'spoke' models) observed by Gavin et al. [15], and bait-prey relationships (the 'spoke' model) observed by Krogan et al. [28]. We combined these interactions to build the unscored Gavin+Krogan network (all edge weights set to 1). We then applied the Iterative-CD^k [40, 64] and FS Weight^k [39] scorings (with $k = 10$ iterations) on the Gavin+Krogan network, and selected all interactions with non-zero scores. This resulted in the ICD(Gavin+Krogan) and FSW(Gavin+Krogan) networks, respectively. In addition, we downloaded the Consolidated$_{3.19}$ and Consolidated$_{0.623}$ networks (with PE cut-off 3.19 recommended by Collins et al. [36], and 0.623, the average PE score) from `http://interactome-cmp.ucsf.edu/`. We also downloaded the Bootstrap$_{0.094}$ network [41] (with BT cut-off: 0.094) from `http://www.bio.ifi.lmu.de/Complexes/ProCope/`. The Consolidated network was derived from the matrix model relationships of the Gavin and Krogan datasets using the PE system [36]. The Bootstrap network was derived from the matrix model relationships using bootstrapped scores [41]. These two networks comprised of additional prey-prey interactions that were missed in the original Krogan-Core dataset. Table 4.2 summarizes some properties of these networks.

The benchmark (reference or 'gold standard') set of complexes was built from three independent sources: 408 complexes of the Wodak lab CYC2008 catalogue [92], 313 complexes of MIPS [90], and 101 complexes curated by Aloy et

PPI Network	# Proteins	# Interactions	Avg node degree
Gavin	1430	7592	10.62
Krogan 'Core'	2708	7123	5.26
Gavin+Krogan	2964	13507	9.12
ICD(Gavin+Krogan)	1628	8707	10.69
FSW(Gavin+Krogan)	1628	8688	10.67
$\text{Consolidated}_{3.19}$	1622	9704	11.96
$\text{Consolidated}_{0.623}$	5423	102393	37.76
$\text{Bootstrap}_{0.094}$	2719	10290	7.56

Table 4.2: Properties of the PPI networks used for the evaluation of MCL-CAw

al. [93]. The properties of these reference sets are shown in Table 4.3. We considered each of these reference sets independently for the evaluation of MCL-CAw. We did not merge them into one comprehensive list of complexes because the individual complex compositions are different across the three sources and some complexes may also get double-counted (because of different names used for the same complex). An alternative strategy was adopted by Wang et al. [69] by integrating the complexes from three sources (MIPS [90], SGD [94] and their own in-house curated complexes) using the Jaccard score: two complexes overlapping with a Jaccard score of at least 0.70 were merged together - the proteins to be included into the resultant complex were chosen based on a voting scheme.

			# Complexes of size				
Benchmark	#Complexes	# Proteins	< 3	3-10	11-25	> 25	Avg density
Wodak	408	1627	172	204	27	5	0.639
MIPS	313	1225	106	138	42	27	0.412
Aloy	101	630	23	58	19	1	0.747

Table 4.3: Properties of hand-curated (verified and *bona fide*) yeast complexes from Wodak lab [92], MIPS [90] and Aloy [93]

To be accurate (as well as fair) while evaluating our method on these benchmark sets, we considered only the set of *derivable benchmark complexes* from each of the PPI networks: if a protein is not present in a PPI network, we remove it from the set of benchmark complexes; by repeated removals, if the size of a benchmark complex shrinks below 3, we remove the complex from our benchmark set to generate the final set of derivable benchmark complexes for each of the PPI networks.

In order to evaluate the biological coherence of our predicted complexes, we

downloaded the list of cellular localizations (GO terms under "Cellular Component") of proteins from Gene Ontology (GO) [37]. We selected only the *informative* GO terms. A GO term is informative if: (a) the term contains more than 30 proteins annotated to it; and (b) each of the term's descendants contains less than 30 proteins annotated to it [95].

4.3.2 Metrics for evaluating the predicted complexes

Let $\mathcal{B} = \{B_1, B_2, ..., B_m\}$ and $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ be the sets of benchmark and predicted complexes, respectively. We use the Jaccard coefficient J to quantify the overlap between a benchmark complex B_i and a predicted complex C_j:

$$J(B_i, C_j) = \frac{|B_i \cap C_j|}{|B_i \cup C_j|}. \tag{4.4}$$

We consider B_i to be covered by C_j, if $J(B_i, C_j) \geq$ *overlap threshold* t. In our experiments, we set the threshold $t = 0.5$, which requires $|B_i \cap C_j| \geq \frac{|B_i|+|C_j|}{3}$. For example, if $|B_i| = |C_j| = 8$, the overlap between B_i and C_j should be at least 6.

We use previously reported [64] definitions of *recall* (coverage) and *precision* (sensitivity) of the set of predicted complexes:

$$Recall = \frac{|\{B_i | B_i \in \mathcal{B} \wedge \exists C_j \in \mathcal{C}; J(B_i, C_j) \geq t\}|}{|\mathcal{B}|} \tag{4.5}$$

Here, $|\{B_i | B_i \in \mathcal{B} \wedge \exists C_j \in \mathcal{C}; J(B_i, C_j) \geq t\}|$ gives the number of *derived benchmarks*.

$$Precision = \frac{|\{C_j | C_j \in \mathcal{C} \wedge \exists B_i \in \mathcal{B}; J(B_i, C_j) \geq t\}|}{|\mathcal{C}|} \tag{4.6}$$

Here, $|\{C_j | C_j \in \mathcal{C} \wedge \exists B_i \in \mathcal{B}; J(B_i, C_j) \geq t\}|$ gives the number of *matched predictions*.

We also evaluate the performance of our method by plotting *precision versus recall curves* for the predicted complexes. These curves are plotted by tuning a threshold on the number of predicted complexes considered for the evaluation. The predicted complexes are considered in decreasing order of their weighted densities (that is, in increasing order of their complex ranks).

4.3.3 Metrics for evaluating the biological coherence

A complex can be formed if its proteins are localized within the same compartment of the cell. So, we use the localization coherence of the predicted complexes as a measure their quality. Let $\mathcal{L} = \{L_1, L_2, ..., L_k\}$ be the set of known localization groups, where each L_i contains a set of proteins with similar localization annotations. The *co-localization score* of a predicted complex C_j is defined as the maximal fraction of its constituent proteins that are co-localized within the same localization group among the proteins that have annotations. This is given as follows [64]:

$$Loc\ score(C_j) = \frac{\max\{|C_j \cap L_i| : i = 1, 2, ..., k\}}{|p : p \in C_j \wedge \exists L_i \in \mathcal{L}, p \in L_i|}. \tag{4.7}$$

Therefore, the co-localization score for the set of predicted complexes $\mathcal{C}$ is just the weighted average over all complexes [64]:

$$Loc\ score(C) = \frac{\sum_{C_j \in \mathcal{C}} \max\{|C_j \cap L_i| : i = 1, 2, ..., k\}}{\sum_{C_j \in \mathcal{C}} |p : p \in C_j \wedge \exists L_i \in \mathcal{L}, p \in L_i|}. \tag{4.8}$$

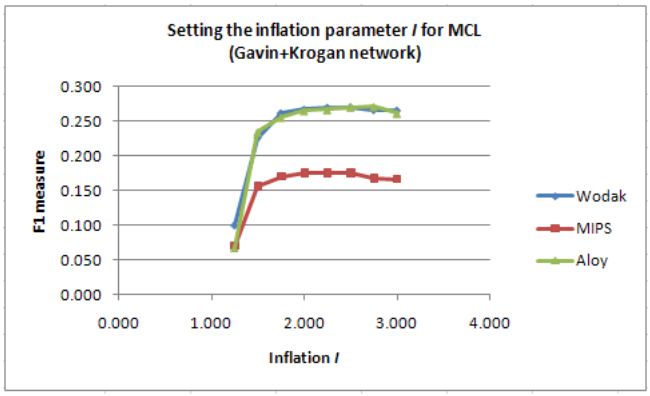
4.3.4 Setting the parameters in MCL-CAw: I, α and γ

Before evaluating the performance of MCL-CAw, we describe the procedure used for setting inflation parameter I for MCL, and α and γ for core-attachment refinement in order to determine a good combination of parameters for MCL-CAw in practice. Only the predicted complexes of size ≥ 4 from MCL and MCL-CAw were considered for setting the parameters as well as in further experiments. We used F1 (harmonic mean of precision and recall) measured against the MIPS [90], Wodak lab [92] and Aloy [93] benchmarks as our basis for choosing the best values for these parameters. Similar procedures based on benchmark complexes were adopted by Brohee and van Helden [86] and Vlasblom et al. [87] to set parameters in their methods.

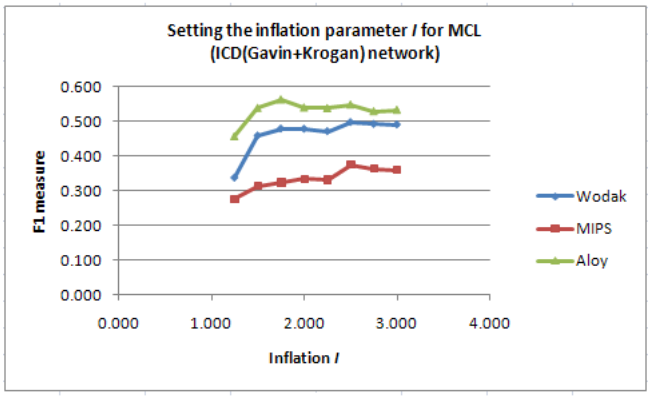
We adopted the following four-step procedure on each PPI network:

- Run MCL for a range of I values and choose the I that offers the best F1 measure;
- Set I to the chosen value, set a certain α for MCL-CAw, and choose γ from a range of values that offers the best F1 measure;

- Set I and γ to the chosen values, and choose α for MCL-CAw from a range of values that offers the best F1 measure;
- Set α and γ for MCL-CAw to the chosen values, and reconfirm the value chosen for I.



(a)



(b)

Figure 4.2: Setting the inflation I in MCL. We measured F1 against Wodak, MIPS and Aloy complexes for a range of $I = 1.25$ to 3.0. We noticed that $I = 2.5$ gave the best F1 for both unscored and scored G+K networks. This figure shows sample F1-versus-I curves for the (a) unscored G+K and (b) ICD(G+K) networks.

Setting I for MCL

Inflation I in MCL determines the granularity of the clustering - the higher the value more finer are the clusters produced. Typical values used for clustering PPI networks are $I = 1.8$ and 1.9 [62,64,86]. For each PPI network, we ran MCL over a range of I, and measured F1 against the three benchmark sets. We then calculated

normalized F1 values across all three benchmarks to obtain the I offering the best F1 measure. In Figure 4.2, we show sample F1 *versus* I plots for the unscored Gavin+Krogan and scored ICD(Gavin+Krogan) networks for the range of $I = 1.25$ to 3.0. We noticed that inflation $I = 2.5$ gave the best F1 on both unscored and scored networks. The F1 obtained at $I = 1.8$ and 1.9 was only marginally less than that at $I = 2.5$.

Setting α and γ for CA refinement

For each PPI network, we set I to the chosen value, fixed a certain α, and ran MCL-CAw over a range of γ. We adopted the same method as above to choose the value of γ offering the best F1 measure. Figure 4.3 shows sample F1 *versus* γ plots on the unscored Gavin+Krogan and scored ICD(Gavin+Krogan) networks for $I = 2.5$, $\alpha = 1.00$ and $\gamma = 0.15$ to 1.50. We noticed that $\gamma = 0.75$ gave the best F1 on both unscored and scored networks.

Next, we set I and γ to the chosen values, and ran MCL-CAw over a range of α. Figure 4.3 shows sample F1 *versus* α plots on the unscored Gavin+Krogan and scored ICD(Gavin+Krogan) networks for $I = 2.5$, $\gamma = 0.75$ and $\alpha = 0.50$ to 1.75. We noticed that $\alpha = 1.50$ gave the best F1 on the unscored network, while $\alpha = 1.0$ gave the best F1 on the scored networks.

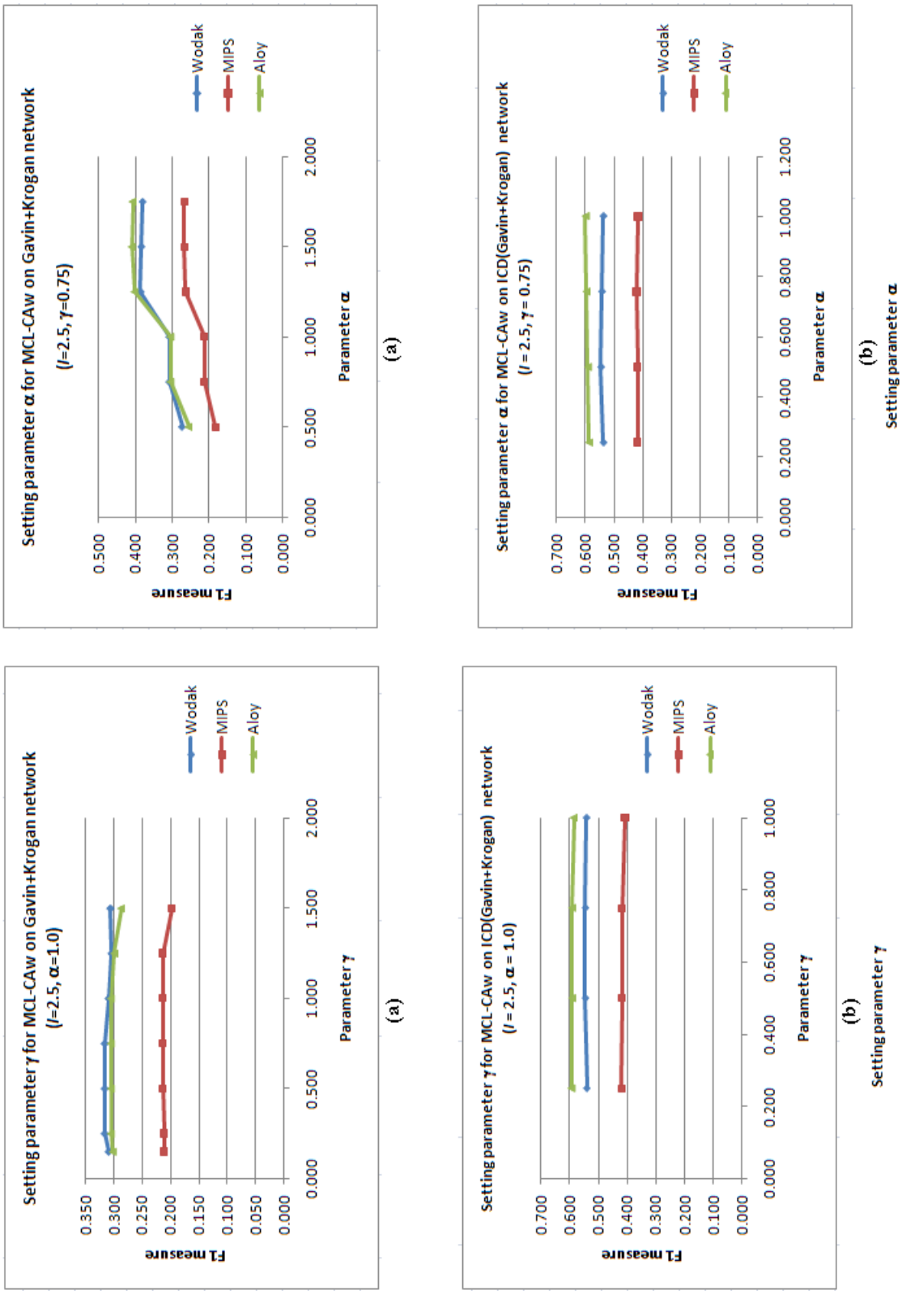


Figure 4.3: Setting parameter γ and α in MCL-CAw. We fixed $I = 2.5$ and varied γ and α over a range of values to obtain the best combination of γ and α that offered the maximum F1. These figures show F1-versus-α / γ plots for the G+K and ICD(G+K) networks. For the G+K network, $I = 2.5$, $\alpha = 1.50$ and $\gamma = 0.75$, and for ICD(G+K), $I = 2.5$, $\alpha = 1.00$ and $\gamma = 0.75$ gave the best F1 measures.

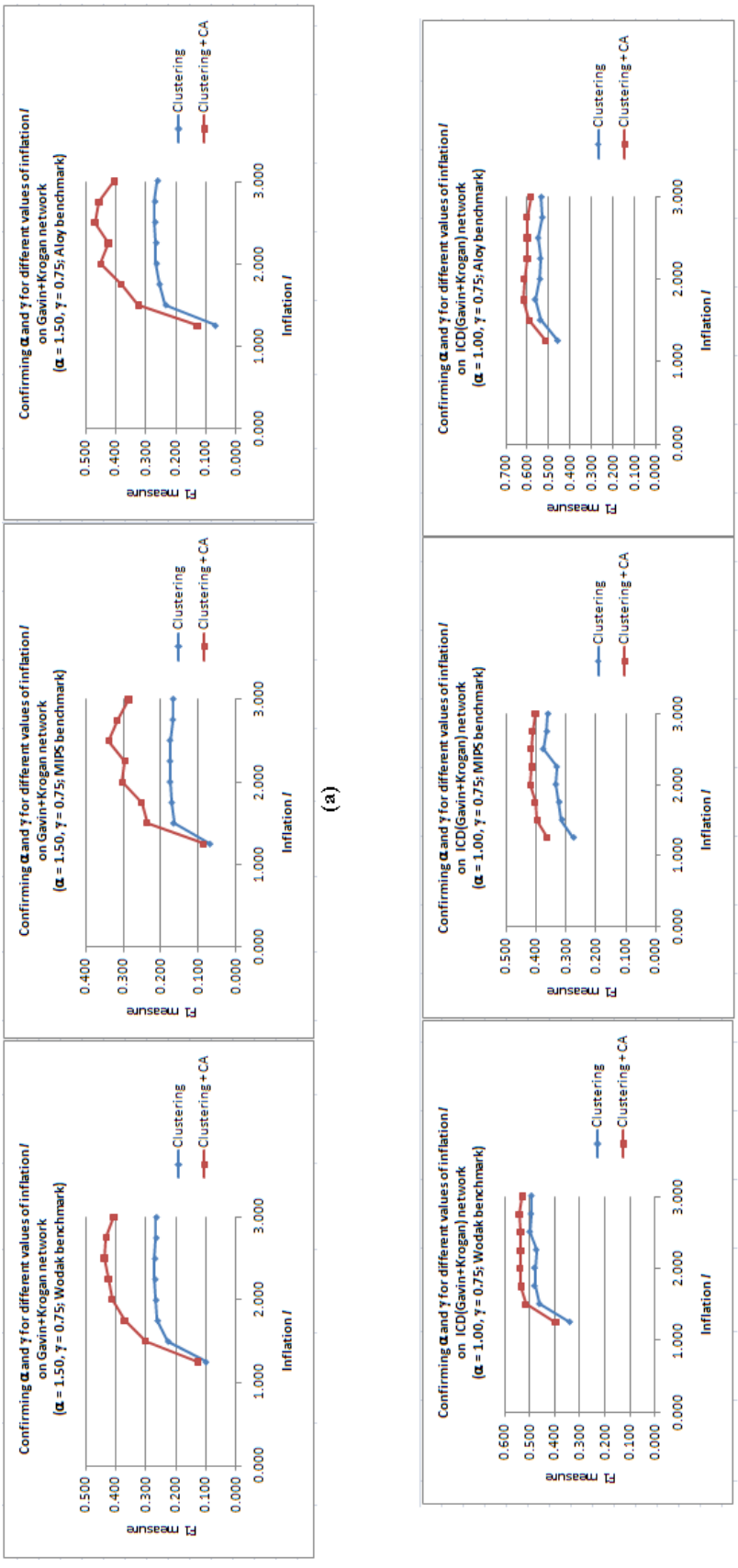


Figure 4.4: Reconfirming the chosen value of I for α and γ. We ran MCL and MCL followed by CA for the chosen α and γ values over a range of I = 1.25 to 3.00. This reconfirmed that $I = 2.5$ gave the best F1 measure. The figure shows these results for the G+K and ICD(G+K) networks.

Reconfirming I for the chosen values of α and γ

Finally, for each PPI network, we ran core-attachment refinement with the chosen values of α and γ over a range of I for MCL. Figure 4.4 compares the F1 *versus* I plots for plain-MCL and MCL followed by CA refinement on the unscored Gavin+Krogan and scored ICD(Gavin+Krogan) networks for range $I = 1.25$ to 3.0. The plots reconfirmed that the chosen values for α and γ gave the best performance for CA refinement when $I = 2.5$ (except for the Aloy benchmark, the smallest benchmark among the three, for which F1 was best at $I = 1.75$ and was marginally lower for $I = 2.5$). We settled on $I = 2.5$, $\alpha = 1.50$ and $\gamma = 0.75$ for the unscored Gavin+Krogan network, and $I = 2.5$, $\alpha = 1.0$ and $\gamma = 0.75$ for the scored networks as our final combination of parameters for MCL-CAw.

4.3.5 Evaluating the performance of MCL-CAw

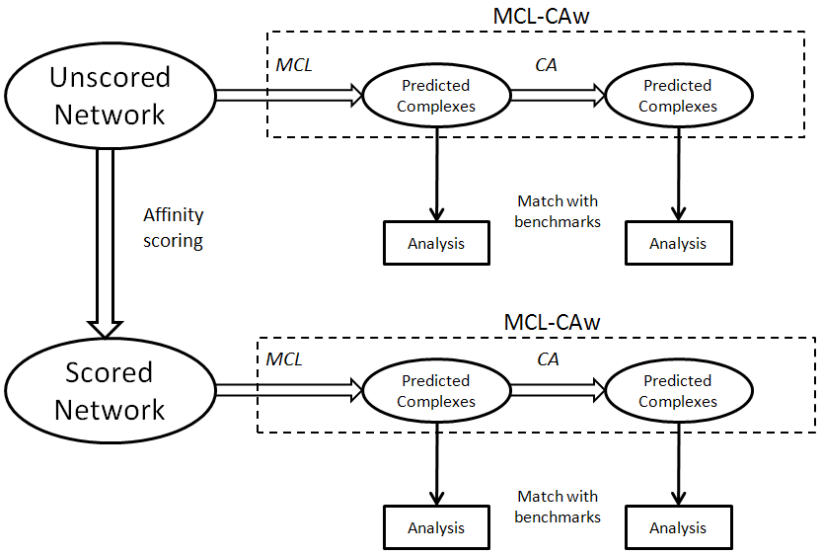


Figure 4.5: Workflow for the evaluation of MCL-CAw.

Figure 4.5 shows the *workflow* considered for the evaluation of MCL-CAw. The predicted complexes were tapped at two successive stages:

- After clustering using MCL;
- After core-attachment refinement using MCL-CAw.

The effect of filtering noisy clusters and segregrating large clusters

Table 4.4 shows the number of complexes predicted at each step of the MCL-CAw algorithm. The core-attachment refinement discarded all noisy clusters (those without any core proteins). We analysed each of these noisy clusters and found most to be the artifacts of MCL. These noisy clusters included proteins that had higher external interactions than internal, yet were grouped together arbitrarily. When we matched these clusters to benchmark complexes, we found that just 7 (out of 40) proteins belonged some real complex, indicating either that these proteins were arbrarily assigned to these noisy clusters though they deserved to belong to non-noisy clusters, or some of these clusters in fact represented real complexes but there was no sufficient topological information to adjudge that. Either way, our investigations suggested, if not rigorously proved, that our filtering procedure was "safe" - did not discard any valuable clusters.

PPI Network	#Clusters from MCL		#Clusters from MCL-CAw		
	Total	Size ≥ 25	After breaking down large clusters	Noisy clusters	After CA refinement
G+K	242	7	246	116	130
ICD(G+K)	136	10	181	16	165
FSW(G+K)	120	14	178	17	161
$\text{Cons}_{3.19}$	116	9	147	17	130
$\text{Boot}_{0.094}$	203	12	223	37	186

Table 4.4: Number of clusters produced at each stage of the MCL-CAw algorithm. Noisy clusters were the clusters without cores.

Next, we considered the clusters of size ≥ 25, and measured the number of clusters that correctly matched benchmark complexes before and after the hierarchical breakdown (segregation) (Table 4.5). Many small benchmark complexes were embedded within these large clusters, but could not be identified. But, when these large clusters were broken down into smaller clusters, some of the benchmark complexes were identified with higher accuracies. However, this process also created several redundant clusters that went on to marginally reduce the final precision.

PPI network	#Clusters		#Benchmarks derived by	
	Large (size ≥ 25)	After segregation	Large clusters (size ≥ 25)	Segregated clusters
G+K	7	11	1	4
ICD(G+K)	10	29	1	9
FSW(G+K)	14	32	3	9
$\text{Cons}_{3.19}$	9	15	3	7
$\text{Boot}_{0.094}$	12	20	4	9

Table 4.5: Impact of breaking down of large clusters (of size ≥ 25) into smaller clusters in MCL-CAw.

The effect of core-attachment refinement on the predictions of MCL

The *topmost rows* for MCL and MCL-CAw in Table 4.6 compares the two methods on the unscored Gavin+Krogan network. MCL-CAw achieved significantly higher recall compared to MCL on Gavin+Krogan - on an average 25.76% higher number of complexes derived than MCL.

In order to further analyse this improvement, we considered two sets of complexes derived from Gavin+Krogan. (a) Set $A = \{\text{MCL} \cap \text{MCL-CAw}\}$, consisting of all complexes derived by both MCL and MCL-CAw, but with different Jaccard accuracies; (b) Set $B = \{\text{MCL-CAw} \setminus \text{MCL}\}$, consisting of all complexes derived by MCL-CAw, but not by MCL. There was no complex derived by MCL that was missed by MCL-CAw. We calculated the increase in accuracies from MCL to MCL-CAw for complexes in A and B. This increase for A was noticably high, the average being 7.53% on the Wodak set. The increase for B was significantly high, the average being 62.26% on the Wodak set. This shows: (a) CA-refinement was successful in improving the accuracies of MCL clusters; (b) This improvement was particularly high for low quality clusters of MCL (that is, set B). MCL-CAw was successful in elevating the accuracies above the threshold $t = 0.50$ for those clusters that could not be matched to known complexes using MCL alone. Consequently, MCL-CAw derived significantly higher number of benchmark complexes than MCL.

Impact of affinity scoring on the performance of MCL and MCL-CAw

Table 4.6 compares different evaluation metrics for MCL and MCL-CAw on the unscored Gavin+Krogan with the four scored PPI networks. Very clearly, both MCL and MCL-CAw showed significant improvement in recall on the scored net-

works - MCL achieved 51.34%, while MCL-CAw achieved 38.53% higher recall on average on the Wodak benchmark from the four scored networks compared to the unscored network. MCL also showed significant improvement in precision on the scored networks. However, the precision for MCL-CAw dropped marginally for ICD(Gavin+Krogan) and FSW(Gavin+Krogan) networks, while for the Consolidated$_{3.19}$ and Bootstrap$_{0.094}$ networks, there was considerable improvement in precision.

		Evaluation on Wodak					
Method	PPI Network	#Predicted complexes	#Matched predictions	Precision	#Derivable benchmarks	#Derived benchmarks	Recall
MCL	G+K	242	55	0.226	182	62	0.338
	ICD(G+K)	136	68	0.500	153	76	0.497
	FSW(G+K)	120	69	0.575	153	78	0.510
	Cons$_{3.19}$	116	70	0.603	145	79	0.545
	Boot$_{0.094}$	203	76	0.374	172	85	0.494
MCL-CAw	G+K	130	69	0.531	182	75	0.412
	ICD(G+K)	165	76	0.461	153	84	0.549
	FSW(G+K)	161	72	0.447	153	84	0.549
	Cons$_{3.19}$	130	83	0.638	145	90	0.621
	Boot$_{0.094}$	186	93	0.500	172	97	0.564

		Evaluation on MIPS					
Method	PPI Network	#Predicted complexes	#Matched predictions	Precision	#Derivable benchmarks	#Derived benchmarks	Recall
MCL	G+K	242	35	0.143	177	40	0.226
	ICD(G+K)	136	47	0.346	151	60	0.397
	FSW(G+K)	120	46	0.383	151	61	0.404
	Cons$_{3.19}$	116	48	0.414	157	63	0.401
	Boot$_{0.094}$	203	44	0.271	168	56	0.333
MCL-CAw	G+K	130	42	0.323	177	53	0.300
	ICD(G+K)	165	49	0.297	151	67	0.444
	FSW(G+K)	161	47	0.292	151	66	0.437
	Cons$_{3.19}$	130	53	0.408	157	67	0.427
	Boot$_{0.094}$	186	53	0.285	168	62	0.369

Among the four scored networks, both MCL and MCL-CAw showed significantly high precision and recall on the Consolidated$_{3.19}$ network, directly attributable to the high quality of this network. However, this high quality of Consolidated$_{3.19}$ came at the expense of lower protein coverage (see Table 4.2; also noted in [31]), resulting in reduced number of derivable complexes (145 Wodak complexes). Therefore, we lowered the PE cut-off to 0.623 (the average PE score) to gather a larger subset of the Consolidated network, which accounted for a higher protein coverage (224 Wodak complexes) (see Table 4.7). We noticed the improvement of MCL-CAw over MCL was significantly higher on Consolidated$_{0.623}$ compared to that seen on

		Evaluation on Aloy					
Method	PPI Network	#Predicted complexes	#Matched predictions	Precision	#Derivable benchmarks	#Derived benchmarks	Recall
MCL	G+K	242	43	0.179	76	42	0.556
	ICD(G+K)	136	58	0.426	75	56	0.747
	FSW(G+K)	120	57	0.475	75	57	0.760
	$Cons_{3.19}$	116	54	0.466	76	55	0.724
	$Boot_{0.094}$	203	56	0.276	76	55	0.724
MCL-CAw	G+K	130	47	0.362	76	52	0.684
	ICD(G+K)	165	63	0.382	75	61	0.813
	FSW(G+K)	161	61	0.379	75	61	0.813
	$Cons_{3.19}$	130	57	0.438	76	55	0.724
	$Boot_{0.094}$	186	64	0.344	76	62	0.816

Table 4.6: (i) Impact of core-attachment refinement on MCL; (ii) Role of affinity-scoring in reducing the impact of natural noise on MCL and MCL-CAw.

