Discovering dynamic protein complexes from static interactomes: Three challenges

Yong Chern Han

A thesis submitted
for the degree of Doctor of Philosophy
Graduate School for Integrative Sciences and Engineering
National University of Singapore

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Yong Chern Han

March 20, 2015

Acknowledgements

This dissertation would not have been possible without the mentorship and motivation of my supervisor, Professor Wong Limsoon, who patiently encouraged and navigated me through six years of repeated experiments, backtracked ideas, re-considered hypotheses, contested causations, and unwarranted conclusions, to finally arrive at the completion of this work.

I also owe a great debt to my parents, who gave me the means to remain a student well into my middle-aged years, for which I am forever grateful.

This dissertation is dedicated to my ever-patient Jenny, who waited—without holding her breath—for the completion of this work so that we can finally commence our honeymoon.

Contents

Summary

Protein complexes are stoichiometrically-stable structures consisting of multiple proteins that bind (interact) together. Protein complexes perform a wide variety of molecular functions in many processes in the cell. Thus it is important to determine the set of existing complexes to gain an understanding of the mechanism, organization, and regulation of cellular processes.

Many algorithms have been proposed to discover protein complexes from protein-protein interaction (PPI) data, which has been made available in large amounts by high-throughput experimental techniques. The general strategy underlying most complex-discovery algorithms is to find clusters of highly-interconnected proteins within the PPI network as protein complexes. However, the performance of most of these approaches still leaves room for improvement. One stumbling block is that the representations and analyses of PPIs for the purpose of complex prediction have been overwhelmingly static, even though proteins and complexes exhibit a sophisticated dynamism in behavior.

In this dissertation we identify three challenges in complex discovery that arise from, or are exacerbated by, this static view of PPIs and protein complexes. First, many complexes are sparsely-connected in the PPI network, so that complex-discovery algorithms cannot pick them out as dense clusters. Second, many complexes are embedded within densely-connected regions in the PPI network, with many extraneous PPIs connecting them to external proteins, so their boundaries cannot be accurately delimited. Third, many complexes are small (consisting of two or three proteins), so that important topological features like density become ineffectual.

We describe three approaches that address each of these challenges. First, Supervised Weighting of Composite Networks (SWC) integrates diverse data sources with supervised learning to weight edges in the PPI network with their probabilities of being co-complex. This successfully fills in missing edges in sparse complexes, allowing them to be predicted. Second, PPI-network decomposition splits the PPI network into spatially- and temporally-coherent subnetworks. This allows complexes embed-

ded within dense regions to be extracted from their respective subnetworks. Third, Size-Specific Supervised Weighting (SSS) integrates diverse data sources, and weights edges with their probabilities of being in a small complex versus a large complex, using a supervised approach. Small complexes are extracted and scored using the edges surrounding each candidate complex. This size-specific approach allows small complexes to be found more accurately than conventional clustering approaches.

We also integrate all three approaches into a single system, which demonstrates superior performance in complex prediction compared to conventional approaches, or compared to each of our approaches individually. This integrated system improves the prediction of all three types of complexes that we identified as challenging—sparse, embedded, and small complexes.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Most cellular processes are performed not by individual proteins acting alone, but by protein complexes consisting of multiple proteins that interact (bind) physically. Protein complexes comprise the modular machinery of the cell, performing a wide variety of molecular functions and participating in many biological processes, so determining the set of existing complexes is important for understanding the mechanism, organization, and regulation of cellular processes. Since proteins in a complex interact physically, many methods have been proposed to discover complexes from protein-protein interaction (PPI) data, which has been made available in large amounts by high-throughput experimental techniques. PPI data is frequently represented as a PPI network (PPIN), where vertices represent proteins and edges represent interactions between proteins.

The general strategy underlying most complex-discovery algorithms is to find clusters of highly-interconnected proteins within the PPI network as protein complexes. Over the past decade, these algorithms have grown in sophistication and variety, and have incorporated increasing amounts of useful biological insights in their designs. However, the performance of most of these approaches still leaves room for improvement: for example, even in yeast with decently-comprehensive PPI data, accurate prediction of complexes at fine resolution remains difficult. One main stumbling block is that the representations and analyses of PPIs for the purpose of complex prediction have been overwhelmingly static, even though it has been well understood that proteins and complexes exhibit a sophisticated dynamism in behavior.

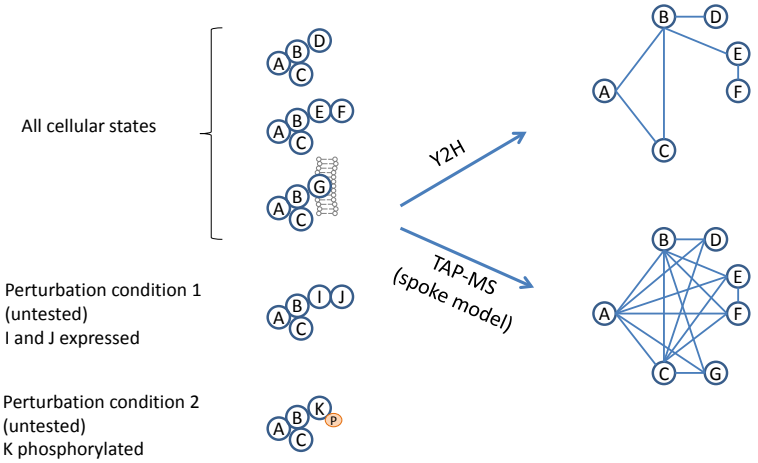

Figure 1.1: Dynamism of protein complexes is lost after PPI screening and representation in the PPI network. Moreover, this dynamism hinders an accurate screening of PPIs.

1.2 Dynamism of PPIs and complexes

Proteins interact in a dynamic fashion, with a variety of interaction timings, locations, and affinities. These are mediated by a wide range of factors from cellular state (such as different cell-cycle phases or perturbation conditions), to biological processes (such as expression, translation, modification, transport, and degradation of the interactor proteins), to the physiochemical environment in the interaction locale (such as the concentration of effector molecules like ATP) [1]. Correspondingly, protein complexes exhibit dynamic behavior which are in fact important functional mechanisms, for example to allow complexes to be formed only at certain times, or to vary the composition of complexes to modulate or activate their functions. However, due to limitations in PPI-detection methodologies, it is difficult to capture the dynamism of PPIs (i.e. when, where, and how a protein interacts with others). Furthermore, this dynamism also precludes a faithful interrogation of PPIs in the cell (e.g. condition-specific PPIs may be missed, or spurious PPIs may be detected in non-physiological experimental systems). Moreover, the representation of PPIs in the PPI network does not preserve any information about the dynamics of PPIs. Thus there exists a disparity between the dynamic nature of PPIs and protein complexes on the one hand, and the static representation and analysis of the PPI network on the other hand.

Figure 1.1 illustrates this problem in a simplified fashion via a made-up complex consisting of an A-B-C core, which forms distinct complexes with either protein D, or proteins E-F, or membrane protein G; additionally, it complexes with proteins I-J which are only expressed during perturbation condition 1, and with protein K only after phosphorylation during perturbation condition 2. With the yeast two-hybrid (Y2H)

screening method, the interaction with membrane protein G is undetected, while the mutually-exclusive interactions with proteins D and E-F are detected and represented as undifferentiated edges. Since the cells interrogated are never in perturbation conditions 1 or 2, proteins I, J, and K are never found to interact with A-B-C. Another common screening method, tandem affinity purification coupled to mass spectrometry (TAP-MS), conflates the three distinct complexes as one large, densely-connected graph (while it appears here that the three complexes can be discerned as separate cliques in the graph, in reality the additional spurious and missing edges due to noise make this task difficult).

1.3 Three challenges in complex discovery

We identify three challenges in protein-complex discovery that arise from, or are exacerbated by, this static view of PPIs and protein complexes.

1. Many complexes exist in sparse regions of the network, so that proteins within the complexes are not densely interconnected. This arises from undetected condition-specific, location-specific, or transient PPIs.

2. Many complexes are embedded within highly-connected regions of the PPI network, with many extraneous edges connecting its member proteins to other proteins outside the complex. This arises from proteins that participate in multiple distinct complexes which correspond to dense overlapping regions in the PPI network, or from spuriously-detected interactions.

3. Many complexes are small (that is, composed of two or three proteins), making measures of important topological features, such as density, ineffectual. This is further exacerbated by extraneous or missing interactions which can embed the small complex in a larger clique, or disconnect it entirely.

Figure 1.2 illustrates some of these challenges in real complexes. The Cdc28p yeast protein (figure 1.2) complexes with various cyclin proteins (Cln1p to Cln3p, Clb1p to Clb6p) to regulate the cell cycle. While the abundance of Cdc28p is constant throughout the cell cycle, the activity of the cyclin proteins are regulated via sophisticated gene-expression and post-translational controls, so that the proper complexes are formed at each point of the cell cycle [2, 3]. Figure 1.2a shows the interactome around these proteins and their neighbours, with the nine different complexes formed by Cdc28p circled. Although these interactions occur at different times during the

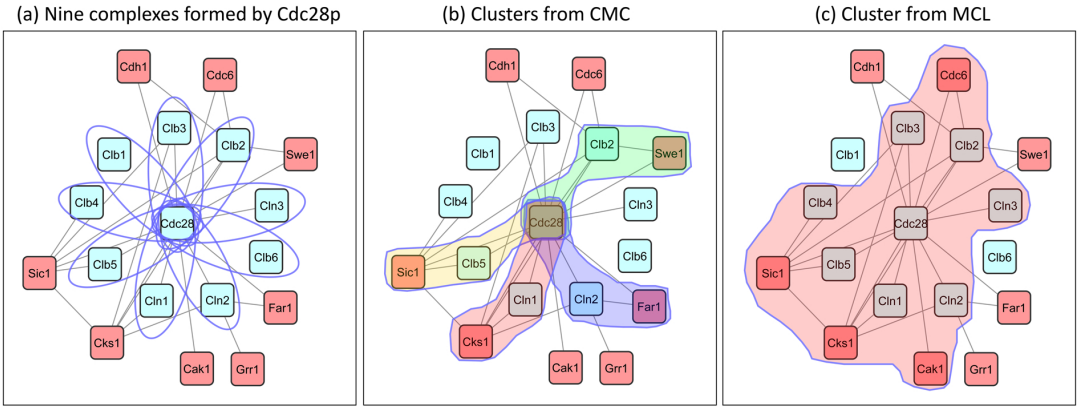

Figure 1.2: (a) Cdc28p is involved in nine distinct complexes, which overlap and have many extraneous edges. Three of the complexes are disconnected. (b) CMC includes extraneous proteins in its clusters. (c) MCL merges the complexes.

cell cycle (e.g. Cdc28p-Cln1p and Cdc28p-Cln2p in G_1 phase, Cdc28p-Clb2p in G_2M phase), they are collapsed into the same static interactome, resulting in a highly-connected region around Cdc28p and its cyclin partners. Furthermore, PPIs are missing between CDC28p and some of its cyclin partners (Clb1p, Clb4p, Clb6p). In fact, some of these PPIs exist in our source datasets, but are filtered as they have fewer experimental evidences to back them up. While it is possible to lower our reliability score cutoff to include these PPIs, this would also include many spurious PPIs and make the discovery of other complexes even more difficult.

Figure 1.2b and c show the clusters predicted by two popular clustering algorithms, CMC and MCL. CMC found four clusters that overlap with four Cdc28p complexes, but with one extraneous protein in each case, while MCL found one large cluster that covered Cdc28p, seven of the nine cyclin proteins, and four extraneous proteins. Note that MCL does not allow overlaps in its predicted clusters, so here it predicts clusters that merge the overlapping and highly-connected complexes together. While CMC allows overlapping clusters, the many extraneous edges and high connectivity to external proteins make it difficult to delimit the overlapping complexes precisely.

1.4 Contribution: Three approaches

In this dissertation, we propose three approaches that help to address these problems in complex discovery.

1. We propose an approach called Supervised Weighting of Composite Networks (SWC [4]) which can address the problem of sparse complexes. SWC integrates PPI data with two additional data sources, functional associations and

co-occurrence in literature, using a supervised approach to weight edges with their posterior probability of belonging to a complex. By integrating diverse data sources that may support co-complex relationships between proteins, SWC fills in the missing edges in many sparse complexes; while supervised weighting leverages on the characteristics of known complexes to reducing the amount of spurious non-co-complex edges. Using this approach, improvements are obtained in both yeast and human complex discovery, especially among the sparse complexes.

2. We propose an approach to decompose the PPI network into spatially- and temporally-coherent subnetworks, which can address the problem of complexes embedded within dense regions of the PPI network [5]. Hub proteins with large numbers of interaction partners are first removed before complex discovery, as they tend to correspond to date hubs with non-simultaneous interactions. Next, cellular-location Gene Ontology terms [6] are used to decompose the PPI network into spatially-coherent subnetworks. The complexes are derived from these subnetworks, and the hubs are re-added to their highly-connected complexes. This allows multiple overlapping complexes to be disambiguated into separate subnetworks, from which they can be more easily extracted. This approach improves the performance of complex discovery, with the biggest improvements among complexes in highly-connected regions.

3. We propose an approach called Size-Specific Supervised Weighting (SSS [7]) to address the problem of predicting small complexes. SSS integrates PPI data with two additional data sources, functional associations and co-occurrence in literature, along with their topological features, using a supervised approach to weight edges with their posterior probabilities of belonging to small complexes versus large complexes. SSS then extracts small complexes from the weighted network, and scores them using the probabilistic weights of edges within, as well as surrounding, the complexes. This approach achieves significant improvements in discovering small complexes.

1.5 Publications

This dissertation is based in part on work published in various venues:

1. The exploration of the dynamism of PPIs and complexes, and the identification of the three challenges in complex discovery, is based on work published in Yong

CH, Wong L, "From the static interactome to dynamic protein complexes: Three challenges", J Bioinform Comput Biol 2015, 13(2):15710018 [8].

2. Supervised Weighting of Composite Networks (SWC) is based on work published in Yong CH, Liu G, Chua HN, Wong L, "Supervised maximum-likelihood weighting of composite protein networks for complex prediction", BMC Syst Biol 2012, 6(Suppl 2):S13 [4].

3. The decomposition of PPI networks for complex discovery is based on work published in Liu G, Yong CH, Chua HN, Wong L, "Decomposing PPI networks for complex discovery", Proteome Sci 2011, 9(Suppl 1):S15 [5].

4. Size-Specific Supervised Weighting (SSS) is based on work published in Yong CH, Maruyama O, Wong L, "Discovery of small protein complexes from PPI networks with size-specific supervised weighting", BMC Syst Biol 2014, 8(Suppl 5):S3 [7].

1.6 Thesis organization

Chapter 2 provides a background on PPIs and protein complexes with an emphasis on their dynamic nature, and describes how this dynamism is not captured and represented in PPI data, and moreover hinders the accurate screening of PPIs. It highlights the three challenges related to the analysis of static PPI data for complex discovery: discovering sparsely-connected complexes, discovering complexes embedded within dense regions, and discovering small complexes. Chapter 3 describes our approach to address the discovery of sparse complexes, supervised weighting of composite networks (SWC). Chapter 4 describes our approach to address the discovery of complexes embedded within dense regions, via the decomposition of PPI networks. Chapter 5 describes our approach to address the discovery of small complexes, size-specific supervised weighting (SSS). Chapter 6 describes our integration of these three approaches into a single complex-discovery system. Finally, Chapter 7 concludes this dissertation with a short summary and lays out potential directions for future work.

Chapter 2

Background and Motivation

2.1 Introduction

In the cell, many proteins bind physically to form stoichiometrically-stable multiprotein structures called protein complexes. Protein complexes perform a wide variety of molecular functions in many cellular processes. Thus it is important to determine the set of complexes in the cell to gain an understanding of the mechanism, organization, and regulation of these processes. Since proteins in a complex interact physically, many algorithms have been proposed to analyze protein-protein interaction (PPI) data to discover protein complexes.

The general strategy underlying most complex-discovery algorithms is to represent PPI data as a PPI network, where vertices represent proteins and edges represent interactions between proteins, and then find clusters of highly-interconnected proteins within the PPI network as protein complexes. Over the past decade, these algorithms have grown in sophistication and variety, and have incorporated increasing amounts of useful biological insights in their designs. However, the performance of most of these approaches still leaves room for improvement: for example, even in yeast with decently-comprehensive PPI data, accurate prediction of complexes at fine resolution remains difficult.

One main stumbling block is that the representations and analyses of PPIs for the purpose of complex prediction have been overwhelmingly static, even though it has been well understood that proteins and complexes exhibit a sophisticated dynamism in behavior. Proteins interact in a dynamic fashion, with a variety of interaction timings, locations, and affinities; these are mediated by a wide range of factors from cellular state (such as different cell cycle phases or perturbation conditions), to biological processes (such as expression, translation, modification, transport, and degradation of the interactor proteins), to the physiochemical environment in the interaction lo-

cale (such as the concentration of effector molecules like ATP) [1]. Correspondingly, protein complexes exhibit dynamic behaviors which are in fact important functional mechanisms, for example to allow complexes to be formed only at certain times, or to vary the composition of complexes to modulate or activate their functions. However, due to limitations in PPI-detection methodologies, it is difficult to interrogate the dynamism of PPIs (i.e. when, where, and how a protein interacts with others). Furthermore, this dynamism also precludes a faithful interrogation of PPIs in the cell (e.g. condition-specific PPIs may be missed, or spurious PPIs may be detected in non-physiological experimental systems). Moreover, the representation of PPIs in the PPI network does not preserve any information about the dynamics of PPIs. Thus there exists a disparity between the dynamic nature of PPIs and protein complexes on the one hand, and the static representation and analysis of the PPI network on the other hand.

We identify three challenges in protein-complex discovery that arise from, or are exacerbated by, this static view of PPIs and protein complexes. First, many complexes are embedded within highly-connected regions of the PPI network, with many extraneous edges connecting a complex's member proteins to other proteins outside the complex. This arises from proteins that participate in multiple distinct complexes which correspond to dense overlapping regions in the PPI network, or from spuriously-detected interactions. Second, many complexes exist in sparse regions of the network, so that proteins within the complexes are not densely interconnected. This arises from undetected condition-specific, location-specific, or transient PPIs. Third, many complexes are small (that is, composed of two or three proteins), making measures of important topological features, such as density, ineffectual. This is further exacerbated by extraneous or missing interactions which can embed the small complex in a larger clique, or disconnect it entirely.

In this chapter, we evaluate the performance of ten complex-discovery algorithms, covering different types of approaches, in the prediction of yeast and human complexes. In particular, we highlight the unsatisfactory performance in predicting complexes embedded within highly-connected regions, complexes within sparse regions, and small complexes, and discuss how an understanding of the dynamics of protein interactions may be used to address the shortcomings of these algorithms with respect to these specific challenges.

A number of surveys on complex discovery have been published in recent years. Li *et al.* [9] in 2010 surveyed a number of complex-discovery algorithms, and categorized

them according to the types of data used and the features of the algorithms. Srihari *et al.* [10] in 2013 further showed that complex-discovery algorithms have evolved to incorporate increasing amounts of biological information in their designs, leading to improved performance and new biological insights. Most recently, Chen *et al.* [11] also surveyed and categorized various complex-discovery algorithms, with a distinct category for algorithms that explicitly model the dynamism of PPIs. Since descriptions and taxonomies of complex-discovery algorithms are already covered in these surveys, here we emphasize specific challenges raised by the dynamism of PPIs, and evaluate a few classic and recent algorithms with respect to these challenges.

In Section 2.2, we elaborate on protein interactions and protein complexes in the cell, with an emphasis on the dynamism of their behaviors. We give a brief background on PPI-screening technologies and their inadequacies, particularly in capturing such dynamism. Then we show how the three challenges that we address in complex discovery follow from the analysis of static PPIs. In Section 2.4, we give a brief taxonomy of clustering-based complex-discovery algorithms, and highlight the ten algorithms that we evaluate in this chapter. In Section 2.5, we describe our experiments to evaluate the ten algorithms in yeast and human complex discovery, with an emphasis on their shortcomings with respect to the three challenges. Finally, in Section 2.6, we look ahead to our proposed solutions to these three challenges, which we discuss in further detail in the following chapters.

2.2 Background: From interactome to complexome

The *interactome* describes the landscape of physical interactions between all molecules in a cell, such as protein-protein, protein-DNA, or protein-RNA interactions. In the study of protein complexes, the interactome is commonly used to refer specifically to physical protein-protein interactions (PPIs), which is the definition that we adopt. The *complexome* describes the set of complexes that exist in an organism, and is of great value in understanding the modular machinery that drives almost all processes in the cell. The link between an organism's interactome and complexome is intuitive: since complexes consist of physically-interacting proteins, they correspond to groups of proteins with high degrees of co-interaction in the interactome. Thus, deriving the complexome from the interactome is a fruitful strategy that has been well researched over the past decade. Many challenges have been acknowledged in this strategy, a significant portion of which we distil as the 'disparity' between the static interactome and the complexome: due to limitations in detection technologies and methodologies

(which have only recently begun to be surpassed), the views and analyses of the interactome and complexome have been overwhelmingly static, without consideration of the dynamic nature of PPIs and the corresponding dynamism of protein complexes.

2.2.1 Dynamism of protein interactions

In fact, the static interactome, understood as the set of PPIs that exist in a cell, is a mere shadow of the dynamic and complex lives of PPIs in reality, which involve a wide range of interaction timings, locations, and binding affinities.

The timing of an interaction is an essential aspect of its dynamism. Frequently, a protein with multiple interaction partners does not interact with all of them simultaneously: it may contain an interacting domain that binds with different partners, one at a time; or it may contain multiple overlapping interacting domains which prevent more than one interaction from occurring at the same time. A study of protein hubs (proteins with a large number of interaction partners) with gene expression data has led to a proposed distinction between date hubs and party hubs [12, 13]: party hubs interact with all of their partners simultaneously as a large complex, while date hubs interact with its partners in mutually exclusive times, and are believed to link diverse biological processes together in the PPI network.

Whether a protein interacts, and which partner it interacts with, can be controlled by different cellular mechanisms. For example, different partners may be expressed at different conditions, may reside in different subcellular locations, or may have different modified states that allow or disallow their binding. Various methods of cellular control of PPIs have been identified [1]: co-localization of the interactors in time and space, as well as the local concentration of the interactors, are controlled by expression, mRNA degradation, protein transport, protein secretion, protein degradation; the binding affinities of different interactors can be controlled through post-translational modification of the interactors, or changes to the physiochemical environment, for example by the concentration of effector molecules like ATP that may change binding affinity.