$Consolidated_{3.19}$. We also noticed that ICD and FSW scoring of $Consolidated_{0.623}$ drastically reduced the size of this network, reconfirming that this larger subset included significant amount ($\sim 81\%$) of false positives (noise). These experiments indicate that any reasonably good algorithm like MCL can perform well on high quality networks like $Consolidated_{3.19}$. However, due to the lack of protein coverage as well as scarcity of such high quality networks, we need to consider larger networks for complex detection (particularly to be able to detect novel complexes). This in turn exposes the algorithms to higher amount of natural noise. Therefore, the need is to develop algorithms that can detect larger number of complexes in the presence of such noise. In this scenario, our results show that MCL-CAw is able to derive considerably higher number of complexes than MCL.

PPI Network	#Proteins	#Interactions	Avg node deg	#Derived complexes (Recall)	
				MCL	MCL-CAw
$Cons_{3.19}$	1622	9704	11.96	79 (0.545)	90 (0.621)
$Cons_{0.623}$	5423	102393	37.76	74 (0.330)	94 (0.419)
ICD($Cons_{3.19}$)	1161	8688	14.96	58 (0.408)	63 (0.443)
ICD($Cons_{0.623}$)	1273	19996	31.41	52 (0.353)	56 (0.381)
FSW($Cons_{3.19}$)	1123	8694	15.48	59 (0.401)	65 (0.442)
FSW($Cons_{0.623}$)	1341	20696	30.87	54 (0.360)	57 (0.380)

Table 4.7: The $Consolidated_{3.19}$ and $Consolidated_{0.623}$ networks were subsets of the Consolidated network [36] derived with PE cut-offs 3.19 and 0.623, respectively. We ran ICD and FSW schemes on these networks. $Consolidated_{0.623}$ had significant amount of false positives ($\sim 81\%$) that were discarded by the scoring. MCL-CAw performed considerably better than MCL on the "more noisy" $Consolidated_{0.623}$.

PPI Network	Co-localization scores MCL clusters	MCL-CAw cores	MCL-CAw complexes
G+K	0.730	0.890	0.866
ICD(G+K)	0.830	0.936	0.912
FSW(G+K)	0.830	0.931	0.912
$\text{Cons}_{3.19}$	0.790	0.923	0.908
$\text{Boot}_{0.094}$	0.788	0.895	0.874

Table 4.8: Co-localization scores of MCL-CAw complex components.

Biological coherence of predicted complex components

The co-localization scores for the various predicted components (cores and whole complexes) of MCL-CAw are shown in Table 4.8. The table shows that: (a) The predicted complexes of MCL-CAw showed high co-localization scores compared to MCL on both the unscored and scored PPI networks. MCL included several noisy proteins into the predicted clusters, thereby reducing their biological coherence; (b) The predicted cores of MCL-CAw displayed higher scores compared to complexes, indicating that proteins within cores were highly localized; (c) The complexes of both MCL and MCL-CAw displayed higher scores on the four scored networks compared to the Gavin+Krogan network, reaffirming the role of scoring.

An analysis of false positive predictions

MCL and MCL-CAw predicted on average 55% false positives from the four scored networks. About 15% of these false positive predictions matched some benchmark complexes with low accuracies (between 0.35 and 0.49) due to inclusion of a few noisy proteins or exclusion of a few complexed proteins from the predictions. Some instances of such "narrowly missed predictions" are discussed later. Among the remaining false positives, about 3% showed high ($\geq$ 0.80) coherence in terms of GO localization and function scores indicating that these might be novel putative complexes absent in the benchmark sets. One such example comprising of four proteins {Oca4, Oca5, Siw14, Oca1} is discussed later.

4.3.6 Comparisons with existing complex detection methods

In order to gauge the performance of MCL-CAw relative to some of the other existing techniques, we selected the following recent algorithms proposed for complex

detection:

- On the unscored Gavin+Krogan network, we compared against MCL [16, 62], MCL-CA - a preliminary and unweighted version of MCL-CAw (2009), CORE by Leung et al. (2009) [76], COACH by Wu Min et al. (2009) [75], CMC by Liu et al. (2009) [64], and HACO by Wang et al. (2009) [69];
- On the affinity scored networks, we compared against MCL, MCL incorporated with cluster overlaps by Pu et al. (2007) [31] (our implementation of this, called MCLO), CMC and HACO.

Property	Method						
	MCL	**MCL-CA**	**MCLO**	**CORE**	**COACH**	**CMC**	**HACO**
Principle	Flow simulation by random walks	Core-attach refinement over MCL	MCL with cluster overlaps	Core-attach by probablity and sampling	Core-attach by dense neighborhood	Maximal clique merging	Hier agglo clustering with overlaps
Scored networks	Yes	No	Yes	No	No	Yes	Yes
Unassigned proteins	No	Yes	No	Yes	Yes	Yes	Yes
Parameters (default)	Inflation I ($I = 2.5$)	Inflation I ($I = 2.5$)	Inflation I, Overlap a, b (2.5,1.0,0.5)	/	Filter t (t=0.225)	Merge m, Overlap t, Min clust size (0.5, 0.25, 4)	UPGMA cutoff (0.2)
References	van Dongen 2000 [16]	Preliminary version	Pu et al. 2007 [31]	Leung et al. 2009 [76]	Wu Min et al. 2009 [75]	Liu et al. 2009 [64]	Wang et al. 2009 [69]

Table 4.9: Methods selected for comparisons with MCL-CAw: CORE (2009), COACH (2009), MCL-CA (2009) were compared against MCL-CAw only on the unscored Gavin+Krogan network, while MCL (2000, 2002), MCLO (2007), CMC (2009) and HACO (2009) were evaluated also on the scored networks.

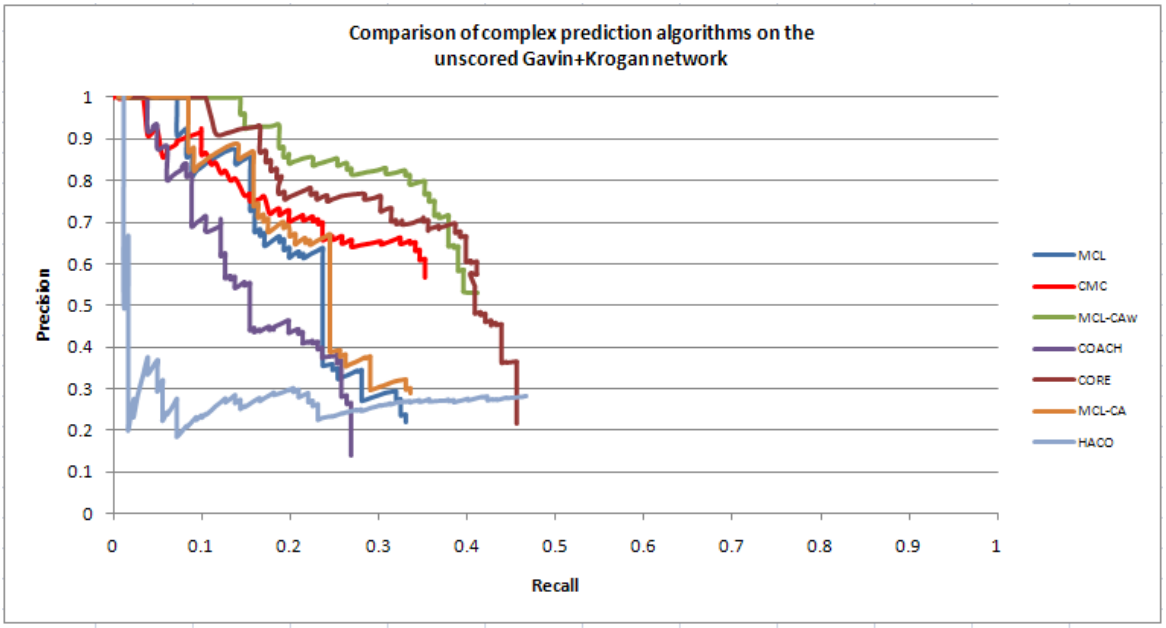
Table 4.9 summarizes some of the properties and the parameter values used in these methods. We considered only complexes of size at least 4 from all algorithms in this entire evaluation. We dropped MCL-CA, CORE and COACH for the comparisons on the affinity-scored networks because these methods assume unweighted networks as inputs. Further, we do not show results for older methods namely MCODE (2003) [14] and RNSC (2004) [77], instead include MCL into all our comparisons, because MCL has been shown to significantly outperform these methods [86–88].

Tables 4.10, 4.11, 4.12, 4.13 and 4.14 show detailed comparisons between complex detection algorithms on the unscored and scored networks. Figures 4.6, 4.7 and 4.8 substantiate these results with precision *versus* recall curves on these networks, while Table 4.15 shows the area under the curve (AUC) values for the curves. Considering ±5% error in AUC values, the table shows that CORE attained the highest AUC followed by MCL-CAw and CMC on the unscored network, while MCL-CAw and CMC achieved the overall highest AUC on the scored networks.

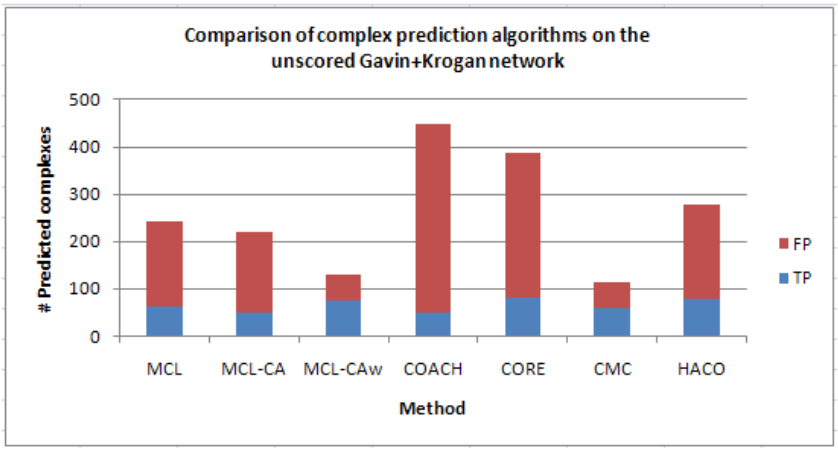
The unscored Gavin+Krogan network
#Proteins 2964; #Interactions 13507

		Method						
		MCL	MCL-CA	MCL-CAw	COACH	CORE	CMC	HACO
	#Predicted	242	219	130	447	386	113	278
Wodak (#182)	#Matched	55	49	69	62	83	60	78
	Precision	0.226	0.224	0.531	0.139	0.215	0.531	0.281
	#Derived	62	49	75	49	83	60	85
	Recall	0.338	0.269	0.412	0.269	0.456	0.330	0.467
MIPS (#177)	#Matched	35	42	42	45	59	41	45
	Precision	0.143	0.192	0.323	0.101	0.153	0.363	0.162
	#Derived	40	42	53	38	59	41	57
	Recall	0.226	0.237	0.300	0.215	0.333	0.232	0.322
Aloy (#76)	#Matched	43	41	47	54	59	43	59
	Precision	0.179	0.187	0.362	0.121	0.153	0.381	0.212
	#Derived	42	41	52	37	59	43	59
	Recall	0.556	0.539	0.684	0.487	0.776	0.566	0.776

Table 4.10: Comparisons between different methods on the unscored Gavin+Krogan network. CORE showed the best recall followed by HACO and MCL-CAw.



(a)



(b)

Figure 4.6: Comparison of different methods on the unscored Gavin+Krogan network: (a) Precision vs. recall curves using the Wodak benchmark; (b) Proportion of TP and FP complexes predicted from the methods.

The ICD(Gavin+Krogan) network
#Proteins 1628; #Interactions 8707

		Method				
		MCL	MCLO	MCL-CAw	CMC	HACO
	#Predicted	136	121	165	171	104
Wodak (#153)	#Matched	68	73	76	86	68
	Precision	0.500	0.603	0.461	0.503	0.654
	#Derived	76	73	84	86	76
	Recall	0.497	0.477	0.549	0.562	0.497
MIPS (#151)	#Matched	47	56	49	65	41
	Precision	0.346	0.463	0.297	0.380	0.394
	#Derived	60	56	67	65	55
	Recall	0.397	0.371	0.444	0.430	0.364
Aloy (#75)	#Matched	58	56	63	59	53
	Precision	0.426	0.463	0.382	0.345	0.510
	#Derived	56	56	61	59	53
	Recall	0.747	0.747	0.813	0.787	0.707

Table 4.11: Comparisons between the different methods on the ICD(Gavin+Krogan) network. CMC and MCL-CAw showed the best recall values.

The FSW(Gavin+Krogan) network
#Proteins 1628; #Interactions 8688

		Method				
		MCL	MCLO	MCL-CAw	CMC	HACO
	#Predicted	120	108	161	176	99
Wodak (#153)	#Matched	69	61	72	76	68
	Precision	0.575	0.564	0.447	0.432	0.687
	#Derived	78	72	84	84	77
	Recall	0.510	0.471	0.549	0.549	0.503
MIPS (#151)	#Matched	46	42	47	49	42
	Precision	0.383	0.388	0.292	0.278	0.424
	#Derived	61	55	66	65	56
	Recall	0.404	0.364	0.437	0.430	0.371
Aloy (#75)	#Matched	57	56	61	59	53
	Precision	0.475	0.518	0.379	0.335	0.535
	#Derived	57	56	61	57	53
	Recall	0.760	0.747	0.813	0.760	0.707

Table 4.12: Comparisons between the different methods on the FSW(Gavin+Krogan) network. MCL-CAw showed the best recall followed by CMC.

The Consolidated$_{3.19}$ network
#Proteins 1622; #Interactions 9704

		Method				
		MCL	MCLO	MCL-CAw	CMC	HACO
	#Predicted	116	119	130	77	101
Wodak (#145)	#Matched	70	80	83	67	57
	Precision	0.603	0.672	0.638	0.870	0.564
	#Derived	79	80	90	67	64
	Recall	0.545	0.552	0.621	0.462	0.441
MIPS (#157)	#Matched	48	65	53	56	40
	Precision	0.414	0.546	0.408	0.727	0.396
	#Derived	63	65	67	56	57
	Recall	0.401	0.414	0.427	0.357	0.363
Aloy (#76)	#Matched	54	56	57	45	44
	Precision	0.466	0.471	0.438	0.584	0.436
	#Derived	55	56	55	45	45
	Recall	0.724	0.737	0.724	0.592	0.592

Table 4.13: Comparisons between the different methods on the Consolidated$_{3.19}$ network. MCL-CAw showed the best recall followed by CMC.

The Bootstrap$_{0.094}$ network
#Proteins 2719; #Interactions 10290

		Method				
		MCL	MCLO	MCL-CAw	CMC	HACO
	#Predicted	203	204	186	203	127
Wodak (#172)	#Matched	76	76	93	110	80
	Precision	0.374	0.372	0.500	0.542	0.630
	#Derived	85	85	97	106	90
	Recall	0.494	0.494	0.564	0.616	0.523
MIPS (#168)	#Matched	44	45	53	67	49
	Precision	0.271	0.220	0.285	0.330	0.386
	#Derived	56	57	62	69	63
	Recall	0.333	0.339	0.369	0.411	0.375
Aloy (#76)	#Matched	56	55	64	76	59
	Precision	0.276	0.269	0.344	0.374	0.465
	#Derived	55	55	62	63	60
	Recall	0.724	0.723	0.816	0.829	0.789

Table 4.14: Comparisons between the different methods on the Bootstrap$_{0.094}$ network. CMC showed the best recall followed by MCL-CAw.

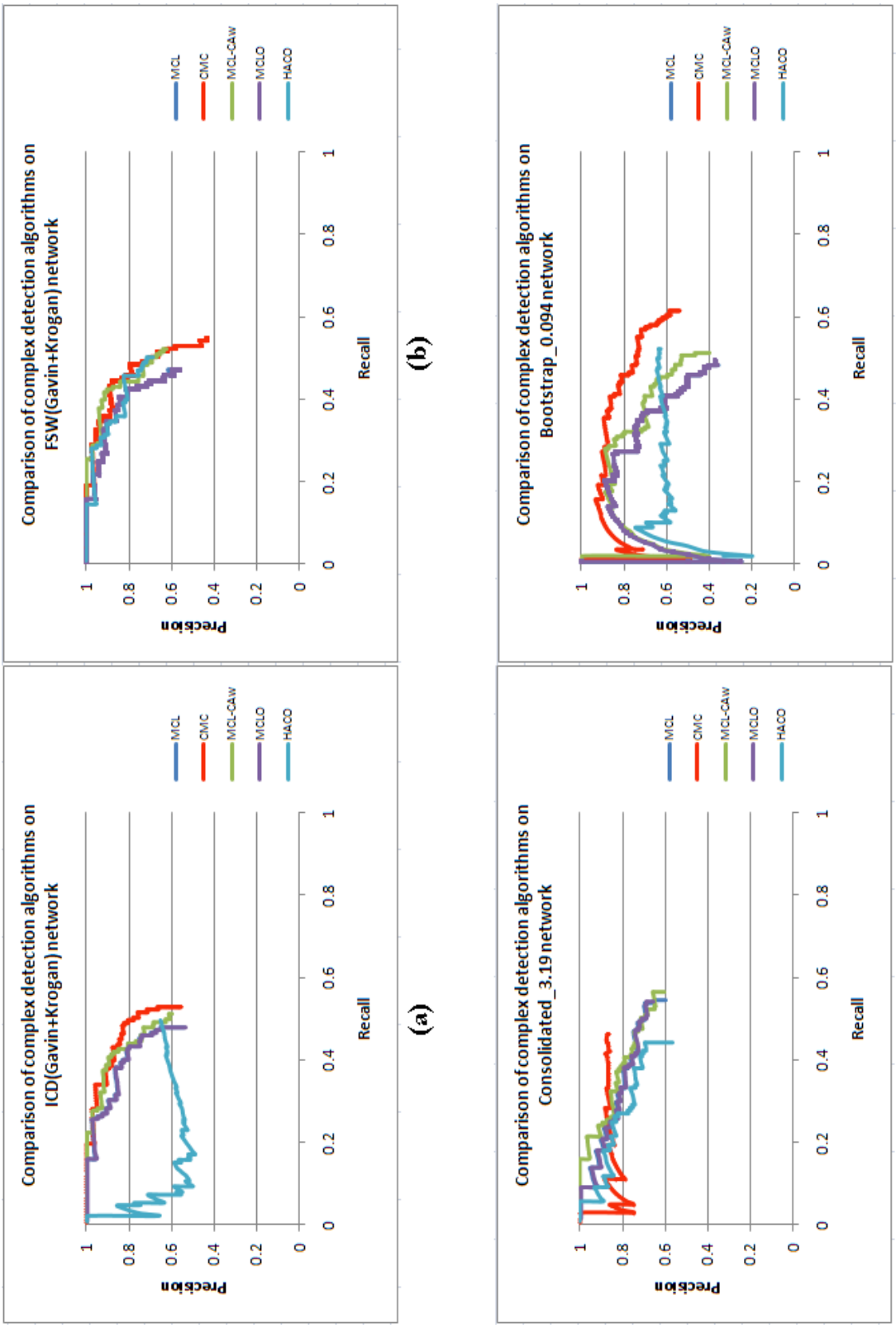


Figure 4.7: Comparative performance of complex detection algorithms on the four scored networks. The figures show the precision vs. recall curves for the Wodak benchmark set on (a) ICD(G+K), (b) FSW(G+K), (c) Consolidated$_{3.19}$ and (d) Bootstrap$_{0.094}$ networks. The curves for MCL-CAw have been drawn after "switching OFF" segregration of large clusters.

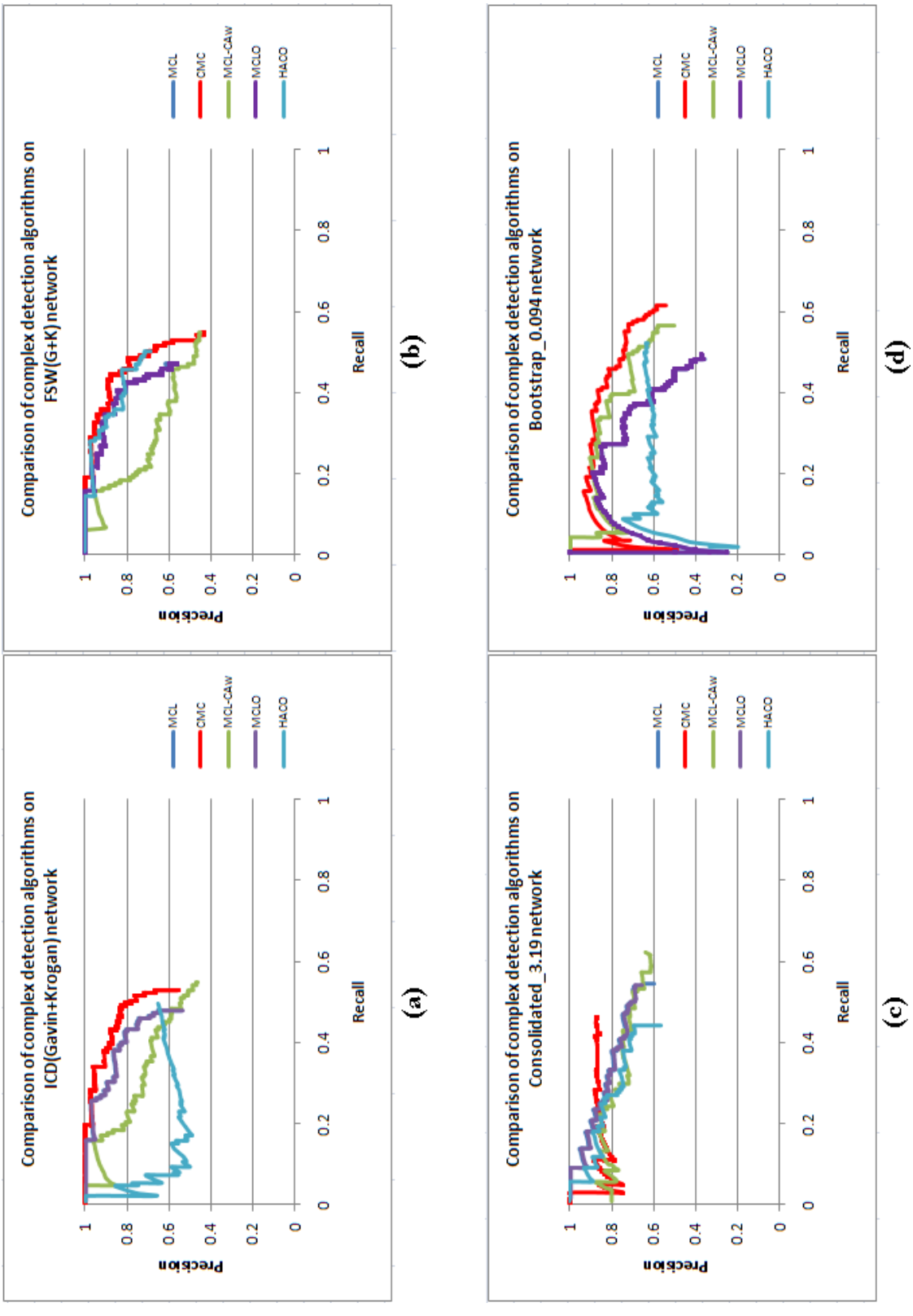


Figure 4.8: Comparative performance of complex detection algorithms on the four scored networks. The figures show the precision vs. recall curves for the Wodak benchmark set on (a) ICD(G+K), (b) FSW(G+K), (c) Consolidated$_{3.19}$ and (d) Bootstrap$_{0.094}$ networks. The curves for MCL-CAw have been drawn after "switching ON" segregation of large clusters. Segregation of large clusters reduces the precision of MCL-CAw, but improves the recall.

PPI network	MCL	MCLO	MCL-CAw	CMC	HACO	COACH	CORE
G+K	0.225	NA	0.323	0.271	0.136	0.169	**0.361**
ICD(G+K)	0.436	0.435	0.472	**0.494**	0.305		
FSW(G+K)	0.431	0.430	**0.487**	0.481	0.461	NA	NA
$\text{Consol}_{3.19}$	0.469	0.463	**0.488**	0.399	0.367		
$\text{Boot}_{0.094}$	0.349	0.353	0.389	**0.513**	0.317		

Table 4.15: Area under the curve (AUC) values of precision versus recall curves for complex detection methods on the unscored and scored PPI networks.

4.3.7 Ranking complex detection methods

The relative performance of the algorithms is not the same over all the networks, hence it is difficult to directly pick a clear winner. To offer a reasonable procedure for ranking, on each network we ranked the algorithms based on their normalized F1 values (with respect to the best performing algorithm on that network), as shown in Table 4.16. We then summed up the normalized F1 values for each algorithm across all the networks to obtain an overall ranking of the algorithms as shown in Table 4.17.

On the unscored network, CMC and HACO performed better than MCL-CAw in terms of F1. On the affinity-scored networks, the algorithms showed varied performance with MCL-CAw displaying the best *overall* performance in terms of F1. In particular, MCL-CAw performed the best on ICD(Gavin+Krogan), FSW(Gavin+Krogan) and $\text{Consolidated}_{3.19}$ networks, while HACO performed the best on $\text{Bootstrap}_{0.094}$ network. There was no single algorithm which performed relatively best on all the scored networks. Having said that, we note that MCL-CAw was always ranked among the top three on each of the scored networks indicating that MCL-CAw responded reasonably well to all the four scoring schemes used here. These results more or less agree with relative ranking obtained using the AUC curves (Table 4.15).

PPI network	Method	Wodak F1	Wodak Norm	MIPS F1	MIPS Norm	Aloy F1	Aloy Norm	Total	Norm
G+K	CMC	0.407	1.000	0.283	1.000	0.455	1.000	3.000	1.000
	HACO	0.351	0.862	0.216	0.761	0.333	0.731	2.355	0.785
	MCL-CAw	0.313	0.768	0.218	0.770	0.270	0.592	2.130	0.710
	CORE	0.292	0.718	0.210	0.741	0.256	0.561	2.020	0.673
	MCL	0.271	0.665	0.175	0.619	0.271	0.595	1.879	0.626
	MCL-CA	0.244	0.601	0.212	0.749	0.278	0.610	1.960	0.653
	COACH	0.183	0.450	0.137	0.486	0.194	0.426	1.361	0.454
ICD(G+K)	MCL-CAw	0.567	1.000	0.450	1.000	0.578	0.976	2.976	1.000
	HACO	0.565	0.995	0.378	0.841	0.593	1.000	2.837	0.953
	MCLO	0.533	0.939	0.412	0.916	0.572	0.965	2.820	0.947
	CMC	0.531	0.936	0.403	0.897	0.480	0.810	2.642	0.888
	MCL	0.498	0.879	0.370	0.822	0.543	0.916	2.616	0.879
FSW(G+K)	MCL-CAw	0.576	0.992	0.423	1.000	0.625	1.000	2.992	1.000
	HACO	0.581	1.000	0.396	0.935	0.609	0.974	2.910	0.972
	MCL	0.541	0.931	0.393	0.929	0.585	0.935	2.795	0.934
	MCLO	0.513	0.884	0.376	0.888	0.612	0.979	2.750	0.919
	CMC	0.484	0.833	0.338	0.798	0.465	0.744	2.375	0.794
$Cons_{3.19}$	MCL-CAw	0.614	1.000	0.487	1.000	0.576	0.979	2.979	1.000
	MCLO	0.606	0.986	0.471	0.967	0.575	0.977	2.930	0.984
	CMC	0.604	0.982	0.479	0.983	0.588	1.000	2.965	0.995
	MCL	0.573	0.932	0.407	0.836	0.567	0.964	2.732	0.917
	HACO	0.475	0.774	0.379	0.777	0.502	0.854	2.405	0.807
$Boot_{0.094}$	HACO	0.572	0.991	0.380	1.000	0.585	1.000	2.991	1.000
	CMC	0.577	1.000	0.367	0.965	0.515	0.881	2.846	0.952
	MCL-CAw	0.447	0.776	0.282	0.742	0.416	0.711	2.229	0.745
	MCL	0.426	0.738	0.299	0.785	0.400	0.683	2.207	0.738
	MCLO	0.424	0.736	0.267	0.701	0.392	0.670	2.108	0.705

Table 4.16: Relative ranking of complex detection algorithms based on F1 on each of the PPI networks. The normalized F1 values were obtained by normalizing the F1 values against the best.

Category	Method	Relative score	Normalized score
Unscored	CMC	3.000	1.000
	HACO	2.355	0.785
	MCL-CAw	2.130	0.710
	CORE	2.020	0.673
	MCL	1.879	0.626
	MCL-CA	1.960	0.653
	COACH	1.361	0.454
Scored	MCL-CAw	3.745	1.000
	HACO	3.733	0.997
	CMC	3.628	0.969
	MCLO	3.555	0.949
	MCL	3.468	0.926

Table 4.17: Overall ranking of the complex detection algorithms based on F1 for the unscored and scored categories of networks.

Taking this further, we ranked the affinity scored networks based on the performance offered to the complex detection algorithms, as shown in Table 4.18. We used the same ranking methodology as above - using normalized F1 scores to rank the networks. The Table 4.19 shows that the Consolidated$_{3.19}$ network offered the best performance to the algorithms, followed by the ICD(Gavin+Krogan), FSW(Gavin+Krogan) and Bootstrap$_{0.094}$ networks.