PPIs have been classified into three categories according to their binding affinities [1, 14, 15]: permanent interactions, with the strongest binding affinity, are irreversible; weak transient interactions, with the weakest binding affinity, are reversible, and involve proteins that switch between both bound and unbound states *in vivo*; strong transient interactions have a binding affinity that lie in the continuum between those of permanent interactions and weak transient interactions, and are reversible when triggered, for example by ligand binding. PPIs can also be characterized as

obligate or non-obligate: proteins with obligate interactions cannot exist as stable structures on their own, and are frequently bound to their partners upon translation and folding; conversely, proteins with non-obligate interactions can exist as stable structures both in bound and unbound states. Obligate interactions are generally permanent, while non-obligate interactions can be permanent or transient.

2.2.2 Dynamism of protein complexes

Consequently, complexes display a range of dynamism in their formation, composition, and stability, which impart important functional mechanisms to the complexes' activities. In a well-known example, the highly conserved Cdc28p (a cyclin-dependent kinase or CDK) yeast protein regulates the cell cycle by forming complexes with different cyclin proteins that phosphorylate different substrates to promote entry into different cell-cycle phases [2,3]: progressing through the cell cycle phases, these include Cdc28p forming complexes with Cln3p to enter the cycle, with Cln1,2p in G_1 phase, with Clb5,6p to begin replication in S phase, and with Clb1,2,3,4p to enter M phase. These complexes are themselves regulated through binding with cyclin-dependent-kinase inhibitors (CKIs) such as Sic1p.

In another example of dynamism in a complex involved in cell cycle regulation, the yeast SCF complex is a ubiquitin ligase consisting of a catalytic core of three proteins (Skp1p, Cul1p, Hrt1p), and a fourth protein that contains an F-box domain [16]. The identity of the F-box-containing protein can vary to produce different SCF ligases that attach ubiquitin to different sets of proteins, depending on the substrate specificity of the F-box-containing protein. For example, the SCF complex with the F-box-containing Cdc4p protein ubiquitinates cell-cycle- and transcription-related proteins, and thus regulates both cell cycle and transcription processes. Furthermore, the SCF complex binds to some substrates only after they have been phosphorylated, thereby increasing its specificity while still allowing involvement in diverse processes.

An integrated analysis of protein complexes with cell-cycle expression data revealed "just-in-time" assembly of most cell-cycle-related complexes in yeast [17]: some subunits of complexes are constitutively expressed (static proteins), while other subunits are expressed only when needed (dynamic proteins), so that the entire complex can be assembled only in specific cell-cycle phases without having to transcriptionally regulate all the subunits of the complex. An example is the prereplication complex, composed of a set of static proteins and other dynamic proteins which are produced and recruited only during the G_1 phase.

In the above examples of complex dynamism, bindings are frequently mediated by strong transient interactions (interactions that associate and disassociate through molecular triggers), for example by binding only after an interactor is phosphorylated. A further example is the heterotrimeric G protein signaling complex, whose α subunit dissociates upon GTP binding. On the other hand, other complexes are made up of permanent, obligate interactions, such as the human chorionic gonadotropin complex and the reverse transcriptase complex [14].

The dynamism of complexes also gives them a modular architecture in function and composition, which has been described with the core-attachment model of complexes [18]. Here, the core of a complex consists of proteins that interact permanently, while attachment proteins are recruited to the core via less permanent interactions, which may modulate or activate the function of the complex.

2.2.3 Interactome screening technologies

The dynamism of PPIs, which provides such important functional mechanisms for complexes, is not captured in the static interactome. A chief reason for this is the technological limitations of past high-throughput PPI screening experiments, which has only recently begun to be surpassed.

In the past decade, the two commonly used methods for high-throughput screening of PPIs are based on the yeast two-hybrid assay (Y2H), which detects binary interactions, and the tandem affinity purification with mass spectrometry (TAP-MS) method, which detects co-complex interactions. The Y2H method, proposed by Fields and Song in 1989 [19], uses a fragmented transcription factor to detect the interaction between a bait protein and a prey protein. The transcription factor of a reporter gene is split into two fragments, the binding domain (BD) and the activating domain (AD). The former is fused with the bait protein, and the latter is fused with the prey protein. When the BD-bait fusion binds to the promoter region of the reporter gene, and the bait and prey interact, both domains of the transcription factor are co-localized at the promoter and the reporter gene is transcribed. Y2H thus detects a binary interaction between the bait and prey proteins. This procedure is scalable to provide high-throughput proteome-wide interaction screening. A recent survey of advances in Y2H technology is provided by Bruckner *et al.* [20].

The Y2H assay is able to detect transient or weak interactions, but is limited to only direct physical PPIs: the interactions between co-complex proteins (proteins in the same complex) that do not physically interact with each other are not detected.

A major drawback of Y2H is that the interactions are assayed at non-physiological conditions: the bait and prey fusion proteins' cDNA, inserted via plasmids, may be overexpressed beyond physiological levels, may be co-expressed whereas they are not co-expressed *in vivo*, or may not undergo the same post-translational modifications as *in vivo*. Furthermore, since they are interrogated in a controlled homogeneous cellular state, interactions that occur in other condition-specific states (such as different cell-cycle or perturbation states) may not be captured.

The classic Y2H assay tests for interactions only in the nucleus, thus interactions are not detected between bait and prey proteins that are unable to interact in the nucleus due to its physiochemical environment, or are unable to localize into the nucleus after translation, even if they do interact *in vivo* in another subcellular compartment—this includes most membrane proteins. Conversely, proteins that are never co-localized *in vivo* and are thus unable to interact might be wrongly detected as interacting in the nucleus. Furthermore, trans-activating proteins, or proteins that activate transcription directly, cannot be used as prey as they would always activate transcription of the reporter gene. However, recent advances in Y2H technology have surpassed some of these limitations [20]. For example, the repressed transactivator (RTA) system allows interrogation of trans-activating baits; the SOS- and RAS-recruitment systems, the G-protein fusion system, and the spit-ubiquitin system allow interrogation of interacting membrane and/or cytosolic proteins; and the SCINEX-P system allows interrogation of interacting proteins in the endoplasmic reticulum.

Aside from the above problems, Y2H also suffers from the variability inherent in interrogating biological systems, leading to poor reproducibility across multiple screens.

TAP-MS, developed in 1999 by Rigaut *et al.* [21], involves tagging a bait protein with an affinity tag (the TAP tag), allowing it to complex with other proteins under physiological conditions, and finally washing it through two affinity columns to detect its co-complex proteins (the prey proteins) via mass spectrometry. This approach is scalable to high-throughput, proteome-wide interrogation of an organism's interactome. A survey of recent advances in MS-based methods is provided by Gavin *et al.* [22].

In TAP-MS, typically only strong interactions are captured, due to the double-purification step. Unlike the Y2H assay which tests for direct interactions, TAP-MS retrieves proteins co-complexed with the bait protein, including those that are only indirectly associated via bridging proteins. Furthermore, for bait proteins that form

multiple distinct complexes, all the proteins that form the union of these complexes may be purified and detected. To uncover the PPIs from the purified complexes, either a spoke model or a matrix model may be used: the spoke model assumes that the bait interacts directly with all the purified proteins, though this leads to a few false positives (direct interactions imputed between indirectly-associated proteins) and false negatives (interactions between prey proteins are not imputed); the matrix model assumes that the bait protein and all the prey proteins interact directly with each other, eliminating false negatives but giving a large number of false positives (interactions imputed between co-complexed but indirectly associated proteins, or between proteins in distinct complexes shared by the prey). More sophisticated models can be utilized: for example, both the socio-affinity index [18] and the purification-enrichment score (PE score [23]) incorporate probabilistic models to take into account both the spoke model (as direct interactions) and the matrix model (as co-occurrence of proteins in the same purification).

In two high-throughput yeast PPI studies based on TAP-MS [18, 24], the TAP tag was fused directly into the bait protein's gene in the chromosome, so that its expression was controlled by its natural promoters, allowing physiological expression levels of the baits. However, in other organisms the TAP-bait fusion protein is largely expressed by non-natural promoters, leading to its over-expression over physiological levels [22].

Under TAP-MS, protein complexes in any subcellular location can be purified. Furthermore, since a heterogeneous collection of cells are purified, complexes present in multiple cellular conditions may be retrieved: for example, the purification of yeast cells growing in a medium may lead to the retrieval of complexes present in various cell-cycle and growth states [18, 24]. Nevertheless, complexes present only in other conditions, such as specific perturbation states, are not retrieved. Only recently have researchers begun interrogating the composition of complexes under different perturbation states, using quantitative AP-MS approaches: affinity purification with selected reaction monitoring (AP-SRM [25]) was proposed to probe quantitative changes in interactions of the Grb2 protein after stimulation with various growth factors; while affinity purification combined with sequential window acquisition of all theoretical spectra (AP-SWATH [26]) was used to study changes in the 14-3-3β protein interactome following stimulation of the insulin-PI3K-AKT pathway. Both works represent key advances in methodologies that will allow dynamic and condition-specific views and analyses of interactomes in the near future; but for now, the range of the proteins and PPIs probed, as well as the conditions tested, remain limited.

2.2.4 The static interactome

As described above, the Y2H and TAP-MS methods do not capture timing (i.e. simultaneity) or localization information about the PPIs. While Y2H detects interactions with a wide range of binding affinities, for interactions whose affinities are dependent on molecular trigger events such as phosphorylation (i.e. strong transient interactions), information about such molecular triggers is lost, and moreover interactions whose triggers are not activated are not captured. Neither Y2H nor TAP-MS interrogate interactions with respect to cellular states: under Y2H, interactions are assayed in a homogeneous cellular state which is frequently non-physiological; while under TAP-MS, interactions are frequently interrogated in heterogeneous cellular growth states, so that proteins present in complexes from various growth states are retrieved as an undifferentiated set. Moreover, complexes present only in specific perturbation conditions, which are absent from the cells, are not found. Although more recent AP methods have investigated the interactions of specific proteins under some specific conditions, the range of proteins and conditions tested is still limited. The PPIs obtained thus represent a static interactome, lacking the dynamism that imparts important functional mechanisms to the PPIs and the complexes that they comprise.

The interactome is frequently represented as a PPI network (PPIN), with vertices representing proteins and edges representing interactions. This representation itself is a simplification of the cellular organization and behavior of PPIs: aside from missing information about interaction timing, location, affinity, and cellular state, the representation of each protein as a single vertex conflates the multiple copies of each protein that exist in the cell into a single entity: in the cell, different copies of the protein may be simultaneously interacting with different partners, may exist in different cellular locations, and may be in different post-translational states, but in the PPIN all these are represented by a single vertex, and all its disparate interactions are represented as undifferentiated outgoing edges from that vertex.

Figure 1.1 illustrates these shortcomings of the Y2H and TAP-MS methods for detecting PPIs via a simple example; we ignore the effects of other factors such as experimental or biological variability, which in reality would lead to additional false positives (spurious edges) and false negatives (missing edges). Here, we use a simple made-up complex consisting of an A-B-C core, which forms distinct complexes with either protein D, or proteins E-F, or membrane protein G; additionally, it complexes with proteins I-J which are only expressed during perturbation condition 1, and with protein K only after phosphorylation during perturbation condition 2. We assume

that all proteins are used as baits in both Y2H and TAP-MS, and in the latter we use the spoke model to obtain individual PPIs. Since the cells interrogated are never in perturbation conditions 1 or 2, proteins I, J, and K are never found to interact with A-B-C. Y2H is unable to detect the interaction with membrane protein G, while the mutually exclusive interactions with proteins D and E-F are detected and represented as undifferentiated edges. TAP-MS likewise conflates the three distinct complexes as one large, densely-connected graph. While it appears here that the three complexes can be discerned as separate cliques in the graph, in reality the additional spurious and missing edges make this task difficult.

2.2.5 Augmenting the static interactome with dynamism

Many researchers have recognized that, while the static interactome is a superficial representation of cellular protein interactions, it is still the only proteome-wide and experimentally-replicated resource of PPIs that is readily available for computational analysis, and so have attempted to augment it with some degree of dynamism using other information sources.

For example, de Lichtenberg *et al.* [17] integrated yeast PPI data with gene expression data from various cell-cycle time-points to analyze the dynamism of complex formation during the cell cycle, and found both constitutively expressed and periodically expressed subunits of most complexes. Likewise, Sriganesh *et al.* [27] also analyzed yeast complexes with cell-cycle expression data, and proposed that constitutively-expressed proteins are more likely to be reused across different complexes.

Other researchers have integrated PPI data with protein-domain information to identify simultaneous or mutually-exclusive interactions. Jung *et al.* [28] decomposed the PPI network into simultaneous protein interaction networks (SPINs), in which all interactions can occur simultaneously, by excluding mutually-exclusive interactions in each SPIN, and then performed complex discovery on each SPIN. Ozawa *et al.* [29] refined complexes predicted by complex-discovery algorithms by eliminating those that included mutually-exclusive interactions.

A major shortcoming of such analyses is that they are based on the PPIN derived from high-throughput experiments such as Y2H and TAP-MS, so they cannot reveal interactions that are only active in untested conditions [30]. Nevertheless, these approaches show that incorporating this aspect of dynamism in PPIs produces complexes that match known complexes more precisely, and may even elucidate novel functional mechanisms in some complexes. However, the limitations of inferring PPI

dynamism indirectly must be noted: for example, gene-expression data does not reflect post-transcriptional activities that further affect complex dynamism, such as protein degradation, transportation, or modification; while using protein-domain information to infer simultaneous or mutually-exclusive interactions is heavily reliant on the coverage and accuracy of protein-domain databases.

2.3 Three challenges in complex discovery

To discover the set of protein complexes in an organism (its complexome), researchers have proposed a wide variety of methods to analyze its interactome, derived from high-throughput PPI-screening technologies. A typical strategy is to impute regions of high inter-connectedness in the interactome as putative complexes, since proteins within complexes interact with each other (a summary of such clustering algorithms is given in the next section). However, since the basis of this analysis is the static interactome, which as described above lacks crucial information about the dynamism of PPIs, including interaction timing, location, binding affinity, and cellular state, a comprehensive and accurate derivation of complexes becomes problematic.

First, a complex may exist within a highly-connected region of the PPI network, with many extraneous outgoing edges connecting it to other proteins outside the complex. Such a complex is challenging to find, as it is difficult to delimit its boundaries accurately. A particular protein in the complex may have many extraneous PPI edges because it participates in other complexes as well, and the extraneous edges correspond to its interactions with the proteins in these other complexes. These distinct but overlapping (in composition) complexes may exist in different cellular locations, or may form in different cellular states which were detected by the PPI-screening technology, or may even exist in the same location and time as distinct complexes, but this information is not captured in the PPI network. These non-simultaneous interactions corresponding to distinct complexes are active in different copies of the protein, but in the PPI network these multiple copies of the protein are conflated into a single vertex, with all its non-simultaneous interactions corresponding to outgoing edges from that vertex, leading to the many extraneous edges.

The extraneous edges may also correspond to false positives due to a non-physiological environment of the assay, for example through over-expression of bait or prey proteins, or through detected interactions due to post-translational modifications that is different *in vivo*, or through Y2H-detected interactions in the nucleus where the interactors would not localize *in vivo*. Finally, the extraneous edges might

simply be an artifact of experimental or other biological variability that is inherent in dealing with biological systems.

Second, a complex may be sparsely connected in the PPI network, with few PPI edges detected between its proteins. Such a complex does not constitute a dense cluster which can be picked out by clustering algorithms. A complex may be sparse because it is condition-specific: only in certain conditions are its proteins expressed, or modified to enable binding, or co-localized, or the physiochemical environment appropriate for complex formation. If the complex only exists in a condition that was not tested during PPI screening, its proteins' co-complex interactions are not detected. PPIs could also be missing due to technological limitations. Under Y2H, proteins in the complex may not localize in the nucleus or interact in the nucleus where the interaction is assayed—in particular, PPIs in most membrane complexes are not detected. Since Y2H assays interactions in a non-physiological environment, the proteins might not have undergone post-translational modification required for binding, or the environment might be inappropriate for complex formation. Under TAP-MS, weaker interactions may not survive the double-washing step, though they may constitute important interactions within the complex. Finally, as with spurious interactions, missing interactions might also be due to variability in the experimental or biological system.

The third challenge, that of finding small complexes (defined as composed of two or three distinct proteins), is an intrinsic challenge which is exacerbated by the shortcomings of a static interactome. It has been noted that the distribution of complex sizes follows a power law distribution [31], meaning that a large majority of complexes are small. Thus the discovery of small complexes is an important subtask within complex discovery. An inherent difficulty in this task is that the strategy of searching for dense clusters becomes problematic: fully-dense (i.e. cliques) size-2 and size-3 clusters correspond to edges and triangles respectively, and only a few among the abundant edges and triangles of the PPI network represent actual small complexes. Furthermore, small complexes are much more sensitive to extraneous or missing edges: for a size-2 complex, a missing co-complex interaction disconnects its two member proteins, while only two extraneous interactions are sufficient to embed it within a larger clique (a triangle).

It is apparent that the challenge of small-complex discovery is exacerbated by the two problems of highly-connected regions with many extraneous edges, and sparse regions with many missing edges, in the PPI network. These problems, as described

above, owe a great deal to the analysis of a static interactome to derive complexes that are dynamic in nature.

2.4 Clustering algorithms for protein-complex discovery

To organize the wide variety of approaches that have been proposed to discover protein complexes from PPI data, we employ a taxonomy composed of five (possibly overlapping) categories: clique-based approaches, seed-and-grow approaches, simulation approaches, hierarchical clustering approaches, and core-attachment approaches.

Clique-based approaches

Broadly speaking, clique-based approaches first search for cliques (fully-connected sets of vertices) in the PPI network, then merge those cliques based on some criteria. CFinder [32] is a classic approach which finds the set of k-clique percolation clusters using the Clique Percolation Method (CPM [33]). For k = 3,4,..., it first searches for the set of all k-cliques (cliques composed of k vertices), then merges all k-cliques that are reachable to each other via adjacency, where two k-cliques are adjacent if they share exactly k-1 vertices. An updated version in 2008 uses CPM with weights (CPMw [34]) to handle weighted graphs as well, by requiring that a clique's intensity, or geometric mean of its edge weights, meets a given threshold.

Clustering by Maximal Cliques (CMC [35]) is another widely-used clique-based approach. Instead of searching for cliques of a given size (as in CFinder), CMC searches for the set of maximal cliques (cliques that are not contained within a larger clique). Then, for overlapping cliques whose overlap exceeds a threshold, CMC either merges them if they are highly interconnected, or removes the clique with the lower density. Another similar clique-based approach is Local Clique Merging Algorithm (LCMA [36]), which merges highly-overlapping local cliques that are found around every vertex.

Seed-and-grow approaches

Seed-and-grow approaches generally initialize each cluster as a seed corresponding to a vertex or a set of vertices, then grow the seeds by adding vertices to obtain the final clusters. MCODE [37], one of the earliest computational methods for finding complexes, is one such approach. It first weights each vertex with its local neighbourhood density, selects the highest weighted vertex as a seed, and grows it by adding highly-weighted neighbouring vertices to it until a threshold density is reached. This is repeated, by finding and growing the next seed from the un-added vertices. Recently,

Rhrissorrakrai proposed Module Identification in Networks (MINE [38]), a similar algorithm to MCODE with a modified vertex weighting strategy and the incorporation of a measure of network modularity during the growing phase.

The Density-Periphery Based Graph Clustering algorithm (DPClus [39]) is another classic seed-and-grow approach. It defines the weight of an edge as the number of common neighbours between the two vertices of the edge, the weight of a vertex as the sum of its incident edges, and the cluster property of a node with respect to a cluster which indicates whether the node is part of the cluster's periphery. A cluster is seeded from the vertex with the highest weight, and a neighbouring vertex is added based on two conditions: that it does not cause the cluster density to drop below a given threshold, and the cluster property of the vertex meets a given threshold, ensuring that the cluster's periphery is reasonably connected to the rest of the cluster. Li *et al.* proposed a modification of DPClus called IPCA [40] which grows clusters based on two novel conditions: cluster diameter, and a cluster connectivity-density requirement.

More recently, the algorithm ClusterOne [41] was proposed, which introduced a novel cohesiveness function of a cluster, the ratio of the sum of edge weights within the cluster versus the sum of edge weights within the cluster as well as outgoing edges from the cluster. ClusterOne selects seeds based on the vertices' degrees, and grows clusters greedily to maximize the cohesiveness function. Furthermore, highly-overlapping clusters are merged.

Optimization or simulation approaches

Optimization approaches search for a clustering or partitioning of the PPI network that optimizes some global function, and frequently model the PPI network as a random (typically Markovian) process. A classic approach is Markov Clustering (MCL [42]), which is based on the principle that a random walker in the PPI network will spend more time traversing a dense region before leaving it. The PPI network is represented as a transition matrix, and the probability of each node visiting every other node at each successive time step is calculated iteratively via matrix multiplication. An inflation step accentuates the differences in probabilities by raising them to a power and then re-normalizing. Regions that are densely connected, with sparse outgoing edges, are found as clusters.

Restricted Neighborhood Search Clustering (RNSC [43]) is a local-search algorithm that explores the solution space to minimize a cost function, calculated according to the number of intra-cluster and inter-cluster edges. RNSC first composes an initial

random clustering, and then iteratively moves nodes between clusters to reduce the clustering cost. It also makes diversification moves to avoid local minima. RNSC performs several runs, and reports the clustering from the best run.

PPSampler 2.3 [44] employs Markov Chain Monte Carlo to find a partition state of the PPI network that minimizes an objective function. A novelty of this method is the inclusion in the objective function of a term that specifies the size distribution of complexes found, which is observed to follow a power-law distribution.

Another optimization-based approach is Super Paramagnetic Clustering (SPC [45]), which models the PPI network as a network of interacting magnetic spins and finds clusters among spins with correlated fluctuating states.

Hierarchical clustering approaches

Hierarchical clustering algorithms create a dendogram (tree representation) of the hierarchical structure of the PPI network, and are frequently used to identify and organize functional modules in general rather than protein complexes specifically. However, the generated dendogram can be cut at a given level of granularity to obtain a set of clusters that correspond to complexes. Hierarchical clustering algorithms can either be agglomerative, which constructs the tree from leaves to root by merging subgraphs; or divisive, which constructs from root to leaves by splitting subgraphs. Hierarchical Agglomerative Clustering with Overlap (HACO [46]) is an extension of the common Hierarchical Agglomerative Clustering algorithm to allow overlaps in its clusters. It first considers all vertices as individual clusters, then iteratively merges pairs of clusters with high connectivity between them. At each merge, the two constituting clusters are remembered; when the merged cluster A is later merged with another cluster B, it also tries to merge the remembered constituting clusters of A with the cluster B, and keeps the (possibly overlapping) resultant clusters if they are highly connected.

Other hierarchical clustering approaches include the G-N algorithm [47], a divisive algorithm which iteratively removes edges with the highest betweenness centrality to obtain a hierarchy of modules; and MoNet [48], an agglomerative algorithm which also uses the betweeness centrality and a refined definition of modules.

Core-attachment approaches

Some complexes exhibit core-attachment functionality *in vivo*, where a subset of proteins in the complex forms a stable core which is functionally modulated or activated by the remaining proteins, called attachments, which may furthermore be

Algorithm	Category	Weighted edges	Overlapping clusters	Parameters
CFinder (CPMw)	Clique-based	Yes	No	Yeast: -k 4 -w .9 -I .92 Human: -k 4 -w .8 -I .84
CMC	Clique-based	Yes	Yes	Yeast: overlap=.5, merge=.5 Human: overlap=.5, merge=.75
IPCA	Seed-and-grow	No	Yes	Yeast: -P2 -T.4 Human: -P2 -T.6
ClusterOne	Seed-and-grow	Yes	Yes	Yeast and human: default
MCL	Optimization	Yes	No	Yeast: -I 2.5 Human: -I 4
RNSC	Optimization	No	No	Yeast and human: default
PPSampler	Optimization	Yes	No	Yeast and human: default
HACO	Hierarchical	Yes	Yes	Yeast: -c c .75 -g .1 Human: -c c .75 -g .5
Coach	Core-attachment	Yes	Yes	Yeast and human: default
MCL-CAw	Core-attachment	Yes	Yes	Yeast: -I 2, α=1, β=.4 Human: -I 2.5, α=1, β=.4

Table 2.1: Summary of the ten clustering algorithms tested, and their optimal parameters found for yeast and human complex discovery.

shared between multiple complexes [18]. Recently, researchers have proposed that such core-attachment structures may be discerned topologically in the PPI network as well, leading to the development of core-attachment approaches for finding complexes. Coach [49] detects complexes in two stages: core detection, and complex formation. In the first stage, neighbourhood subgraphs are induced around each vertex and its neighbours, and cores are found as vertices in each neighbourhood subgraph that have higher-than-average local degree, and whose induced subgraph is dense. In the second stage, proteins that are connected to at least some proportion of each core's vertices are recruited as attachments to the core.

MCL-CAw [50] incorporates a core-attachment model to refine clusters found by MCL, producing overlapping clusters that exhibit core-attachment structures. Given clusters found by MCL, it selects the core proteins within each cluster as those vertices that are highly interconnected, and discards clusters without any cores. Next, it recruits attachment proteins to cores as those remaining proteins from clusters that are highly connected to those cores, allowing attachments to be shared among multiple cores.

In this review we evaluate ten clustering algorithms representative of the different approaches: CFinder, CMC, IPCA, ClusterOne, MCL, RNSC, PPSampler, HACO, Coach, and MCL-CAw. Table 2.1 summarizes the features of these algorithms, and the best parameter settings found for prediction of yeast and human complexes.

2.5 Poor performance of current methods

In this section we evaluate the ten clustering algorithms listed in Table 2.1 for the prediction of yeast and human complexes. In particular, we highlight the three challenges of complex discovery that we described earlier: the prediction of complexes within highly-connected regions of the PPI network, the prediction of sparsely-connected complexes, and the prediction of small complexes. To approach these challenges individually, we first study the initial two challenges (complexes in highly-connected regions and sparsely-connected complexes) among large complexes only; finally we study small complexes, with an emphasis on those that are in highly-connected regions and those that are sparsely connected.

2.5.1 Data sources

PPI data

A number of repositories for PPI data are available, covering a range of organisms, interactions types (genetic interactions or physical PPIs), interactions sources (such as curated PPIs, experimental PPIs, or predicted PPIs), and experimental detection methods. A recent survey of PPIs in [51] includes a comprehensive summary and statistics of these repositories. In our work, we obtain our yeast and human PPIs by taking the union of physical PPIs from three repositories: BioGRID [52], IntAct [53], and MINT [54]. In addition, in yeast we also incorporate the widely-used Consolidated PPI dataset [23]. This dataset is a union of two high-throughput TAP-MS datasets from Krogan *et al.* [24] and Gavin *et al.* [18], scored and filtered by a sophisticated probabilistic framework called Purification Enrichment (PE) which was designed for TAP-MS data (and these two datasets in particular).

We unite these datasets, and score and filter the PPIs, using a simple reliability metric based on the Noisy-Or model to combine experimental evidences (also used in [55]). For each experimental detection method e, we estimate its reliability as the fraction of interactions detected where both interacting proteins share at least one high-level cellular-component Gene Ontology term. Then the score of an interaction (a, b) is estimated as:

$$score(a, b) = 1 - \prod_{i \in E_{a,b}} (1 - rel_i)^{n_{i,a,b}}$$

where rel_i is the estimated reliability of experimental method i, $E_{a,b}$ is the set of experimental methods that detected interaction (a, b), and $n_{i,a,b}$ is the number of

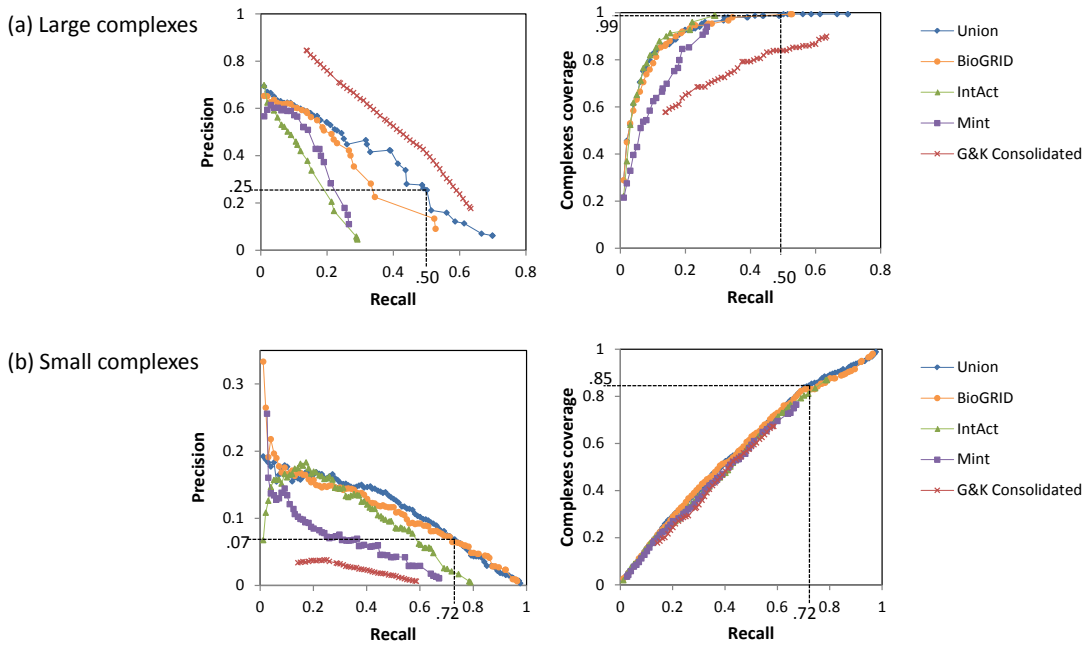

Figure 2.1: Precision-recall and complex-coverage graphs for classification of co-complex edges in yeast using different PPI datasets, for (a) large complexes, (b) small complexes.

times that experimental method i detected interaction (a, b). The scaled PE scores from the Consolidated dataset are discretized into ten equally-spaced bins $(0-0.1, 0.1-0.2, \ldots$, each of which is considered as a separate experimental method in our scoring scheme. We avoid duplicate counting of evidences across the datasets by using their publication IDs (in particular, PPIs from the Krogan and Gavin publications, which are represented in the Consolidated dataset, are removed from the BioGRID, IntAct, and MINT datasets).

Most clustering algorithms perform better when a smaller subset of high-quality PPIs are used. In our preliminary experiments (not shown), we found that taking the top $20,000$ edges gave decent performance in most clustering algorithms for discovering large complexes; for small complexes, taking the top $10,000$ gave decent performance.

Reference complexes for yeast and human

To evaluate the performance of complex-discovery algorithms, we use reference complexes that have been manually validated via literature curation. For yeast, we use the CYC2008 set, which consists of 408 yeast complexes [56]. For human, we use the CORUM set, which consists of 1829 human complexes [57].

To check how well our scored yeast and human PPIs correspond to actual co-complex protein pairs (two proteins within the same complex), we plot their precision-recall graphs. First, given a set of reference complexes $\mathbf{C}$, define CP as the set of

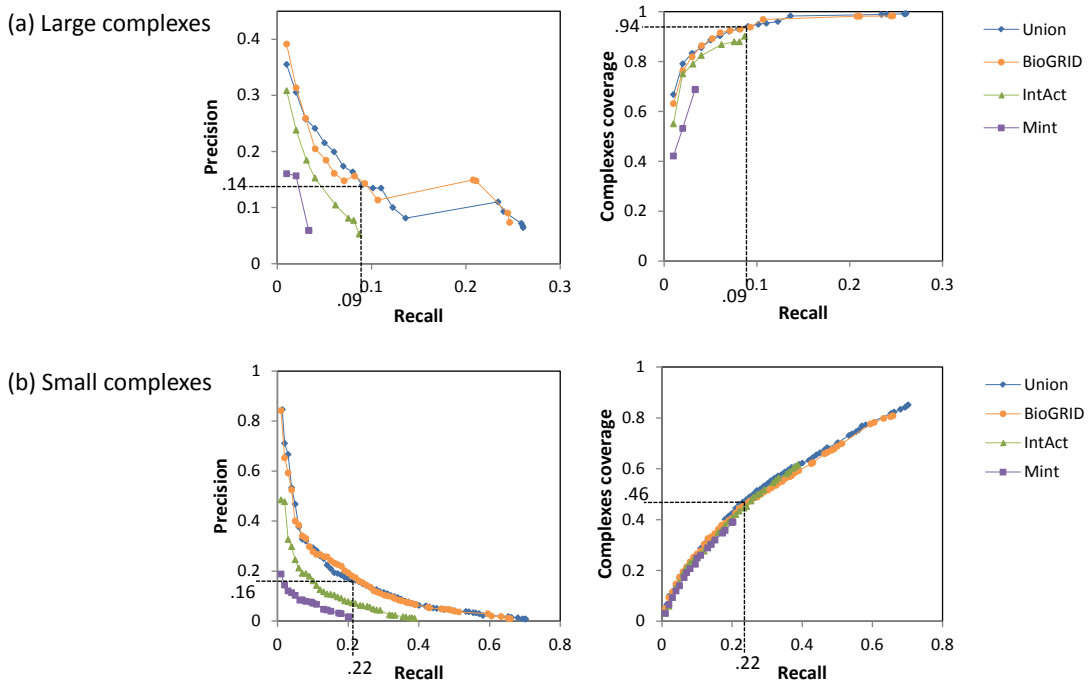

Figure 2.2: Precision-recall and complex-coverage graphs for classification of co-complex edges in human using different PPI datasets, for (a) large complexes, (b) small complexes.

co-complex pairs from $\mathbf{C}$:

$$CP = \{(a,b)|a \in C_i \wedge b \in C_i \wedge C_i \in \mathbf{C}\}$$

Given a set of scored PPIs I, the precision and recall at a score threshold t are given as:

$$precision_t = \frac{|\{(a,b) \in I|score(a,b) \geq t \wedge (a,b) \in CP\}|}{|\{(a,b) \in I|score(a,b) \geq t\}|}$$

$$recall_t = \frac{|\{(a,b) \in CP|(a,b) \in I \wedge score(a,b) \geq t\}|}{|CP|}$$

To quantify how well a set of PPIs are distributed among the reference complexes $\mathbf{C}$, we also define the coverage of complexes of the PPIs at score threshold t as:

$$coverage_t = \frac{|\{C_i \in \mathbf{C}|\exists(a,b) \in I \wedge score(a,b) \geq t \wedge a \in C_i \wedge b \in C_i\}|}{|\mathbf{C}|}$$

We can plot a precision-recall graph and a coverage-recall graph from the set of precision, recall, and coverage points obtained by varying the score threshold t. Figure 2.1 show the precision-recall graphs (left charts) and coverage-recall graphs (right charts) for yeast PPIs from the four source datasets separately (BioGRID, IntAct, MINT, and Consolidated) as well from our union dataset, in predicting co-complex protein

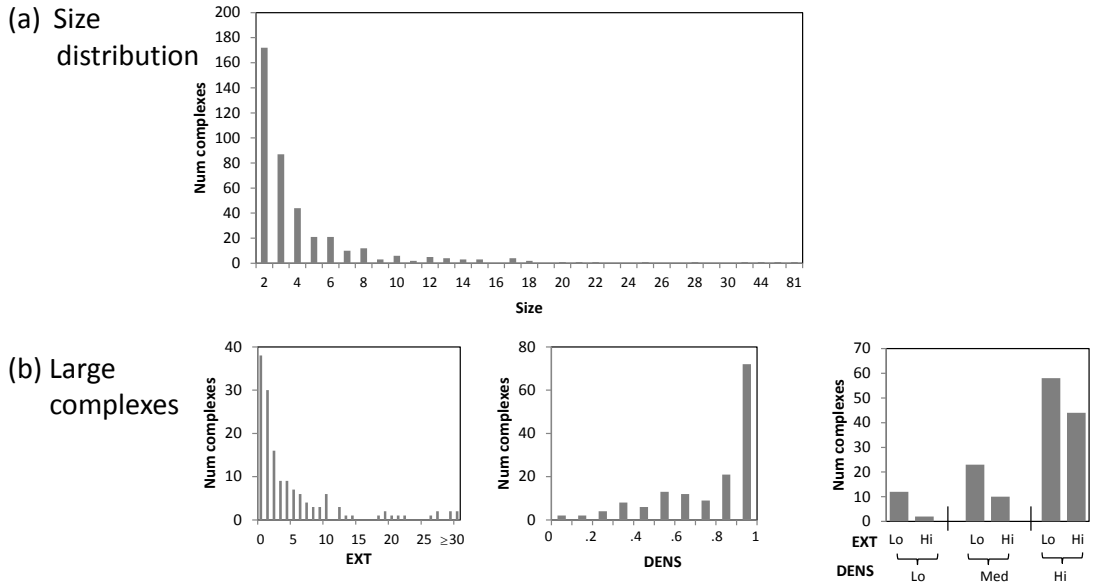

Figure 2.3: Statistics of the yeast reference complexes, from the CYC2008 database. (a) The size distribution of the complexes. (b) EXT (number of highly-connected external proteins) and DENS (density) distributions of large complexes.

pairs from large and small complexes separately. For large complexes (Figure 2.1a), our union dataset achieves higher recall and precision compared to using BioGRID, IntAct, or MINT, but has lower precision compared to the Consolidated dataset. However, the coverage-recall graph shows that the PPIs from the Consolidated dataset cover much fewer complexes. Furthermore, among small complexes (Figure 2.1b), the Consolidated dataset has the lowest recall, precision, and complexes coverage. Thus, we conclude that the widely-used Consolidated dataset is of higher quality only among a subset of large complexes: its PPIs are clustered together in fewer complexes, and moreover do not correspond well to protein pairs in small complexes. Thus we use our Union PPIs in our experiments to cover a wide range of both large and small complexes with decent quality.

Figure 2.2 shows the corresponding graphs for human PPIs. Here our Union dataset has similar quality as the BioGRID dataset alone, but for consistency we use the Union PPIs in our experiments for human complexes.

As mentioned above, taking the top $20,000$ and $10,000$ edges gave decent performance for most clustering algorithms, in large and small complex discovery respectively. The corresponding precision, recall, and coverage obtained from taking these cutoffs are shown in Figures 2.1 and 2.2.

To investigate the performance of the clustering algorithms with respect to the three highlighted challenges, we stratify the reference complexes in terms of their sizes, extraneous edges, and densities. First, to quantify whether a complex is embedded

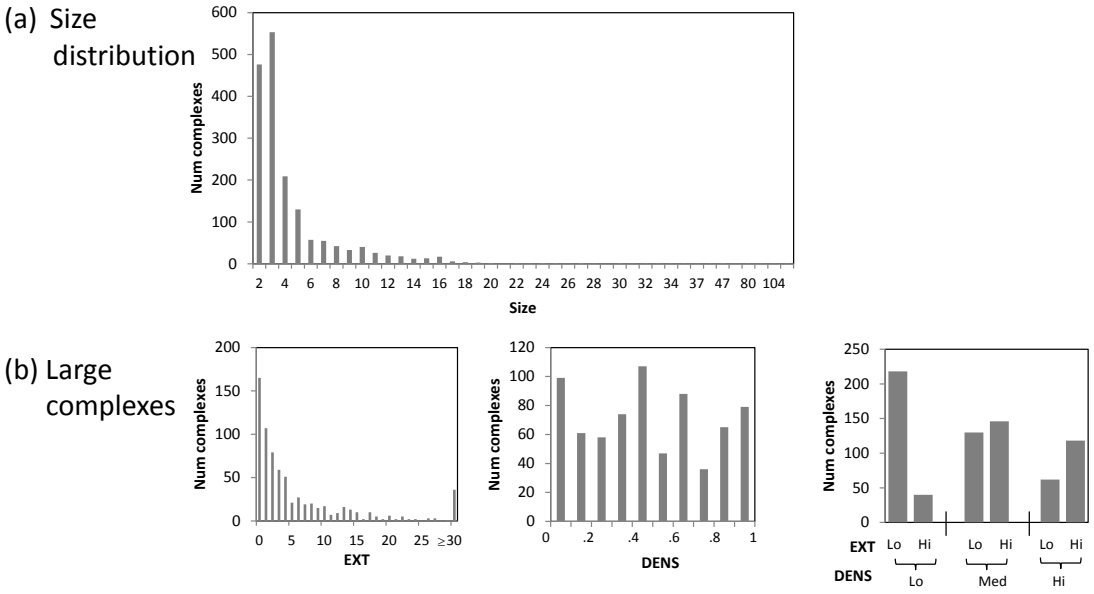

Figure 2.4: Statistics of the human reference complexes, from the CORUM database. (a) The size distribution of the complexes. (b) EXT (number of highly-connected external proteins) and DENS (density) distributions of large complexes.

within a highly-connected region of the PPI network, we derive EXT, the number of external proteins that are highly connected to it, defined as being connected to at least half of the proteins in the complex. Second, to quantify how sparse a complex is, we derive DENS, the density of each complex, defined as the number of PPI edges in the complex divided by the total number of possible edges in the complex. In our analysis, we stratify the complexes into large and small complexes, and further stratify the large complexes into low, medium, and high DENS (corresponding to DENS of $[0, .35]$, $(.35, .7]$, and $(.7, 1]$ respectively), and low and high EXT (corresponding to EXT ≤ 3 and > 3 respectively), to give seven total strata (one for small complexes, and six for large complexes).

Figure 2.3 illustrates the size distribution of the yeast complexes, and the distributions of EXT, DENS, and our six analysis strata (stratified by EXT and DENS), among the large yeast complexes. Figure 2.4 shows the corresponding distributions for human complexes. In both yeast and human, the sizes of complexes follow the power-law distribution [31], which highlights the important subtask of predicting small complexes (of size two and three): among both yeast and human complexes, about 60% are small complexes (259 out of 408 in yeast, 1029 out of 1829 in human).

Among large complexes in both yeast and human, about 40% of complexes have high EXT. We expect the prediction of these complexes to be extremely challenging, as it would be difficult to accurately delimit their borders from their highly-connected surroundings (the highly-connected external proteins are likely to be recruited into the

predicted complexes). Only 10% of large complexes in yeast have low density. On the other hand, in human about 35% of large complexes are sparsely-connected with low DENS. We expect these sparsely-connected complexes to also be difficult to predict, as they do not form dense clusters that are picked out by most clustering algorithms.

2.5.2 Evaluation methods

For any cluster P produced by any clustering algorithm, we define its score as its weighted density:

$$score(P) = dens(P) = \frac{\sum_{u \in P, v \in P} w(u,v)}{|P|.(|P|-1)}$$

where $w(u,v)$ is the weight of edge (u,v).

We say that a cluster (i.e. a predicted complex) P matches a known complex C at a given match threshold $match\_thresh$ if $Jaccard(P,C) \geq match\_thresh$, where $Jaccard(P,C)$ is the Jaccard similarity between the proteins contained in P and C:

$$Jaccard(P,C) = \frac{|V_P \cap V_C|}{|V_P \cup V_C|}$$

where V_X is the set of proteins contained in X. For large complexes, we use a stringent matching criteria of $match\_thresh = 0.75$ in matching yeast complexes, and a rougher matching criteria of $match\_thresh = 0.5$ in matching human complexes, as the latter is much more difficult. For small complexes, we use the most stringent criteria of $match\_thresh = 1$, as it is easier for a small cluster to match a small complex by chance. Given a set of clusters $\mathbf{P} = \{P_1, P_2, \ldots\}$, and a set of reference complexes $\mathbf{C} = \{C_1, C_2, \ldots\}$, we define the precision and recall of the clusters at score threshold d as:

$$precision_d = \frac{|\{P_i \in \mathbf{P} | dens(P_i) \geq d \wedge \exists C_j \in \mathbf{C}, P_i \; matches \; C_j\}|}{|\{P_k \in \mathbf{P} | dens(P_i) \geq d\}|}$$

$$recall_d = \frac{|\{C_i \in \mathbf{C} | \exists P_j \in \mathbf{P}, dens(P_j) \geq d, P_j \; matches \; C_i\}|}{|\mathbf{C}|}$$

We can plot the precision-recall graph of a set of predicted clusters, by using the precision-recall points obtained by varying the cluster score threshold d.

We also use four statistics to summarize the performance of each complex-discovery algorithm: the area-under-curve (AUC) of its precision-recall graph; the precision of all its predicted clusters (without any cluster score threshold); likewise, the recall of all

its predicted clusters; and the F-measure of all its predicted clusters, which is defined as the harmonic mean of the precision and recall:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

2.5.3 Results

Figure 2.5a shows the performance of the ten clustering algorithms on prediction of large yeast complexes, at a fine-resolution matching level of $match\_thresh = 0.75$. Five algorithms achieve substantially higher recall than the others: CMC, IPCA, Coach, HACO, and RNSC have recalls of $35\% - 45\%$. Of these five algorithms, IPCA, Coach, HACO, and CMC also suffer from low precision levels (although CMC's precision is ranked third, it is still markedly lower than the two highest precision levels of RNSC and CFinder). Thus it is apparent that the prediction of the yeast complexes at this fine resolution is a difficult task, as the algorithms that best manage to predict these complexes also tend to generate many false positive clusters at the same time. An exception is RNSC, which achieves a balance between precision and recall, attaining the highest F-measure as a result, although its recall is almost 10% lower than CMC's.