PPI network	Method	Wodak F1	Wodak Norm	MIPS F1	MIPS Norm	Aloy F1	Aloy Norm	Total	Norm
MCL	Cons$_{3.19}$	0.573	1.000	0.407	1.000	0.567	0.970	2.970	1.000
	FSW(G+K)	0.541	0.944	0.393	0.965	0.585	1.000	2.909	0.980
	ICD(G+K)	0.498	0.871	0.370	0.908	0.543	0.928	2.706	0.911
	Boot$_{0.094}$	0.426	0.744	0.299	0.733	0.400	0.684	2.161	0.728
MCLO	Cons$_{3.19}$	0.606	1.000	0.471	1.000	0.575	0.939	2.939	1.000
	ICD(G+K)	0.533	0.879	0.412	0.875	0.572	0.934	2.688	0.914
	FSW(G+K)	0.513	0.847	0.376	0.798	0.612	1.000	2.645	0.900
	Boot$_{0.094}$	0.424	0.700	0.267	0.567	0.392	0.641	1.908	0.649
MCL-CAw	Cons$_{3.19}$	0.629	1.000	0.417	1.000	0.546	1.000	3.000	1.000
	ICD(G+K)	0.506	0.805	0.365	0.875	0.535	0.981	2.660	0.887
	FSW(G+K)	0.485	0.770	0.348	0.834	0.514	0.942	2.546	0.849
	Boot$_{0.094}$	0.530	0.842	0.322	0.711	0.484	0.887	2.500	0.833
CMC	Cons$_{3.19}$	0.604	1.000	0.479	1.000	0.588	1.000	3.000	1.000
	Boot$_{0.094}$	0.577	0.955	0.366	0.764	0.515	0.877	2.597	0.866
	ICD(G+K)	0.531	0.880	0.403	0.843	0.480	0.816	2.538	0.846
	FSW(G+K)	0.484	0.801	0.338	0.705	0.465	0.791	2.297	0.766
HACO	FSW(G+K)	0.581	1.000	0.396	1.000	0.609	1.000	3.000	1.000
	Boot$_{0.094}$	0.572	0.984	0.380	0.961	0.585	0.961	2.906	0.969
	ICD(G+K)	0.565	0.972	0.378	0.956	0.593	0.973	2.902	0.967
	Cons$_{3.19}$	0.495	0.852	0.379	0.957	0.502	0.824	2.634	0.878

Table 4.18: Relative ranking of affinity scored networks for each complex detection algorithm based on F1 measures. The normalized F1 scores were obtained by normalizing the F1 measures against the best.

Scored network	Relative score	Normalized score
Cons$_{3.19}$	4.878	1.000
ICD(G+K)	4.526	0.928
FSW(G+K)	4.494	0.921
Boot$_{0.094}$	4.044	0.829

Table 4.19: Overall ranking of affinity scored networks for complex detection based on F1 measures.

4.3.8 In-depth analysis of predicted complexes

To facilitate the analysis of our individual predicted complexes, we mapped the complexes back to the PPI networks and examined the interactions between components of the same complex, as well as between components of a given complex and other proteins in the network. We visualized these using Cytoscape

(`http://www.cytoscape.org/`) [97].

Instances of correctly predicted complexes of MCL-CAw

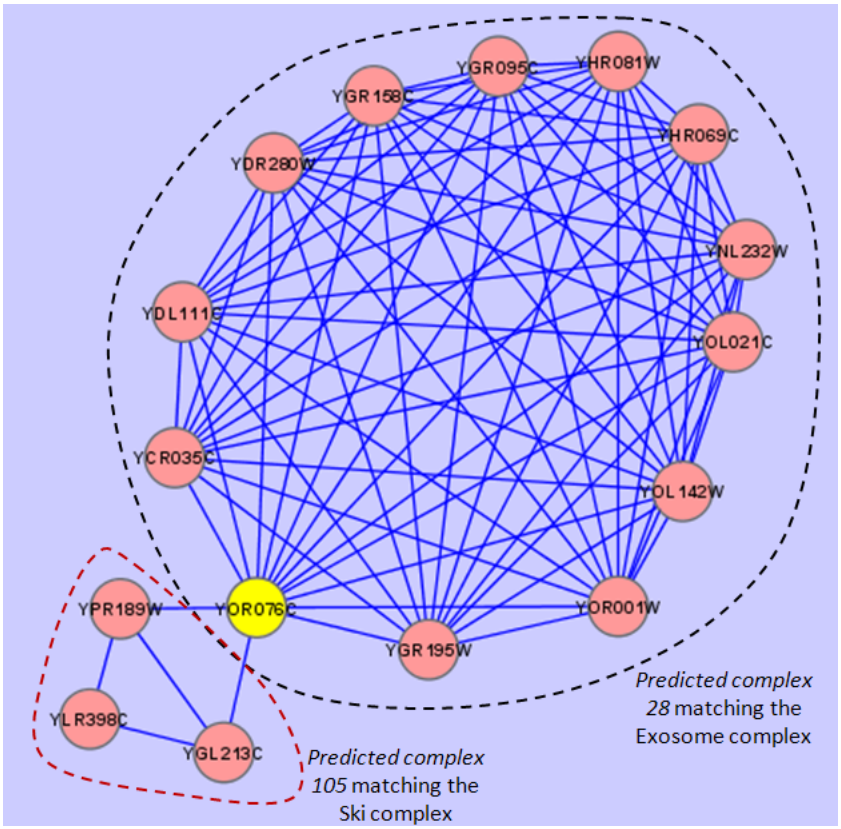


Figure 4.9: Ski7 (Yor076c) predicted as part of two complexes, the exosome and Ski complexes, in agreement with available evidence [102].

The first example is of an attachment protein shared between two predicted complexes of MCL-CAw. The subunits of these predicted complexes made up the Compass complex, involved in telomeric silencing of gene expression [98], and the mRNA cleavage and polyadenylation specificity factor, involved in RNAP II transcription termination [99]. The shared attachment Swd2 (Ykl018w) formed high-confidence connections with the subunits of both predicted complexes. On this basis, the post-processing procedure assigned Swd2 (Ykl018w) to both predicted complexes, in agreement with available evidence [100] that Swd2 (Ykl018w) belongs to both Compass and mRNA cleavage complexes.

The next example illustrates the case where a new protein was predicted as a subunit of a known complex. The attachment protein Ski7 (Yor076c) was included

into a predicted complex that matched the Exosome complex involved in RNA processing and degradation [101]. Additionally, Ski7 (Yor076c) was also included into a prediction matching the Ski complex (see Figure 4.9). However, the Ski complex in the Wodak lab catalogue [92] did not include this new protein. A literature search suggested that Ski7 acts as a mediator between the Ski and Exosome complexes for 3'-to-5' mRNA decay in yeast [102].

The RNA polymerase I, II, and III complexes (also called Pol I, II, and III, respectively) are required for the generation of RNA chains [103]. As per Wodak lab [92], all the three complexes share subunits: Yor224c, Ybr154c, Yor210w and Ypr187w, while Pol I and Pol III share Ynl113w and Ypr110c. Due to the extensive sharing of subunits, the corresponding predictions were grouped together into one large cluster by MCL. On the other hand, MCL-CAw was successful in segregating the large cluster into three independent clusters that matched the individual complexes (Pol I - J =0.714, Pol II - J =0.732 and Pol III - J =0.824).

In addition to these cases, a good fraction of already known core-attachment structures (reported in the supplementary materials of Gavin et al. [15]) were confirmed; some examples are worth quoting here. A predicted complex covering the HOPS complex had all five cores {Ylr148w, Ylr396c, Ymr231w, Ypl045w, Yal002w} and two attachments {Ydr080w, Ydl077c} matching those reported in Gavin et al. Experiments show that the cores have the function of vacuole protein sorting, and with the help of attachments, the complex can perform homotypic vacuole fusion [104]. Next, we identified the ubiquitin ligase ERAD-L complex comprising of subunits {Yos9(Ydr057w), Hrd3 (Ylr207w), Usa1 (Yml029w), Hrd1 (Yol013c)} that is involved in the degradation of ER proteins [105]. This matched the Hrd1/Hrd3 (complex m11) purified in Gavin et al.

A novel complex: Finally, four subunits {Oca4, Oca5, Siw14, Oca1} of a predicted novel complex showed high similarity in functions (oxidant-induced cell-cycle arrest) and localization (cytoplasmic) when verified in SGD [94]. This complex exactly matched the putative complex 490 reported in Gavin et al. [15].

Instances depicting mistakes in the predictions of MCL-CAw

Here we discuss an interesting case in which the sharing of subunits was so extensive and the web of interactions was so dense that separating out the smaller subsumed complexes purely on the basis of the interaction information was much harder. It was the amalgamation of the clusters matching the SAGA, SAGA-like (SLIK), ADA and TFIID complexes. Based on the Wodak lab catalogue [92], the 20 subunits making up the SAGA complex involved in transcriptional regulation [106] include four subunits (Ygr252w, Ydr176w, Ydr448w, Ypl254w) that are members of the ADA complex [107] as well. Sixteen components of the SAGA complex including the four shared with the ADA complex, are also the components of the SLIK complex [108]. Additionally, five subunits (Ybr198c, Ygl112c, Ymr236w, Ydr167w, Ydr145w) of the SAGA complex also belong to the TFIID complex [106]. Because of such extensive sharing of subunits involved in a dense web of interactions (436 interactions among 31 constituent proteins, as seen on the ICD(Gavin+Krogan) network), MCL-CAw was able to segregrate out only two distinct complexes - SAGA (accuracy - 0.708) and SLIK (accuracy - 0.625). The clusters matching TFIID and ADA remained amalgamated together leading to low accuracies (TFIID - 0.370 and ADA - 0.430).

Matched benchmark complex		#Incorrect proteins in predictions from G+K		ICD(G+K)		Accuracy J	
Name	#Proteins	Missed	Addl	Missed	Addnl	G+K	ICD(G+K)
Kornbergs SRB	25	1	0	2	0	0.960	0.920
SWI/SNF	12	3	0	4	0	0.769	0.667
TRAPP	10	0	0	1	0	1.000	0.900
19/22S reg	22	0	4	0	5	0.909	0.815
TRAMP	3	0	1	0	4	0.750	0.429
Alpha-1,6	5	0	4	0	6	0.556	0.455
eIF3	7	2	3	1	8	0.500	0.400
Protein phosp	3	0	2	0	4	0.600	0.333
Cdc73p/Paf1p	7	1	3	0	11	0.556	0.388
Chs5p/Arf-1	6	2	0	2	6	0.556	0.400

Table 4.20: Complexes derived with lesser accuracy or missed by MCL-CAw due to affinity scoring. The upper half shows sample complexes from Wodak lab derived with lower accuracies from the ICD(Gavin+Krogan) network compared to those from the Gavin+Krogan network. The lower half shows those missed from the ICD(Gavin+Krogan) network.

Instances of complexes missed by MCL-CAw due to affinity scoring: In the next set of analysis, we compared the derived complexes from the Gavin+Krogan and the ICD(Gavin+Krogan) networks, and identified cases where MCL-CAw had missed

a few proteins or whole complexes due to affinity scoring. From the Wodak, MIPS and Aloy reference sets, there were 13, 18 and 16 complexes, respectively, that were derived with better accuracies from the Gavin+Krogan network than from the ICD(Gavin+Krogan) network. And, there were 6, 2 and 2 complexes, respectively, that were derived from the Gavin+Krogan network, but missed totally from the ICD(Gavin+Krogan) network. Table 4.20 shows a sample of such complexes from the Wodak reference set. For the complexes that were derived with lower accuracies (upper half of Table 4.20), MCL-CAw had missed a few proteins due to low scores assigned to the corresponding interactions. For example, in the predicted complex from the ICD(Gavin+Krogan) network matching the SWI/SNF complex, two proteins (Ymr033w and Ypr034w) out of the four missed ones were absent due to their weak connections with the rest of the members; instead, these proteins were present in the prediction matching the RSC complex. In the Gavin+Krogan network, these two proteins were shared between two complexes matching the SWI/SNF and RSC complexes, which also agreed with the Wodak catalogue [92].

In the cases where MCL-CAw had completely missed some complexes from the scored network (lower half of Table 4.20), it is interesting to note that MCL-CAw had pulled-in many additional (noisy) proteins as attachments into the predicted complexes, which caused the accuracies to drop below 0.5. One such case is of the predicted complex matching the eIF3 complex with a low Jaccard score of 0.4. The eIF3 complex from Wodak lab consisted of 7 proteins: Yor361c, Ylr192c, Ybr079c, Ymr309c, Ydr429c, Ymr012w and Ymr146c. The corresponding complex predicted from the Gavin+Krogan network consisted of 8 proteins (Figure 4.10): 5 cores (Yor361c, Ylr192c, Ybr079c, Ymr309c, Ydr429c) and 3 attachments (Yor096w, Yal035w, Ydr091c). Therefore, there were 2 missed and 3 additional proteins in the prediction, leading to an accuracy of 0.5. The corresponding complex predicted from the ICD(Gavin+Krogan) network consisted of 14 proteins: 6 cores (Yor361c, Ylr192c, Ybr079c, Ymr309c, Ydr429c, Yor096w) and 8 attachments (Yal035w, Ydr091c, Yjl190c, Yml063w, Ymr146c, Ynl244c, Yor204w, Ypr041w). Therefore, there were 1 missed and 8 additional proteins in the prediction, leading to an even lower accuracy of 0.4. All the core proteins had same or similar GO annotations (involvement in translation, localized in cytoplasm or ribosomal subunit) [37]. Upon

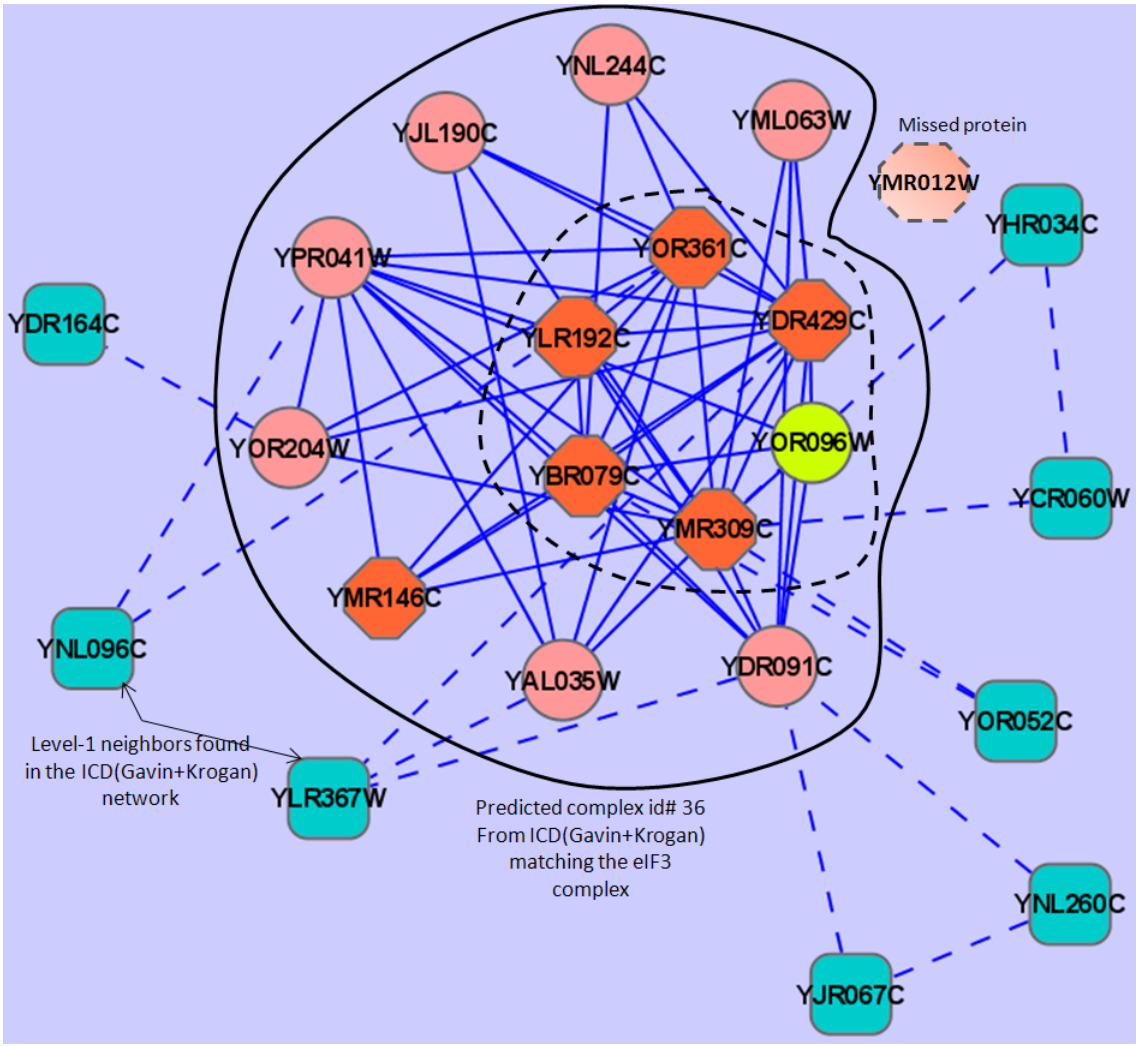


Figure 4.10: Example of a complex missed by MCL-CAw from the ICD(Gavin+Krogan) network, but found from the Gavin+Krogan network. The eIF3 complex from Wodak lab consisted of 7 proteins: Yor361c, Ylr192c, Ybr079c, Ymr309c, Ydr429c, Ymr012w and Ymr146c. The predicted complex id#36 from the ICD(Gavin+Krogan) network consisted of 14 proteins: 6 cores (Yor361c, Ylr192c, Ybr079c, Ymr309c, Ydr429c, Yor096w) and 8 attachments (Yal035w, Ydr091c, Yjl190c, Yml063w, Ymr146c, Ynl244c, Yor204w, Ypr041w). Therefore, there were 1 missed and 8 additional proteins in the prediction, leading to a low accuracy of 0.4. Orange: eIF3 from Wodak lab; Orange, Yellow and Pink: predicted complex; Turquoise: Level-1 neighbors.

analysing the GO annotations of the 8 attachment proteins, we noticed that only one (Ymr146c) had the *same* annotation as the core proteins. This was also part of the eIF3 complex from Wodak lab [92]. Out of the remaining 7 attachment proteins, five (Ypr041w, Ynl244c, Yml063w, Yjl190c, Ydr091c) had *similar or related* GO annotations (translation initiation, GTPase activity, cytoplasmic, ribosomal subunit) as the core proteins. A literature search revealed that these proteins belonged to the multi-eIF initiation factor conglomerate (containing eIF1, eIF2, eIF3 and eIF5) and the 40S ribosomal subunit involved in translation [109]. The remaining two (Yal035w, Yor204w) were involved in translation activity, but were absent in the Wodak lab catalogue. These might be potentially new proteins belonging to the eIF3 or related complexes, and need to be further investigated. We also analysed the GO annotations of the level-1 neighbors to the predicted complex seen in the network, none of them had annotations similar to the proteins within the network.

Instances of narrowly-missed complexes by MCL-CAw: We analysed the predicted complexes of MCL-CAw that matched benchmark complexes with accuracies between 0.35 and 0.50. This analysis revealed that most of these predictions in fact included several additional proteins instead of missing a few, thereby lowering the accuracies. Further investigation revealed that these were amalgamated clusters that were not successfully segregated by MCL-CAw, and therefore embedded multiple complexes within them. For example, the Swr1p (#proteins: 13) and Ino80p (#proteins: 12) from Wodak lab catalogue [92] share four proteins: Ydr190c, Yfl039c, Yjl081c and Ypl235w. From the $\text{Consolidated}_{3.19}$ network, MCL-CAw generated a large cluster (#proteins: 19) containing the "internal" proteins of these two complexes and these four shared proteins. This large cluster matched the two real complexes with low accuracies of 0.455 and 0.50, respectively. Upon analysis we found that these four shared proteins interacted densely with the "internal" proteins of these two complexes, leading to the amalgamation. Separating the cluster using only topological information was difficult.

4.4 Lessons from MCL-CAw

Harkening back to our "bin-and-stack" chronology-based classification introduced in Chapter 3, we position MCL-CAw into it, as shown in Figure 4.11. Doing so reconfirms that incorporating core-attachment structure followed by affinity scoring has indeed improved complex detection performance.

Though we have moved a step forward in improving the performance, a glance through Tables 4.10 to 4.14 reveals that all the methods considered for comparison in this work achieve very low recall on the MIPS reference set compared to the Wodak and Aloy sets. Table 4.3 shows that the average density of complexes in MIPS is much lower than that of Wodak and Aloy sets. Only 52 out of 137 (37.95%) derivable MIPS complexes of size ≥ 5 could be detected from the Gavin+Krogan network by all methods put together. We analysed the remaining 85 MIPS complexes and found most of them to have very low densities (average about 0.217) in the Gavin+Krogan network. For example, the MIPS complex 440.30.10 (involved in mRNA splicing) went undetected by all the methods even though 40 of its 42 proteins were present in Gavin+Krogan. There were 144 interactions among these 40 proteins, giving a low density of 0.184 to the complex in this network. This shows that complex detection methods generally do not perform well when the embedded complexes are of low densities. Apart from this limitation, we already saw that existing methods tend to amalgamate smaller complexes into larger modules causing them to be missed. These limitations are also seen in MCL-CAw; we list them as follows in decreasing order of seriousness:

1. Missing complexes of low densities;
2. Amalgamation of densely-interacting complexes;
3. Missing of small complexes (size ≤ 3).

The focus of the next two chapters will be to overcome some of these limitations to further improve complex detection performance.

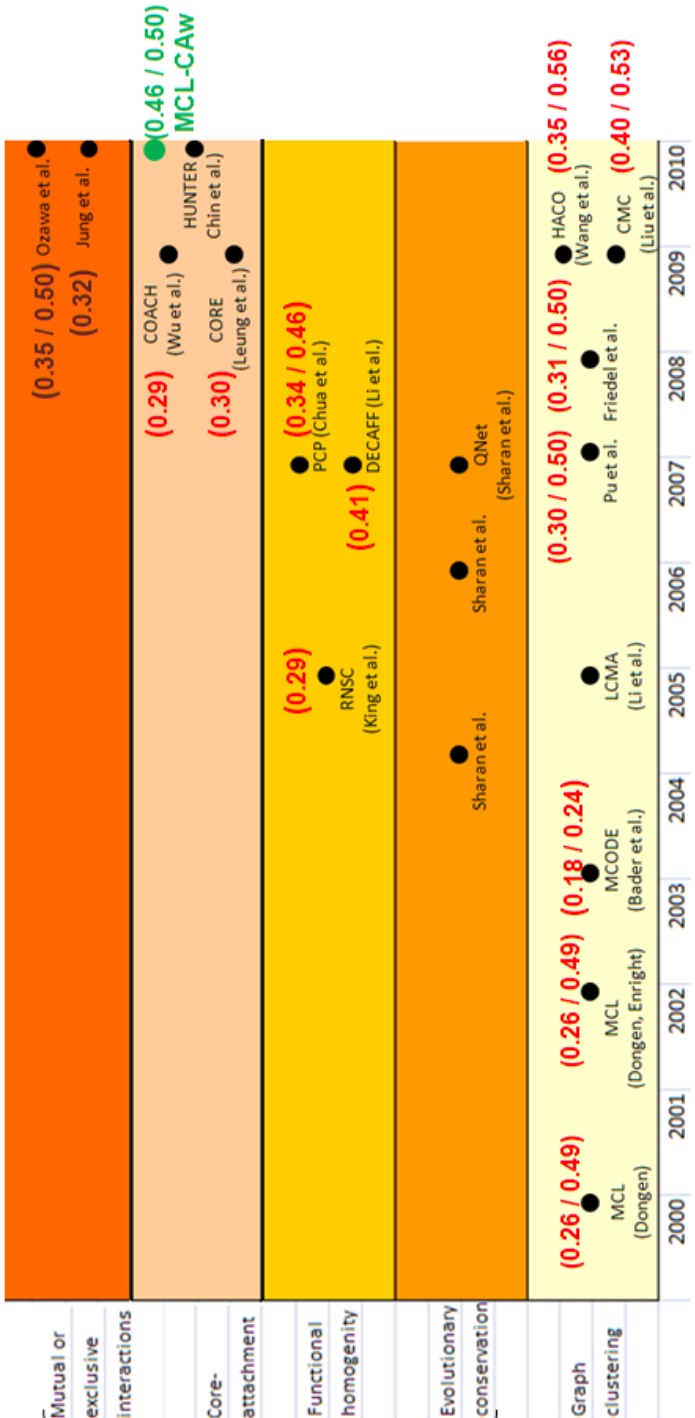


Figure 4.11: Positioning MCL-CAw into the "Bin-and-Stack" classification (all data points with respect to the Gavin + Krogan network scored using Purification Enrichment [36]). Incorporating core-attachment structure followed by affinity scoring has helped to improve performance.

CHAPTER 5

Characterization and detection of sparse complexes

Euclid taught me that without assumptions there is no proof. Therefore, in any argument, examine the assumptions.

- *Eric Temple Bell*, as quoted in [110]

In the previous chapter, we designed and developed MCL-CAw, a method for complex detection by incorporating core-attachment structure into MCL. Our detailed evaluation of MCL-CAw showed that MCL-CAw performed better or at least as good as recent methods, and also showed consistent performance across multiple scoring schemes. At the same time this evaluation also revealed many crucial limitations in complex detection methods. In particular, we noticed that all methods failed to detect many known complexes, especially those that had low densities in the networks. For example, MCL missed 65 out of the 123 MIPS complexes present in the Consolidated$_{3.19}$ network from Collins et al. [36]. Even the "union" of four methods, MCL, MCL-CAw, CMC and HACO, missed 52 out of the 123 complexes. Since the goal in this thesis is to study genome-wide compositions of complexes (the 'complexosome'), failure to detect even the known complexes reflects severe limitations in current methods.

5.1 Insights into the topologies of undetected complexes

In order to understand the characteristics of these missed complexes, we "superimposed" yeast complexes taken from MIPS [90] onto the high-confidence $\text{Consolidated}_{3.19}$ yeast PPI network [36] (#proteins: 1622, #interactions: 9704, average node degree: 11.187). This "superimposition" involves identifying the proteins of a benchmark complex in the PPI network, and extracting out the subnetwork induced by those proteins. Figure 5.1 shows this "superimposition" visualized using *Cytoscape* [97].

The immediate observation, which is of course typical to most PPI networks, was that the network comprised of one main large component and multiple *disjoint* smaller components of sizes 2 to 50. Out of the 123 MIPS complexes containing at least four proteins in the network, 89 were completely embedded in the main component, and the remaining 34 were "scattered" among more than one components. When we ran MCL on this network, it was able to recover only 58 of these 123 complexes. Of the 65 undetected complexes, 27 complexes were the ones that were "scattered", and 34 complexes, though intact, had very low interaction densities (< 0.50) in the network. In fact, some of these complexes lacked internal connectivities to an extent that it was impossible for *any* method to assemble back these disconnected pieces into whole complexes solely based on topological information. For example, the MIPS complex 510.190.110 (CCR4 complex) had seven proteins in the network scattered among four disjoint components (shown within ellipses in Figure 5.1). This complex remained disconnected with a low density of 0.1905, and naturally went undetected by all the methods.

Further, most MIPS complexes being small (sizes $\leq$ 10-15), lacking in just a few proteins or interactions easily rendered many complexes disconnected or with low interaction densities, resulting in them going undetected (see Figure 5.2). All these findings revealed that a potentially strong correlation existed between the "network constitution" of a complex (the number of member proteins in the network and their connectivities) and the possibility of it being detected using existing methods.

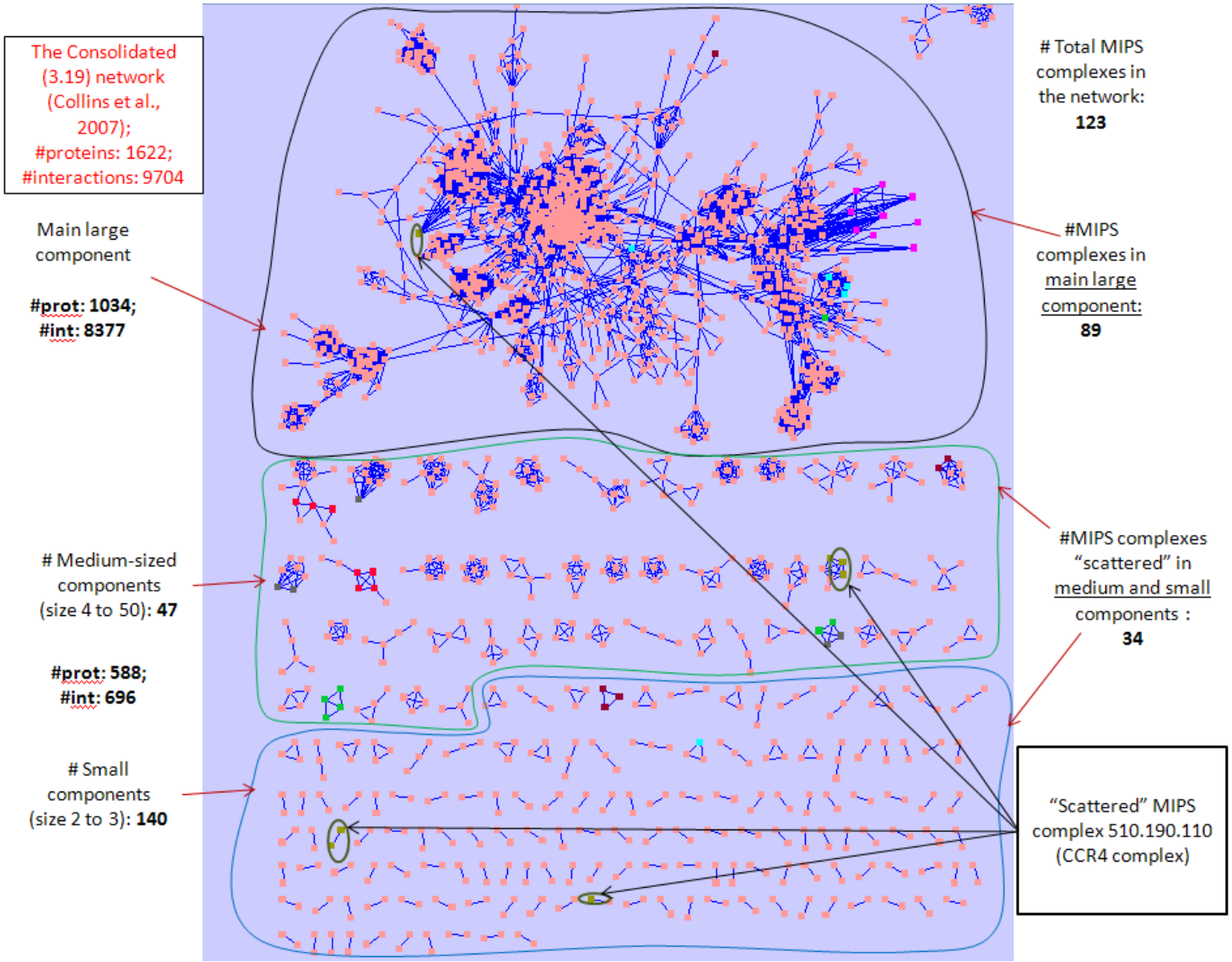


Figure 5.1: The figure shows the "superimposition" of MIPS complexes onto the Consolidated yeast network visualized using *Cytoscape*. The MIPS complex 510.190.110 (CCR4 complex) had seven proteins (marked within ellipses) that were "scattered" among four disjoint components resulting in a low density of 0.1905. This complex went undetected by the considered methods.