Figure 2.5b shows the performance of the clustering algorithms on the prediction of small yeast complexes, at a perfect matching requirement of $match\_thresh = 1.0$. CFinder, Coach, and MCL-CAw perform poorly, predicting fewer than 5% of small complexes. It is clear that the core-attachment models (of Coach and MCL-CAw) is challenging for such small complexes, as it is problematic to define tightly-connected cores with less-connected attachments when only two or three vertices are available. While HACO and IPCA achieve the highest recall of almost 50%, they also attain the lowest precision levels, showing that the algorithms that predict the most complexes also suffer from many false positives.

Figure 2.6a shows the performance of the ten clustering algorithms on prediction of large human complexes, using a rougher-resolution matching level of $match\_thresh = 0.5$, as prediction of human complexes is a more difficult task (at $match\_thresh = 0.75$, the highest recall achieved is only about 10%, not shown). Even at the lowered $match\_thresh$, only IPCA manages recall of over 40%, while it suffers from low precision of 20%. Similarly, Coach and HACO achieve recalls of $30\% - 35\%$, with low precisions of $15\% - 20\%$. The highest F scores are attained by CMC, with precision and recall of 35% and 27%, and RNSC, which achieves the highest precision of 37% but a low recall of 23%. It can be seen that most human complexes cannot be predicted

even at low matching resolution (and at higher matching resolution the vast majority cannot be predicted), and moreoever those algorithms that do predict some complexes also predict many false positive clusters.

Figure 2.6b shows the performance of the clustering algorithms on small human complexes, which is a much more difficult task: here the highest recall and precision attained are both slightly above 10%. Again, CFinder, Coach, and MCL-CAw perform poorly, predicting less than 1% of complexes. HACO is able to achieve both recall and precision to give the highest F measure, while CMC and IPCA achieve relatively high recall (around 10%), but also suffer from the two lowest precision levels.

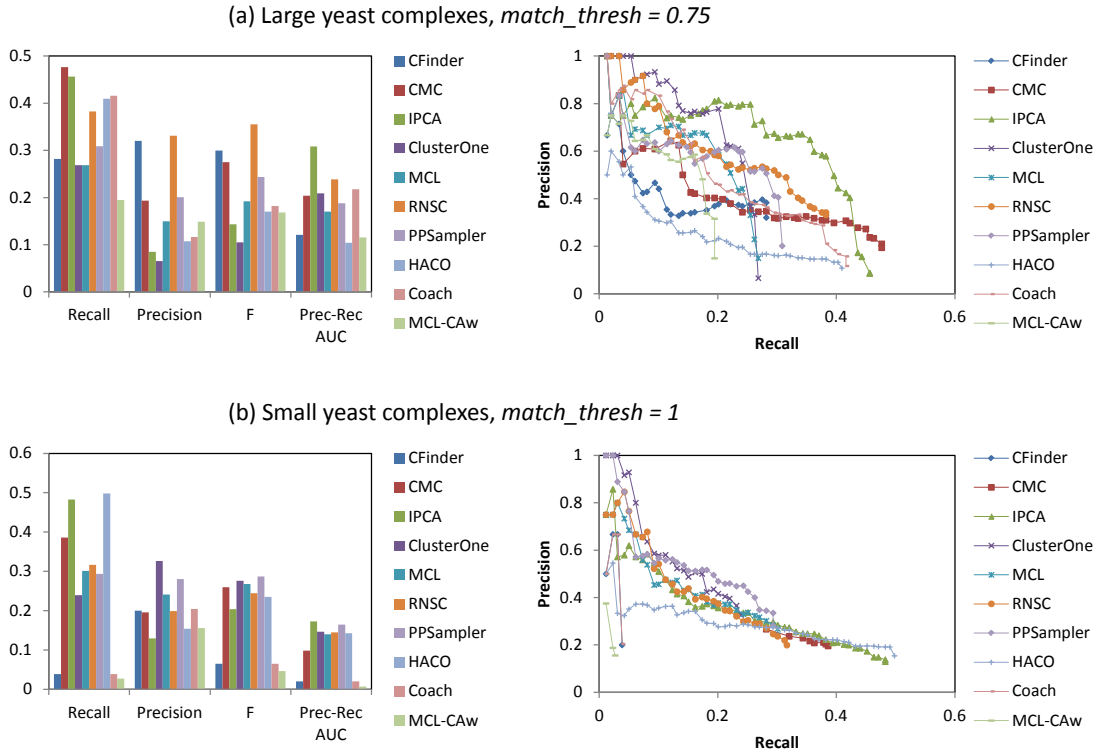

Figure 2.5: Performance of the ten clustering algorithms on prediction of yeast complexes, with (a) $match\_thresh = 0.75$ for large complexes, (b) $match\_thresh = 1$ for small complexes. The left chart shows the precision, recall, F score, and AUC of the precision-recall graph. The right chart shows the precision-recall graph.

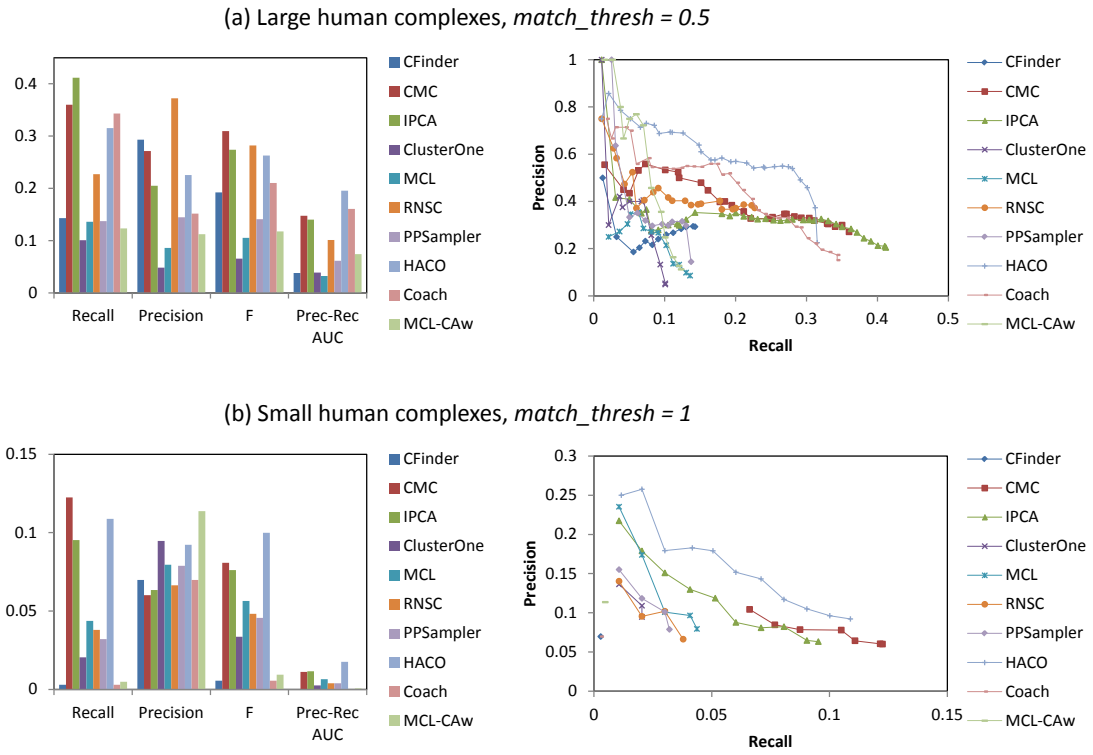

Figure 2.6: Performance of the ten clustering algorithms on prediction of human complexes, with (a) $match\_thresh = 0.5$ for large complexes, (b) $match\_thresh = 1$ for small complexes. The left chart shows the precision, recall, F score, and AUC of the precision-recall graph. The right chart shows the precision-recall graph.

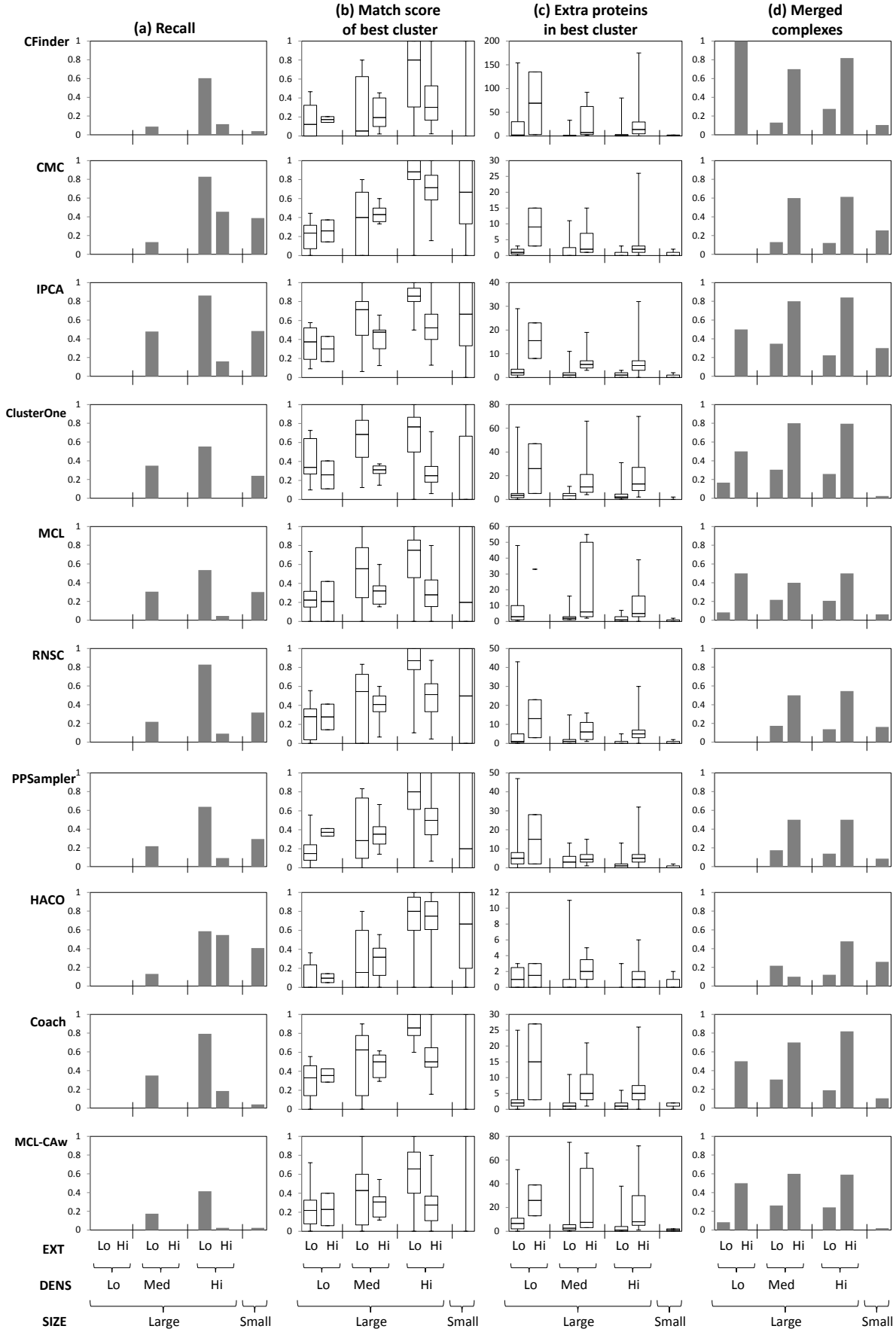

Figure 2.7: Performance of complex-discovery algorithms on yeast complexes, stratified by size, DENS, and EXT. The x-axis of each chart corresponds to the different stratified groups of complexes, given at the bottom of the figure.

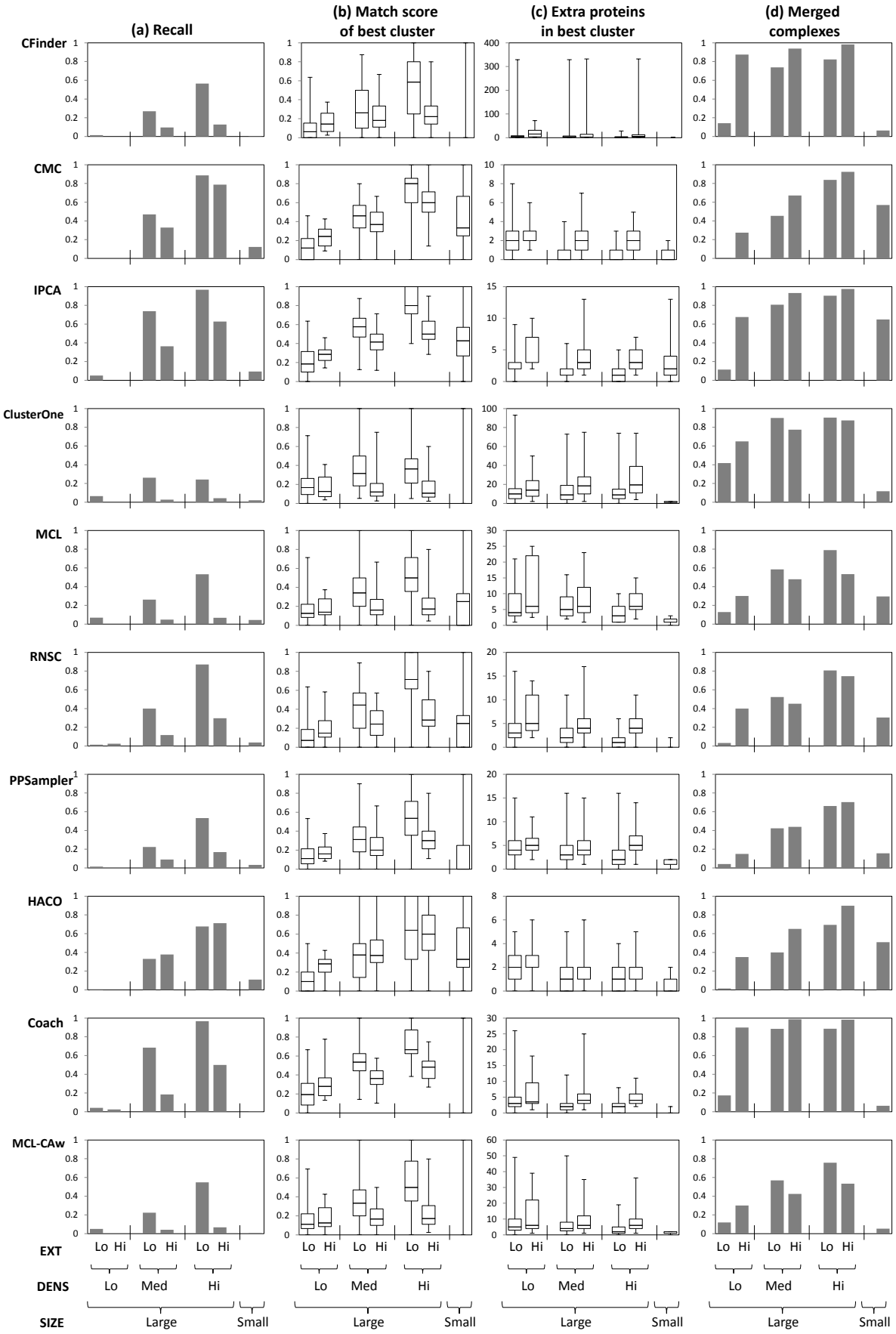

Figure 2.8: Performance of complex-discovery algorithms on human complexes, stratified by size, DENS, and EXT. The x-axis of each chart corresponds to the different stratified groups of complexes, given at the bottom of the figure.

To investigate which complexes are problematic to predict, we study the performance of the complex-discovery algorithms on the complexes stratified in terms of their sizes, extraneous edges, and densities. As described above, the complexes are stratified into small and large complexes, and large complexes are further stratified by density (DENS) and number of highly-connected external proteins (EXT), to give seven groups of complexes (see Figures 2.3 and 2.4 for the distribution of size, DENS, and EXT of yeast and human complexes).

Figure 2.7a shows that yeast complexes with lower density are much harder to predict than those with higher density: no complex with low DENS are predicted at all by any clustering algorithm, while complexes with high DENS are predicted much more frequently. Furthermore, complexes with higher EXT are harder to predict than those with lower EXT: in each density strata,complexes with high EXT have lower recall than those with low EXT. Small complexes are also challenging to predict: most clustering algorithms do not predict more than 40% of small complexes. As expected, the easiest complexes to predict are the large complexes with high DENS and low EXT.

Figure 2.7b shows that complexes with higher density can be predicted with better-matching clusters: within each EXT strata, the match score increases with density. Furthermore, complexes with lower EXT are predicted with better-matching clusters: among complexes with medium or high DENS, match score is higher among those with low EXT than high EXT (in the low-DENS stratum, only 2 complexes have high EXT, making comparisons here difficult).

Figures 2.7c and d reveal why complexes with higher EXT are difficult to predict. Figure 2.7c shows that clustering algorithms tend to include many extraneous proteins when predicting complexes with higher EXT: across all DENS strata, complexes with higher EXT have greater number of extra proteins in their best-matched clusters (intuitively, the extraneous proteins are likely to be those highly-connected external proteins). Figure 2.7d shows that clustering algorithms tend to merge together complexes with higher EXT: across all DENS stratas, complexes with higher EXT tend to be found in clusters merged with other complexes.

Figure 2.8 shows the corresponding performance of the clustering algorithms on the stratified human complexes. Similar conclusions can be drawn here as from yeast complexes. Small complexes are challenging to predict, with most clustering algorithms predicting less than 10% of them. Complexes with lower density are harder to predict than those with higher density, and are predicted with clusters that match

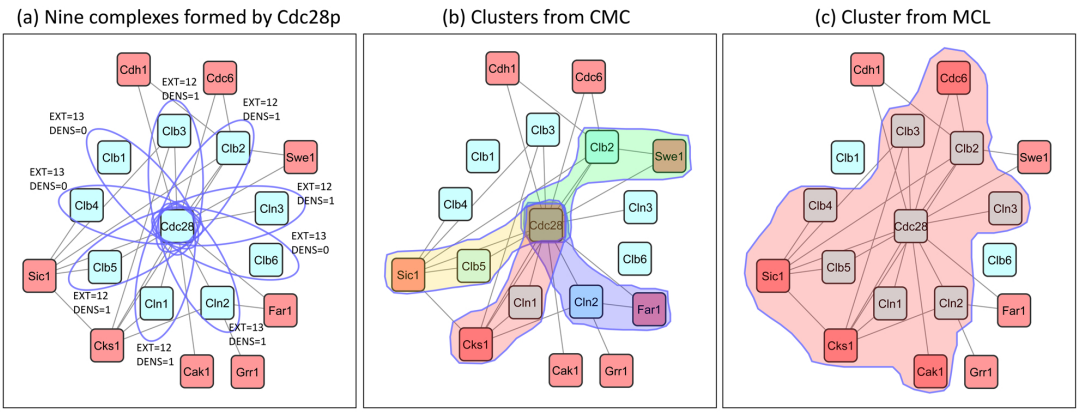

Figure 2.9: (a) Cdc28p is involved in nine distinct complexes, which overlap and have many highly-connected external proteins (EXT). Three of the complexes are disconnected (DENS= 0). (b) CMC includes extraneous proteins in its clusters. (c) MCL merges the complexes.

them less well; likewise, complexes with higher EXT are also harder to predict than those with lower EXT, and are also predicted with clusters that match them less well (Figures 2.8a and b). However, Figure 2.8b shows that, within the low-DENS stratum, complexes with high EXT attain slightly higher match scores than those with low EXT, because these low-density complexes with high EXT are likely to slightly overlap with clusters consisting of complex proteins with the external proteins that they are highly-connected to; indeed, in these cases the match scores are mostly under 0.5.

Figure 2.8c shows that, as in yeast, human complexes with high EXT are predicted with clusters that include many more extraneous proteins. Figure 2.8d shows that complexes with higher EXT tend to be merged together in clusters (although this is not seen for clusters predicted by ClusterOne, RNCS, MCL, and MCL-CAw).

2.5.4 Example Complexes

Here we highlight some example complexes that are known to behave dynamically, and show how their static interactomes exhibit characteristics (such as high EXT and low DENS) which result from their static representation, and which make them difficult to predict.

The Cdc28p yeast protein, as described above, complexes with various cyclin proteins (Cln1p to Cln3p, Clb1p to Clb6p) to regulate the cell cycle. While the abundance of Cdc28p is constant throughout the cell cycle, the activity of the cyclin proteins are regulated via sophisticated gene-expression and post-translational controls, so that the proper complexes are formed at each point of the cell cycle [2,3]. Figure 2.9a shows the interactome around these proteins and their neighbours, with the nine different com-

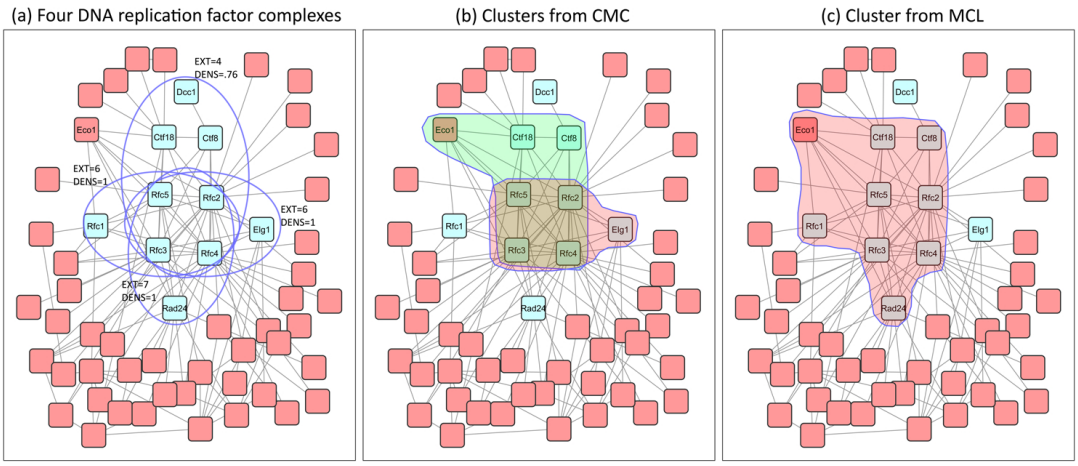

Figure 2.10: (a) A common core is shared among four DNA replication factor complexes, which contributes to a high number of external proteins (EXT) in each complex. (b) CMC finds only one of the four complexes. (c) MCL merges three of the four complexes.

plexes formed by Cdc28p circled. Although these interactions occur at different times during the cell cycle (e.g. Cdc28p-Cln1p and Cdc28p-Cln2p in G_1 phase, Cdc28p-Clb2p in G_2M phase), they are collapsed into the same static interactome, resulting in a highly-connected region around Cdc28p and its cyclin partners: note that the EXT for each of the complexes range from 12 to 13. Furthermore, PPIs are missing between CDC28p and some of its cyclin partners (Clb1p, Clb4p, Clb6p), giving a density of 0 to these complexes. In fact, these PPIs exist in our source datasets, but with slightly fewer experimental evidences to back them up compared to the other Cdc28p PPIs; thus they scored slightly lower in reliability and they were filtered from our PPI network. While it is possible to lower our reliability score cutoff to include these PPIs, this would also include many spurious PPIs and make the discovery of other complexes even more difficult.

Figure 2.9b and c show the clusters predicted by CMC and MCL respectively. CMC found four clusters that overlap with four Cdc28p complexes, but with one extraneous protein in each case, while MCL found one large cluster that covered Cdc28p, seven of the nine cyclin proteins, and four extraneous proteins.

The four Replication Factor C (RFC) complexes in yeast are structurally similar complexes involved in DNA metabolism. Each of these complexes consist of four subunits (Rfc2p to Rfc5p), and distinct attachment proteins: the first one with Rfc1p, involved in DNA metabolism; the second with Ctf8p, Ctf18p, and Dcc1p, involved in sister chromatid cohesion; the third with Elg1p, involved in maintaining genome integrity; and the fourth with Rad24p, involved in checking for DNA damage [58]. The interactome of the RFC complexes and their neighbours are shown in Figure 2.10a,

with the four complexes circled. Here again, conflating the four distinct complexes in the static interactome results in many extraneous edges and high connectivity to proteins outside each complex: the EXT for the four complexes range from 4 to 7.

Figure 2.10b and c show the clusters predicted by CMC and MCL respectively. CMC predicted one of the RFC complexes perfectly, while predicting a second cluster that matched another complex less well; MCL predicted a large cluster that overlapped with three of the RFC complexes.

Note that MCL does not allow overlaps in its predicted clusters, so in the above examples it predicts clusters that merge the overlapping and highly-connected complexes together. While CMC allows overlapping clusters, the many extraneous edges and high connectivity to external proteins make it difficult to delimit the overlapping complexes precisely.

2.6 Discussion

Protein interactions behave in a dynamic fashion, with a variety of interaction timings, locations, and affinities. The cellular control of this dynamism gives important functional mechanisms to protein complexes, allowing complexes to assemble at specific times, or to vary in composition to activate or modulate their functions. Interaction detection technologies are limited in their ability to capture such dynamics; furthermore, this dynamism also impedes accurate and comprehensive screening of interactions. Moreover, the representation of interactions in a PPI network does not preserve any information about interaction dynamism, allowing only a static analysis of a dynamic reality.

In Section 2.3 we identified three challenges in complex prediction that result from, and are exacerbated by, the analysis of the static interactome to derive complexes that behave dynamically in nature. First, many proteins participate in multiple complexes, leading to overlapping complexes embedded within highly-connected regions of the PPI network with many extraneous edges connecting them to external proteins. This makes it difficult to accurately delimit the boundaries of such complexes. Second, many condition- and location-specific PPIs are not detected, leading to sparsely-connected complexes that cannot be picked out by clustering algorithms. Third, the majority of complexes are small complexes (made up of two or three proteins), which are extra sensitive to the effects of extraneous edges and missing co-complex edges.

In Section 2.5 we presented results of ten clustering algorithms for prediction of large and small complexes in yeast and human, and showed that only large complexes

with high density and few highly-connected external proteins can be consistently predicted: more than 80% of such large complexes can be predicted in yeast and human (with $match\_thresh = 0.75$ and 0.5 respectively).Complexes with low density frequently could not be predicted at all, while those with many highly-connected external proteins tended to be predicted in clusters with many extraneous proteins or merged complexes. Small complexes are also challenging to predict, particularly in human for which recall rates are extremely low; given that the majority of complexes are small, this means that a sizable number of all complexes cannot be predicted.

Drawing on our insight into the causes of these challenges, we propose an approach for each of these problems that can improve the performance of complex discovery. To discover sparse complexes, we use a naive-Bayes supervised learning approach to integrate multiple sources of data besides PPIs, which adds missing co-complex edges to sparse complexes as well as reduces the amount of spurious edges. This method is described in Chapter 3.

To discover complexes within highly-connected regions, we decompose the PPI network into subnetworks of PPIs that are localized in separate cellular locations, and furthermore remove hub proteins (proteins with high degree) that may participate in multiple non-simultaneous interactions, before performing complex discovery. This method is described in Chapter 4.

To discover small complexes, we integrate PPI data with additional data sources along with their topological features, using a supervised approach to weight edges with their posterior probabilities of belonging to small complexes versus large complexes. Small complexes extracted from the weighted network are scored using the probabilistic weights of edges within, as well as surrounding, the complexes. This method is described in Chapter 5.

Finally in Chapter 6 we combine all our proposed methods for the prediction of both large and small complexes, and show that this ameliorates many of the difficulties in discovering dynamic protein complexes from an analysis of the static interactome.

Chapter 3

Supervised Weighting of Composite Protein Networks

3.1 Introduction

Protein complexes are typically predicted based on topological characteristics in the PPI network. For example, many approaches search for regions of high density or connectivity [35, 41–43, 46]. Other approaches further incorporate subgraph diameters of known complexes [40], and core-attachment models of connected clusters [49, 50]. Qi *et al.* used a set of topological features including density, degree, edge weight, and graph eigenvalues, with a supervised naive-Bayes approach to learn these feature parameters from training complexes [59].

The performance of these complex-discovery algorithms is reliant on the quality of the protein interaction data, which is often associated with substantial numbers of spuriously-detected interactions (false positives) and missing interactions (false negatives). In particular, sparse complexes, with many missing PPIs between their constituent proteins, cannot be picked out by most complex-discovery algorithms as they do not constitute dense clusters in the PPI network. The sparseness of such complexes could be because they are condition-specific: only in certain conditions are their proteins expressed, or modified to enable binding, or co-localized, or the physiochemical environment appropriate for complex formation. If the complexes only exist in conditions that were not tested during the PPI screening assay, their proteins' co-complex interactions are not detected. PPIs could also be missing due to technological limitations. Under the yeast two-hybrid assay (Y2H), proteins in complexes might not be able to localize or interact in the nucleus where the interaction is assayed; in particular, PPIs in most membrane complexes are not detected. Since Y2H assays interactions in a non-physiological environment, the proteins might not have undergone post-translational modification required for binding, or the environment might

be inappropriate for complex formation. Under Tandem-affinity purification (TAP), weaker interactions may not survive the double washing step, though they may constitute important interactions within a complex. Finally, missing interactions might also be due to variability in the experimental or biological system.

Spurious interactions also present a challenge for complex discovery: a complex with many extraneous outgoing edges is challenging to find, as it is difficult to delimit its boundaries accurately. Such interactions may be due to extremely transient, non-specific binding, in processes such as ubiquitination. Spurious interactions may be caused by a non-physiological environment of the assay, for example through over-expression of bait or prey proteins, or through detected interactions due to post-translational modifications that is different from *in vivo*, or through Y2H-detected interactions in the nucleus where the interactors would not localize *in vivo*. Alternatively, the extraneous edges might simply be an artifact of experimental or other biological variability that is inherent in dealing with biological systems.

Figure 3.1 provides an illustrative example of these challenges. The mitochondrial cytochrome bc1 complex is a well-known complex involved in the electron-transport chain in the mitochondrial inner membrane. In *Saccharomyces cerevisiae* (yeast), this complex is composed of ten proteins. Figure 3.1 shows the PPI subgraph around these ten proteins, using PPI data obtained from BioGRID [52], IntAct [53], MINT [54], and the Consolidated [23] datasets. Nineteen PPIs (out of a possible 45) were detected between these ten proteins; the rest remain undetected, likely due to the difficulty of detecting interactions between membrane proteins, or because not all proteins in this complex interact with each other. 145 extraneous interactions were detected between the proteins from this complex and 94 proteins outside the complex. While some of these extraneous interactions might be spuriously detected, others constitute non-specific interactions. Five proteins likely involved in such non-specific interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination respectively, and bind to many proteins to perform their functions; PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain, and interact with proteins of the complex, although they are not part of it. The density of the complex is lost amidst the noise of the extraneous interactions, making the discovery of this complex from PPI data extremely difficult: none of the six complex-discovery algorithms we use here successfully detected it.

Many algorithms have been developed to assess the reliability of high-throughput protein interactions [61–63] or predict new protein interactions [35,64–67], using various

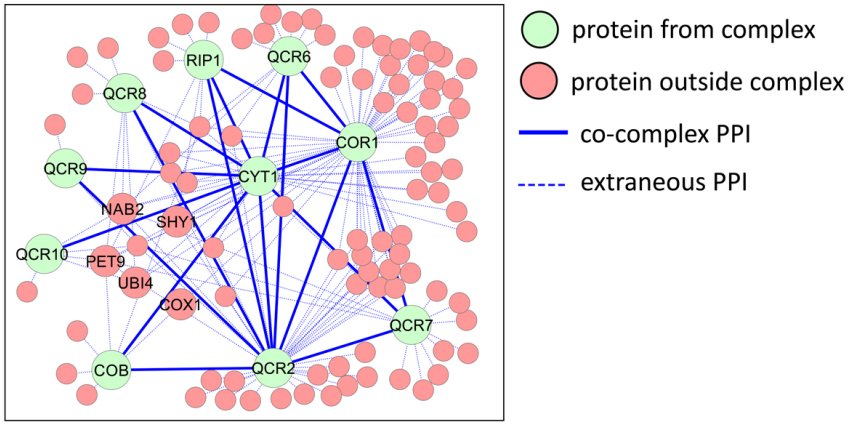

Figure 3.1: PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [60].

information such as gene sequences, annotations, interacting domains, 3D structures, experimental repeatability, or topological characteristics of PPI networks. These approaches have been shown to be effective in reducing false positives or false negatives.

Researchers have also proposed integrating heterogeneous data sources with supervised approaches to predict co-complex protein pairs (protein pairs that belong to the same complex), using a reference set of training complexes. Data integration leverages on the fact that diverse data sources other than PPI can also reveal co-complex relationships, while a supervised approach targeted at predicting co-complex protein pairs can be trained to discriminate between actual co-complex interactions and spuriously-detected or non-specific interactions. Qiu and Noble [68] integrated PPI, protein sequences, gene expression, interologs, and functional information, to train kernel-based models, and achieved high classification accuracy in predicting co-complex protein pairs. However, they did not apply or test their method on reconstructing and predicting complexes. Wang *et al.* [46] integrated PPI, gene expression, localization annotations, and transmembrane features, and applied a boosting method to predict co-complex protein pairs. They showed that this approach, combined with their proposed clustering method HACO, achieved higher sensitivity in recovering reference complexes compared to unsupervised approaches. However, they did not explore how well their classification approach works when used in conjunction with other clustering methods: while sensitivity was improved, many reference complexes were still unable to be predicted in part due to limitations of HACO, thus raising the question of whether other clustering methods may also see an improvement when used with their

co-complex predictions. Furthermore, these approaches directly produce co-complex affinity scores between protein pairs, without providing measurements of the predictive strengths of the different data sources, nor how the different score values of each data source indicate co-complex relationships. In our view, this is important when integrating different data sources: while using PPI for complex prediction is biologically reasonable because proteins in a complex interact and bind with each other, using other data sources such as sequences, expression, or literature co-occurence is not as biologically intuitive, even if they do reveal co-complex relationships. Providing a measurement of how these data sources contribute to co-complex predictions allows human judgment of the validity and credibility of novel predicted complexes.

We propose a method to address these challenges of complex discovery: first, the PPI network is integrated with other heterogeneous data sources that specify relationships between proteins, such as functional association and co-occurrence in literature, to form an expanded, composite network. Next, each edge is weighted based on its posterior probability of belonging to a protein complex, using a naive-Bayes maximum-likelihood model learned from a set of training complexes. A complex-discovery algorithm can then be used on this weighted composite network to predict protein complexes. Our method offers several advantages over current unsupervised or non-integrative weighting approaches. First, a composite protein network constructed from multiple data sources is more likely to have denser subgraphs for protein complexes, as it not only reduces the number of missing interactions, but also adds edges between non-interacting proteins from the same complex, because such proteins are likely to be related in ways other than by physical interactions. Second, learning a model from training complexes not only provides a powerful method to assess the reliability of interactions, but also allows the discrimination between non-specific and co-complex interactions. Third, utilizing multiple data sources to assess the reliability of interactions is likely to be more accurate than using just PPI data.

Our choice of a naive-Bayes maximum-likelihood model also offers several advantages over other supervised data-integration approaches. Firstly our model is transparent, in that learned parameters can be validated and analyzed, for example to reveal the predictive strengths of the different data sources. Furthermore, for a predicted complex, the learned parameters can then be used to visualize the component evidences from the different data sources, allowing human judgment of the credibility of the prediction. Second, maximum-likelihood models are known to be robust and have low variance, even when few training samples are available. Although we describe our

experiments using yeast and human, this is important when we apply our approach to less-studied organisms with fewer known complexes available for training. Finally, we utilize different clustering algorithms as well as a simple aggregative clustering strategy to evaluate the performance of our method, and show that we improve the performance of complex prediction compared to other weighting methods.

3.2 Methods

3.2.1 Building the composite network

Heterogeneous data sources are combined to build the composite network. Each data source provides a list of scored protein pairs: for each pair of proteins (u, v) with score s, u is related to v with score s, according to that data source. For both yeast and human, the following data sources are used:

- PPI data is as described in Chapter 2.5.1, obtained by combining physical interactions from multiple databases, then scored by reliability, and filtered to take the top $20,000$ edges.

- PPI topological data is obtained by scoring the PPIs using a topological function, Iterative AdjustCD (with two iterations), which has been shown to improve the performance of complex discovery [35]. Iterative AdjustCD uses expectation maximization to score each interaction (u, v) based on the number of shared neighbors of u and v. Interactions between proteins that have no shared neighbors are regarded as unreliable and are discarded. Protein pairs that do not directly interact but have shared neighbors are also scored, with pairs scored above 0.1 kept.

- Predicted functional-association data is obtained from the STRING database [69] (data downloaded in January 2012). STRING predicts each association between two proteins u and v (or their respective genes) using the following evidence types: gene co-occurrence across genomes; gene-fusion events; gene proximity in the genome; homology; coexpression; physical interactions; co-occurrence in literature; and orthologs of the latter five evidence types transferred from other organisms (STRING also includes evidence obtained from databases, which we discard as this may include co-complex relationships which we are trying to predict). Each evidence type is associated with quantitative information (e.g. the number of gene-fusion events), which STRING maps to a confidence score of

		YEAST			HUMAN		
Data source	Description	# pairs	# distinct proteins	% complex edges	# pairs	# distinct proteins	% complex edges
PPIREL	PPIs, scored by reliability	48,286	5,030	13.6%	44,636	9,535	10.8%
PPITOPO	Topological score of PPI edges	274,277	5,469	3.4%	298,399	9,771	6.1%
STRING	Predicted functional association	175,712	5,964	5.7%	311,435	14,784	3.1%
PubMed	Literature co-occurrence	161,213	5,109	4.9%	91,751	10,659	4.3%
All		518,417	6,099	2.1%	636,966	17,945	3.4%

Table 3.1: Statistics of data sources.

functional association based on co-occurrence in KEGG pathways. The confidence scores of the different evidence types are then combined probabilistically to give a final functional-association score for (u, v). Only pairs with score greater than 0.5 are kept.

- Co-occurrence of proteins or genes in PubMed literature (data downloaded in January 2012). Each pair (u, v) is scored by the Jaccard similarity of the sets of papers that u and v appear in:

$$s = \frac{|A_u \cap A_v|}{|A_u \cup A_v|}$$

where A_x is the set of PubMed papers that contain protein x. For yeast, that would be the papers that contain the gene name or open reading frame (ORF) ID of x as well as the word "cerevisiae"; for human that would be the papers that contain the gene name or Uniprot ID of x as well as the words "human" or "sapiens".

While there seems to be overlap between STRING's use of PPI and literature co-occurrence data with our use of them as separate data sources, note that STRING uses these data as only as component evidences for functional association and scores them accordingly. Thus we treat the STRING data as a representation of functional association between proteins, regardless of how this association was derived. Table 3.1 gives some summarizing statistics for these data sources.

In the composite network, vertices represent proteins and edges represent relationships between proteins. The composite network has an edge between proteins u and v if and only if there is a relationship between u and v according to any of the data sources.

3.2.2 Edge-weighting by posterior probability

Next, each edge (u, v) is weighted based on its posterior probability of being a co-complex edge (i.e. both u and v are in the same complex), given the scores of the data source relationships between u and v.

We use a naive-Bayes maximum-likelihood model to derive the posterior probability. Each edge (u, v) between proteins u and v of the composite network is cast as a data instance. The set of features is the set of data sources, and for each instance (u, v), feature F has value f if proteins u and v are related by data source F with score f. If u and v are not related by data source F, then feature F is given a score of 0. Using a reference set of protein complexes, each instance (u, v) in the training set is given a class label *co-complex* if both u and v are in the same complex; otherwise its class label is *non-co-complex*.

Learning proceeds in two steps:

1. Minimum description length (MDL) supervised discretization [70] is performed to discretize the features. MDL discretization recursively partitions the range of each feature to minimize the information entropy of the classes. If a feature cannot be discretized, that means it is not possible to find a partition that reduces the information entropy, so the feature is removed. Thus this step also serves as simple feature selection.

2. The maximum-likelihood parameters are learned for the two classes *co-complex* and *non-co-complex*:

$$P(F = f|\textit{co-comp}) = \frac{n_{c,F=f}}{n_c}$$

$$P(F = f|\textit{non-co-comp}) = \frac{n_{\neg c,F=f}}{n_{\neg c}}$$

   for each discretized value f of each feature F. n_c is the number of edges with class label *co-complex*, $n_{c,F=f}$ is the number of edges with class label *co-complex* and whose feature F has value f, $n_{\neg c}$ is the number of edges with class label *non-co-complex*, and $n_{\neg c,F=f}$ is the number of edges with class label *non-co-complex* and whose feature F has value f.

After learning the maximum-likelihood model, the weight for each edge e with feature values $F_1 = f_1, F_2 = f_2, \ldots$ is calculated as its posterior probability of being a co-complex edge:

$$
\begin{aligned}
& weight(e) \\
= \; & P(\textit{co-comp}|F_1 = f_1, F_2 = f_2, \ldots) \\
= \; & \frac{P(F_1 = f_1, F_2 = f_2, \ldots|\textit{co-comp})P(\textit{co-comp})}{Z} \\
= \; & \frac{\prod_i P(F_i = f_i|\textit{co-comp})P(\textit{co-comp})}{Z} \\
= \; & \frac{\prod_i P(F_i = f_i|\textit{co-comp})P(\textit{co-comp})}{\prod_i P(F_i = f_i|\textit{co-comp})P(\textit{co-comp}) + \prod_i P(F_i = f_i|\textit{non-co-comp})P(\textit{non-co-comp})}
\end{aligned}
$$

where Z is a normalizing factor to ensure the probabilities sum to 1. Although the second last equality makes the assumption that the features are independent, naive-Bayes classifiers have been found to perform well even when this assumption is false [71]. Specifically, while the probability estimates are frequently inaccurate, their rank orders usually remain correct, so that edges with likelier co-complex feature values are assigned higher scores than edges with likelier non-co-complex feature values.

3.2.3 Complex discovery

After the composite network is weighted, the top k edges are used by a clustering algorithm to predict protein complexes. We use the following clustering algorithms in our study: MCL, RNSC, IPCA, CMC, HACO, and ClusterONE (these are described in Chapter 2.4).

CMC, MCL, HACO, and ClusterONE are able to utilize edge weights in their input networks, whereas RNSC and IPCA do not; in this case, the selection of the top k edges provides less noisy networks as inputs to the algorithms.

CMC, MCL, IPCA, and HACO utilize parameters whose optimal values are at least partly dependent on the input networks' distribution of edge weights. For example, given an input network with high edge weights, using CMC with too low a $merge\_thres$ produces too many clusters consisting of merged cliques. Thus, we run these algorithms with a range of values for their respective parameters, so as to obtain a more comprehensive picture of their performances across different weighting approaches. We run ClusterONE, RNSC, and IPCA with mostly default or recommended parameters. The parameter settings used in our experiments for the six clustering algorithms are given in Table 3.2.

We also use a simple voting-based aggregative strategy **COMBINED**, which takes the union of the clusters produced by the six algorithms above. If two or more clusters are found to be similar to each other, then only the cluster with the highest

Clustering algorithm	Parameter settings
CMC	min_deg_ratio=1, min_size=4, overlap_thres=0.5, merge_thres=0.25
	min_deg_ratio=1, min_size=4, overlap_thres=0.5, merge_thres=0.5
	min_deg_ratio=1, min_size=4, overlap_thres=0.5, merge_thres=0.75
MCL	-I 2
	-I 3
	-I 4
HACO	-l ave -c c 0.75 -g 0.1
	-l ave -c c 0.9 -g 0.1
IPCA	-S4 -P2 -T0.4
	-S4 -P2 -T0.6
ClusterONE	-s 4 -d 0
RNSC	-e10 -D50 -d10 -t20 -T3

Table 3.2: Summary of the six clustering algorithms used, and their parameters tested for yeast and human complex discovery.

weighted density is kept, and its score is defined as its weighted density multiplied by the number of algorithms that produced the group of similar clusters; otherwise its score is its weighted density as usual. We define two clusters C and D to be similar if $Jaccard(C, D) >= 0.75$, where $Jaccard(C, D)$ is the Jaccard similarity between the proteins contained in C and D.