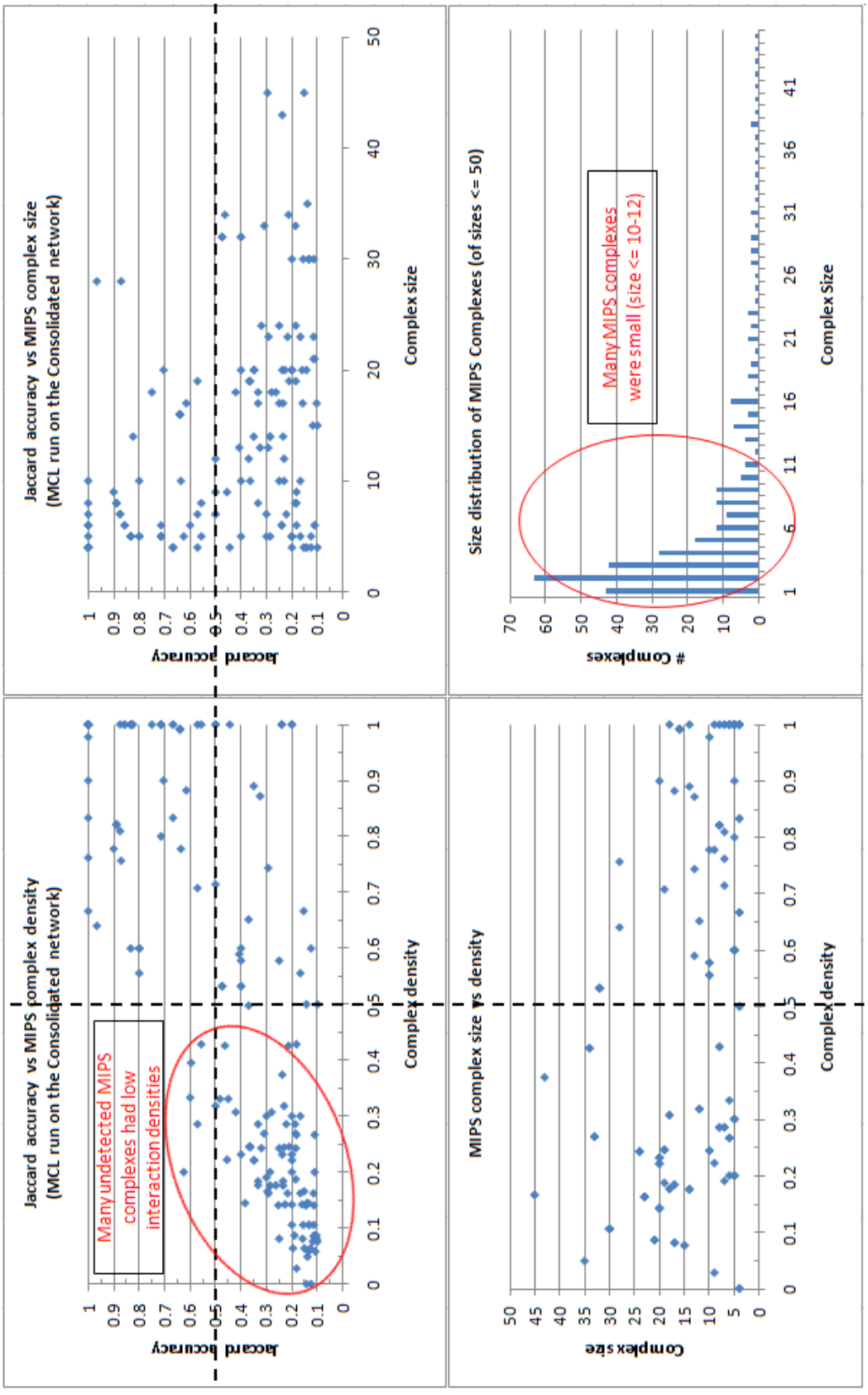


Figure 5.2: The plot of Jaccard accuracy (with which the complexes were derived) *versus* edge density of MIPS complexes in the Consolidated network shows that many MIPS complexes derived with low accuracies had in fact low densities (< 0.50) in the network. This pointed towards a potentially strong correlation between the "network constitution" of a benchmark complex in the PPI network and the possibility of it being detected using existing methods.

A natural thing is to question the underlying assumption: How accurate is this "denseness" assumption of complexes for computational prediction from PPI networks? Or alternately, to what extent can we rely on this "denseness" assumption to predict complexes? It is perfectly appropriate to ask this because, as we saw, overly relying on this assumption in the wake of insufficient credible PPI data can cause low density or disconnected complexes to be totally missed. Of course we could go for devising more "sensitive" models that can cover such low density complexes (one such attempt is the work by Habibi et al. (2010) [111] that models complexes as k-connected subnetworks). However, there is a limit to how "sensitive" these models can get. Too sensitive models can also result in too many false positive predictions (as noted in Habibi et al.'s work [111]). Therefore, it is also important to look at other "work-arounds" to detect these low density complexes.

The aim of this chapter is two-fold: (i) to topologically *characterize* these undetected complexes, that is, to quantitatively measure their "network constitution"; and (ii) to propose a novel "work-around" to *aid* existing methods in detecting them satisfactorily. A simple yet elegant "work-around" we propose here is to *non-randomly* "fill the gaps" in PPI networks by looking beyond physical interactions to handle the low density regions of the networks.

5.2 Characterizing sparse complexes

Sticking to our previously adopted terminologies, we represent our PPI network as $G = (V, E)$, where V is the set of proteins and E is the set of interactions between the proteins. Each interaction $e = (u, v) \in E$ is assigned a weight $0 \leq w(u, v) \leq 1$ that reflects the confidence of the interaction, which is usually determined using an affinity weighting scheme (the weight it is set to 1 if no scheme is used). For any $u \in V$, $\mathcal{N}(u)$ refers to the set of neighbors of u. Let $\mathcal{B} = \{B_1, B_2, ..., B_m\}$ be the set of benchmark complexes. We propose the term *sparse complexes* for the undetected complexes and "very broadly" define them as follows:

Definition 5.1 SPARSE COMPLEXES: *Given a PPI network G and a set of benchmark complexes $\mathcal{B}$ known to be embedded in G, the subset $\mathcal{B}' \subseteq \mathcal{B}$ of complexes that cannot be satisfactorily detected from G by existing methods are called sparse complexes.*

5.2.1 Indices for complex derivability from PPI networks

We next propose *indices* that measure the "derivability" of a benchmark complex from a given PPI network. These indices capture whether or not a benchmark complex is derivable from a given PPI network, and if so, to what extent. We propose two kinds of indices here. The first kind defines definitive criteria to categorize a given benchmark complex as derivable or not from the PPI network, and provides *derivability bounds* on the number of such complexes in the network. The second kind does not strictly categorize the benchmark complex as derivable or not, but instead assigns a *derivability score* to the complex.

Derivability indices with bounds

To begin with, a naive yet natural way to categorize a benchmark complex as *derivable* from a PPI network is if it satisfies two criteria: (i) it has sufficient number of proteins in the network; and (ii) it is connected within the network.

We consider a benchmark complex $B_i \in \mathcal{B}$ to be *k-protein-derivable* from G if at least $k > 0$ of its member proteins are present in G. We consider a k-protein-derivable complex to be *k-network-derivable* from G if these member proteins form a connected subnetwork within G.

Definition 5.2 k-PROTEIN-DERIVABLE COMPLEX: *A benchmark complex $B_i \in \mathcal{B}$ is k-protein-derivable from network $G = (V, E)$ if $|B_i \cap V| \geq k$, for some $k > 0$.*

The set of k-protein-derivable complexes in G is represented by $D_P(\mathcal{B}, G, k)$, and the *k-protein-derivability index* of G is $|D_P(\mathcal{B}, G, k)|$.

Definition 5.3 k-NETWORK-DERIVABLE COMPLEX: *A benchmark complex $B_i \in \mathcal{B}$ is k-network-derivable from $G = (V, E)$ if $|B_i \cap V| \geq k$ for some $k > 0$, and $B_i \cap V$ forms a connected subnetwork in G.*

The set of k-network-derivable complexes in G is represented by $D_N(\mathcal{B}, G, k)$, and the *k-network-derivability index* of G is $|D_N(\mathcal{B}, G, k)|$.

Derivability indices with scores

From our systematic experiments (see the "side note" below), we found that two factors strongly contributed to the "derivability" of a given complex from the network

- the presence of a significant fraction of complex proteins within the same connected component, and the density of the complex relative to its local neighborhood. Based on these two factors we next define indices that assign *derivability scores* to each benchmark complex to reflect the confidence or extent to which the complex is derivable from the network.

Component Score $CS(B_i, G)$: In the network G, let any k-protein-derivable complex B_i be decomposed into several connected components, $\{S_1(B_i, G), S_2(B_i, G), ..., S_r(B_i, G)\}$, ordered in non-increasing order of size. We define $CS(B_i, G)$ as the fraction of proteins within the maximal component $S_1(B_i, G)$ among all *non-isolated proteins* in B_i:

$$CS(B_i, G) = \frac{|S_1(B_i, G)|}{|B_i'|} \text{ for } |B_i'| > 0, \text{ else } CS(B_i, G) = 0, \tag{5.1}$$

where $B_i' = \{p : p \in B_i, \exists q \in B_i, (p, q) \in E\}$.

Edge Score $ES(B_i, G)$: We define $ES(B_i, G)$ as the ratio of the weight of interactions within B_i to the total weight of interactions within B_i and its immediate neighborhood in G:

$$ES(B_i, G) = \frac{\sum_{e \in E(B_i)} w(e)}{\sum_{e \in E(NB_i)} w(e)} \text{ for } E(NB_i) \neq \emptyset, \text{ else } ES(B_i, G) = 0. \tag{5.2}$$

The denominator is the weight of interactions in the subnetwork of G induced by the member proteins of B_i and their direct neighbors, given by: $V(NB_i) = \{p : p \in B_i\} \bigcup \{q : q \in \mathcal{N}(p), p \in B_i\}$ and $E(NB_i) = \{(p, q) : p, q \in V(NB_i), (p, q) \in E\}$. Note that the edge score is different from the absolute *edge density* of B_i which is not relative to the neighborhood, defined as: $d(B_i, G) = \sum_{e \in E(B_i)} w(e)/(|V(B_i)|.(|V(B_i)| - 1))$.

We define the *Component-Edge score* $CE(B_i, G)$ as the product of the component and edge scores of B_i:

$$CE(B_i, G) = CS(B_i, G) * ES(B_i, G). \tag{5.3}$$

Definition 5.4 k-CE-DERIVABLE COMPLEX: *Given a threshold* $0 \leq t_{ce} \leq 1$*, a* k*-protein-derivable complex* B_i *is* k*-CE-derivable if* $CE(B_i, G) \geq t_{ce}$.

Therefore, the set of k-CE-derivable complexes in G is given by: $D_{CE}(\mathcal{B}, G, k, t_{ce}) = \{B_i : B_i \in D_P(\mathcal{B}, G, k), CE(B_i, G) \geq t_{ce}\}$, and the *$k$-CE-derivability index* of G is $|D_{CE}(\mathcal{B}, G, k, t_{ce})|$.

A side note: Here, we give a broad idea of the experiments we performed to observe the two factors influencing complex derivability. We first constructed an "ideal" network G' from the PPI network G by considering only the proteins $V \cap \mathcal{B}$ and their interactions. We tested the performance of several existing methods on G' and analysed how many benchmark complexes of $\mathcal{B}$ were reconstructed successfully. As expected, all methods performed well. But, a noticeable pattern was that the methods were able to reconstruct those complexes better that had a significant fraction of the proteins within a single connected component. We next we constructed a "slightly hazy" network G'' by adding the remaining proteins $V \setminus \mathcal{B}$ and their interactions to G', and repeated our analysis. We noticed that the methods were beginning to get "confused": in cases where the boundary between the embedded complex and its neighborhood was too obscure to discern clearly. This indicated that local neighborhood played a vital role in complex identification. Finally we added the remaining interactions and repeated our analysis, and found that these additional interactions further "confused" the methods. These findings led us to define our derivability scores based on the two factors - the presence of a significant fraction of complex proteins within the same connected component, and the density of the complex relative to its local neighborhood.

Relationships among the derivability indices

For any $k > 0$, by definition $D_N(\mathcal{B}, G, k) \subseteq D_P(\mathcal{B}, G, k)$. Given a threshold $0 \leq t_{ce} \leq 1$, the relationships between $D_P(\mathcal{B}, G, k)$ and $D_N(\mathcal{B}, G, k)$ with $D_{CE}(\mathcal{B}, G, k, t_{ce})$ are as follows. When $t_{ce} = 0$, all k-CE-derivable complexes are also k-protein-derivable, but because they may not be connected we can say, $D_N(\mathcal{B}, G, k) \subseteq D_{CE}(\mathcal{B}, G, k, t_{ce} = 0) \subseteq D_P(\mathcal{B}, G, k)$. When $t_{ce} = 1$, all k-CE-derivable complexes are connected complexes that are disjoint, therefore $D_{CE}(\mathcal{B}, G, k, t_{ce} = 1) \subseteq D_N(\mathcal{B}, G, k) \subseteq D_P(\mathcal{B}, G, k)$ (see Figure 5.3). Intuitively, t_{ce} can be varied in the entire range $[0, 1]$ to include the "hardest" complexes to detect (without any internal connectivities) to only the "easiest" complexes to de-

tect (disjoint connected complexes). These "hardest" complexes to detect can form "holes" in the network by having zero interactions among their member proteins but having interactions with their immediate neighbors

5.2.2 Validating the derivability indices against ground truth

We now validate the derivability scores (CS, ES, CE scores and absolute edge density) of benchmark complexes with respect to the PPI network against the accuracies with which these complexes are actually derived using existing methods.

We use two PPI networks for this validation, the Consolidated$_{3.19}$ network (a weighted network) from Collins et al. [36], and the 'Filtered Yeast Interaction' (FYI) network (a literature-validated but unweighted network) from Han et al. [119]. Tables 5.1 and 5.2 show the Pearson correlation values between these indices and the *Jaccard* accuracies of complexes derived from these networks using four complex detection methods, MCL, MCL-CAw, CMC and HACO, and evaluated against MIPS and Wodak catalogues. The corresponding correlation plots for MCL-CAw and CMC are shown in Figures 5.4 and 5.5 (the other two methods also displayed similar plots). The results show the CE scores are *strongly correlated* with Jaccard accuracies. This is followed by the ES, CS and edge density scores. This means our proposed CE-score is a *stronger* indicator of actual complex derivability compared to the traditionally adopted indicators like edge density. (Note: There are a few other indices like Newman and Girvan's global and local modularity [112], but these do not capture the notion of proteins being part of the same connected component, and they perform similar to our edge-score ES).

5.2.3 A measure of sparse complexes

We can now employ our proposed CE-score to give a more quantitative definition for sparse complexes.

Definition 5.5 SPARSE COMPLEXES: *Given a PPI network G, a benchmark complex B_i and a threshold $0 \leq t_{ce} \leq 1$, the complex B_i is called sparse with respect to G if $CE(B_i, G) < t_{ce}$.*

Notice how the two definitions 5.1 and 5.5 can be "linked" using our CE-score and threshold t_{ce}, which offer a quantitative value to the derivability of complexes.

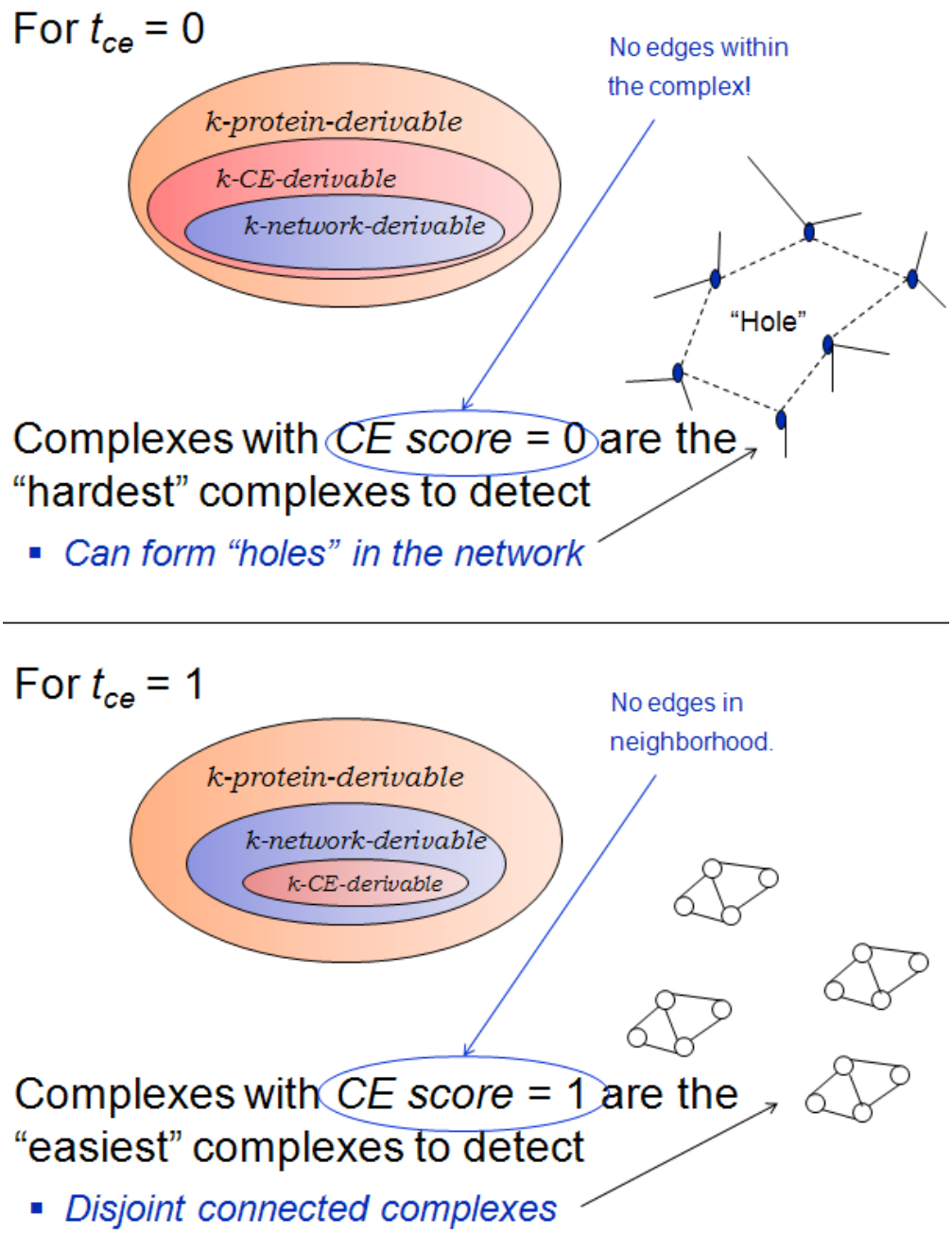


Figure 5.3: Relationships among the derivability indices for $t_{ce} = 0$ and $t_{ce} = 1$. From the "hardest" to the "easiest" complexes to detect.

The Consolidated$_{3.19}$ network: #proteins 1622, #interactions 9704

	Pearson correlation with Jaccard accuracy				
Benchmark	Method	Edge density	CE-score	Comp score	Edge score
MIPS (#313)	MCL	0.101	0.719	0.511	0.518
	MCL-CAw	0.196	0.785	0.492	0.628
	CMC	0.174	0.649	0.471	0.477
	HACO	0.159	0.786	0.472	0.608
Wodak (#405)	MCL	0.141	0.734	0.597	0.623
	MCL-CAw	0.152	0.792	0.611	0.638
	CMC	0.196	0.709	0.479	0.442
	HACO	0.168	0.789	0.523	0.612

Table 5.1: Pearson correlation between the derivability indices and Jaccard accuracies (on the Consolidated network). The CE-scores show the strongest correlation with the accuracies.

The Filtered Yeast Interaction (FYI) network: #proteins 1379, #interactions 2493

	Pearson correlation with Jaccard accuracy				
Benchmark	Method	Edge density	CE-score	Comp score	Edge score
MIPS (#313)	MCL	0.097	0.699	0.423	0.507
	MCL-CAw	0.116	0.746	0.501	0.621
	CMC	0.198	0.718	0.527	0.649
	HACO	0.173	0.772	0.412	0.648
Wodak (#405)	MCL	0.126	0.708	0.554	0.599
	MCL-CAw	0.153	0.718	0.597	0.605
	CMC	0.188	0.689	0.407	0.412
	HACO	0.160	0.701	0.512	0.602

Table 5.2: Pearson correlation between the derivability indices and Jaccard accuracies (on the Filtered Yeast Interaction network). The CE-scores show the strongest correlation with the accuracies.

If this value is less than a certain threshold, the complex is highly likely to go undetected from existing methods and therefore it is *sparse*, else it is highly likely to be detected and therefore it is *dense*. In general, for the benchmark complexes $\mathcal{B}$, the set of sparse complexes is given by $\mathcal{S}(\mathcal{B}, G, k, t_{ce}) = \{B_i : B_i \in D_P(\mathcal{B}, G, k), CE(B_i, G) < t_{ce}\}$, and its complementary set $\mathcal{D}(\mathcal{B}, G, k, t_{ce}) = \{B_i : B_i \in D_P(\mathcal{B}, G, k), CE(B_i, G) \geq t_{ce}\}$ forms the dense complexes. The threshold t_{ce} defines this "boundary" between the sparse and dense benchmark complexes in the network. Since we do not know at which value of t_{ce} existing methods operate, we propose an approach that "packs" higher number of dense complexes for all values of $t_{ce} \in [0, 1]$ or at least for the larger values of t_{ce}.

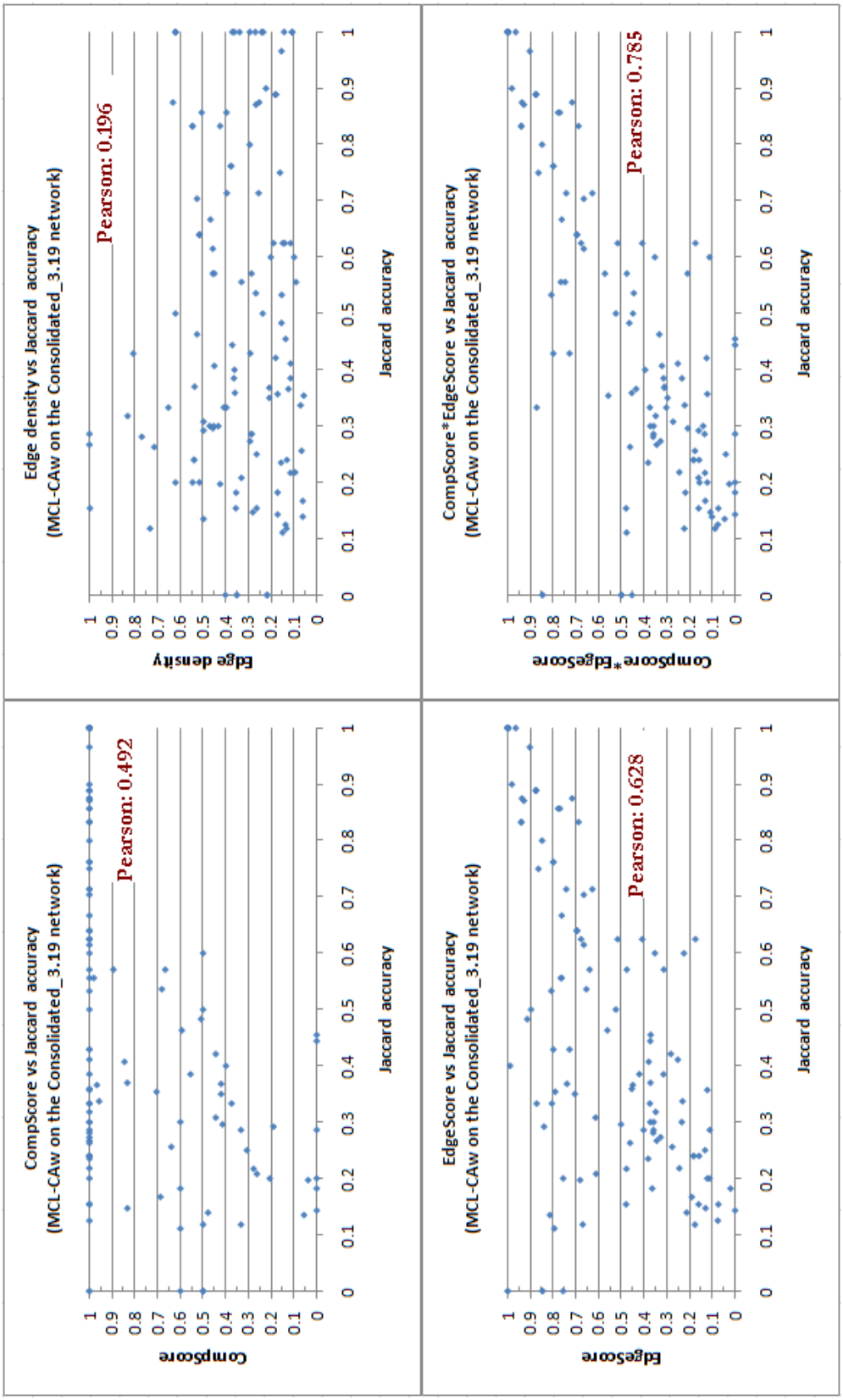


Figure 5.4: Validating the derivability indices against ground truth: scatter plot for MCL-CAw. The CE-scores showed strong correlation with Jaccard accuracies.

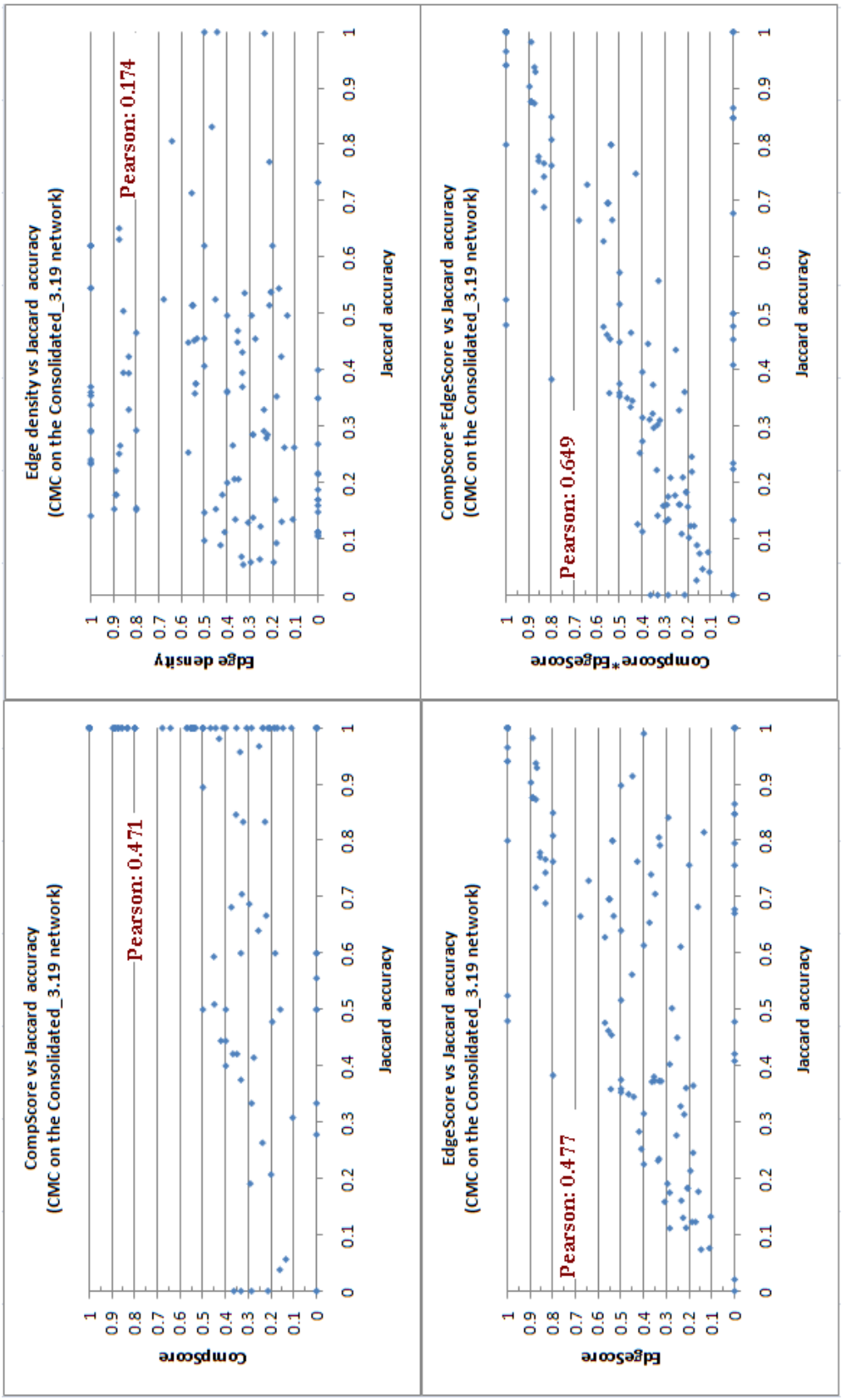


Figure 5.5: Validating the derivability indices against ground truth: scatter plot for CMC. The CE-scores showed strong correlation with Jaccard accuracies.

5.3 Detecting sparse complexes

We noted in Section 5.1 that existing methods are severely constrained by "gaps" in crucial topological information required to ensure the two required criteria for complex derivability namely, component-based connectivity and relative edge density. In fact, any new method based solely on PPI networks would also face these constraints. Due to these reasons, a natural approach to aid existing methods or devise new methods would be to first fill these "topological gaps" in existing PPI networks.

Even though this seems like a simple enough solution to pursue, we are severely lacking in the interaction data required to fill these gaps. Current estimates on yeast [34], put the verified fraction of the physical interactome to ∼70%, which means we are still lacking in ∼30% reliable interaction data, mainly due to limitations in existing experimental and computational techniques. Consequently, a novel solution is to look beyond physical interactions to fill these topological gaps. In our work, we propose to use *functional interactions* for this purpose, specifically aimed at improving complex prediction.

5.3.1 Employing functional interactions to detect sparse complexes

Functional interactions or associations are logical interactions among proteins that share similar functions [55]. These interactions can be inferred among proteins participating in the same multi-protein assemblies (complexes, functional modules and pathways), or annotated to similar biological functions and processes, or encoded by genes maintained and regulated together or genes having the same 'phylogenetic profile' (present or absent together across several genomes), etc. [55]. Therefore, these interactions "encode" information beyond just direct physical interactions. In fact many of the computational methods developed to predict protein interactions mainly manage to predict functional interactions.

Functional interactions can be considered more "general" or a "superset" of direct physical interactions: two proteins involved in a stable physical interaction are functionally related, but two proteins involved in a functional interaction may not

necessarily interact physically. This means functional interactions have a potential to effectively *complement* physical interactions. We capitalize on this complementarity by non-randomly adding functional interactions to ensure the two required criteria: (i) Some functional interactions may be direct physical interactions missing in the physical datasets - these are directly useful to "pull-in" disconnected proteins; and (ii) Even if some functional interactions do not correspond to direct physical interactions, if they fall within the same complex, they can "artificially" increase the density of that complex.

5.3.2 The SPARC algorithm for employing functional interactions

Here, we propose a post-processing based algorithm SPARC to empower existing methods (provide them the "spark") to detect SPARse Complexes by using functional interactions. SPARC works as follows (see Algorithm 2). Let $G_P = (V_P, E_P)$ be the PPI network and $G_F = (V_F, E_F)$ be the functional network.

Step 1: The input to the algorithm is the set of physical clusters $\mathcal{C}_P$ from network G_P generated using an existing method. It then calculates the CE-score $CE(G_P, C_i)$ for each cluster $C_i \in \mathcal{C}_P$. All clusters with CE-scores above a threshold δ, that is, $\{C_i \in \mathcal{C}_P : CE(C_i, G_P) \geq \delta\}$, are output as predicted complexes, while the remaining are reserved for further processing.

Step 2: We then add-in the interactions of G_F to G_P to produce a larger network $G_A = (V_A, E_A)$, where $V_A = V_P \cup V_F$ and $E_A = E_P \cup E_F$.

Step 3 (iterative): For each reserved cluster C_j, the CE-score is recalculated with respect to G_A. If for the cluster C_j, the CE-score improves beyond δ, that is, $CE(C_j, G_A) \geq \delta$, it is output as a predicted complex. If not, we explore in the neighborhood of C_j to include proteins that can potentially improve $CE(C_j, G_A)$. We consider the set of direct neighbors $\mathcal{N}(C_j, G_A)$, and sort them in non-increasing order of their interaction weights to C_j. We then repeatedly consider a protein $p \in \mathcal{N}(C_j, G_A)$ in that order such that $CE(C_j \cup \{p\}, G_A) > CE(C_j, G_A)$ and add it to C_j, till the CE-score cannot be improved any further. If the improved CE-score manages to cross δ, we output the cluster C_j as a predicted complex.

The key idea behind SPARC is as follows. Many complexes have low CE-

scores in the PPI network. If adding functional interactions can either increase their internal connectivities or "pull in" the disconnected proteins, we can increase the CE-scores of these complexes. However, blindly adding functional interactions can result in many false positive predictions. Therefore, here we selectively utilize functional interactions only to improve the CE-scores of clusters predicted out of the physical network. Those clusters that show the improvement correspond to real complexes.

Algorithm 2 SPARC($G_P, G_F, \mathcal{C}_P, t$)

```
for each C_i ∈ C_P do
  if CE(C_i, G_P) ≥ δ then
    Output C_i;
  end if
end for

Augment the networks: G_A = (V_A, E_A), where V_A = V_P ∪ V_F, E_A = E_P ∪ E_F.

for each remaining C_j do
  if CE(C_j, G_A) ≥ δ then
    Output C_j;
  else
    Sort N(C_j, G_A) in non-increasing order of interaction weights to C_j;
    while ΔCE(C_j, G_A) > 0 do
      Choose the next p ∈ N(C_j, G_A);
      C_j := C_j ∪ {p};
      Recalculate CE(C_j, G_A);
    end while
    if CE(C_j, G_A) ≥ δ then
      Output C_j;
    end if
  end if
end for

Output the final set of predicted complexes;
```

5.4 Experimental results

5.4.1 Preparation of experimental data

We gathered physical interactions from *Saccharomyces cerevisiae* (budding yeast) inferred from the following yeast two-hybrid and affinity purification experiments, deposited in Biogrid [54]: Uetz et al. [12], Ito et al. [13], Gavin et al. [15,27], Krogan et al. [28] and Collins et al. [36], to build the protein interaction network, which we call the *Physical network* P (therefore, P comprises of the Gavin+Krogan network

(of Chapter 4) together with Y2H interactions from a few other experiments). The interactions of P are not scored.

Next, high-confidence functional interactions from yeast were gathered from the String database [55] to build the *Functional network* F. These functional interactions showed confidence scores ≥ 0.90 in at least two of the following evidences: gene neighborhood, co-occurrence, co-expression and text mining (these scores are available from String).

We combined the two networks to generate a larger network which we call the *Augmented* Physical+Functional network $P + F$. Table 5.3 shows some properties of these networks. The overlaps between the two networks is shown in Figure 5.6.

Network	# Proteins	# Interactions	Avg node degree
Physical (P)	4113	26518	12.89
Functional (F)	3960	18683	10.12
Augmented ($P + F$)	5145	43905	17.07

Table 5.3: Properties of the physical and functional networks obtained from yeast.

The presence of *noise* (false positives) is a severe limiting factor in publicly available interaction datasets in spite of gathering only high-confidence datasets. Therefore, we further *filtered* these datasets, which involves assigning each interaction a confidence score (between 0 and 1) that reflects its reliability, and discarding interactions with low scores (< 0.20). Here, we (re)scored the networks using three scoring schemes, two of which were based on network topology namely, *FS-Weight* [39] and *Iterative-CD* [64], while the third was based on evidences from

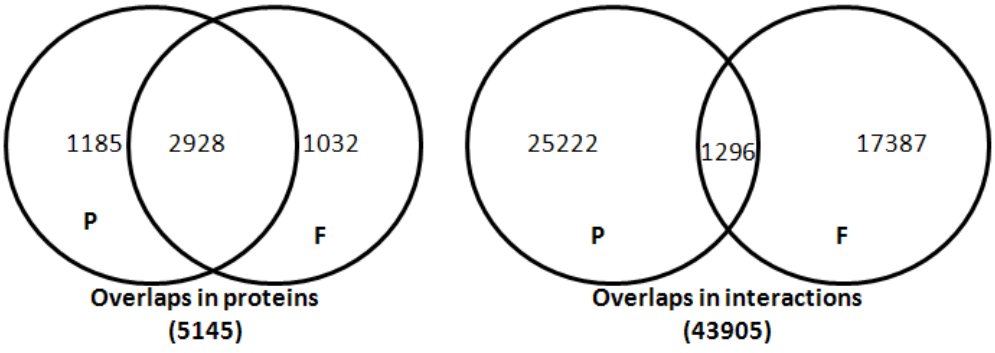


Figure 5.6: Overlaps between the physical and functional datasets

Gene Ontology (GO), called *TCSS* [46].

Benchmark complexes and GO annotations

The benchmark (or reference or 'gold standard') set of complexes was assembled from two independent sources: 313 complexes of MIPS [90] and 408 complexes of the Wodak lab CYC2008 catalogue [92]. The properties of these benchmark sets are shown in Table 5.4. For the evaluation, we considered only the 4-protein-derivable complexes out of these sets. This is because it is typically difficult to predict very small complexes (size < 4) with high accuracy by using primarily topological information [18, 64].

		Size distribution			
Benchmark	#Complexes	< 3	3-10	11-25	> 25
MIPS	313	106	138	42	27
Wodak	408	172	204	27	5

Table 5.4: Properties of hand-curated (benchmark) yeast complexes from the MIPS and Wodak CYC2008 catalogues.

The GO annotations for yeast proteins were downloaded from the *Saccharomyces* Genome Database (SGD) [94], which include the annotations (not considering the Inferred from Electronic Annotations or IEA) for three ontologies - Cellular Component (CC), Biological Process (BP) and Molecular Function (MF). These annotations were used as evidences in the TCSS scheme [46]. We excluded the branch corresponding to the GO term 'macromolecular complex' (GO:0032991) to avoid any bias coming from the GO complexes.

5.4.2 Complex detection algorithms and evaluation metrics

We used four complex detecting algorithms mentioned previously, MCL [63], CMC [64], HACO [69] and MCL-CAw (Chapter 4). Some of their properties and the preset parameter values are summarized in Table 5.5. These methods are different from one another in the algorithmic techniques employed, and therefore form a good mix of methods for our evaluation.

Usually, recall Rc (coverage) and precision Pr (sensitivity) are used to evaluate the performance of methods against benchmark complexes. Here, we use previously reported [64] definitions for these measures. Let $\mathcal{B} = \{B_1, B_2, ..., B_m\}$ and $\mathcal{C} =$

Property	**MCL**	**MCL-CAw**	**CMC**	**HACO**
Principle	Flow simulation	Core-attach refinement over MCL	Maximal clique merging	Hier agglo cluster with overlaps
Parameters (preset values)	I (2.5)	I, α, γ (2.5, 1.5, 0.75)	Merge m, Overlap t, Min clust size (0.5, 0.4, 4)	UPGMA cutoff (0.2)

Table 5.5: Existing complex detection methods used in the evaluation.

$\{C_1, C_2, ..., C_n\}$ be the sets of benchmark and predicted complexes, respectively. We use the Jaccard coefficient J to quantify the overlap between a B_i and a C_j: $J(B_i, C_j) = |B_i \cap C_j|/|B_i \cup C_j|$.

We consider B_i to be covered by C_j, if $J(B_i, C_j) \geq$ *overlap threshold* J_{min}. In our experiments, we set the threshold $J_{min} = 0.50$, which requires $|B_i \cap C_j| \geq \frac{|B_i|+|C_j|}{3}$. For example, if $|B_i| = |C_j| = 8$, then the overlap between B_i and C_j should be at least 6. Based on this the recall Rc is given by:

$$Rc(\mathcal{B}, \mathcal{P}) = \frac{|\{B_i | B_i \in \mathcal{B} \wedge \exists C_j \in \mathcal{C}; J(B_i, C_j) \geq J_{min}\}|}{|\mathcal{B}|}. \tag{5.4}$$

Here, $|\{B_i | B_i \in \mathcal{B} \wedge \exists C_j \in \mathcal{C}; J(B_i, C_j) \geq J_{min}\}|$ gives the number of *derived benchmarks*. And the precision Pr is given by:

$$Pr(\mathcal{B}, \mathcal{P}) = \frac{|\{C_j | C_j \in \mathcal{C} \wedge \exists B_i \in \mathcal{B}; J(B_i, C_j) \geq J_{min}\}|}{|\mathcal{C}|}. \tag{5.5}$$

Here, $|\{C_j | C_j \in \mathcal{C} \wedge \exists B_i \in \mathcal{B}; J(B_i, C_j) \geq J_{min}\}|$ gives the number of *matched predictions*.

5.4.3 Impact of adding functional interactions on complex derivability

To begin with, we measured the number of derivable benchmark complexes from the Physical (P), Functional (F), Augmented ($P + F$) networks and their scored versions, $ICD(P+F)$, $FSW(P+F)$ and $TCSS(P+F)$, using our proposed derivability indices.

Table 5.6 shows the number of protein-derivable and network-derivable bench-

mark complexes from these networks. The findings can be summarized as follows: (a) The network-derivable complexes were significantly fewer than the protein-derivable complexes further supporting the claim that many benchmark complexes remained disconnected within the networks. (b) The number of protein-derivable and network-derivable complexes were higher for the $P + F$ network than the individual P and F networks. The significance of this increase was gauged against a random network R built using the same set of proteins and the average node degree in F. The $P + R$ network showed fewer network-derivable complexes compared to $P + F$. This indicated that F added more interactions to "complexed" regions in P compared to what the R network added. (c) The number of protein-derivable and network-derivable complexes in the scored networks, $ICD(P + F)$, $FSW(P + F)$ and $TCSS(P + F)$, were fewer than the $P + F$ network. This is not a concern because filtering usually discards interaction data leading to smaller networks. (d) Even though protein-derivable complexes in the scored networks were fewer than the $P + F$ network, the corresponding decrease in network-derivable complexes was relatively marginal. This indicated that the scoring schemes retained most interactions among complexed proteins, and discarded mainly the noisy ones.

	MIPS (# 313)		Wodak CYC2008 (# 408)	
Network	#Protein-derivable	#Network-derivable	#Protein-derivable	#Network-derivable
P	155	59	135	81
F	153	28	127	37
P+R	164	61	147	82
P+F	164	68	147	92
ICD(P+F)	122	67	124	91
FSW(P+F)	119	67	95	78
TCSS(P+F)	158	68	143	75

Table 5.6: Impact of augmenting functional interactions on protein-derivability and network-derivability for $k = 4$.

Next, Table 5.7 shows the number of CE-derivable benchmark complexes from these networks for all threshold values $t_{ce} \in [0, 1]$. This table does a more fine-scale dissection of the improvement shown before. For lower values of t_{ce}, the number of CE-derivable complexes was higher for $P + F$ compared to P. But, for higher values of t_{ce}, the number was lower compared to P. Similarly, for lower values of t_{ce}, the number of CE-derivable complexes was higher for $P + F$ compared to

the three scored networks. But, for higher values of t_{ce}, the three scored networks showed considerably higher CE-derivable complexes than both the P and $P + F$ networks. These findings indicate that noise had a sizable impact on the CE-scores of complexes: the improvement obtained by adding functional interactions was completely canceled out by noise, leading to lower performance of the $P + F$ network. But, affinity scoring (filtering) considerably alleviated this impact of noise, thereby improving the CE-derivability of the networks.

MIPS (#313)

	# Complexes with CE-score $\geq t_{ce}$					
Threshold t_{ce}	P	F	P+F	ICD(P+F)	FSW(P+F)	TCSS(P+F)
0.00	155	153	164	152	119	162
0.10	153	151	162	148	116	160
0.20	149	136	158	145	113	157
0.30	140	108	149	142	110	154
0.40	129	81	135	137	108	148
0.50	101	54	102	112	101	126
0.60	81	21	70	93	87	101
0.70	62	9	55	71	69	86
0.80	39	0	34	44	42	59
0.90	19	0	14	21	21	35
1.00	6	0	3	11	10	18

Table 5.7: Impact of augmenting functional interactions on CE-derivability for $k = 4$ (MIPS benchmark).

Wodak CYC2008 (#408)

	# Complexes with CE-score $\geq t_{ce}$					
Threshold t_{ce}	P	F	P+F	ICD(P+F)	FSW(P+F)	TCSS(P+F)
0.00	135	127	147	124	95	143
0.10	131	112	144	121	93	141
0.20	123	93	129	116	87	135
0.30	112	66	114	113	84	126
0.40	99	31	101	102	77	109
0.50	83	8	75	91	59	94
0.60	71	1	62	78	43	83
0.70	59	0	41	61	39	67
0.80	34	0	21	42	26	44
0.90	12	0	6	29	13	31
1.00	8	0	0	18	8	20

Table 5.8: Impact of augmenting functional interactions on CE-derivability for $k = 4$ (Wodak benchmark).

5.4.4 Improvement in complex detection using SPARC

Table 5.9 shows the performance of the four methods MCL, MCL-CAw, CMC and HACO on the raw and scored physical networks (we do not show the results on F because functional interactions are only used to improve the physical clusters, and not for complex detection by themselves - many of the functional clusters do not correspond to physical complexes). It shows that scoring helped to reconstruct significantly more complexes and with better accuracies (similar results were observed on the Wodak catalogue) over raw datasets.

		Matched against MIPS complexes. Jaccard threshold $J_{min} = 0.50$.					
Method	Network	#Predicted	#Matched	#Derivable	#Derived	Pr	Rc
MCL	Physical P	294	29	155	38	0.098	0.245
	FSW(P)	156	31	102	40	0.198	0.333
	ICD(P)	167	32	109	40	0.191	0.293
	TCSS(P)	172	39	112	41	0.226	0.366
MCL -CAw	Physical P	297	39	155	49	0.131	0.316
	FSW(P)	149	38	102	51	0.255	0.392
	ICD(P)	162	41	109	52	0.253	0.376
	TCSS(P)	168	41	112	54	0.244	0.366
CMC	Physical P	156	41	155	56	0.263	0.361
	FSW(P)	144	31	102	59	0.215	0.313
	ICD(P)	165	43	109	60	0.260	0.394
	TCSS(P)	128	39	112	59	0.304	0.357
HACO	Physical P	414	34	155	41	0.082	0.264
	FSW(P)	221	32	102	44	0.144	0.313
	ICD(P)	248	37	109	45	0.149	0.339
	TCSS(P)	253	46	112	45	0.181	0.410

Table 5.9: Impact of scoring on complex detection methods (evaluation against MIPS). 'Derivable' refers to 4-protein-derivable complexes.

Next, Table 5.10 shows the performance after refining these physical clusters using functional interactions through SPARC (at $\delta = 0.40$). It shows that post-processing using raw functional interactions (P+F) led to many noisy clusters, resulting in lower precision and recall. But, adding filtered (scored) functional interactions to scored physical datasets (denoted as FSW(P+F), ICD(P+F) and TCSS(P+F)) through SPARC helped to reconstruct significantly more benchmark complexes. This shows that scoring combined with SPARC-based refinement significantly boosted the performance of all methods.

Table 5.11 does a more finescale analysis of the complexes reconstructed from the sparse phsyical clusters before and after SPARC-based post-processing. It shows that many of the "initial" physical clusters that were sparse (CE-score < 0.40)

		Matched against MIPS complexes. Jaccard threshold $J_{min} = 0.50$.						
Method	Network	#Predicted	Size	#Matched	#Derivable	#Derived	Pr	Rc
MCL	P	294	7.96	29	155	38	0.098	0.245
	P+F	338	8.66	19	164	23	0.056	0.140
	FSW(P+F)	102	15.88	29	119	38	0.284	0.319
	ICD(P+F)	138	17.14	33	122	44	0.239	0.361
	TCSS(P+F)	261	10.52	42	158	54	0.161	0.342
	Consensus	429	13.01	57	164	56	0.133	0.341
MCL -CAw	P	297	7.94	39	155	49	0.131	0.316
	P+F	342	8.34	25	164	29	0.073	0.177
	FSW(P+F)	136	9.46	41	119	57	0.301	0.479
	ICD(P+F)	141	7.44	48	122	61	0.340	0.500
	TCSS(P+F)	296	9.98	49	158	61	0.166	0.386
	Consensus	484	8.72	81	164	71	0.167	0.432
CMC	P	156	11.42	41	155	56	0.263	0.361
	P+F	306	14.39	33	164	41	0.108	0.250
	FSW(P+F)	136	12.44	36	119	48	0.265	0.403
	ICD(P+F)	252	8.91	51	122	63	0.202	0.516
	TCSS(P+F)	127	11.66	45	158	60	0.354	0.380
	Consensus	429	9.80	80	164	66	0.186	0.402
HACO	P	414	5.98	34	155	41	0.082	0.264
	P+F	510	6.68	28	164	34	0.055	0.207
	FSW(P+F)	111	10.17	39	119	54	0.351	0.454
	ICD(P+F)	131	8.90	43	122	60	0.328	0.492
	TCSS(P+F)	269	7.49	55	158	67	0.204	0.424
	Consensus	419	7.61	79	164	74	0.189	0.451

Table 5.10: Impact of adding functional interactions using SPARC on complex detection methods (evaluation against MIPS). 'Derivable' refers to 4-protein-derivable complexes.

		#Predicted clusters				#Derived benchmarks	
Method	Network	Initial	Sparse ($CE < 0.40$)	Processed by SPARC	Final (Size ≥ 4)	Before	After
MCL	P+F	638	269	8	338	0	2
	FSW(P+F)	188	42	16	102	1	9
	ICD(P+F)	258	57	18	138	2	9
	TCSS(P+F)	380	102	19	261	2	10
MCL-CAw	P+F	472	212	8	342	0	2
	FSW(P+F)	255	37	19	136	2	11
	ICD(P+F)	258	39	21	141	2	13
	TCSS(P+F)	408	97	26	296	3	16
CMC	P+F	424	186	20	306	0	8
	FSW(P+F)	251	32	23	136	2	18
	ICD(P+F)	354	44	36	252	2	21
	TCSS(P+F)	224	56	41	127	4	27
HACO	P+F	389	25	510	338	1	10
	FSW(P+F)	53	29	111	102	2	21
	ICD(P+F)	59	31	131	138	3	23
	TCSS(P+F)	66	43	269	261	6	36

Table 5.11: The number of benchmark complexes recovered by sparse clusters before and after the SPARC-based processing.

underwent SPARC post-processing. These post-processed clusters were able to reconstruct significantly higher number of benchmark complexes. Their CE-scores showed a huge improvement, and the correlation between this improvement and the improvement in their Jaccard accuracies (when matched to benchmark complexes) is shown in Figure 5.7.

Note: One interesting point to note in Table 5.10 is that the compositions of predicted complexes vary based on the scoring scheme used, and therefore we had to construct a *consensus set* of complexes from the three scoring schemes for each of the methods. To do this, we employed a three-way agreement scheme based on Jaccard overlaps. Let $\{A, B, C\}$ be a complex triplet, each complex predicted from a different scored network by the same method. If at least two complex pairs from $\{(A, B), (B, C), (C, A)\}$ achieve significant Jaccard overlaps (≥ 0.70), then the proteins of A, B and C are merged together into a single consensus complex T. Only the proteins originating from at least two complexes are included in T. We noticed that this consensus operation further improves the accuracies of the predictions leading to better reconstruction of benchmark complexes.

An edge density-wise break up study of improvement: Figures 5.8 and 5.9 show an edge density wise break up of complexes derived before and after SPARC-refinement. We exclude MCL to draw any conclusions, and considering the other three methods, we note that there are two "bands of impact" (marked in circles) due to SPARC: (i) The first band is around low density complexes - there is improvement seen for complexes of densities as low as 0.10, which is due to increase in their densities and also pulling-in of disconnected proteins. (ii) Even interestingly, there is improvement seen around 0.70, which is the second band, which is mainly due to pulling-in of disconnected proteins into the (denser) complexes. This shows that SPARC has two distinct "bands of impact", each serving the purpose SPARC was devised for.

Further, we notice that there are still a large number of very low density (0.10 and less) complexes that are untouched by SPARC, which fall into the "twilight zone". These complexes lack significantly many interactions or proteins, and therefore call for more effective methods that look beyond interaction networks by combining a wider variety of biological information effectively.

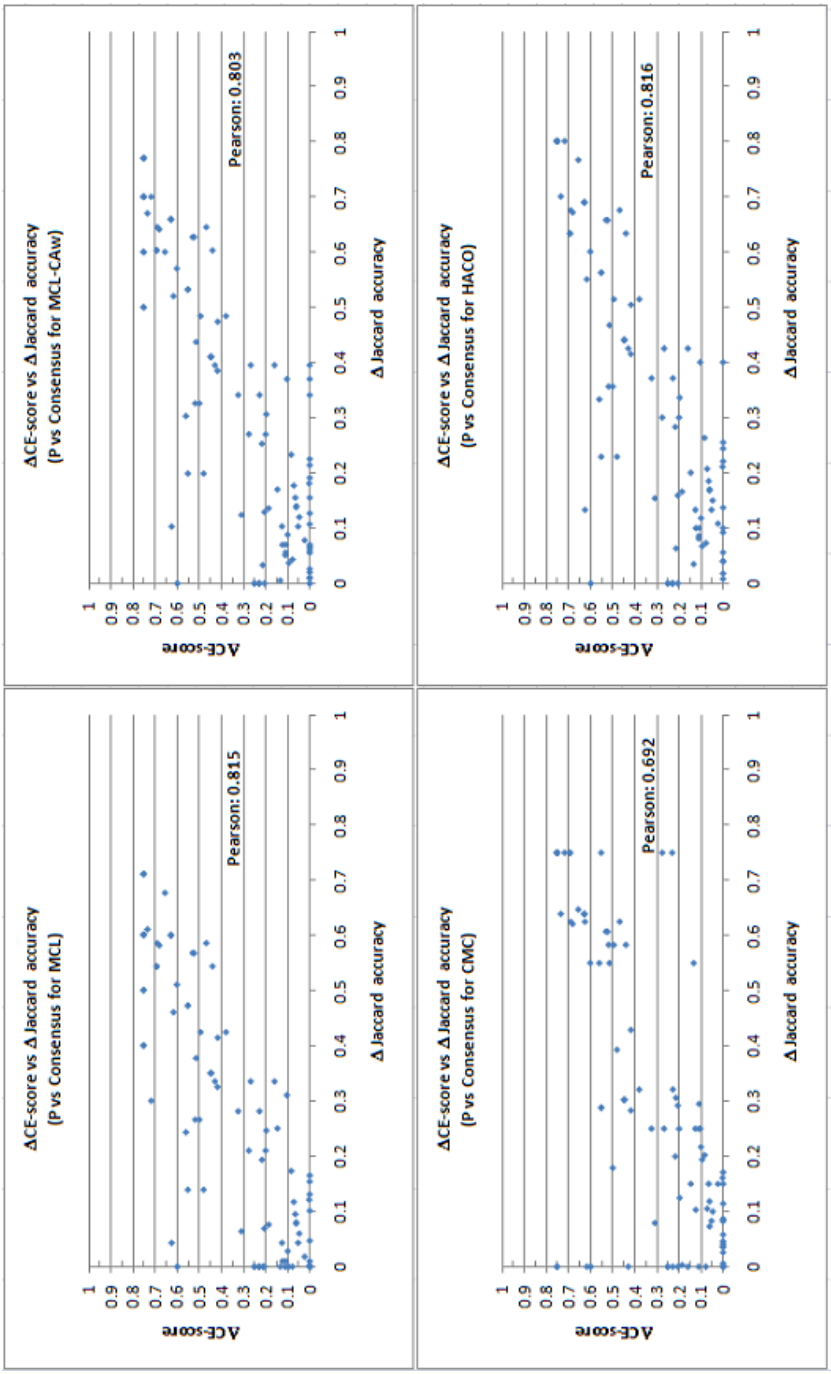


Figure 5.7: Increase in CE-scores of predicted complexes using SPARC-based refinement translates into increase in Jaccard accuracies when matched to benchmark complexes.

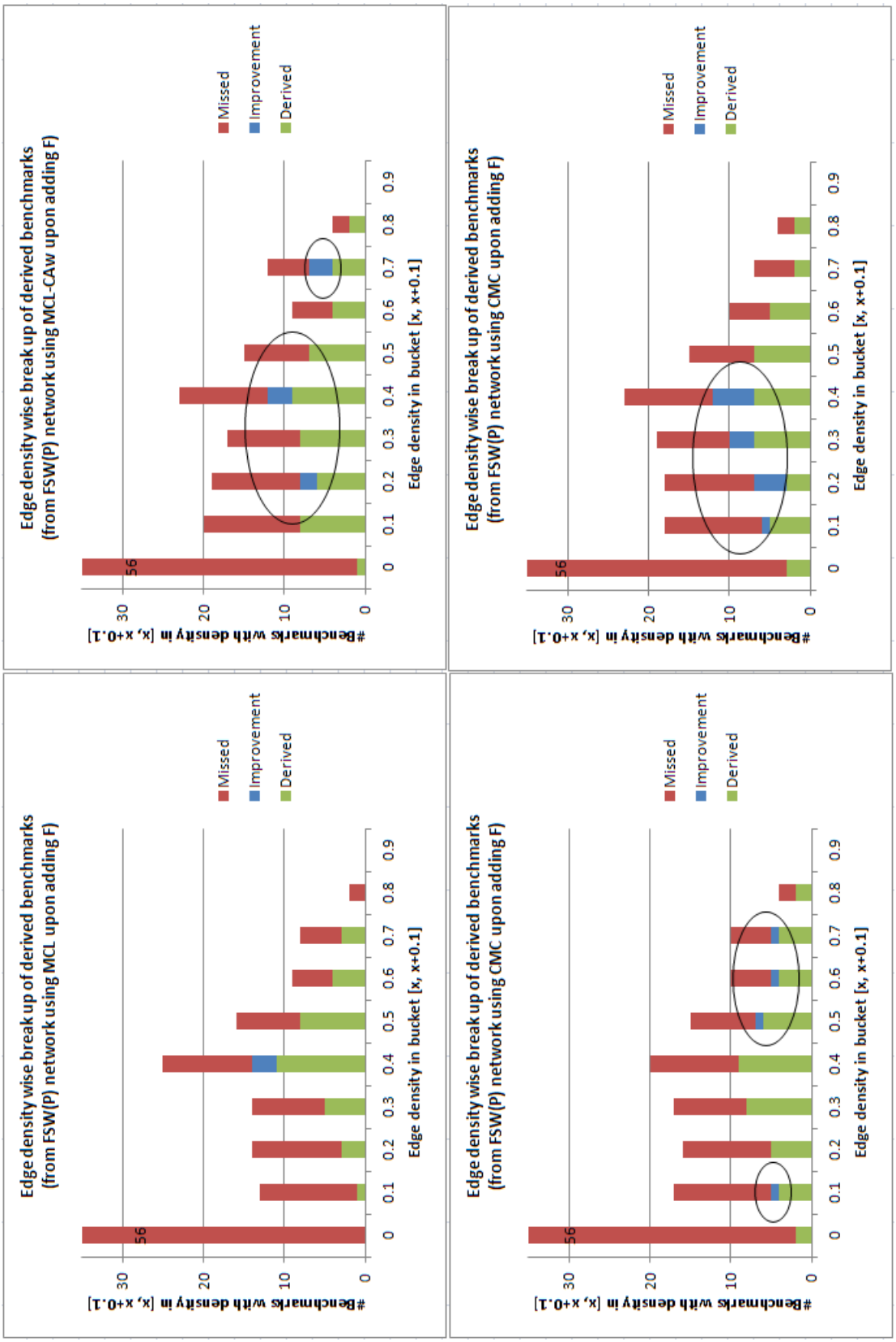


Figure 5.8: An edge density break up of derived complexes from the FSW (P+F) network. There are approximately two distinct "bands of impact" (shown as circles) of SPARC - around the low (0.20) and relatively high (0.70) density complexes.