3.3 Results

3.3.1 Experimental setup

In our main experiment, we compare the performance of five weighting approaches:

1. SWC: supervised weighting of composite network (our proposed method)
2. BOOST: supervised weighting of composite network using LogitBoost [46]
3. PPIREL: PPI network weighted by reliability (these weights are equivalent to the PPI reliability feature in our composite network)
4. TOPO: unsupervised topological weighting of PPI network with Iterative AdjustCD [35], including level-2 PPIs (these weights are equivalent to the PPI topological feature in our composite network)
5. STR: network of predicted and scored functional associations from STRING [69] (these weights are equivalent to the STRING feature in our composite network)

We perform random sub-sampling cross-validation, repeated over ten rounds, using manually-curated complexes as reference complexes for training and testing. For yeast, we use the CYC2008 [56] set which consists of 408 complexes. Only complexes of size greater than three proteins are used for testing; there are 149 such complexes in

CYC2008. For human, we use the CORUM [57] set which consists of 1829 complexes, of which 714 are of size greater than three. In each cross-validation round, $t\%$ of the complexes of size greater than three are selected for testing, while all the remaining complexes are used for training. Each edge (u, v) in the network is given a class label *co-complex* if u and v are in the same training complex, otherwise its class label is *non-co-complex*. For SWC and BOOST, learning is performed using these labels, and the edges of the entire network are then weighted using the learned models. TOPO, STRING, and PPIREL require no learning, so the labels are not used; instead, for TOPO the edges of the network are weighted with topological scores, for STRING the edges are weighted with functional-association scores, and for PPIREL the edges are weighted with PPI reliability scores. The top-weighted k edges from the network are then used by the clustering algorithms to predict complexes. In our experiments we use $k = 10000, 20000$. We do not use all edges for these methods, because weighting enriches the network in dense clusters, which causes some of the clustering algorithms to require too much time to run when all edges are used; moreover, our experiments indicate that the performance of these methods drop when more than 20000 edges are used. The predicted clusters are evaluated on how well they match the test complexes.

We designed our experiment to simulate a real-use scenario of complex prediction in an organism where a few complexes might already be known, and novel complexes are to be predicted: in each round of cross-validation, the training complexes are those that are known and leveraged for learning to discover new complexes, while the test complexes are used to evaluate the performance of each approach at this task. Thus we use a large percentage of test complexes $t = 90\%$. In yeast, this gives 134 test complexes (among the 149 complexes of size greater than three), and 274 training complexes (only 15 of size greater than three); in human, this gives 643 test complexes (among the 714 of size greater than three), and 1186 training complexes (71 of size greater than three).

3.3.2 Evaluation methods

We use precision-recall graphs to evaluate the predicted clusters. First, a cluster P is said to match a complex C at a given match threshold $match\_thres$ if $Jaccard(P, C) \geq match\_thres$. Each cluster P is ranked by its score. To obtain a precision-recall graph, we calculate and plot the precision and recall of the predicted clusters at various cluster-score thresholds. The precision and recall differ slightly from that of Chapter 2.5.2, to account for the complexes used for training and testing. Given a set of predicted

clusters $P = \{P_1, P_2, \ldots\}$, a set of test reference complexes $C = \{C_1, C_2, \ldots\}$, and a set of training reference complexes $T = \{T_1, T_2, \ldots\}$, the recall and precision at score threshold d are defined as follows:

$$Recall_d = \frac{|\{C_i | C_i \in C \wedge \exists P_j \in P, dens(P_j) \geq d, P_j \; matches \; C_i\}|}{|C|}$$

$$Precision_d = \frac{|\{P_j | P_j \in P, dens(P_j) \geq d \wedge \exists C_i \in C, C_i \; matches \; P_j\}|}{|\{P_k | P_k \in P, dens(P_k) \geq d \wedge (\nexists T_i \in T, T_i \; matches \; P_k \vee \exists C_i \in C, C_i \; matches \; P_k)\}|}$$

The precision of clusters is calculated only among those clusters that do not match a training complex, to eliminate the bias of the supervised approaches (SWC and BOOST) for predicting training complexes well. As a summarizing statistic of a precision-recall graph, we also calculate the area under the curve (AUC) of a precision-recall graph. Besides evaluating the performance of complex prediction, we also evaluate the performance of edge classification, in which the edge weights are used to classify edges as co-complex or non-co-complex edges.

To evaluate the quality of novel predicted complexes, we define three measures of semantic coherence for each complex: its biological process (BP), cellular compartment (CC), and molecular function (MF) semantic coherence. These are calculated from the proteins' annotations to Gene Ontology (GO) terms, which span the three classes BP, CC, and MF [6]. We use the most informative common ancestor method of calculating the semantic similarity between two proteins, as outlined in [72]. Briefly, the semantic similarity of two GO terms is first defined as the information content of their most informative common ancestor. Next, the BP semantic similarity of two proteins is defined as the highest semantic similarity between their two sets of annotated BP terms. Then, we define the BP semantic coherence of a predicted complex as the average BP semantic similarity between every pair of proteins in that complex (likewise for CC and MF).

3.3.3 Classification of co-complex edges

Yeast

We first evaluate each approach in classification of co-complex edges. Here, each weighting approach is used to weight the network edges, and the edges are classified as co-complex by taking a threshold on their weights. We obtain precision-recall graphs

(solid markers, left axis) by taking a series of decreasing thresholds; at each recall level, we also indicate the proportion of test complexes covered by at least one predicted edge (hollow markers, right axis).

Figure 3.2a shows the performance of the five weighting approaches for classification of co-complex edges in yeast, and demonstrates that SWC achieves decent precision levels, while covering a large number of complexes. BOOST integrates the same data sources as SWC, but uses LogitBoost instead to learn to classify co-complex edges. Its points in the graph are clustered in two regions: one set of edges are given high scores, achieving about 40% recall and 35% precision (lower than SWC's precision of 50% at this recall level), while the remaining edges are given low scores. Thus BOOST performs classification in a categorical manner, whereas SWC produces co-complex scores that reflect a wide range of confidence.

PPIREL gives lower precision than SWC, and moreover reaches a maximum recall of around 70% only. This shows that even a union of PPI's from multiple databases misses out on a large number of co-complex interactions, and demonstrates the value of integrating non-PPI data sources to cover more co-complex edges.

TOPO has higher precision than SWC among the highly-weighted edges, indicating that edges with high topological scores are more likely to be co-complex compared to edges with high SWC scores. However, these edges are clustered in a few test complexes, giving lower complex coverage. When more edges are included to predict co-complex edges in a wider range of complexes, TOPO's precision drops well below that of SWC. Thus, topological weighting can only accurately predict edges in a few densely-connected complexes whose edges have high topological scores; for less-dense complexes, SWC performs better by using multiple data sources and supervised learning.

On the other hand, SWC is more accurate than STR in predicting co-complex edges. This is because many proteins that are highly functionally associated are not co-complex. In contrast, SWC's supervised-learning approach produces weights that are targeted at predicting co-complex edges; so highly-weighted edges are more likely to be co-complex.

Human

Figure 3.2b shows the corresponding precision-recall graphs for classification of co-complex edges in human. Compared to yeast, the coverage of co-complex edges is much lower in human.

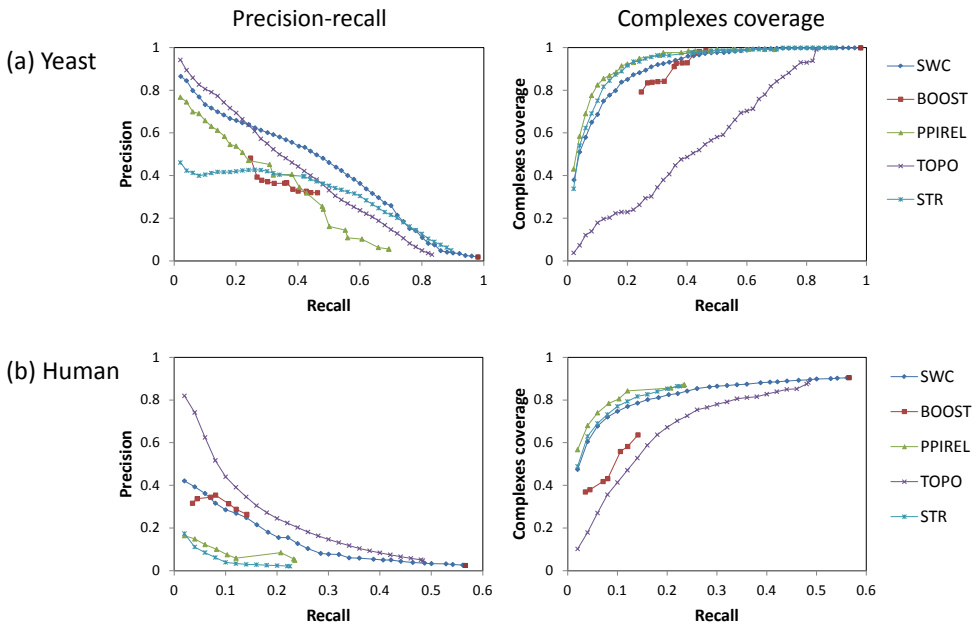

Figure 3.2: Precision-recall and complex-coverage graphs for classification of co-complex edges using the five weighting schemes, for (a) yeast, (b) human. Only TOPO has higher precision than SWC, but its edges are clustered in a few complexes (complexes coverage graph on right).

Just like in yeast, BOOST performs classification in a categorical manner: a set of edges are predicted as co-complex with high scores, achieving 12% recall and similar precision levels as SWC, while the remaining edges are predicted as non-co-complex with low scores.

In human, PPIREL gives very low precision in co-complex edge classification, and moreoever only achieves a maximum recall of under 25%, showing (as in yeast) that PPIs from multiple databases do not cover enough co-complex edges, and integrating diverse data sources can overcome this problem.

Compared to TOPO, SWC has lower precision along TOPO's entire recall range. However, once again TOPO's predicted edges are clustered in fewer complexes, giving lower complex coverage: for example, to cover 80% of complexes requires TOPO to recall 34% of edges at a precision of 12%; SWC has to recall only 16% of edges at a higher precision of 22% to cover the same amount of complexes. Thus, for human as well as yeast, SWC is able to predict co-complex edges for a wider range of complexes compared to TOPO, whose range is limited to fewer complexes that are densely connected.

For human, STR's functional-association scores are the least accurate for predicting co-complex edges, giving the lowest precision among all the weighting approaches.

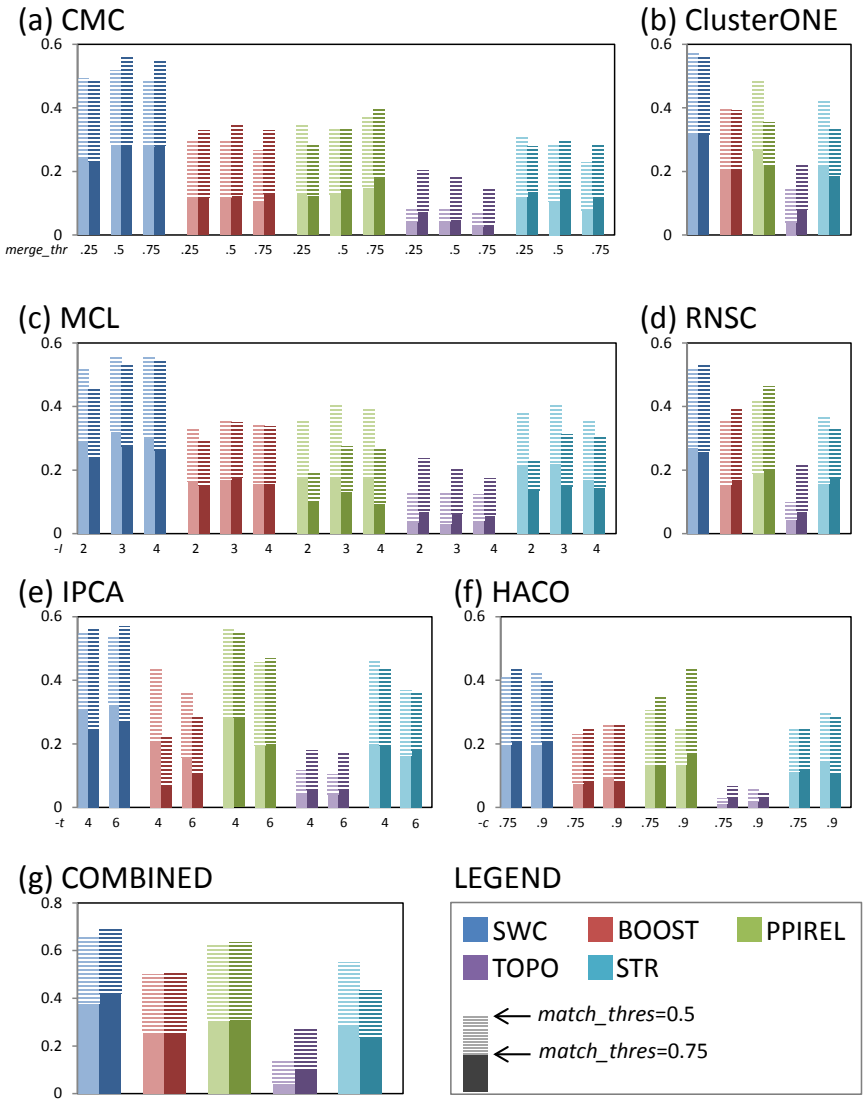

Figure 3.3: Precision-recall AUC for yeast complex prediction, using the five weighting approaches for each of the six clustering algorithms and the COMBINED clustering strategy, for $k = 10000$ (lighter shade), and $k = 20000$ (darker shade). For CMC, MCL, IPCA, and HACO, different sets of clustering parameters are tried. The AUC for $match\_thres = 0.5$ and $match\_thres = 0.75$ are shown in each bar. SWC achieves highest precision-recall AUC for all clustering algorithms except IPCA and HACO, where it performs about evenly with PPIREL at $match\_thres = 0.5$ but better at $match\_thres = 0.75$. The COMBINED strategy achieves higher AUC compared to using any single clustering algorithm alone.

3.3.4 Prediction of complexes

Yeast

We compare the performance of the five weighting approaches in complex prediction, when each of the six clustering algorithms is used separately, and when all the clustering algorithms are used together with the COMBINED strategy. Figure 3.3 shows the precision-recall AUC for prediction of yeast complexes, and demonstrates that SWC outperforms the other weighting approaches in most cases: using the best clustering parameter settings for each approach, SWC achieves the highest AUC with all clus-

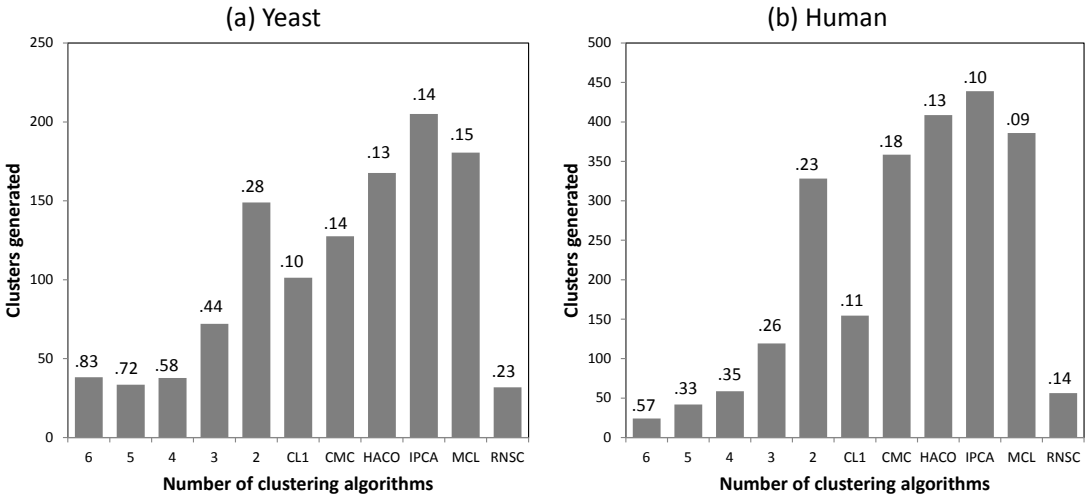

Figure 3.4: Distribution of clusters from the COMBINED strategy among different numbers of clustering algorithms that generated them using the SWC network, and their precision (proportion of clusters that match test complexes), in (a) yeast, (b) human. Different clustering algorithms produce different sets of clusters: in either yeast or human, about 70% of clusters are generated by a single unique algorithm, while less than 10% of clusters are generated by four or more algorithms. Thus aggregating clusters from different algorithms increases the recall of complex prediction. Furthermore, precision increases as clusters are generated by a greater number of algorithms: the highest precision of clusters generated by a single algorithm is 23% and 18% in yeast and human respectively, increasing to 83% and 57% for clusters generated by all algorithms.

tering algorithms except for IPCA and HACO (where SWC performs about evenly with PPIREL at $match\_thres = 0.5$ but better at $match\_thres = 0.75$). PPIREL outperforms the remaining weighting approaches with all clustering algorithms, while BOOST and STR perform at similar levels, and finally TOPO achieves the lowest AUCs. The COMBINED strategy achieves higher AUC compared to using each individual clustering algorithm, for all weighting approaches. Using the COMBINED strategy, SWC achieves the highest AUC, followed by PPIREL, STR, BOOST, and finally TOPO.

We analyze the clusters from the COMBINED strategy to determine how it achieves greater complex-prediction performance by aggregating clusters from the different clustering algorithms with simple voting. Figure 3.4a shows how clusters from the COMBINED strategy are distributed among any single or multiple number of clustering algorithms that generated them, as well as their precision (the percentage of clusters that match test complexes), in yeast. For brevity we present only the figures for the SWC weighting approach. It reveals that the different algorithms produce different sets of clusters: around 71% of clusters are uniquely generated by a single algorithm, 13% of clusters are generated by two algorithms, and the remaining 6% of clusters are generated by three or more algorithms. Thus, taking their union increases the

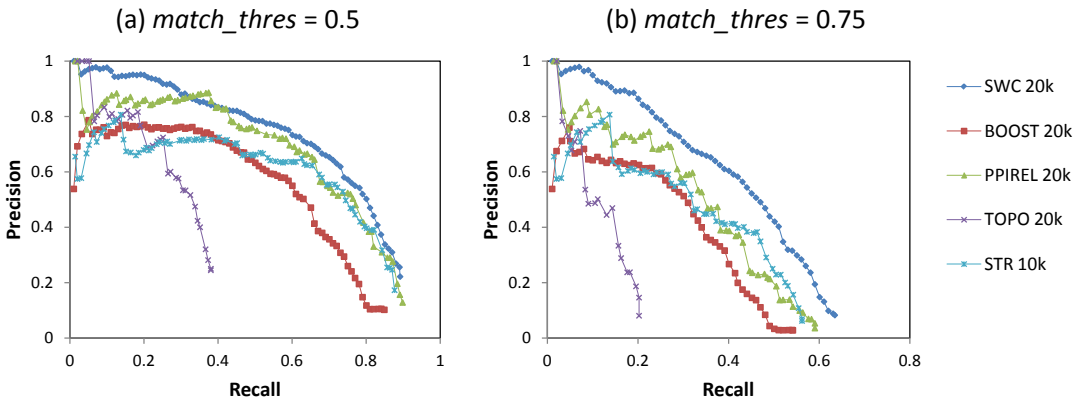

Figure 3.5: Precision-recall graphs for yeast complex prediction using the five weighting approaches with the COMBINED clustering strategy, using $k = 20000$ for SWC, BOOST, PPIREL, and TOPO, and $k = 10000$ for STR, at (a) $match\_thres = 0.5$, (b) $match\_thres = 0.75$. At $match\_thres = 0.5$, SWC achieves similar recall as BOOST, PPPIREL, and STR, but with the higher precision at almost all recall levels. At the stricter $match\_thres = 0.75$, SWC achieves the highest recall with the highest precision at almost all recall levels. Thus it outperforms all other weighting approaches, especially at predicting complexes with fine granularity.

recall substantially. Furthermore, the precision of clusters increases with the number of algorithms that generated them: among clusters generated by a single algorithm, the highest precision is 23%; clusters generated by two algorithms have a precision of 28%; the precision increases to 83% among the clusters generated by all six algorithms. Thus, voting helps to increase precision by giving greater scores to those clusters predicted by multiple clustering algorithms.

Figure 3.5 shows the precision-recall graphs for prediction of yeast complexes for the five weighting approaches, using the COMBINED clustering strategy. For brevity, for each approach we show and discuss only the graph for the value of k that achieves the highest AUC ($k = 20000$ for SWC, BOOST, PPIREL, and TOPO, $k = 10000$ for STR). At $match\_thres = 0.5$, SWC achieves similar recall as BOOST, PPPIREL, and STR, but with the higher precision at almost all recall levels. At the stricter $match\_thres = 0.75$, SWC achieves the highest recall with the highest precision at almost all recall levels. Thus it outperforms all other weighting approaches, especially at predicting complexes with fine granularity.

PPIREL achieves just slightly lower recall and precision than SWC at $match\_thres = 0.5$, but its performance drops substantially at a higher $match\_thres = 0.75$. While experiment-derived PPIs are adequate to predict the test complexes at a rough granularity, the missing and spurious interactions cause many clusters to miss real proteins or include extra proteins, so that they cannot match the test complexes at a finer granularity. Similarly, at the lower $match\_thres$, STR achieves almost the same

recall as SWC at lower precision levels, but its recall and precision are much worse at a higher *match_thres*. Since STR classifies co-complex edges across a large range of clusters, it is able to recall many test complexes; but its lower accuracy in edge classification means that many of its clusters also include extra or missing proteins, causing them not to be matched at a stricter matching threshold. BOOST achieves similar recall as STR but with substantially lower precision levels at both match thresholds. Since it classifies edges categorically, many edges have similar scores that do not vary with classification accuracy; thus the ranking of clusters (based on their weighted-densities) does not correlate as well with their correctness, giving lower precision levels. TOPO achieves the lowest recall of all approaches. While its precision for its highest-scoring clusters is comparable to SWC's at $match\_thres = 0.5$ (at the extreme left end of the graph), it drops rapidly for the remaining clusters. This is because TOPO classifies co-complex edges accurately for a limited number of complexes which are dense and thus easy to predict, while the remaining complexes' edges are not as accurately classified, creating many false positive clusters and low recall.

Human

Figure 3.6 shows the precision-recall AUC of the five weighting approaches for the prediction of human complexes. The AUC here is considerably lower than for prediction of yeast complexes, especially at $match\_thres = 0.75$. Nevertheless, it is clear that SWC outperforms all the other weighting approaches. Using each clustering algorithm's best parameter settings for each approach, SWC achieves substantially higher AUC than all the other approaches, for all clustering algorithms. After SWC, TOPO and PPIREL perform the next best, followed by BOOST. STR performs the worst in all clustering algorithms.

The COMBINED strategy shows less clear benefits for human complexes, in terms of AUC: it actually gives worse performance for STR and TOPO compared to using CMC, IPCA, or HACO alone. Figure 3.4b shows the distribution of clusters from the COMBINED strategy for SWC in human. As in yeast, around 70% of clusters are uniquely generated by any single clustering algorithm. The precision of the clusters increases as they are generated by more clustering algorithms: from a maximum of 18% when generated by a single algorithm, to 57% when generated by all six algorithms.