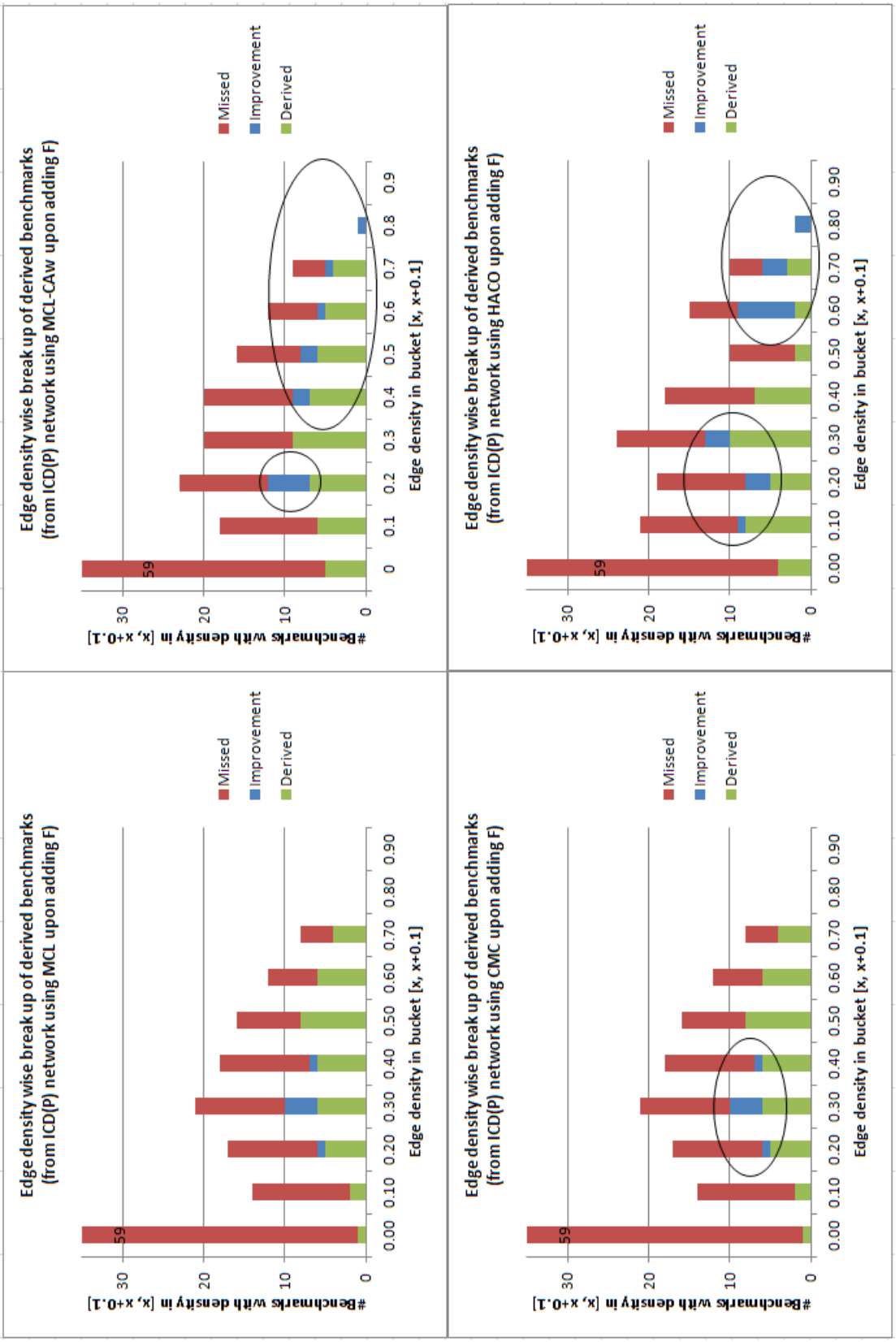


Figure 5.9: An edge density break up of derived complexes from the ICD (P+F) network. There are approximately two distinct "bands of impact" (shown as circles) of SPARC - around the low (0.20) and relatively high (0.70) density complexes.

5.4.5 Sensitivity ranking of complex detection methods

Apart from measuring the qualities of clusters, the CE-score can used in an interesting way to measure the sensitivities of methods for complex detection - the more sensitive a method is the more effective it is on low density networks as well as in countering noise. This can be done as follows. For any given method, we calculate the average CE-score of all complexes detected at "borderline", say with Jaccard accuracies in the range $[0.45, 0.55]$. The lower this average CE-score ($AvgCE$) the more "sensitive" the method is for detecting low density complexes and in countering noise. We can then compare the relative sensitivies of the methods across different networks.

		MIPS			**Wodak**				
Network	Method	$AvgCE$	$1/(AvgCE)$	Norm	$AvgCE$	$1/(AvgCE)$	Norm	Total	Norm
P	HACO	0.35	2.86	1.00	0.32	3.13	1.00	2.00	1.00
	CMC	0.39	2.56	0.90	0.37	2.70	0.86	1.76	0.88
	MCL-CAw	0.41	2.44	0.85	0.40	2.50	0.80	1.65	0.83
	MCL	0.44	2.27	0.80	0.43	2.33	0.74	1.54	0.77
P+F	HACO	0.41	2.44	1.00	0.41	2.44	1.00	2.00	1.00
	CMC	0.44	2.27	0.93	0.43	2.33	0.95	1.89	0.94
	MCL-CAw	0.49	2.04	0.84	0.48	2.08	0.85	1.69	0.85
	MCL	0.56	1.79	0.73	0.55	1.82	0.75	1.48	0.74
ICD(P+F)	CMC	0.31	3.23	1.00	0.31	3.23	1.00	2.00	1.00
	MCL-CAw	0.34	2.94	0.91	0.34	2.94	0.91	1.82	0.91
	HACO	0.36	2.78	0.86	0.35	2.86	0.89	1.75	0.87
	MCL	0.37	2.70	0.84	0.36	2.78	0.86	1.70	0.85
FSW(P+F)	MCL-CAw	0.32	3.13	1.00	0.31	3.23	1.00	2.00	1.00
	HACO	0.36	2.78	0.89	0.36	2.78	0.86	1.75	0.88
	CMC	0.36	2.78	0.89	0.36	2.78	0.86	1.75	0.88
	MCL	0.37	2.70	0.86	0.37	2.70	0.84	1.70	0.85
TCSS(P+F)	MCL-CAw	0.29	3.45	1.00	0.27	3.70	1.00	2.00	1.00
	HACO	0.32	3.13	0.91	0.31	3.23	0.87	1.78	0.89
	CMC	0.36	2.78	0.81	0.35	2.86	0.77	1.58	0.79
	MCL	0.41	2.44	0.71	0.41	2.44	0.66	1.37	0.68

Table 5.12: Relative ranking of methods based on their sensitivities.

Category	Method	Relative score	Normalized score
Unscored	HACO	2.00	1.00
	CMC	1.82	0.91
	MCL-CA	1.67	0.84
	MCL	1.51	0.75
Scored	MCL-CAw	2.91	1.00
	HACO	2.64	0.91
	CMC	2.56	0.88
	MCL	2.49	0.85

Table 5.13: Overall ranking of the methods based on sensitivities.

To go about this we calculated the inverse of average CE-scores ($1/AvgCE$)

of the "borderline" complexes detected from each of the methods on each of the networks, and ranked the methods by normalizing these inverse averages against the best (as done previously in Chapter 4). Table 5.12 shows the relative ranking of the four methods on the P, $P+F$ and the scored($P+F$) networks, while Table 5.13 gives the overall ranking. The tables show that HACO is ranked the best on the P and $P+F$ networks, while MCL-CAw is ranked the best on the scored networks. In other words, HACO is more effective in detecting low density complexes and also in countering noise on raw networks, but when the networks are scored, MCL-CAw is more effective in detecting low density complexes and in effectively making use of the scoring. This agrees with the findings from Chapter 4 (see Tables 4.16, 4.17).

5.4.6 In-depth analysis of detected complexes

We performed in-depth analysis of some of the predicted complexes using *Cytoscape* [97]. For example, the CCR4-NOT complex is a multifunctional complex that regulates transcription, plays a role in mRNA degradation, and also regulates cellular functions in response to changes in environmental signals in yeast [114]. This complex was "scattered" among multiple disjoint components of the Physical network, and therefore went undetected from all four methods. The addition of functional interactions facilitated linking together of these components, enabling the methods to detect it successfully (see Figure 5.10).

While many additional complexes were detected upon employing functional interactions, there were a few complexes that were missed as well. For example, the RNA polymerase complexes I, II and III, that are involved in the formation of RNA chains during transcription [103], were bundled into a large dense module together with some of the TBP-associated factors and TFIID complexes, which are also involved in transcription [115]. Due to the functional similarity between the subunits of all these complexes, several functional interactions were added among them. Consequently, the methods recovered a large dense module housing all these complexes from which the individual complexes could not be segregated. The same was the case with the multi-eIF complexes and the SAGA-SLIK-ADA-TFIID complexes.

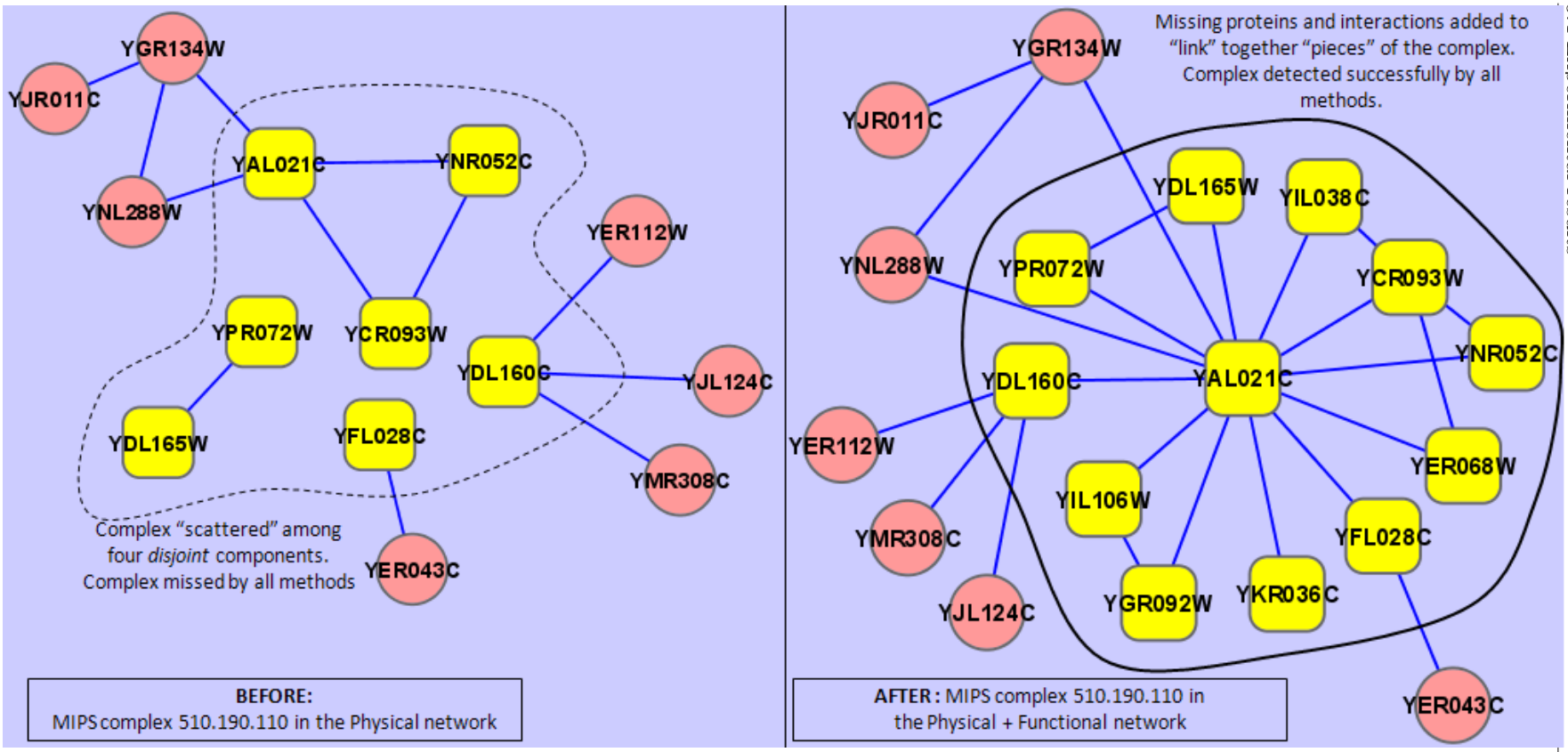


Figure 5.10: MIPS 510.190.110 complex before and after refinement using functional interactions by SPARC, and the effect on its detection using existing methods. BEFORE: The complex was "scattered" among four components; CE-score = 0.1905. AFTER: The four components were linked together into a single component; CE-score = 0.623.

Segregating the amalgamated complexes

The amalgamated clusters do not match benchmark complexes with high Jaccard accuracies causing difficulty in identifying the individual complexes. One way to identify these individual complexes is replace to the Jaccard match criteria by a different criteria as follows. For any amalgamated cluster C and a benchmark complex B, we just measure the proportion of proteins in B covered by C, that is, $P(C, B) = |C \cap B|/|B|$. If $P(C, B) \geq 0.50$, we consider B to be covered by C. Using this criteria, we can get an idea of the individual complexes bundled together within the cluster C.

However, if we wish to explicitly segregate out the individual complexes, we need to post-process these amalgamated clusters. In order to do such a post-processing, we note that amalgamation is caused when too many functional interactions are added across the individual complexes. Therefore, selective removal of these functional interactions is one way to segregate out the complexes (Note: In an alternative approach, Liu et al. [116] removed hubs from the PPI network to prevent methods from amalgamating complexes. This approach showed reasonable performance improvement in CMC, but not in MCL).

For each fused cluster we arrange its functional interactions in non-decreasing order of their interaction weights. Then we repeately remove the first k interactions and reprocess the cluster using the same four methods (MCL, MCL-CAw, CMC and HACO). We apply this procedure for all clusters of size ≥ 20 that are likely to contain more than one benchmark complex as per the above criteria. There are only a very few such fused clusters, hence such a simple method is sufficient to identify the individual complexes. Table 5.14 shows the results of this procedure.

5.5 Lessons from employing functional interactions

In Figure 5.11, we position the detection of sparse complexes using functional interactions into our "bin-and-stack" chronological classification introduced in Chapter 3. We have added an extra "layer" because functional interactions can be inferred from a variety of biological information apart from those already mentioned in the lower layers. The F1-values clearly show that detecting of sparse complexes has

Amalgamated clusters		Post-processing of clusters	
Cluster	Complexes likely present	#Interactions removed	Complexes identified
Cluster 1 (#p 27, #i 280)	SAGA, ADA, SLIK, TFIID	10	SAGA, ADA
		20	SAGA, ADA
		30	SAGA, ADA, TFIID
Cluster 2 (#p 25, #i 198)	Pol I, II, III	10	Pol I, III
		20	Pol I, III
		30	Pol I, III
Cluster 3 (#p 20, #i 144)	eIF1, eIF2, eIF5	10	eIF3
		20	eIF3
		30	eIF3, eIF2, eIF5

Table 5.14: Segregating the individual complexes from amalgamated clusters by removal of functional interactions. Removal of interactions beyond 30 caused clusters to become too sparse to be processed properly.

indeed been a leap forward in improving complex detection.

In spite of these advantages, there can be some obstacles and limitations in utilizing functional interactions. Functional interactions can be considered a "superset" of physical interactions. However, Figure 5.6 seems to be projecting a different picture: very low overlaps between the Physical and Functional datasets. The differential curation of the two datasets - the Physical dataset is curated from experimental techniques, while the Functional dataset is curated from computational techniques - along with the presence of many missing (true negatives) and spurious (false positives) interactions, give rise to these low overlaps. Though this is an observation from only the two yeast datasets considered here, it is worthwhile investigating how far away are we from the "ideal" picture of physical interactions being a proper subset of functional interactions in order to make most effective use of the two.

In addition to these, employing functional interactions can potentially "lump together" several functionally-similar complexes into functional modules, as we saw in the cases of the Pol-I, II, III, and SAGA-SLIK-ADA-TFIID complexes. In fact, Table 5.10 show quite a large increase in the average sizes of predicted complexes indicating that some complexes might potentially be amalgamated together into larger modules. This is because functional interactions are too "general" for identifying only the physically interacting groups of proteins that correspond to complexes within these functional modules. Therefore, functional interactions will need a different treatment from physical interactions in complex detection studies.

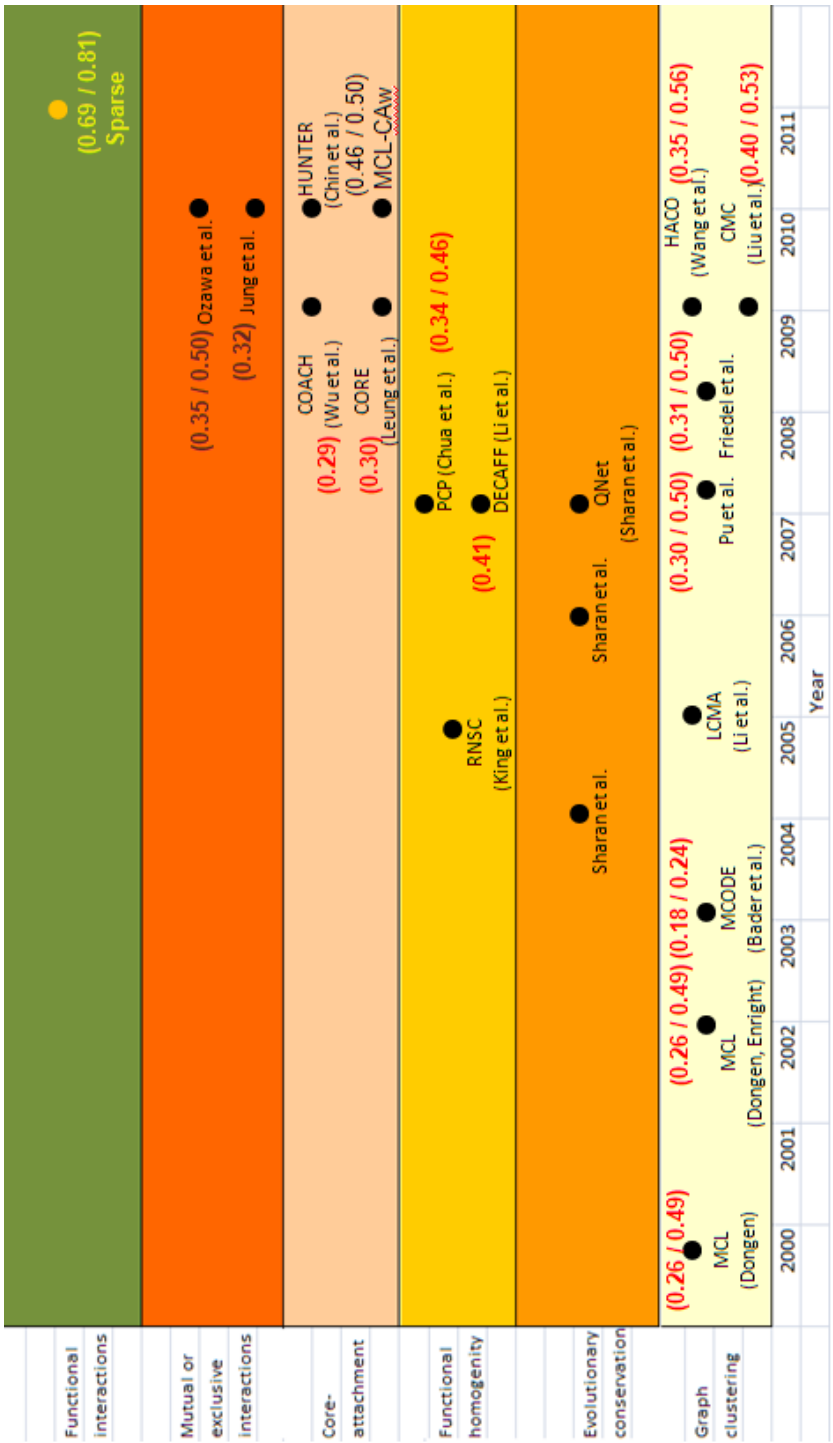


Figure 5.11: Positioning "detection of sparse complexes by adding functional interactions" into the "Bin-and-Stack" chronological classification (all data points with respect to the Gavin + Krogan network scored using Purification Enrichment [36]). Detecting sparse complexes has indeed been a leap forward in complex detection.

The "twilight zone": The very low density complexes, which cannot be detected even with addition of functional interactions, form a "twilight zone", and they call for newer methods that look beyond interaction network topologies by combining a wider variety of biological information effectively.

To conclude here, we say that reasonable progress on complex detection has been done in the previous as well as the current chapter. In the next chapter, we will dwelve into some of the biological insights obtained from deeper analysis of our detected complexes in yeast.

CHAPTER 6

Protein essentiality and periodicity in complex formations

Governing dynamics, gentlemen!

A Beautiful Mind, 2001

Directed by Ron Howard

- *John Nash* played by Russell Crowe

In the previous chapters, we introduced the method MCL-CAw to predict complexes from the yeast physical interactome, and further built upon its capabilities to detect sparse complexes by adding functional interactions using SPARC. We critically evaluated these methods in terms of their precision and recall, and also presented a few case studies on the predicted complexes. However, these evaluations were restricted mainly to the quantitative performance of the methods. In this chapter, we employ the detected complexes for gaining possible novel insights into the cellular machinery, further justifying the applicability of our developed methods.

The PPI network and the complexes predicted from it can provide vital insights into the cellular organization. For example, Wang et al. (2009) [69] utilized the complexes predicted from their method HACO to build a 'ComplexNet', a network of complexes and proteins, in order to study the higher level organization of complexes within the cell. In another study, Vanunu et al. (2010) [117] associated complexes to diseases using physical and functional interactions, and identified a

significant number of disease-related complexes, a study vital to understanding diseases and their cures. More recently, Isoe et al. (2011) [118] found that knock out of individual proteins from the COPI complex disrupts the enzyme secretion process for digestion of blood in mosquitoes. By experimenting on mosquitoes, they found that knocking out of COPI killed 90% of those mosquitoes within two days after feeding on blood, a result very useful to prevent mosquito-borne diseases like dengue, yellow fever and malaria.

An exhaustive study of complexes from the point of view of gaining novel biological insights is out of the scope of this thesis. But, to demonstrate the usefulness of our developed techniques, here we utilize our predicted complexes to understand the roles of protein *essentiality* and *periodicity* in complex formations. These studies will be useful to gain deeper insights into the biological phenomena driving complex formations.

6.1 Role of protein essentiality in complex formations

Some early works by Jeong et al. [10] and Han et al. [119] studied the essentialities of proteins based on pairwise interactions within the interaction network, and concluded that hub (high-degree) proteins are more likely to be essential (the "centrality-lethality" rule [10]). However, a deeper insight can be obtained by studying the essentialities at cluster or group level of proteins rather than pairwise interactions. Recently, Zotenko et al. (2008) [120] argued that essential proteins often group together into densely connected sets of proteins performing essential functions, and thereby get involved in higher number of interactions resulting in their hubness property. Therefore, hubness may just an indirect indicator of protein essentiality. More recently, Kang et al. (2010) [121] studied essentiality of proteins by generating the reverse neighbor (RNN) topology [122] out of protein networks. This topology groups those proteins together that are within the reverse neighborhood of a given protein. Kang et al. concluded that centrality within the RNN topology is a better estimator of essentiality than hubness or degree in the interaction network. Studies by Hart et al. [38] showed that essential proteins are concentrated only in certain complexes, resulting in a dichotomy of essential and non-essential complexes. Wang et al. [69] concluded that the size of the (largest)

recruiting complex of a protein may be a better indicator of protein essentiality than hubness. Pereira-Leal et al. [123] calculated the fraction of essential proteins among proteins found in multiple complexes, and found a consistent trend across different datasets showing a large fraction of multi-complex proteins to be essential.

6.1.1 Our study of protein essentiality in complexes

In our analysis, we try to understand the relationship between the essentiality of proteins and their ability to form complexes. Our analysis is based on the predicted complexes from MCL-CAw from the four PPI networks studied previously (in Chapter 4) namely, the ICD(Gavin+Krogan), FSW (Gavin+Krogan), Consolidated$_{3.19}$ and Bootstrap$_{0.094}$ networks (listed again in Table 6.1).

PPI Network	# Proteins	# Interactions	Avg node degree
ICD(Gavin+Krogan)	1628	8707	10.69
FSW(Gavin+Krogan)	1628	8688	10.67
Consolidated$_{3.19}$	1622	9704	11.96
Bootstrap$_{0.094}$	2719	10290	7.56

Table 6.1: PPI networks used in the analysis of protein essentiality and periodicity

In the first set of analysis, we calculated the proportion of essential proteins present in the complexes, shown in Table 6.2 (the proportion of essential proteins in a complex = #essential proteins/total #proteins in the complex). The table shows that a high proportion (77.65%, 78.03%, 81.34% and 76.35% from the ICD(Gavin+Krogan), FSW (Gavin+Krogan), Consolidated$_{3.19}$ and Bootstrap$_{0.094}$ networks, respecively) of essential proteins present in the networks belonged to at least some complex. This indicated that essential proteins are often members of complexes or co-clustered groups of proteins.

Essential genes in Yeast Genome Deletion Project [124, 125]: 1123

	Number (Proportion) of essential genes present in		
PPI Network	Whole network	Predicted cores	Predicted complexes
ICD(Gavin+Krogan)	604 (0.537)	510 (0.454)	552 (0.491)
FSW(Gavin+Krogan)	604 (0.537)	510 (0.454)	552 (0.491)
Consolidated$_{3.19}$	611 (0.544)	568 (0.506)	576 (0.513)
Bootstrap$_{0.094}$	757 (0.674)	634 (0.564)	676 (0.601)

Table 6.2: Proportion of essential genes in the predicted complexes of MCL-CAw

Next, we binned the complexes based on their sizes and calculated the proportion of essential proteins in all complexes for each bin, shown in Figure 6.1 (a).

The figure shows that essential proteins were present in higher proportions within larger complexes. Next, we calculated the proportion of essential proteins within the top K ranked complexes, shown in Figure 6.1 (b). The figure shows that essential proteins were present in higher proportions within higher-ranked complexes (that is, complexes predicted with higher reliability). Both these figures hint at the same finding: essential proteins come together in large groups, some of which are complexes, to perform essential functions, thereby indicating a strong correlation between the essentiality of complexes and their ability to take part in complexes.

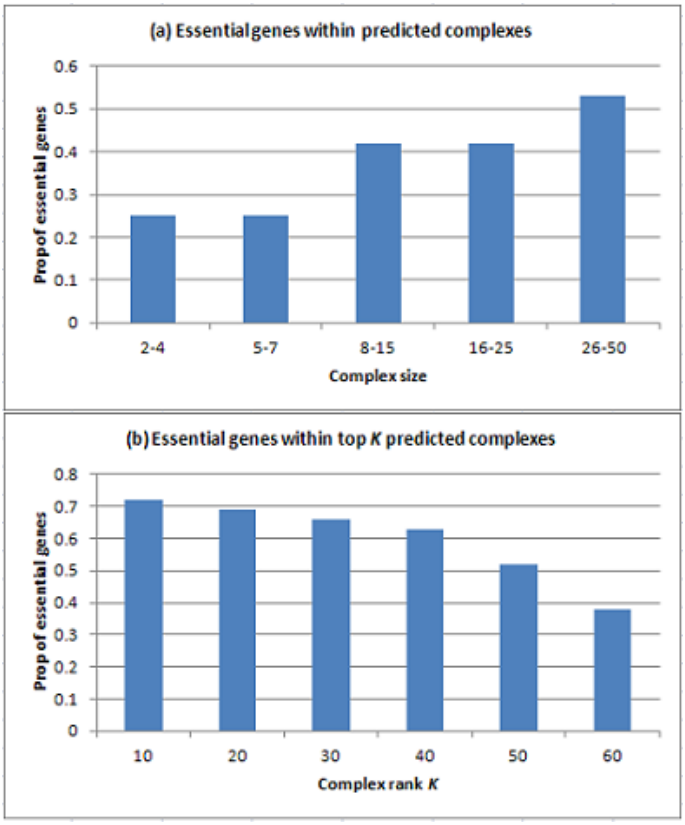


Figure 6.1: Correlation between essentiality of proteins and their abilities to form complexes. Proportion of essential proteins within: (a) complexes of different sizes, predicted from $\text{Consol}_{3.19}$ network; (b) top K ranked complexes.

6.2 Role of protein 'dynamics' in complex formations

In a recent (2010) foresightful survey by Przytycka et al. [85], the application of network *dynamics* (temporal information) into computational analyses is discussed at good lengths, particularly on the analysis of protein interaction networks. The

authors suggest that if sufficient information about the 'timing activities' of proteins can be obtained, the dynamical nature of the underlying organizational principles in interaction networks can be better understood. This shift from static to dynamic network analysis is vital to understanding several cellular processes, some of which may have been wrongly understood due to ignoring dynamic information.

Correlation between topological positioning of proteins in PPI network and their expression profiles

Based on the analysis using a high-confidence yeast PPI network, Han et al. (2004) [119] reported an interesting dichotomy of hubs in PPI networks - 'date' hubs and 'party' hubs. Date hubs interact with a single protein at a given intracellular space and time, while party hubs interact with multiple proteins at the same space and time. Han et al. reported a strong correlation between the topological positioning of these hub proteins in PPI networks and their expression profiles - party hubs are 'modular' and are highly coexpressed with their neighbors, while date hubs are 'central' and are not coexpressed with their neighbors. Though this finding was critically questioned by Batada et al. [126], the existence of such dichotomy is now increasingly being accepted [127, 128], and it paved the way for simultaneous analysis of topologies of networks and their dynamics.

Recently (2007), Komurov et al. [128] studied how proteins with different expression dynamics were positioned in the yeast PPI network. Komurov et al. calculated the statistical expression variance (EV) of each gene in the yeast genome across 272 experiments compiled from SGD [94]. An EV close to 0 indicated a gene with lowest variance (least dynamic), while an EV close to 1 indicated a gene with highest variance (most dynamic). Using a high-confidence PPI network comprising of 5456 interactions among 2315 proteins, Komurov et al. compared the EVs of proteins with their neighbors in the network, and found a strikingly high correlation between EVs of proteins and their neighbor EVs. This suggested that proteins had similar expression dynamics as their immediate neighbors in the network. This confirmed earlier findings (2001) [129] that co-regulated proteins frequently interacted with each other. Carrying this forward, Komurov et al. extended the date-party hub hypothesis of Han et al. [119] by proposing 'family' hubs. Komurov et al. reported that

family hubs were always present in the network and interacted with their neighbors constitutively, while party hubs were dynamically coexpressed with their neighbors with which they interacted. Therefore, family hubs formed 'static modules' and party hubs formed 'dynamic modules', whereas date hubs organized the network. Furthermore, they reported that these static and dynamic modules were enriched with specialized functions.

Yu et al. (2007) [130] studied the topological positioning of hubs in the yeast PPI network, and said 'date' hubs show high betweenness and are therefore inter-modular, while 'party' hubs show high clustering coefficient and therefore intra-modular.

More recently (2011), Patil et al. [131] classified hubs in PPI networks using a combination of gene co-expression correlation and co-expression stability among interacting proteins. The co-expression stability measures the extent to which a pair protein is constitutively co-expressed, that is, how "stable" is the co-expression. Based on these two measures, Patil et al. found that hubs showing high co-expression correlation as well as high stability (which they call Category 1 hubs) with their neighbors were likely to be intra-modular, while hubs showing low co-expression correlation but high stability (Category 2 hubs) with their neighbors were likely to be inter-modular. Many of the Category 2 hubs were involved in transient interactions, and corresponded to 'date' hubs.

The 'dynamics' of complex formation during the yeast cell cycle

de Lichtenberg et al. (2005) [132] studied the dynamics of complex formations during the yeast cell cycle. They constructed a PPI network comprising of 300 proteins (184 dynamic and 116 static) using Y2H and TAP/MS screens. Extraction of complexes from these screens and comparisons with known complexes from MIPS [90] revealed 29 heavily intraconnected modules (complexes or complex variants) that existed at different "time points" during the cell cycle. Further, most complexes contained both constitutively expressed (static) as well as periodically expressed (dynamic) proteins. More interestingly, almost all eukaryotic complexes were *assembled* just-in-time contrary to the just-in-time *synthesis* observed in bacteria. Just-in-time assembly meant that most subunits of complexes were pre-transcribed,

while some subunits were transcribed when required to assemble the final complex (see Figure 6.2). This was more advantageous than just-in-time synthesis because only a few components of entire complexes had to be tightly regulated to control the timing of the final complex assembly. Holding off on the last components enabled the cell to prevent "switching on" of complexes at the wrong times.

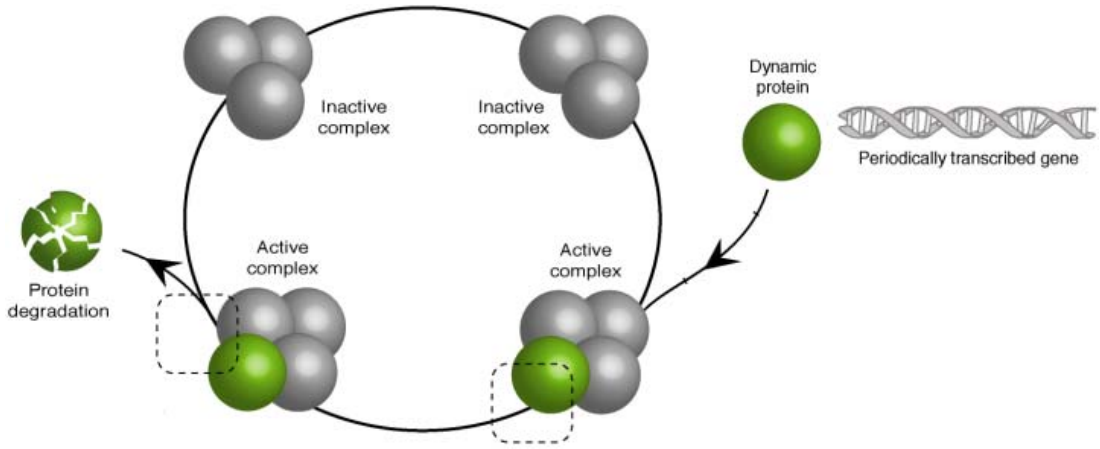


Figure 6.2: "Just-in-time assembly" of eukaryotic complexes, adopted from [132]. The periodically transcribed protein (in green) assembles with static proteins (in grey) to form an active complex.

More recently (2009), Wu et al. [133] partitioned a high-confidence PPI network into four "phase sub-networks" based on the cell-cycle phases ($G1$, S, $G2$ and M) in which the dynamic proteins were transcribed. They analysed the properties of hubs within these sub-networks and found that only 69% of the hubs still acted as hubs in at least one of the four sub-networks. They also investigated the dynamic properties of the anaphase-promoting complex and the chromatin-remodeling complex, and found a network-based explanation for the dynamic assembly of these two complexes during the yeast cell cycle.

6.2.1 Our study of protein 'dynamics' in complexes

It is possible to correlate and study the topological positioning and temporal behavior of proteins by combining PPI network topology and gene expression data, as we saw in the reviewed works above. However, a deeper insight can be obtained by studying proteins in larger groups than just pairs of neighbors in the network. Therefore, here we study the temporal behavior of proteins via their complexes.

To make the analysis simpler, we first "discretize" the expression for each protein based on the yeast cell cycle phase ($G1 \rightarrow S \rightarrow G2 \rightarrow M$) in which the expression is maximum. We call this discretization procedure as *Peak Expression Discretization* (PED). This makes the analysis simpler because we can now assign a single 'phase' to each protein in any given complex, and study the order of assembly and disassembly of that complex - the ordered sequence in which the proteins get together to assemble into the final complex and disassemble after that.

For computing these phases we took the aid of the Cyclebase database (`http://www.cyclebase.org/`) [134]. Cyclebase averages gene expression data obtained from multiple microarray studies to compute the approximate phase of peak expression for each protein (see Figure 6.3). If a protein is expressed maximum in all four phases, that is, it shows constitutive expression, it is labeled 'static', else it is labeled 'dynamic' along with the phase in which it expresses maximum. As of September 2010, the database has 6114 yeast proteins, out of which 5514 are labeled 'static', and the remaining 600 are 'dynamic'. Out of these 'dynamic' proteins, 576 have a peak time, while the remaining 24 are labeled 'uncertain'.

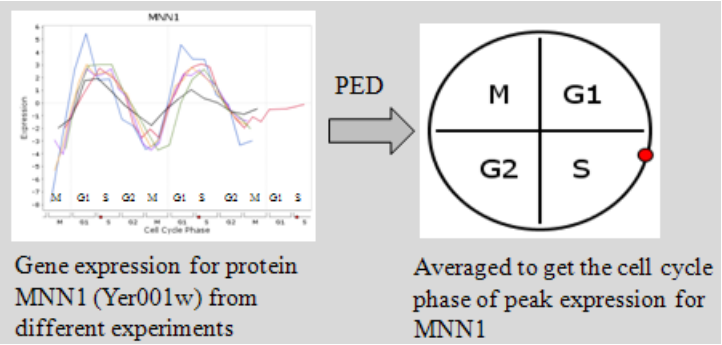


Figure 6.3: Peak Expression Discretization (PED) for a protein with respect to the yeast cell cycle phases (taken from Cyclebase [134])

For a start, we integrated the computed cell cycle phases of proteins onto the PPI network and performed a brief study of network dynamics, as shown in Table 6.3. The table shows that the interactions among static proteins dominated the network (for example, 94.69% in $\text{Consol}_{3.19}$). The static-dynamic and dynamic-dynamic interactions formed comparatively smaller portions of the networks (for example, S-D: 4.6% and D-D: 0.716% in the $\text{Consol}_{3.19}$ network). Among the 64

dynamic-dynamic interactions in the $\text{Consol}_{3.19}$ network, 42 were "intra-phase", that is, among dynamic proteins that peaked during the same phase, while the remaining 22 were "inter-phase", that is, among dynamic proteins that peaked during different phases.

When we examined the static-dynamic interactions in detail, we noticed many of the static proteins were involved in transient interactions with dynamic proteins expressed in different phases. These static proteins were enriched in a variety of GO terms, the prominent ones being signal transduction and transcription. This revealed the "multipurpose" nature of these static proteins. This also indicated that 'staticness' or constitutive expression might be linked to the potential ease in "reusability" of such multipurpose proteins.

Network	# Proteins		# Interactions				
	Total	Annotated	Total	Annotated	S–S	S–D	D–D
ICD(G+K)	1628	1613	8707	8296	7612	363	42
FSW(G+K)	1628	1613	8688	8296	7612	363	42
$\text{Consol}_{3.19}$	1622	1613	9704	8941	8466	411	64
$\text{Boot}_{0.094}$	2719	2142	10290	9723	8997	518	79

Table 6.3: Analysis of 'dynamism' in four yeast PPI networks. "Annotated" refers to labeled as 'static' or 'dynamic' in the Cyclebase database [134].

A workflow for studying 'dynamics' in protein complexes

Next, we performed our intended study on protein complexes using cell cycle phase information. The workflow for this study is shown in Figure 6.4. Essentially, we collated the predicted complexes and integrated the phase data (from PED) with these complexes to study their dynamic assembly and disassembly.

A case study of cyclin-CDK complexes:

We first present a case study illustrating complexes formed by the kinase Cdc28. Upon clustering the Consolidated network using MCL-CAw, we obtained the following cluster containing Cdc28 (Ybr160w): {Ybr160w, Ygr108w, Ypr119w, Ydl155w, Ylr210w, Ypr120c, Ygr109c, Ymr199w, Ypl256c, Yal040c}. When we added the cell cycle phase data to the proteins in this cluster, we noticed that the proteins were expressed during different phases: Ybr160w - Static, Ygr108w - M, Ypr119w - G_2, Ydl155w - S, Ylr210w - S, Ypr120c - G_1, Ygr109c - G_1, Ymr199w - G_1/S, Ypl256c

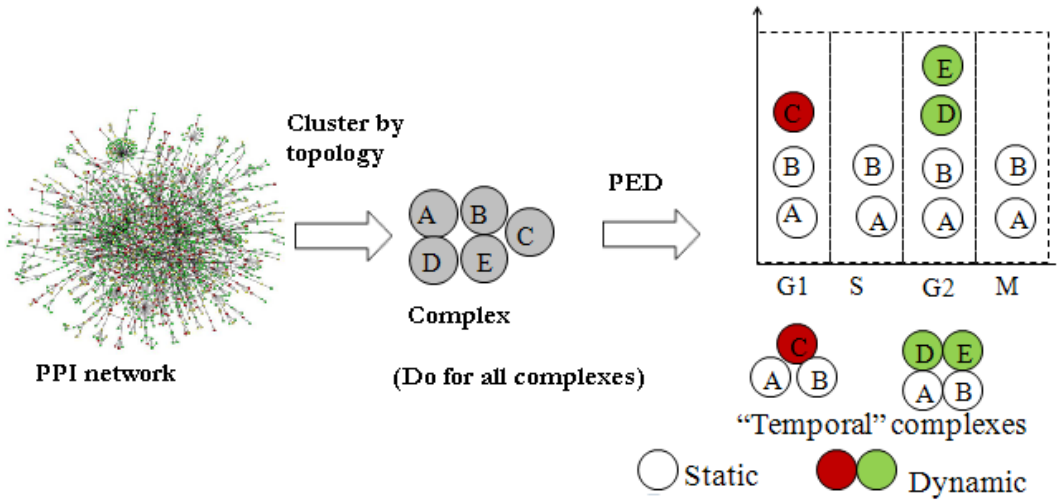


Figure 6.4: A high-level workflow to study dynamics of protein complex formations

- G_1, and Yal040c - M (see Figure 6.5). This revealed the possible existence of multiple 'time-based' complexes within this large cluster. Validation against SGD [94] confirmed that Cdc28, also named Ybr160w, is a *cyclin-dependent kinase* (CDK) that participates in multiple complexes with its *cyclin* partners. And these SGD complexes matched the ones we segregated from the cluster by incorporating cell cycle phase data (see Figure 6.5).

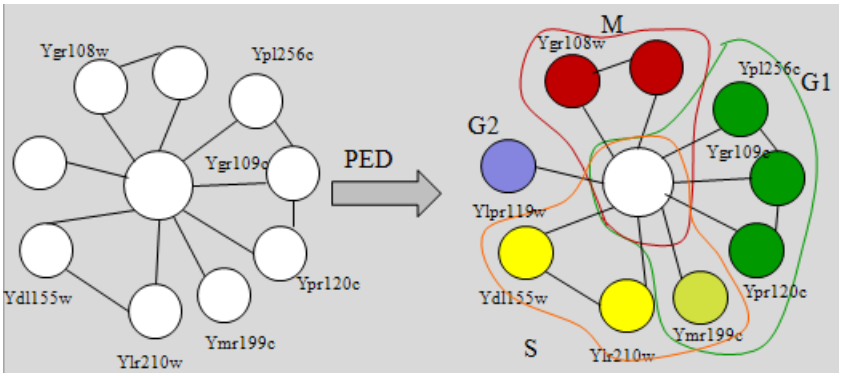


Figure 6.5: Cdc28 and its cyclin-dependent complexes identified by incorporating cell-cycle phase information. Cdc28 is temporally "reused" among the complexes.

The *CDK-cyclin complexes* control the passage through the cell cycle in yeast, and are comprised of cyclins, the regulatory subunits, and CDKs, the catalytic subunits [135]. The concentrations of cyclins increase and decrease as the cell progresses through the cell cycle, while the concentrations of CDKs do not fluctuate

in such a characteristic manner, but they have no kinase activity unless they are associated with a cyclin. The CDKs associate with different cyclins to form cyclin-CDK complexes that activate or inhibit several proteins involved in the cell-cycle progression [135].

This case study revealed two interesting insights:

- Small complexes (such as the Cdc28-cyclin complexes) can be identified by incorporating additional information (in this case, cell cycle phases) into PPI network topology;
- The 'static' Cdc28 kinase is *temporally "reused"* across multiple complexes indicating a possible link between 'staticness' and "reusability" of proteins across complexes.

A global study of temporal "resuability" of proteins in complexes

In order to further understand the temporal "reusability" of proteins, we next performed a global study of all predicted complexes from the yeast PPI network using MCL-CAw. We performed this study using the "core-attachment" model of complexes proposed by Gavin et al. [15] and adopted in MCL-CAw.

As per the "core-attachment" model of complexes (see Chapter 4 for details of the model), the core proteins are the main functional units with which the attachment proteins collaborate to form complexes. These attachment proteins may be shared during the formation of multiple complexes. Among these attachments are tightly-coupled subsets of proteins called "modules" that are shared in groups as a whole across these complexes. Therefore, we expect these shared proteins (that is, attachments and modules) to be enriched higher in 'staticness' compared to cores within complexes. This is to allow for temporal "reusability" of shared proteins among complexes (see Figure 6.6). We state this as our hypothesis,

Hypothesis 6.1 *We expect 'staticness' to be more enriched in attachments compared to cores in complexes.*

Testing our hypothesis: Let $\lambda_s(X)$ denote the number of static proteins in set X, and $\lambda_d(X)$ denote the number of dynamic proteins in X. Using this, we define the *enrichment* E for static (dynamic) proteins among attachments and cores in the set

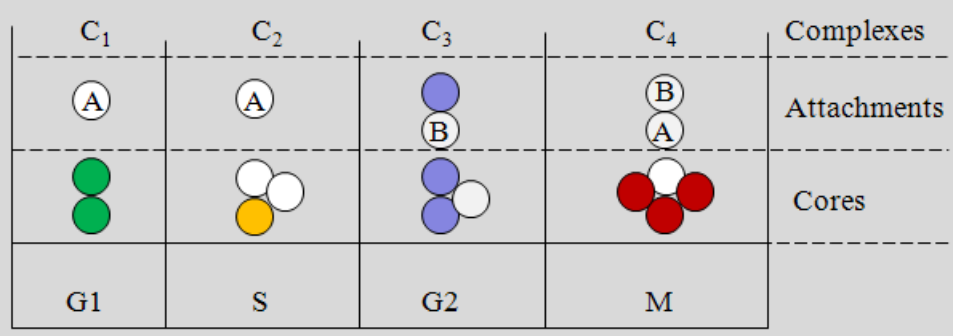


Figure 6.6: Relating the "core-attachment" model to temporal "reusability": we expect the attachment proteins, which are more likely to be shared among complexes, to be more enriched in 'staticness' compared to the core proteins.

of complexes $\mathcal{C}$ as follows. For a complex $C \in \mathcal{C}$ the enrichment in the attachments $Attach(C)$ is,

$$E_s(Attach(C)) = \frac{|\lambda_s(Attach(C))|}{|\lambda_s(C)|}, \tag{6.1}$$

$$E_d(Attach(C)) = \frac{|\lambda_d(Attach(C))|}{|\lambda_d(C)|}. \tag{6.2}$$

Therefore, the *relative enrichment* $RE(Attach(C))$ of static to dynamic proteins in the attachments in C is,

$$RE(Attach(C)) = \frac{E_s(Attach(C)}{E_d(Attach(C)}. \tag{6.3}$$

The enrichment and relative enrichment for cores is defined in a similar way. See an example in Figure 6.7. The overall enrichment and relative enrichment for $\mathcal{C}$ is obtained by averaging over all complexes.

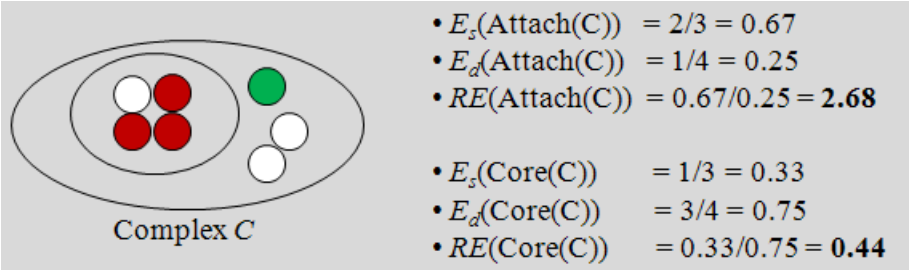


Figure 6.7: Calculating enrichment E and relative enrichment RE.

Table 6.4 shows these values for the predicted (annotated) complexes from

four yeast PPI networks. The relative enrichment RE values for complexes predicted from the highly reliable Consolidated network were $RE(Attach) = 3.402$ and $RE(Core) = 0.839$. This shows that the attachment (the shared) proteins were enriched higher in 'staticness' compared to core proteins, thus supporting our hypothesis. When we mapped some of these complexes back onto the PPI network, we found many of the shared 'static' proteins to be involved in "multiphase" interactions - several dynamic proteins transcribed in different phases interacted with these shared 'static' proteins to form periodic complexes. In other words, the static proteins formed "anchors" for periodic cores to form periodic complexes. These findings supported the biological design principle of temporal "reusability". The sharing of static proteins among complexes instead of the dynamic proteins ensured maintenance of the generic proteins throughout all phases for their "reusability", while only the periodic proteins had to be transcribed just-in-time to assemble the required complexes, which agreed with the findings by de Lichtenberg et al. [132]. We analysed some of these shared 'static' proteins and found many to be *kinases* that were involved in activating or deactivating several complexes (for example, Cdc20 involved in deactivating the Anaphase Promoting Complex/Cyclosome) during the yeast cell cycle.

PPI Network	#Complexes (annotated)	Enrichment E			
		Attach		Core	
		Static	Dynamic	Static	Dynamic
ICD(G+K)	49	0.523	0.179	0.442	0.509
FSW(G+K)	48	0.518	0.177	0.442	0.512
$Consol_{3.19}$	57	0.626	0.184	0.445	0.530
$Boot_{0.094}$	52	0.661	0.192	0.562	0.586

Table 6.4: Enrichment of static and dynamic proteins among attachments and cores of annotated complexes from yeast PPI networks.

Table 6.4 also shows that the enrichments of static and dynamic proteins in cores were almost equal, indicating that both static as well as dynamic proteins were equally capable of being part of cores. These are *specialized* (non-reused) sets of proteins that may be either static or dynamic, and form functional parts of complexes. This agreed with the findings by Komurov et al. [128] that both static as well as dynamic proteins were capable of forming functional modules - the static proteins formed 'static modules' while the dynamic proteins formed 'dynamic

modules', both of which were involved in vital functions of the cell.

Relating our findings to previous studies

Based on the analyses here, we relate our findings to previously discussed studies on combining PPI network and gene expression data by Han et al. [119], Kumorov and White [128], Yu et al. [130] and Patil et al. [131], and the work on essential proteins by Pereira-Leal et al. [123]. We classify proteins based on participation in complexes into static "reused" and static/dynamic specialized (non-resused) proteins, and relate this classification to that of hubs by the previous works, as show in Table 6.5.

	Reused	Specialized	Previous works
Static	'Date' hubs Inter-modular Category 2 Essential	'Family' hubs Intra-modular Category 1	Han et al., 2004; Komurov and White, 2007 Yu et al., 2007 Patil et al., 2011 Pereira-Leal, 2006
Dynamic		'Party' hubs Intra-modular	Han et al., 2004; Komurov and White, 2007 Yu et al., 2007

Table 6.5: Relating our classification of based on participation in complexes into static "reused" and static/dynamic specialized proteins to the classification of hubs by previous works

The hub proteins that Han et al. and Kumorov and White categorized as 'date' and 'party' hubs correspond to the static "reused" proteins and the dynamic specialized proteins within complexes, respectively, in our study. The static "reused" proteins among complexes interact transiently with different sets of proteins to form different complexes (for example, Cdk kinases), and thereby correspond to 'date' hubs. The dynamic proteins get together to form dynamic complexes at a particular time and disintegrate after that correspond to the 'party' hubs (for example, dynamic proteins forming the APC/C complex in G1/S phases). The 'family' hubs of Kumorov and White correspond to the static specialized proteins that form (static) complexes (for example, the ribosomal complexes). Further, the Category 2 and Category 1 hubs of Patil et al.'s studies correspond to our static "reused" and static specialized proteins, respectively. Relating to Yu et al.'s characterization of hubs into inter-modular and intra-modular, we note that the static "reused" hubs are shared among complexes and therefore inter-modular, while the static/dynamic

specialized hubs are found within complexes and therefore intra-modular. Finally, relating to Pereira-Leal et al.'s findings, we note that many of our "reused" proteins are kinases, which are known to be essential proteins.

To summarize, our study provides alternative explanations and additional evidence based on participation in complexes to the classification of hubs from previous studies.

A novel putative role for RAD53 in polarized cell growth

Incorporating phase information into complexes also led us to provide further evidence for a novel putative role for the kinase Rad53. Rad53 is known to be involved in DNA damage/replication response - required for cell-cycle arrest in response to DNA damage, and also plays a role in DNA replication [94]. We combined a recent yeast dataset enriched in interactions involving kinases-phosphatases from Breitkreutz et al. (2010) [47] with the Gavin+Krogan network (from Chapter 4), and analysed the Rad53-mediated complexes derived from this combined network.

MCL-CAw derived a cluster comprising of Rad53 and the Septins, which indicated a possible role of Rad53 in mediating the Septins (see Figure 6.8). Septins are proteins known for their roles in cytokinesis, they form a ring-like scaffold at the mother-bud neck to recruit proteins to form complexes during cytokinesis [94]. However, we could not find any complexes containing Rad53-Septins in the Wodak-lab [92] and MIPS [90] databases, and neither any evidence in SGD [94] or GO [37] for the combined roles. Detailed literature search revealed that very recently Wang et al. (2009) [136] noticed interactions between Rad53 and Sep7 (Ydl225w) and hypothesized the role of Rad53 in polarized growth via the Septins. Our observations provided further evidence to support Wang et al.'s hypothesis.

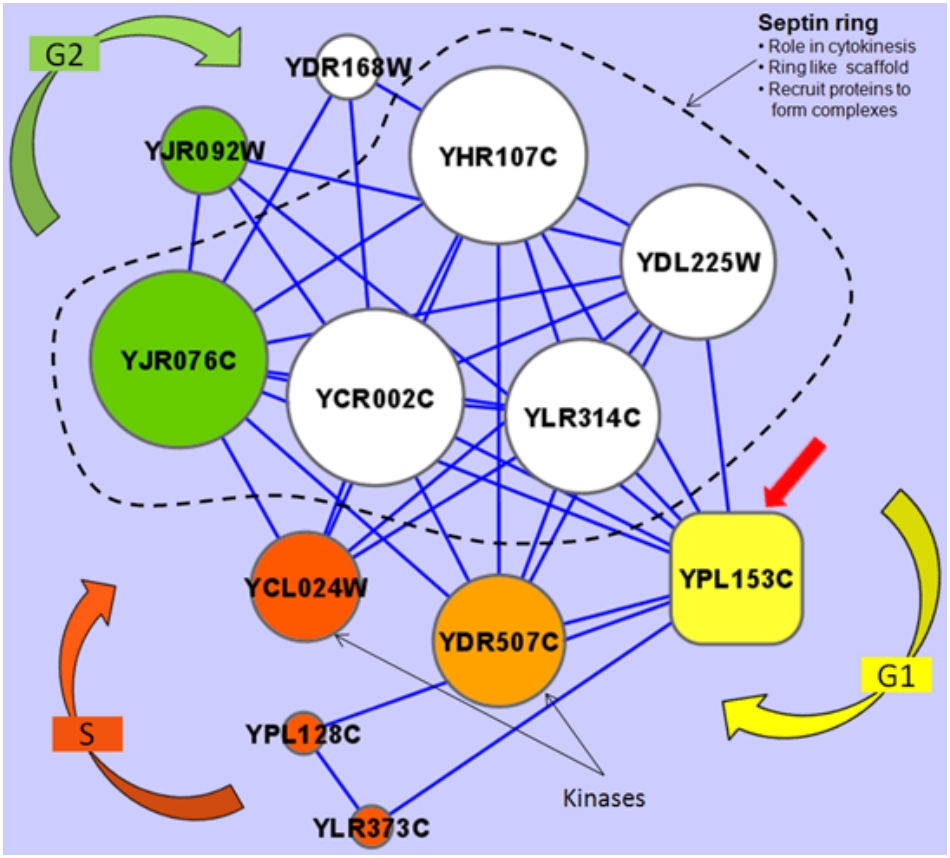


Figure 6.8: A cluster comprising of Rad53 (Ypl153c) and the Septins indicated a possible role of Rad53 in mediating the Septins. This was also observed by Wang et al. [136], who hypothesized that Rad53 may have a role in polarized cell growth via the Septins.

6.3 Concluding remarks

In this chapter, we utilized the predicted complexes from MCL-CAw to gain deeper insights into some of the cellular mechanisms driving complex formations. The investigations in this chapter were basic and non-exhaustive, yet they hinted at interesting biological phenomena driving complex formations, reflecting the applicability of our computational methods in deciphering the cellular machinery.

This chapter provides an apt sign off to MCL-CAw and other techiques developed in this thesis, and paves the way for the final conclusion in the next chapter, where we summarise the significance of main results in this thesis and list possible ventures for further research on related problems.

CHAPTER 7

Conclusion

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but "Thats funny."

- *Issac Asimov*, as quoted in [137]

Protein complexes are one of the fundamental functional units responsible for many biological mechanisms within the cell. Their identification is therefore necessary to understand the cellular organization and machinery. With the advent of "high-throughput" techniques in molecular biology (some of them reviewed in Chapter 2) large-scale identification of interactions among proteins has become feasible, which in turn has paved the way for *in silico* discovery of complexes from protein interaction networks. Over the last few years, many computational methods have been developed for detecting complexes of organisms such as yeast (exhaustively reviewed in Chapter 3). Even though promising, complex detection still requires careful attention in handling errors and noise and reconstructing complexes with high accuracies. In this respect, this thesis focused on devising and developing several techniques and algorithms for accurate complex detection. The results shown in this thesis were motivated by the following desirable properties:

1. Detecting possibly all complexes and with high accuracies.
2. Effective countering of noise observed in experimental datasets.

3. In-depth analyses of detected complexes to gain deeper and possibly novel insights into biological phenomena.

In order to achieve the aforesaid results, the thesis proposed effective methods (Chapters 4, 5 and 6) that integrate a variety of biological information with the topology of PPI networks to detect complexes.

7.1 Significance of the main contributions

Specifically, this thesis contributed several new principles and procedures of inquiry into complex detection, which can be summarized as follows:

1. *A 'foresightful' survey and taxonomy of computational methods:* Chapter 3 presented an elaborate taxonomical survey of techniques developed for complex detection over the last decade. Though there have been several surveys from time-to-time [86–88], a taxonomy that provides a "sense of time" - when the methods were developed and links them to experimental improvements - has not been presented in any of these works. Our taxonomy condensed the history of complex prediction, and we believe has the potential to "drive" future research by providing vital insights. For example, it revealed that incorporating biological information and capitalizing on reliability scoring significantly boosts up performance, an insight that inspired us to develop MCL-CAw (in Chapter 4). Further, as and when we developed new techniques in this thesis, we positioned them as new "data points" and/or "layers" in our proposed taxonomy further reinforcing its usefulness.