Figure 3.7 shows the precision-recall graphs for prediction of human complexes for the five weighting approaches, using the COMBINED clustering strategy. For brevity, for each approach we show and discuss only the graph for the value of k that achieves

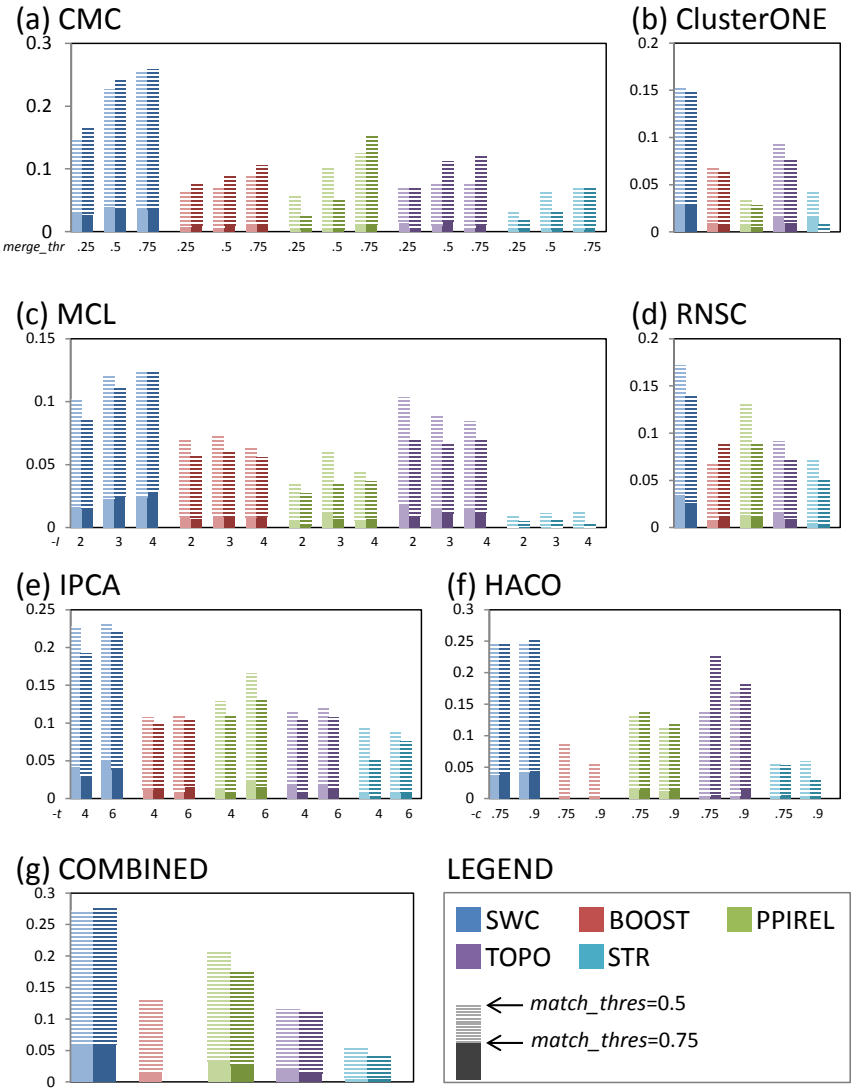

Figure 3.6: Precision-recall AUC for human complex prediction, using the five weighting approaches for each of the six clustering algorithms and the COMBINED clustering strategy, for $k = 10000$ (lighter shade), and $k = 20000$ (darker shade). For CMC, MCL, IPCA, and HACO, different sets of clustering parameters are tried. The AUC for $match\_thres = 0.5$ and $match\_thres = 0.75$ are shown in each bar. SWC consistently achieves highest precision-recall AUC for all clustering algorithms and the COMBINED strategy. The COMBINED strategy achieves higher AUC compared to using any single clustering algorithm alone.

the highest AUC ($k = 20000$ for SWC, TOPO, and BOOST, $k = 10000$ for STR, $k = all$ for NOWEI).

SWC attains the highest recall at both $match\_thres$, with higher precision at all recall levels (except that PPIREL's top-scoring clusters have higher precision at the lowest recall range). The performance advantage is even more pronounced at $match\_thres = 0.75$, where SWC recalls 50% more test complexes compared to the other approaches, and maintains almost twice the precision throughout most of its recall range.

PPIREL achieves higher precision than SWC at the lowest recall range, but its

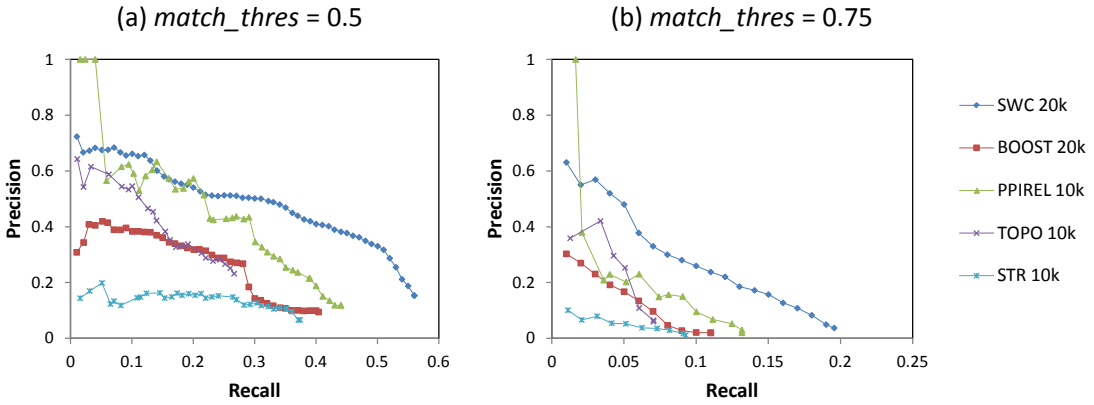

Figure 3.7: Precision-recall graphs for human complex prediction using the five weighting approaches for the COMBINED clustering strategy. SWC achieves the highest recall with the highest precision at almost all recall levels, especially with the stricter $match\_thres = 0.75$, where SWC recalls at least 50% more test complexes compared to the other approaches and maintains almost twice the precision throughout most of its recall range. Thus it outperforms all other weighting approaches, especially at predicting complexes with fine granularity.

precision drops among its remaining clusters, and moreover does not achieve as high recall as SWC. This shows that clusters predicted from highly-reliable PPIs do match real complexes well, but this is limited to only a few top-scoring clusters that match a limited subset of complexes.

TOPO achieves lower recall, but at $match\_thres = 0.5$ its precision for its high-scoring clusters is comparable to SWC's for its highest-scoring clusters. Once again, TOPO's high accuracy in classifying edges for a limited number of dense complexes means it is only able to predict a few complexes well at rough granularity.

Unlike in yeast, here STR performs extremely poorly with the lowest recall and precision levels of all weighting approaches. This is not surprising given that STR performs poorly in edge classification as well.

3.3.5 Performance among stratified complexes

To further investigate how SWC improves the performance of large-complex prediction, we study its effects on predicting large complexes with different degrees of extraneous and missing interactions. As described in Chapter 2.5.1, we stratify the complexes by size, EXT (the number of external proteins that are highly connected to it), and DENS (density). Figures 2.3 and 2.4 show the distribution of the large complexes (containing four or more proteins) in terms of DENS, EXT, and our six analysis groups (stratified by DENS and EXT), for yeast and human.

In both yeast and human, around 40% of complexes have high EXT. We expect the prediction of these complexes to be extremely challenging, as it would be difficult to accurately delimit their borders from their highly-connected surroundings (the highly-

connected external proteins are likely to be recruited into the predicted complexes). Most complexes in yeast have high density: only 10% of complexes have low DENS. On the other hand, in human about 35% of complexes are sparsely connected with low DENS. We expect these sparsely-connected complexes to also be difficult to predict, as they do not form dense clusters that are picked out by most clustering algorithms.

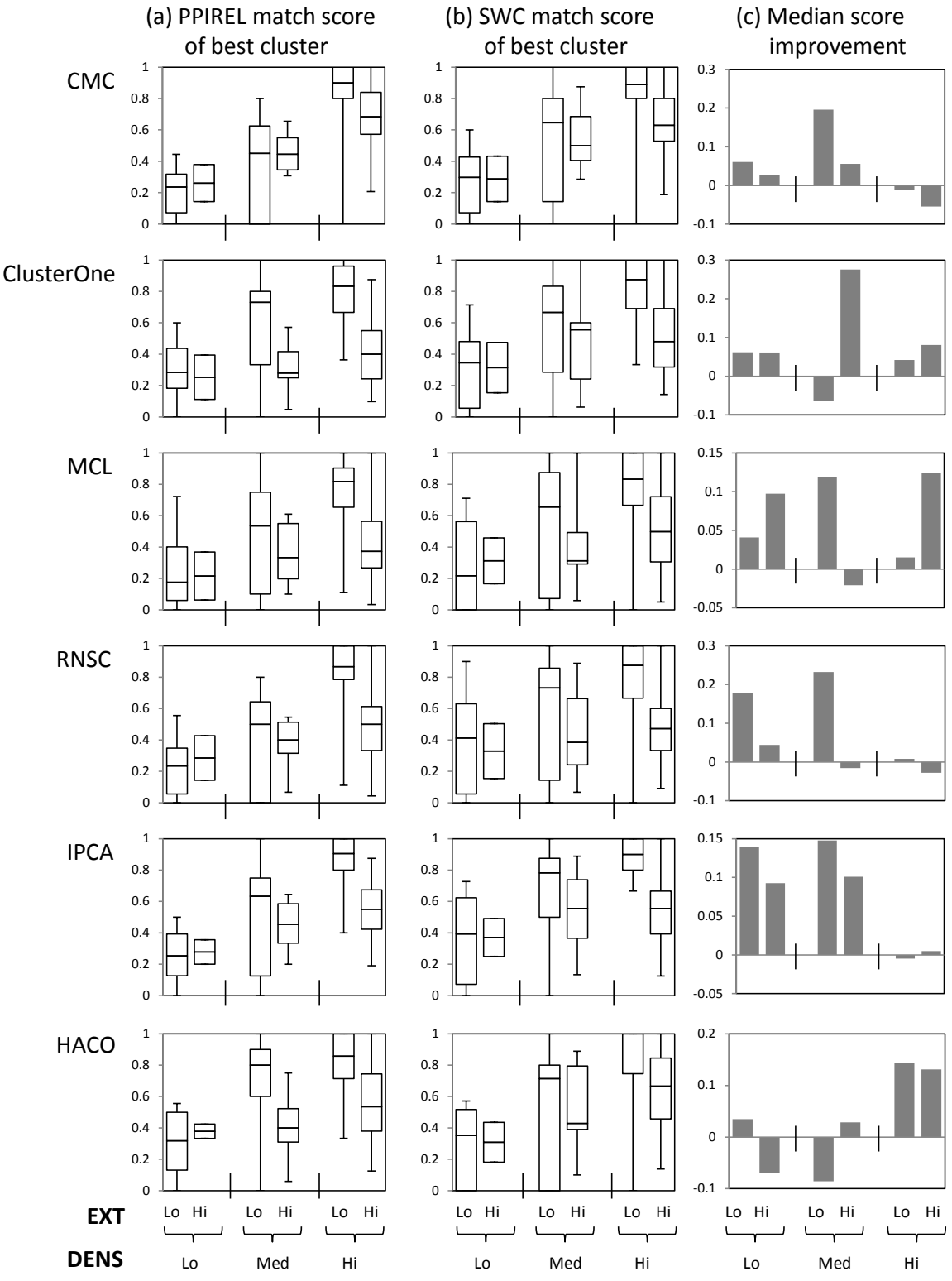

Figure 3.8: Match scores of the best clusters to yeast complexes in the six analysis strata, using (a) PPIREL, and (b) SWC, generated by various clustering algorithms. (c) shows the improvements score medians. SWC gives bigger improvements among low- and medium-density complexes for most clustering algorithms.

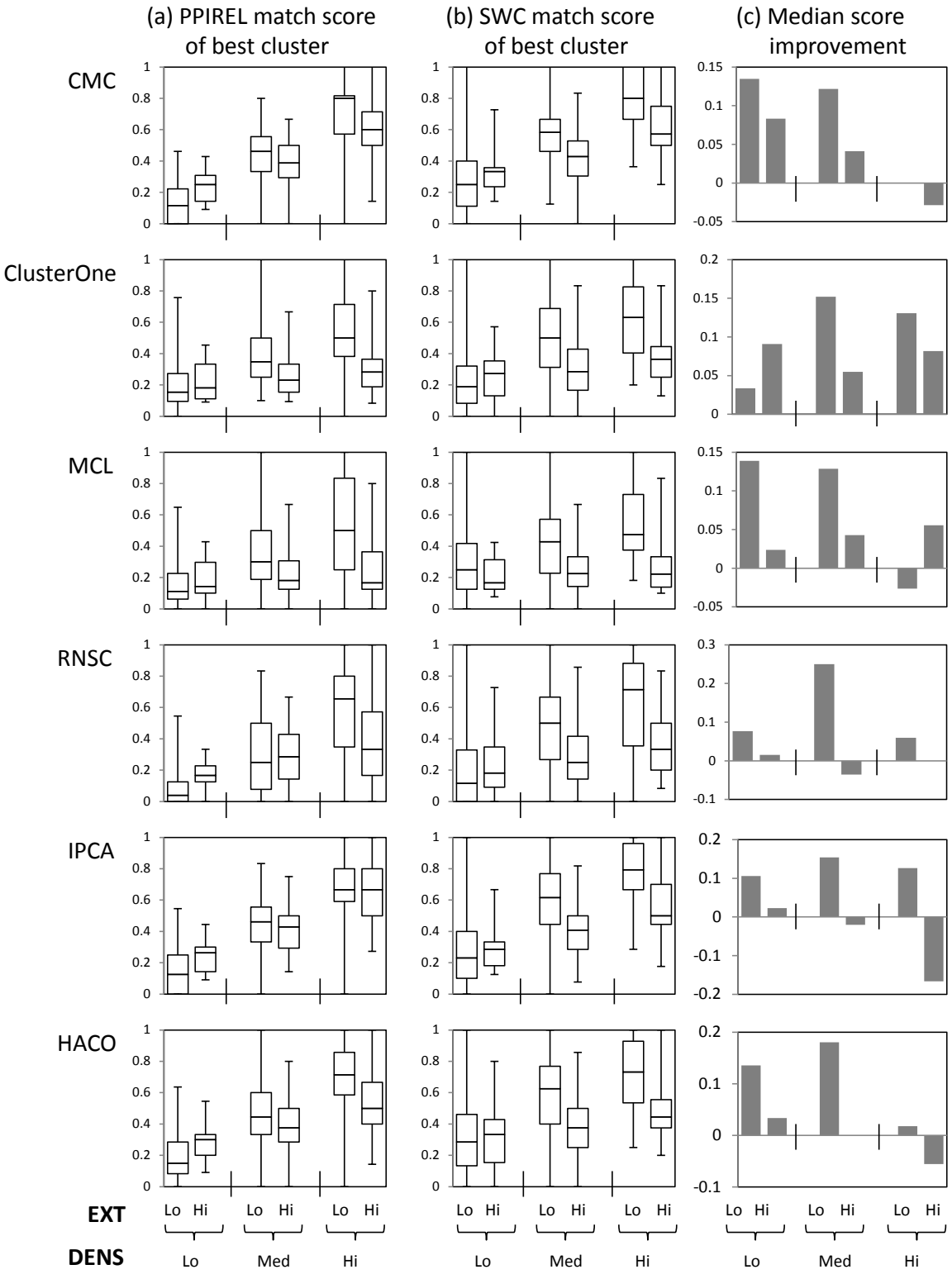

Figure 3.9: Match scores of the best clusters to human complexes in the six analysis strata, using (a) PPIREL, and (b) SWC, generated by various clustering algorithms. (c) shows the improvements score medians. SWC gives bigger improvements among low- and medium-density complexes for most clustering algorithms.

To investigate the benefits of SWC in predicting complexes from the different stratified groups of reference complexes, we compare how well the complexes from the different strata are matched by clusters generated from SWC versus PPIREL. Figure 3.8 shows the match scores of the best-matching clusters found for the yeast complexes in the different strata, using (a) PPIREL for weighting, or (b) SWC for weighting, while (c) shows the improvements in the score medians. We see that SWC gives bigger improvements (in terms of generating more well-matched clusters) among low- and medium-density complexes, for almost all clustering algorithms except MCL (where dense complexes also improve in matches), and HACO (where the improvements lie mostly in the denser complexes, and sparse complexes suffer worse matches under SWC).

Figure 3.9 shows the corresponding charts of complex match improvements among human complexes from the different analysis strata. As in yeast, SWC gives bigger improvements among low- and medium-density complexes for most clustering algorithms. As will be illustrated below with an example yeast complex, this improvement among sparse complexes can be attributed to SWC integrating diverse data sources to fill in the missing interactions, while using supervised weighting to control the amount of noisy edges, which allows such sparse complexes to be discovered despite missing interactions.

3.3.6 Prediction of novel complexes

We evaluate the five weighting approaches (SWC, BOOST, PPIREL, TOPO, and STR) on the number and quality of high-confidence novel complexes predicted in yeast and human. For the supervised approaches (SWC and BOOST), we use the entire reference set of complexes (CYC2008 for yeast, CORUM for human) for training. Next, the edges of the entire network are weighted, and the top k edges are used to predict complexes with the COMBINED clustering strategy, which combines clusters predicted by the six clustering algorithms. For each approach we use the value of k that gave the best performance in cross-validation.

We filter the set of predicted complexes to obtain a set of unique, novel, high-confidence predictions. First, complexes that are too similar are removed: if any two predicted complexes match with $match\_thres = 0.5$, then the complex with the lower score is removed. Next, only novel predictions are kept: if any predicted complex matches any reference complex with $match\_thres = 0.5$, then that predicted complex is removed. Finally, only high-confidence predictions are kept: for each weighting

Biological process	# complexes
Protein metabolic process	39
RNA metabolic process	25
DNA metabolic process	9
Small molecule metabolic process	16
Regulation of metabolic process	20
Regulation of gene expression	13
Organelle organization	33
Transport	44
Response to stress	16
Response to chemical stimulus	5
Cell cycle process	8

Table 3.3: Biological processes of novel predicted yeast complexes.

approach, using the cross-validation results, the score of each predicted complex is benchmarked to a precision value, and predicted complexes whose estimated precision are less than a confidence threshold are removed. For yeast, this confidence threshold is 0.5; for human, since much fewer complexes are predicted with high precision, we use a 0.4 confidence threshold.

Yeast

Figure 3.10a shows the number of novel yeast complexes predicted using the five weighting approaches and the COMBINED clustering strategy. SWC predicts 186 yeast complexes covering 1021 proteins, substantially more than any of the other weighting approaches. Figure 3.10b shows the BP, CC, and MF coherence of the novel predicted yeast complexes. SWC's complexes have higher BP and CC coherence compared to BOOST's ($p = 0.07$), higher BP, CC, and MF coherence compared to PPIREL's ($p < 0.01$), higher BP and MF coherence compared to TOPO's ($p < 0.05$), but similar coherences compared to STR's. However, the reference complexes of CYC2008 still have much higher BP and CC coherence ($p < 0.0005$). Thus, weighting by SWC generates a larger number of novel yeast complexes compared to all the other weighting approaches, with greater semantic coherence compared to the other weighting approaches except for STR.

To explore the functions of the novel predicted complexes, we select a set of eleven high-level BP terms, and annotate a novel complex with a BP if that BP is annotated to the most number and a majority of proteins in the complex. Some complexes may be annotated to more than one high-level term. Table 3.3 shows that almost half of the novel predicted yeast complexes participate in metabolic processes, while the remainder are involved in regulation, cell organization, transport, cellular response, and cell cycle processes.

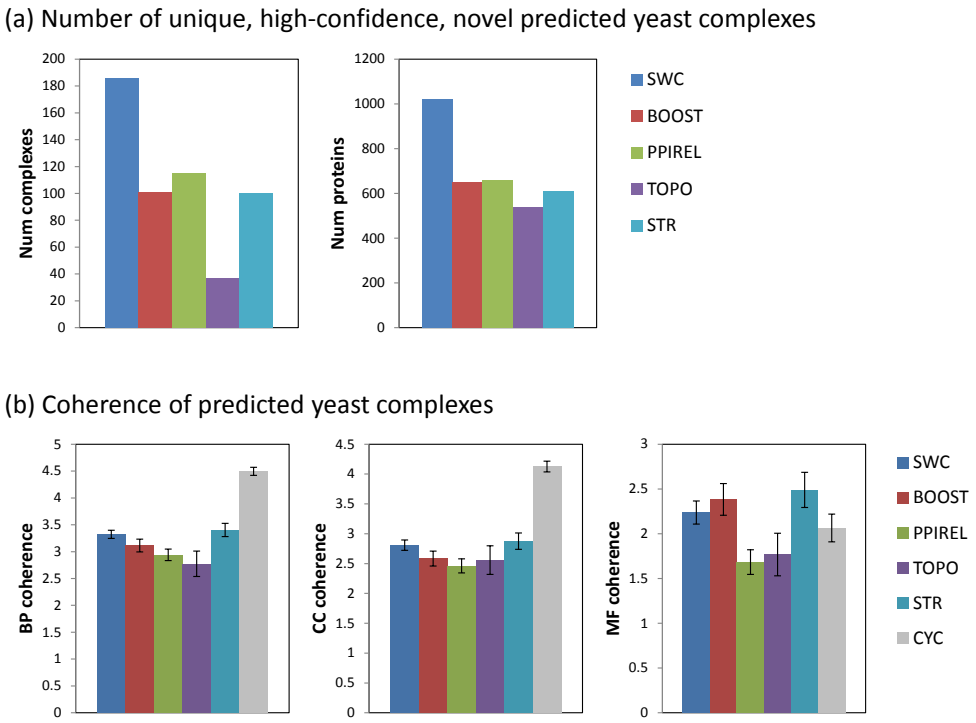

Figure 3.10: Unique, high-confidence, novel predicted yeast complexes. (a) Number of complexes predicted and number of proteins covered. (b) Semantic coherence of predicted complexes. (a) Number of yeast complexes predicted and number of proteins covered, using the five weighting approaches and the COMBINED clustering strategy. SWC generates more novel complexes that cover a greater number of proteins. (b) BP, CC, and MF semantic coherence of the predicted complexes and the reference complexes CYC2008. SWC's complexes have higher BP and CC coherence compared to BOOST's ($p = 0.07$), higher BP, CC, and MF coherence compared to PPIREL's ($p < 0.01$), higher BP and MF coherence compared to TOPO's ($p < 0.05$), but similar coherences compared to STR's. The CYC2008 reference complexes have much higher BP and CC coherence than the predicted complexes from all approaches.