2. *A new complex detection method using core-attachment structure:* Inspired by the core-attachment modularity structure revealed by Gavin and colleagues [15], Chapter 4 presented a novel complex detection method (called MCL-CAw) incorporating the core-attachment insight into the topological structure of PPI networks. We demonstrated that MCL-CAw performed better or at least as good as several recent methods, and showed consistently good performance across a variety of reliability scoring schemes.

   The thesis presented the first ever comprehensive comparison of complex detection methods across networks scored using a variety of scoring schemes.

Doing so demonstrated that scoring boosts the performance of methods.

3. *A quantitative definition to the notion of complex "derivability":* Through our Component-Edge (CE) score (introduced in Chapter 5), the thesis presented a quantitative measure of complex "derivability" that correlates better with actual prediction accuracies compared to several previously adopted measures like absolute edge density. The CE-score says that even if a complex has low absolute edge density, but a significant portion of its proteins are connected within the same component and its edge density is higher relative to its immediate neighborhood, then the complex has a high chance of being detected. The CE-score therefore helps to identify likely factors that influence complex derivability. Such a score certainly has strong applicability in developing future complex detection algorithms.

4. *Detection of sparse complexes and the use of functional interactions:* Sparse complexes have been hardly studied in prior works mainly due to the inherent assumption that complexes are embedded within dense regions of the network, which may be weak in the wake of insufficient PPI data. In Chapter 5, the thesis presented a novel characterization of sparse complexes using the CE-score. Further, it presented a novel algorithm SPARC employing functional (logical) interactions to detect sparse complexes. This is a novel contribution because it looks beyond physical interactions to bolster PPI networks for detecting complexes.

5. *Novel biological insights deciphered on complex formations:* Finally (in Chapter 6), utilizing the complexes detected from MCL-CAw, the thesis presented two interesting insights into complex formations in yeast: (i) A strong correlation between the essentiality of proteins and their ability to form complexes; and (ii) The relatively higher enrichment of 'staticness' (constitutively expressed) in proteins shared among 'time-based' complexes hinting towards the biological design principle of temporal "reusability" of 'static' proteins for complex formations.

Therefore, this thesis is a valuable contribution in the area of computational molecular and systems biology.

7.2 Limitations of the research

All the experimental results, analysis and inferences in this thesis are based on complexes detected for *Saccharomyces cerevisiae* (yeast). This is because yeast is the most widely studied organism with fairly complete data available on interactions for computation, and *bona fide* complexes for evaluation, and also auxiliary information such as literature reports, gene annotation data, cell-cyle phase data, etc. for further analysis. Though the studies on yeast are an important step towards understanding eukaryotic cellular mechanisms, it is vital to perform similar studies on higher eukaryotes such as human. The identification, cataloguing and comparative analysis of human complexes will be vital to understand novel biological phenomena, causes and cures of diseases and in drug development. Based on this we recommend the following avenues for future research.

7.3 Recommendations for further research

1. *Detection of complexes from other organisms, particularly human:* An important avenue for research is to test the current methods and where necessary develop new methods to detect complexes from the human interactome. From a technical aspect, this can involve "retrofitting" of current methods onto human datasets, which as things stand currently, are significantly sparser as well as noisier than yeast datasets. However, the analysis required on the predictions from human datasets can be very different from yeast. To give an essense, the methods on yeast can be evaluated by calculating the recall (sensitivity) against a 'gold standard' set of yeast complexes (like MIPS [90] and Wodak CYC2008 [92]) because this 'gold standard' is fairly complete. However, in the case of human, the 'gold standard' is largely incomplete and hence even a very high recall may not make much sense. Instead alternative validation of the unmatched predictions will be more crucial in order to identify novel complexes that can potentially complete the 'gold standard'.

   Another interesting aspect to study here is the evolutionary conservation of complexes across organisms. Several interesting questions and hypotheses can

be put forth here that can provide vital insights into complex evolution - for example, whether "core" proteins are more conserved than "attachment" proteins?; how do complexes lose or gain components during evolution and how is the 'rewiring' done?; etc.

2. *Prediction of membrane protein complexes:* The focus of this thesis has been the identification of all possible complexes from the interactome. However, interaction networks corresponding to specific proteins can be isolated, and complexes involved in specific functions can be detected and studied. Such focused studies are vital for understanding specific phenomena that may not be general across all complexes.

   For example, the conventional Y2H and TAP/MS screens are not effective in detecting membrane, membrane-associated and extracellular protein interactions (see Chapter 2). This is because Y2H is confined to the nucleus for testing interactions, while TAP screens cannot capture membrane complexes due to the insoluble or hydrophobic nature of membrane proteins as well as the ready dissociation of subunit interactions. To counter these limitations, recently specialized screens like split ubiquitin-based membrane Y2H system (MYTH) have been developed to identify interactions among membrane proteins [138]. The study of the membrane protein subinteractome will be useful to identify membrane complexes. The formation of membrane complexes involves chaperone-assisted ordered assembly of intermediaries, as well as a complicated mechanism of 'dynamic exchange' of proteins among the complexes, phenomena which cannot be understood by studying the entire set of complexes in general.

Bibliography

[1] Ezzel, C.: **Proteins Rule**. *Scientific American* 2002, **286(4)**:40–47.

[2] Alberts, B.: **The cell as a collection of protein machines: preparing the next generation molecular biologists**, *Cell* 1998, **92(3)**:291–294.

[3] Baker, T.A., Bell, S.P.: **Polymerases and the replisome: machines within machines**. *Cell* 1998, **92(3)**:295–305.

[4] Hartwell, L.H., Hopfield, J.J., Leiber, S., Murray, A.W.: **From molecular to modular cell biology**. *Nature* 1999, **402**:C47–C52.

[5] Winther, R.G.: **Varieties of modules: kinds, levels, origins, and behaviors**. *Journal of Experimental Zoology* 2001, **291**:116–129.

[6] Bryson, B.: **A Short History of Nearly Everything**. *Black Swan Books* London, 2003:501.

[7] Barabasi, A-L., Albert, R.: **Emergence of Scaling in Random Networks**. *Science* 1999, **286(5439)**:509–512.

[8] Barabasi, A.L., Oltavi, Z.N.: **Network biology: understanding the cell's functional organization**. *Nature Reviews Genetics* 2004, **5(2)**:101–113.

[9] Freeman, L.: **A set of measures of centrality based upon betweenness**. *Sociometry* 1977, **40(1)**:35–41.

[10] Jeong, H., Mason, S.P., Barabasi, A-L., Oltavi, Z.N.:**Lethality and centrality in protein networks**. *Nature* 2001, **411(6833)**:41–42.

[11] Batada, N., Hurst, L.D., Tyers, M.: **Evolutionary and physiological importance of hub proteins**. *PLoS Computational Biology* 2006, **2(7)**:e88.

[12] Uetz, P., Giot, L., Cagney, G., Traci, A., Judson R., et al.: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403(6770)**:623–627.

[13] Ito T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc. Natl. Acad. Sci.* 2001, **98(8)**:4569–4574.

[14] Bader, G.D., Hogue, C.W.V.: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4(2)**.

[15] Gavin, A.C., Aloy, P., Grandi, P., Krause, R., et al.: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440(7084)**:631–636.

[16] van Dongen S.: **Graph clustering by flow simulation**. *PhD thesis* 2000, University of Utrecht.

[17] Srihari, S., Ning, K., Leong, H.W.: **Refining Markov Clustering for complex detection by incorporating core-attachment structure**. *International Conference on Genome Informatics (GIW)* 2009, **23(1)**:159–168.

[18] Srihari, S., Ning, K., Leong, H.W.: **MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure**. *BMC Bioinformatics* 2010, **11(504)**.

[19] Srihari, S., Leong, H.W.: **Employing functional interactions for the characterization and detection of sparse complexes from yeast PPI networks**. *International Journal of Bioinformatics Research and Applications / Asia Pacific Bioinformatics Conference (APBC)* 2012, To appear.

[20] Srihari, S., Leong, H.W.: **"Reusuability" of 'static' protein complex components during the yeast cell cycle**. *International Conference on Bioinformatics (InCoB)* 2011, Poster 220.

[21] Gleick, J.: **Genius : The Life and Science of Richard Feynman**. *Vintage Books* 1993.

[22] Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., et al.: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415(6868)**:180–183.

[23] Michinck, S.W.: **Protein fragment complementation strategies for biochemical network mapping**. *Current Opinion in Biotech* 2003, **14(6)**:610–617.

[24] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., Seraphin, B.: **A generic protein purification method for protein complex characterization and proteome exploration**. *Nature Biotechnology* 1999, **17(10)**:1030–1032.

[25] Brown, J.A., Sherlock, G., Myers, C.L., Burrows, N.M., Deng, C.: **Global sythetic lethality analysis and yeast functional profiling**. *Trends Genetics*, **22(1)**:56–63.

[26] Shoemaker, B.A., Panchenko, A.R.: **Deciphering protein-protein interactions. Part I: Experimental techniques and databases**. *PLoS Computational Biology* 2007, **30(3)**:e42.

[27] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., et al.: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415(6868)**:141–147.

[28] Krogan, N.J., Cagney, G., Yu, H., Zhong, G., et al.: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature* 2006, **440(7084)**:637–643.

[29] Zhang, B., Park, B.H., Karpinets, T., Samatova, N.: **From pull-down data to protein interaction networks and complexes with biological relevance**. *Systems Biology* 2008, **24(7)**:979–986.

[30] Spirin, V., Mirny, L.: **Protein complexes and functional modules in molecular networks**. *Proc. Natl. Acad. Sci.* 2000, **100(21)**:12123–8.

[31] Pu, S., Vlasblom, J., Emili, A., Greenbalt, J., Wodak, S.: **Identifying functional modules in the physical interactome of Saccharomyces cerevisiae**. *Proteomics* 2007, **7(6)**:944–960.

[32] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: **Comparative assessment of large-scale datasets of protein-protein interactions**. *Nature* 2002, **417(6887)**:399–403.

[33] Bader, G.D., Hogue, C.W.V.: **Analyzing yeast protein-protein interaction data obtained from different sources**. *Nature Biotechnology* 2002, **20(10)**:991–997.

[34] Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A-R., Simonis, N., Rual, J-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P., Vidal, M: **Literature-curated protein interaction datasets**. *Nature Methods* 2008, **6(1)**:39–46.

[35] Mackay, J.P., Sunde, M., Lowry, S., Crossley, M., Matthews, J.M.: **Protein interactions: to believe or not to believe?**. *Trends in Biochemical Sciences* 2008, **30(1)**:242–243.

[36] Collins, S.R., Kemmeren P., Zhao, X.C., Greenbalt, J.F., Spencer F., Holstege, F. et al.: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae**. *Molecular Cellular Proteomics* 2007, **6(3)**:439–450.

[37] Ashburner, M., Ball C.A., Blake J.A., Botstein, D., Butler, H., Cherry, M. et al.: **Gene ontology: a tool for the unification of biology**. *Nature Genetics* 2000, **25(1)**:25–29.

[38] Hart, G., Lee, I., Marcotte, E.R.: **A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality**. *BMC Bioinformatics* 2007, **8(1)**:236–247.

[39] Chua, H., Ning, K., Sung, W., Leong, H., Wong, L.: **Using indirect protein-protein interactions for protein complex prediction**. *J. Bioinformatics and Computational Biology* 2008, **6(3)**:435–466.

[40] Liu, G., Li, J., Wong, L.: **Assessing and predicting protein interactions using both local and global network topological metrics**. *Genome Informatics (GIW)* 2008, **21**:138–149.

[41] Friedel C., Krumsiek, J., Zimmer, R.: **Bootstrapping the interactome: unsupervised identification of protein complexes in yeast**. *Research in Computational Molecular Biology (RECOMB)* 2008, 3–16.

[42] Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., Ideker, T.: **A direct comparison of protein interaction confidence assignment schemes**. *BMC Bioinformatics* 2006, **7(1)**:360.

[43] Kuchaiev, O., Rasajski, M., Higham, D., Przulj, N.: **Geometric de-noising of protein-protein interaction networks**. *PLoS Computational Biology* 2009, **5(8)**:e1000454.

[44] Higham, D., Rasajski, M., Przulj. N.: **Fitting a geometric graph to a protein-protein interaction network**. *Bioinformatics* 2008, **24(8)**:1093–1099.

[45] Voevodski, K., Teng, S-H., Yu, X.: **Spectral affinity in protein networks**. *BMC Systems Biology* 2009, **3(1)**:112.

[46] Jain, S., Bader, G.: **An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology**. *BMC Bioinformatics* 2010, **11**:562.

[47] Brietkreutz, A., Choi, H., Sharom, J.R., Boucher, L., et al.: **A global protein kinase and phosphatase interaction network in yeast**. *Science* 2010, **328(5981)**:1043–1046.

[48] Ermolaeva, M.D., White, O., Salzberg, S.L.: **Prediction of operons in microbial genomes**. *Nucleic Acids Research* 2002, **29(5)**:1216–1221.

[49] Dandekar T., Snel, B., Huynen, M., Bork, P.: **Conservation of gene order: A fingerprint of proteins that physically interact**. *Trends Biochemistry* 1998, **23(9)**: 324–328.

[50] Teichmann, S.A., Babu, M.M.: **Conservation of gene co-regulation in prokaryotes and eukaryotes**. *Trends Biotechology* 2002, **20(10)**:407–410.

[51] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., et al.: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285(5428)**:751–753.

[52] Xenarious, I., Salwinski, L., Duan, J.X., Higney, P., Kim, S-L., Eisenberg, D.: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Research* 2002, **30(1)**:303–304.

[53] Bader, G.D., Donaldson, I., Wolting, C., Quellette, B.F., Pawson, T., Hogue C.W.: **BIND: The Biomolecular Interaction Network Database**. *Nucleic Acids Research* 2001, **29(1)**:242–245.

[54] Breitkreutz, B., Stark, C., Tyers, M.: **The GRID: The General Repository for Interaction Datasets**. *Genome Biology* 2003, **4(3)**:R23.

[55] von Mering, C., Huynen, M., Jaeggi, D., Schimdt, S., Bork, P., Snel, B.: **STRING: a database of predicted functional associations between proteins**. *Nucleic Acids Research* 2003, **31(1)**:258–261.

[56] Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., et al.: **CYGD: the Comprehensive Yeast Genome Database**. *Nucleic Acids Research* 2005, **34**:D364–368.

[57] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., et al.: **The MIPS mammalian protein-protein interaction database**. *Bioinformatics* 2005, **21(6)**:832–834.

[58] Peri, S., Navarro, D., Kristiansen, T., Amanchy, R., Surendranath, V.: **Human protein reference database as a discovery resource for proteomics**. *Nucleic Acids Research* 2004, **32**:D497–501.

[59] Brown, K.R., Jurisica, I.: **Unequal evolutionary conservation of human protein interactions in interologous networks**. *Genome Biology* 2007, **8(5)**:95.

[60] Mellor, J.C., Yanai, I., Karl, H., Mintseris, J., DeLisi, C.: **Predictome: a database of putative functional links between proteins**. *Nucleic Acids Research* 2002, **30(1)**:306–309.

[61] Kalna, G., Higham, D.: **A clustering coefficient for weighted networks, with application to gene expression data**. *AI Communications* 2007, **20(4)**:263–271.

[62] Enright, A.J., van Dongen, S., Ouzounis, C.A.: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Research* 2002, **30(7)**: 1575-1584.

[63] Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: **Detection of functional modules from protein interaction networks**. *Proteins* 2004, **54(1)**:49–57.

[64] Liu, G., Wong, L., Chua, H.N.: **Complex discovery from weighted PPI networks**. *Bioinformatics* 2009, **25(15)**:1891–1897

[65] Adamcsek, B., Palla, G., Farkas, I., Derenyi, I., Vicsek, T.: **CFinder: locating cliques and overlapping modules in biological networks**. *Bioinformatics* 2006, **22(8)**:1021–1023.

[66] Li X.L., Tan, S.H., Foo, C.S., Ng, S.K.: **Interaction Graph mining for protein complexes using local clique merging**. *Genome Informatics (GIW)* 2005, **16(2)**:260–269.

[67] Tomita, E., Tanaka, A., Takahashi, H.: **The worst-case time complexity for generating all maximal cliques and computational experiments**. *Theoretical Computer Science* 2006, **363(1)**:28–42.

[68] Chua H.N., Ning, K., Sung, W.K., Leong, H.W., Wong, L.: **Using indirect protein-protein interactions for protein complex prediction.** *Computational Systems Bioinformatics (CSB)* 2007.

[69] Wang, H., Kakaradov B., Collins S.R., Karotki, L., Fiedler, D., Shales M., Shokat, K.M., Walter, T., Krogan N.J., Koller, D.: **A complex-based reconstruction of the Saccharomyces cerevisiae interactome**. *Molecular Cellular Proteomics* 2009, **8(6)**:1361–1377.

[70] Zhang, J.: **A dynamical method to extract communities induced by low or middle-degree nodes**. *IEEE Int Conf on Systems Biology* 2011: 340–344.

[71] Ma, X., Gao, L.: **Detecting protein complexes in PPI network: roles of interactions**. *IEEE Int Conf on Systems Biology* 2011: 223–239.

[72] Wang, Y., Gao, L., Chen, Z.: **An edge based core-attachment method to detect protein complexes in PPI networks**. *IEEE Int Conf on Systems Biology* 2011: 249–252.

[73] Chin, C.H., Chen, S.H., Ho, C.W., Ko, M.T., Lin, C.Y.: **A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles**. *BMC Bioinformatics* 2010, **11(S25)**.

[74] Dezso, D., Oltavi, Z.D., Barabasi, A.L.: **Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisae**. *Genome Research* 2003, **13(11)**:2450–2454.

[75] Wu, M., Li, X., Ng S.K.: **A core-attachment based method to detect protein complexes in PPI networks**. *BMC Bioinformatics* 2009, **10**:169.

[76] Leung, H., Xiang, Q., Yiu, S.M., Chin, F.Y.: **Predicting protein complexes from PPI data: a core-attachment approach**. *Journal of Computational Biology* 2009, **16(2)**:133–44.

[77] King, A.D., Przulj, N., Jurisca, I: **Protein complex prediction via cost-based clustering**. *Bioinformatics* 2004, **20(17)**:3013–3020.

[78] Li, X.L., Foo C.S., Ng, S.K.: **Discovering protein complexes in dense reliable neighborhoods of protein interaction networks**. *Computational Systems Bioinformatics (CSB)* 2007:157–168.

[79] Sharan, R., Ideker, T., Kelley, B.P., Shamir, R., Karp, R.M.: **Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data**. *Research in Computational Molecular Biology (RECOMB)* 2004:282–289.

[80] Hirsh, R., Sharan, R.: **Identification of conserved protein complexes based on a model of protein network evolution**. *Bioinformatics* 2006: **23(2)**:e170–e176.

[81] Dost, B., Shlomi, T., Gupta, N., Ruppin, E., Bafna, V., Sharan, R.: **QNet: A Tool for Querying Protein Interaction Networks**. *Research in Computational Molecular Biology (RECOMB)* 2007: 1–15.

[82] Kim, P.M., Lu, L.J., Xia, Y., Gerstein, M.B.: **Relating three-dimensional structures to protein networks provides evolutionary insights**. *Science* 2006, **314(5807)**:1938–1941.

[83] Ozawa, Y., Saito, R., Fujimori, S., Kashima, H., Ishizaka, M., Yanagawa, H., Miyamoto-Sato, E., Tomita, M.: **Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions**. *BMC Bioinformatics* 2010, **11**:350.

[84] Jung, S.H., Hyun, B., Jang, W.H., Hur, H.Y., Han, D.S.: **Protein complex prediction based on simultaneous protein interaction network**. *Bioinformatics* 2010, **26(3)**:385–391.

[85] Przytycka, T., Singh, M., Slonim, D.K.: **Toward the dynamic interactome: it's about time**. *Briefings in Bioinformatics* 2010, **2(1)**:15–29.

[86] Brohee, S., van Helden, J.: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.

[87] Vlasblom, J., Wodak, S.: **Markov clustering versus affinity propagation for the partitioning of protein interaction graphs**. *BMC Bioinformatics* 2009, **10**:99.

[88] Li, X.L., Wu, M., Kwoh, C.C., Ng, S.K.: **Computational approaches for detecting protein complexes from protein interaction networks: a survey**. *BMC Genomics* 2010, **11(S3)**.

[89] Blatt, M., Wiseman, S., Domany, E.: **Superparamagnetic clustering of data**. *Physical Review Letters* 1996, **76(18)**:3251-3254.

[90] Mewes, H.W., Amid, C., Arnold, R., Frishman, D. et al.: **MIPS: analysis and annotation of proteins from whole genomes**. *Nucleic Acids Research* 2006, **34**:D169–D172.

[91] Frey, B.J., Dueck, D.: **Clustering by passing messages between data points**. *Science* 2007, **315(5814)**:972–976.

[92] Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S.: **Up-to-date catalogues of yeast protein complexes**. *Nucleic Acids Research* 2009, **37(3)**:825-831.

[93] Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C. Mellwig, C., Fischer, S. et al.: **Structure-based assembly of protein complexes of yeast**. *Science* 2004, **303(5666)**:2026–2029.

[94] Cherry, J.M., Adler, C., Chervitz S.A., Dwight S.S., Jia, Y. et al.: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Research* 1998, **26(1)**:73–79.

[95] Zhou, X., Kao, M.C., Wong, W.H.: **Transitive functional annotation by shortest-path analysis of gene expression data**. *Proc. Natl. Acad. Sci.* 2002, **99(20)**:12783–8.

[96] Andreopoulos, B., An, A., Wang, X., Faloutsos, M., Schroeder, M.: **Clustering by common friends finds locally significant proteins mediating modules**. *Bioinformatics* 2007, **23(9)**:1124-1131.

[97] Shannon, P., Markiel, A., Ozier O., Baliga, N.S., Wang, J. et al.: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Research* 2003, **13(11)**:2498–2504.

[98] Miller, T., Krogan, N.J., Dover, J., Bromage, E.H. et al.: **COMPASS: a complex of proteins associated with a trithorax-related SET domain protein**. *Proc. Natl. Acad. Sci.* 2001, **98(23)**:12902–7.

[99] Zhao, J., Kessler, M., Moore, C.L.: **Cleavage factor II of Saccharomyces cerevisiae contains homologues to subunits of the mammalian cleavage/polyadenylation specificity factor and exhibits sequence-specific, ATP-dependent interaction with precursor RNA**. *J Biological Chemistry* 1997, **272(16)**:10831–8.

[100] Cheng, H., He, X., Moore, C.: **The Essential WD Repeat Protein Swd2 Has Dual Functions in RNA Polymerase II Transcription Termination and Lysine 4 Methylation of Histone H3**. *Molecular Cellular Biology* 2004, **24(7)**:2932–2943.

[101] J.S. Luz, J.R. Tavares, F.A. Gonzales, M.C.T. Santosa, C.C. Oliveira: **Analysis of the Saccharomyces cerevisiae exosome architecture and of the RNA binding activity of Rrp40p**. *J Biochemistry* 2006, **89(5)**:686–691.

[102] Araki, Y., Takahashi, S., Kobaysashi, T., Kajiho, H., Hoshino, S., Katada, T. et al.: **Ski7p G protein interacts with the exosome and the Ski complex for 3'-to-5' mRNA decay in yeast**. *EMBO J* 2001, **20(17)**:4684–4693.

[103] Hurwitz, J.: **The discovery of RNA polymerase**. *J Biological Chemistry* 2005, **280(52)**:42477–42485.

[104] Seals, D.F., Eitzen, G., Margolis, N., Wickner, T., Price, A.: **A Ypt/Rab effector complex containing the Sec1 homolog Vps33p is required for homotypic vacuole fusion**. *Proc. Natl. Acad. Sci.* 2000, **97(17)**:9402–9407.

[105] Carvalho, P., Goder, V., Rapoport, T.A.: **Distinct ubiquitin-ligase complexes define convergent pathways for the degradation of ER proteins**. *Cell* 2006, **126(2)**:361–373.

[106] Grant, P.A., Schieltz D., Pray-Grant, M.G., Reese, J.C., Yates, Wolkman, J.L.: **A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation**. *Cell* 1998, **94(1)**:45–53.

[107] Eberharter A., Sterner D.E., Schieltz, D., Hassan, A. et al.: **The ADA complex is a distinct histone acetyltransferase complex in Saccharomyces cerevisiae**. *Molecular Cellular Biology* 1999, **19(10)**:6621–6631.

[108] Grant, P.A. , Schieltz, D., McMahon, S.J., Wood, J.M. et al.: **The novel SLIK histone acetyltransferase complex functions in the yeast retrograde response pathway**. *Molecular Cellular Biology* 2002, **22(24)**:8774–8786.

[109] Asano, K., Clayton, J., Shalev, A., Hinnebusch, G.: **A multifactor complex of eukaryotic initiation factors, eIF1, eIF2, eIF3, eIF5, and initiator tRNA-Met is an important translation initiation intermediate in vivo**. *Genes and Development* 2000, **14(19)**:2534–2546.

[110] Eves, H.: **Return to Mathematical Circles**. *Prindle, Weber and Schmidt* Boston 1998.

[111] Habibi, M., Eslahchi, C., Wong, L.S.: **Protein complex prediction based on k-connected subgraphs in protein interaction network**. *BMC Systems Biology* 2010, **4**:129.

[112] Newman, M.J., Girvan, M.: **Finding and evaluating community structure in networks**. *Physical Review* 2004, **69(2)**:e69.

[113] Skrabanek, L., Saini, H.K., Bader, G., Enright, A.: **Computational prediction of protein-protein interactions**. *Molecular Biotechnology* 2008, **38(1)**:1–17.

[114] Panasenko, O., Landrieux, E., Feuermann, M., Finka, A., Paquet, N., Collart, M.: **The yeast CCR-NOT complex controls ubiquitination of the nascent-associated polypeptide (NAC-EGD) complex**. *J Biological Chemistry* 2006, **281(42)**:31389–31398.

[115] Green M.R.: **TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes**. *Trends Biochemical Sciences* 2000, **25(2)**:59–63.

[116] Liu, G., Yong, C.H., Chua, H.N., Wong, L.: **Decomposing PPI networks for complex discovery**. *Proteome Science* 2011, **9(1)**:S15.

[117] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: **Associating genes and protein complexes with disease via network propagation**. *PLoS Computational Biology* 2010, **6(1)**:e1000641.

[118] Isoe, J., Collins, J., Badgandi, H., Day, A.W., Miesfeld, R.L.: **Defects in coatomer protein I (COPI) transport cause blood feeding-induced mortality in yellow fever mosquitoes**. *Proc. Natl. Acad. Sci.* 2011, **108(24)**:e211–217.

[119] Han, J.D., Bertin, N., Hao, T., Debra S., Gabriel F., Zhang, V. et al.: **Evidence for dynamically organized modularity in the yeast protein interaction network**. *Nature* 2004, **430(6995)**:88–93.

[120] Zotenko, E., Mestre, J., O'Leary, D.P., Przytycka, T.M.: **Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality**. *PLoS Genetics* 2008, **4(8)**:e1000140

[121] Kang, N., Ng, H.K., Srihari, S., Leong, H.W., Nesvizhskii, A.: **Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology**. *BMC Bioinformatics* 2010, **11**:505.

[122] Tao Y., Yiu, M.L., Mamoulis, N.: **Reverse neighbor search in metric spaces**. *IEEE Trans. Knowledge Data Engineering* 2006, **18**:1239–1252.

[123] Pereira-Leal, J.B., Levy, E.D., Teichmann, S.A.: **The origins and evolution of functional modules: lessons from protein complexes**. *Phil. Trans. R. Soc. B.* 2006, **361**:507–517.

[124] Winzeler E.A., Shoemaker D.D., Astromoff, A., Liang, H., Anderson, K. et al.; **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285(5429)**:901-906.

[125] Giaever G., Chu A.M., Ni, L., Connelly, C., Riles, L. et al.: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418(6896)**:387-391.

[126] Batada, N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B-J., Hurst, L.D., Tyers, M.: **Still stratus not altocumulus: further evidence against the date/party hub distinction**. *PLoS Computational Biology* 2007, **5(6)**:e154.

[127] Qi, Y., Ge, H.: **Modularity and dynamics of cellular networks**. *PLoS Computational Biology* 2006, **2(12)**:e174.

[128] Komuruv, K., White, M.: **Revealing static and dynamic modular architecture of the eukaryotic protein interaction network**. *Molecular Systems Biology* 2007, **3(1)**:110.

[129] Ge, H., Liu, Z., Church, G.M., Vidal, M.: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nature Genetics* 2001, **29(1)**:482–486.

[130] Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M.: **The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics**. *PLoS Computational Biology* 2007, **3(4)**:e59.

[131] Patil, A., Nakai, K., Kinoshita, K.: **Assessing the utility of gene coexpression stability in combination with correlation in the analysis of protein-protein interaction networks**. *BMC Genomics* 2011, **12(3)**:S19.

[132] de Lichtenberg, U., Jensen, L.J., Brunak, S., Bork, P.: **Dynamic complex formation during yeast cell cycle**. *Science* 2005, **307(5710)**:724–727.

[133] Wu, X., Guo, J., Zhang, D.Y., Lin,K.: **The properties of hub proteins in a yeast-aggregated cell cycle network and its phase sub-networks**. *Proteomics* 2009, **9(20)**:4812–4824.

[134] Gauthier, N. P., Jensen, L. J., Wernersson, R., Brunak, S., and Jensen, T. S.: **Cyclebase.org - a comprehensive multi-organism online database of cell-cycle experiments**. *Nucleic Acids Research* 2009, **36**:D854–859.

[135] Lodish, H., Berk, A., Kaiser, C., Krieger, M., Scott, M., Bretscher, A., Ploegh, H., Matsudiara, P.: **Molecular Cell Biology**. *W.H. Freeman and Co.* 2007, Ed. 6.

[136] Wang, Y.: **CDKs and the yeast-hyphal decision**. *Current Opinion in Microbiology* 2009, **12(6)**:644–649.

[137] Cline, R.B.: **Becoming a behavioral science researcher: A guide to producing research that matters**. *Guilford Press New York* 2008: 236.

[138] Miller, J., Lo, R., Ben-Hur, A., Desmarais, A., Stagljar, I., Noble, W.S., Fields, S.: **Large-scale identification of yeast integral membrane protein interactions**. *Proc. Natl. Acad. Sci.* 2005, **102(34)**:12123–12128.