Human

Figure 3.11 shows the corresponding statistics for the novel predicted human complexes. SWC predicts 277 human complexes covering 1285 proteins, substantially more than any of the other weighting approaches. SWC's complexes have higher BP, and MF coherence compared to those of TOPO ($p < 0.05$), but similar coherences compared to STR's and PPIREL's. The CORUM reference complexes have higher BP and CC coherence than the predicted complexes. Thus, weighting by SWC generates a larger number of novel human complexes, with equal or greater semantic coherence than other weighting approaches.

Table 3.4 shows how many of the predicted human complexes participate in eleven high-level BP terms. A large number of the predicted complexes participate in regulation, a quarter participate in metabolic processes, and the remainder in cell organization, transport, cellular response, and cell cycle processes.

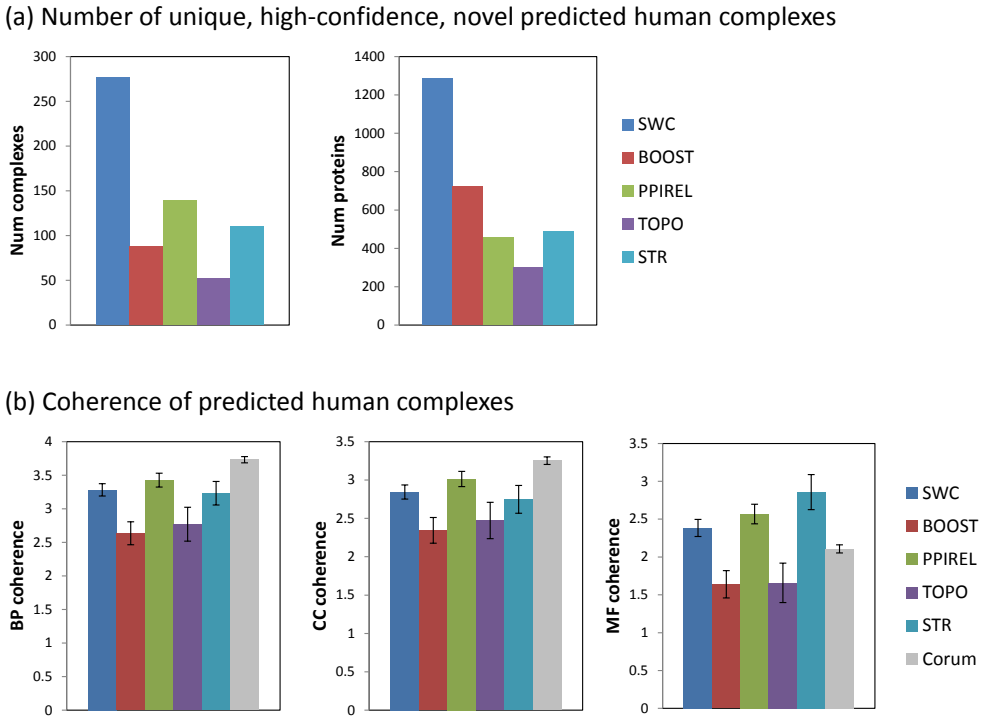

Figure 3.11: Unique, high-confidence, novel predicted human complexes. (a) Number of complexes predicted and number of proteins covered. (b) Semantic coherence of predicted complexes. (a) Number of human complexes predicted and number of proteins covered, using the five weighting approaches and the COMBINED clustering strategy. SWC generates more novel complexes that cover a greater number of proteins. (b) BP, CC, and MF semantic coherence of the predicted complexes and the reference complexes CORUM. SWC's complexes have higher BP and MF coherence compared to TOPO's ($p < 0.05$), but similar coherences compared to STR's and PPIREL's. The CORUM reference complexes have higher BP and CC coherence than the predicted complexes.

3.3.7 Analysis of learned parameters

Figures 3.12a and 3.12b show the learned likelihood parameters for yeast and human respectively, averaged over the cross-validation rounds. The likelihood parameters are expressed as likelihood ratios, or how many times likelier is an edge co-complex rather than not co-complex, given the feature value:

$$likelihood\ ratio = \frac{P(F = f|co\text{-}complex)}{P(F = f|non\text{-}co\text{-}complex)}$$

The likelihood ratio is a reflection of "co-complexness strength". In general, the likelihood ratios increase as the scores for the data sources (i.e. the x-axes) increase. For the PPI and L2-PPI data sources, protein pairs with higher scores have greater number of shared neighbors, and are likelier to be co-complex: when the score of PPIREL is close to 1, indicating that the PPI has very high estimated reliability form repeated detections from high-confidence experiments, the pair is more than 100 times likelier to be co-complex. When the score of PPITOPO is close to 1, indicating that

Biological process	# complexes
Protein metabolic process	46
RNA metabolic process	29
DNA metabolic process	8
Small molecule metabolic process	6
Regulation of metabolic process	96
Regulation of gene expression	49
Organelle organization	23
Transport	33
Response to stress	50
Response to chemical stimulus	42
Cell cycle process	16

Table 3.4: Biological processes of novel predicted human complexes.

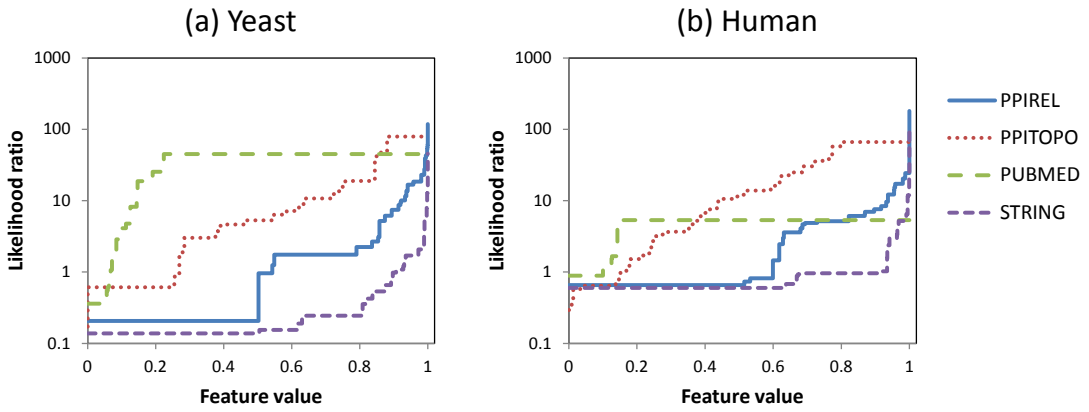

Figure 3.12: Learned likelihood parameters, expressed as likelihood ratios, for (a) yeast, (b) human. For PPIREL data, interacting proteins with higher PPI reliability are likelier to be co-complex. For PPITOPO data, proteins pairs are likelier to be co-complex when they have more shared neighbors. For STRING data, protein pairs with predicted functional associations are very likely to be co-complex when the prediction score is high; at low scores, protein pairs are not much likelier to be co-complex. For PubMed data, protein pairs that co-occur in literature, even infrequently, are already much likelier to be co-complex; however, pairs that co-occur more frequently in literature are not any more likelier to be co-complex compared to pairs that co-occur less frequently.

almost all of the protein pair's neighbors are shared, the pair is also 100 times likelier to be co-complex (even if the proteins do not actually interact according to PPI databases, as this feature includes non-interacting pairs with many shared neighbours).

For the STRING data source, only protein pairs with very high functional-association scores are likelier to be co-complex: those with the highest scores are almost 100 times likelier to be co-complex in yeast and 40 times likelier to be co-complex in human, whereas protein pairs with lower functional-association scores do not seem any likelier to be co-complex. Indeed, protein pairs with STRING scores of less than 0.9 are actually likelier to be non-co-complex.

For PubMed data, protein pairs that co-occur in literature, even infrequently, are already much likelier to be co-complex: about 40 times likelier in yeast and 5 times likelier in human. However, pairs that co-occur more frequently in literature are not

any more likelier to be co-complex compared to pairs that co-occur less frequently.

The likelihood ratios for the different data sources show that the co-complexness strength of each data source does not increase linearly with its score. Moreover, between the different data sources, the relationships between data score and co-complexness are different. Thus, combining data scores across different data sources without factoring their dissimilar co-complexness relationships is evidently unsound, while our supervised approach scales the heterogeneous scores to a uniform co-complexness score in terms of likelihoods, which can then be combined probabilistically using the naive-Bayes formulation.

The high likelihood ratios for the data sources also demonstrate that they are indeed indicative of edges belonging to complexes: during cross-validation for both yeast and human, none of the data sources were removed by feature selection in any round.

3.3.8 Visualization of example complexes

Yeast cytochrome bc1 complex

In this section we use two example complexes to illustrate the power and mechanism of SWC. Figure 3.13a shows the PPI subgraph of the yeast mitochondrial cytochrome bc1 complex discussed earlier, which is involved in the electron-transport chain in the mitochondrial inner membrane. The complex's PPI subgraph has 19 co-complex interactions, and 145 extraneous interactions with 94 external proteins, among which five are labeled: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination respectively, and bind to many proteins to perform their functions; PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain, and interact with proteins of the complex, although they are not part of it. In the composite network (Figure 3.13b), the edges from the other data sources induce a full clique among the complex proteins, although the number of extraneous edges and number of neighbors outside the complex increase to 1735 and 640 respectively. After weighting by SWC and selecting the top $k = 20000$ edges (Figure 3.13c), the complex's subgraph is still relatively dense; furthermore, only 26 extraneous edges and 18 neighboring proteins remain. Note that among the five labeled external proteins, the two involved in unrelated processes (NAB2 and UBI4) have been disconnected at this point, while the three also involved in the electron transport chain with the complex (PET9, SHY1, and COX1) are still connected to the network. With this network, both IPCA and RNSC detect the cluster shaded in gray,

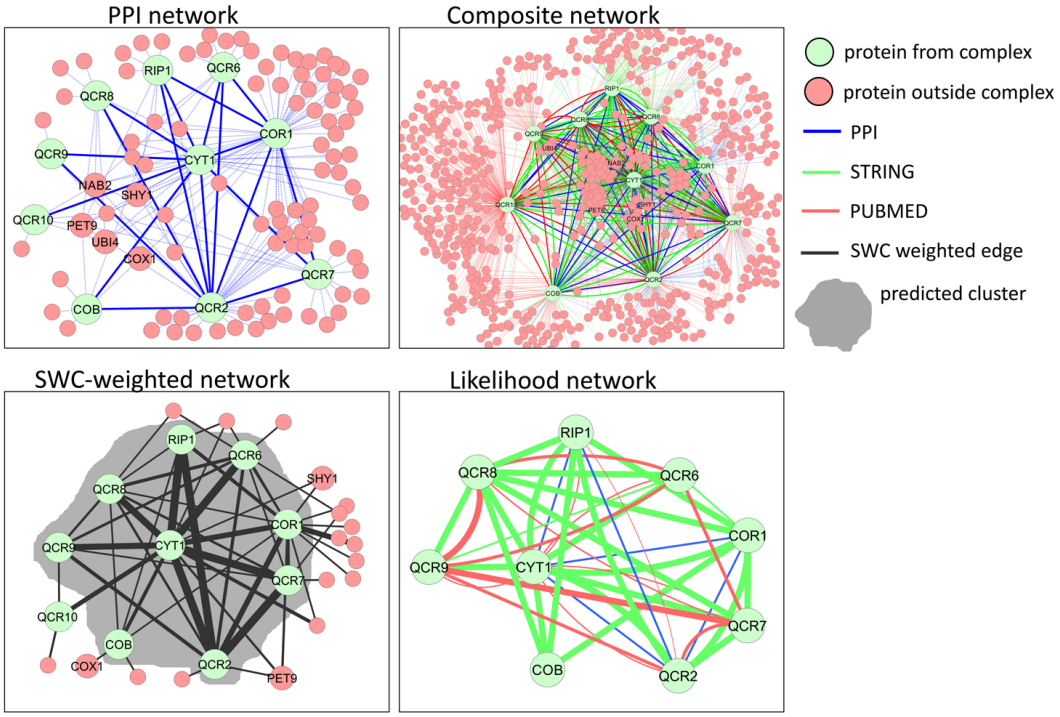

Figure 3.13: Yeast mitochondrial cytochrome bc1 complex: (a) PPI network, (b) composite network, (c) SWC-weighted network, and (d) likelihood network. (a) The PPI subgraph includes many extraneous edges and external neighboring proteins. (b) In the composite network, the extra edges from the other data sources induce a full clique among the complex proteins, although the number of extraneous edges increases dramatically as well. (c) In the SWC-weighted network, the complex's subgraph is still relatively dense, with fewer extraneous edges remaining, allowing the complex to be easily found by both IPCA and RNSC (although missing one protein). (d) In the likelihood network, diverse data sources connect many proteins within the cluster with high SWC scores. CYT1-RIP1-QCR2 are fully connected with each other via all three data sources with moderate to high co-complexness, making them a central triplet within the cluster. CYT1-COR1-QCR2 and CYT1-QCR7-QCR2 are connected via two or more data sources with moderate to high co-complexness, and are deeply embedded in the cluster as well. The other proteins appear less central in the cluster, especially COB, a fringe member which is only connected via functional associations to four proteins.

which matches the complex with Jaccard similarity of 0.9.

The likelihood network for the cluster (Figure 3.13d) visualizes the component evidences for the prediction: the contribution of each data source to an edge's SWC score is reflected in the edge thickness, which is scaled with its likelihood ratio, or co-complexness strength. The likelihood network reveals that diverse data sources connect many proteins within the cluster with high SWC scores. CYT1, RIP1, and QCR2 are fully connected with each other via all three data sources, making them the strongest co-complex triplet that is centrally embedded in the cluster, while CYT1-COR1-QCR2 and CYT1-QCR7-QCR2 are connected with two or more data sources, making them highly co-complex and deeply embedded as well. The other proteins appear less central in the cluster, especially COB, a fringe member which is only connected via functional

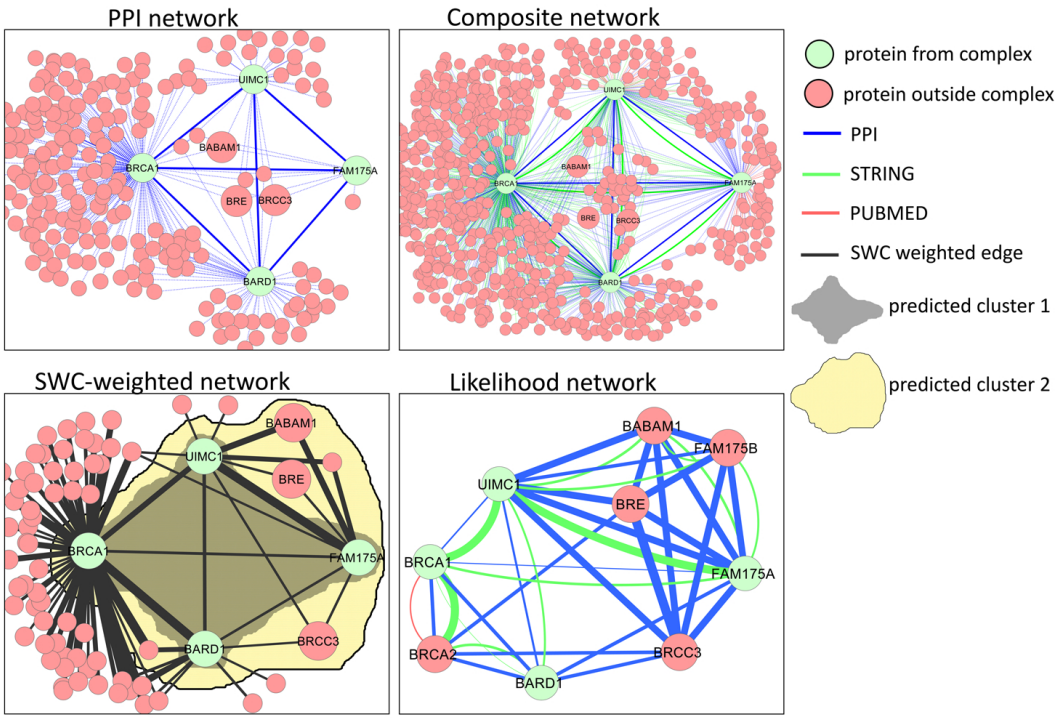

Figure 3.14: Human BRCA1-A complex: (a) PPI network, (b) composite network, (c) SWC-weighted network, and (d) likelihood network. (a) While the PPI network is fully connected, there are large numbers of extraneous edges and neighboring proteins (chiefly because BRCA1 itself is connected to around 180 proteins). (b) Composite network has a large number of extraneous edges. (c) In the SWC network, BRCA1 is still connected to a large number of proteins, but most of them are not connected to the other proteins in the complex, so they are unlikely to be clustered together. Clustering this network produces both the cluster consisting of the four complex proteins (generated by CMC), as well as a larger cluster consisting of the four complex proteins plus five additional proteins BABAM1, BRE, BRCC3, BRCA2, and FAM175B. Recent papers indicate that the former three additional proteins have been included in the BRCA1-A complex. (d) The likelihood network shows that the three additional members are completely connected in a clique with two of the original complex members FAM175A and UIMC1 via PPI edges with strong co-complexness. The four original members themselves are less strongly connected, via two functional associations with high co-complexness and a few low co-complexness PPIs.

associations to four proteins.

Human BRCA1-A complex

Figure 3.14 shows the human BRCA1-A complex, which is involved in DNA repair. The CORUM reference set of complexes specify that complex consists of four proteins, BRCA1, BARD1, FAM175A, and UIMC1, while a survey of current literature reveals that it is composed of at least three more proteins, BRE, BABAM1, and BRCC3. While the PPI network for this complex is fully connected, there are extremely large numbers of extraneous edges and neighboring proteins, chiefly because BRCA1 itself is connected to around 180 proteins. Note that the three new members BRE, BABAM1, and BRCC3 are also connected to the original complex proteins. After weighting the

composite network and keeping the top $k = 20000$ edges, BRCA1 is still connected to a large number of proteins (62), but the majority of them are not connected to the other proteins in the complex, so they are unlikely to be clustered together. Moreover, BRE, BABAM1, and BRCC3 are still highly connected to the original complex proteins. Indeed, clustering this network produces both the cluster consisting of the four CORUM proteins (generated by CMC), as well as a larger cluster consisting of the four CORUM proteins plus the three new members and two extra proteins (generated by IPCA). The likelihood network shows that PPI edges with strong co-complexness induce a full clique between two CORUM complex members FAM175A and UIMC1 with the three new members and an additional protein FAM175B; on the other hand, the four CORUM complex proteins themselves are less strongly connected, via two functional associations with high co-complexness and a few low co-complexness PPIs. This provides ample evidence that the three new proteins belong to this complex, while the inclusion of two extra proteins BRCA2 and FAM175B is likely due to their participation in other complexes that overlap with the BRCA1-A complex.

3.3.9 Two novel predicted complexes

We select two novel complexes predicted with the COMBINED strategy using the SWC network, with the entire reference set of complexes for training.

One high-scoring novel yeast complex, generated by all six clustering algorithms, is composed of four proteins, MMS1, MMS22, RTT101, and RTT107, and is annotated with two high-level BP terms, DNA metabolic process and response to stress. Figure 3.15a shows its likelihood network. The four proteins are fully connected by six literature co-occurrences with strong co-complexness, and six functional associations with strong or moderate co-complexness. Five PPI edges with moderate or weak co-complexness also connect the proteins. The diverse mix of data sources provides convincing evidence for this complex. A scan through the literature reveals that these four proteins form a complex named Cul8-RING ubiquitin ligase complex [73], thought to be involved in DNA repair and regulation of chromatin metabolism, which the yeast reference complexes set CYC2008 has apparently failed to include.

Figure 3.15b shows a high-scoring novel human complex, generated by all six clustering algorithms, made up of four proteins, HCN1, HCN2, HCN3, and HCN4, and annotated with one high-level BP term, transport. These proteins are fully connected by six PPIs with strong co-complexness, while five functional associations with strong to moderate co-complexness and five literature co-occurrences with strong to weak co-

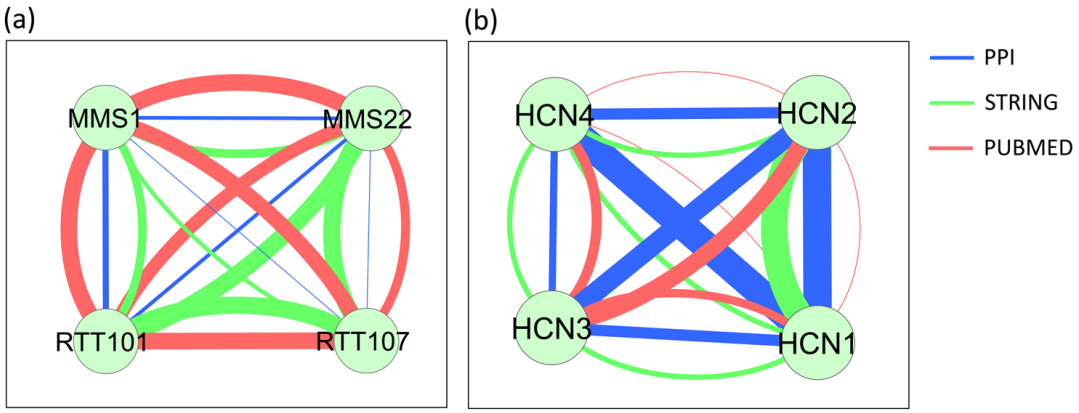

Figure 3.15: Two novel predicted complexes in (a) yeast and (b) human. (a) Novel yeast predicted complex, annotated with DNA metabolic process and response to stress. The four proteins are fully connected by six literature co-occurrences with strong co-complexness, six functional associations with strong or moderate co-complexness, and five PPI edges with moderate or weak co-complexness. The diverse mix of data sources provides convincing evidence for this complex. A scan through the literature reveals that these four proteins form a complex named Cul8-RING ubiquitin ligase complex [73], thought to be involved in DNA repair and regulation of chromatin metabolism, although our set of reference complexes has not been updated to include this complex. (b) Novel human predicted complex annotated with transport process. These proteins are fully connected by six PPIs with strong co-complexness, five functional associations with strong to moderate co-complexness, and five literature co-occurrences with strong to weak co-complexness. The strong PPIs, reinforced by the other data sources, provide high credibility to this prediction. The Uniprot descriptions for these proteins suggest that they may constitute subunits of a potassium channel complex [74].

complexness also connect the proteins. The strong PPIs, reinforced by the other data sources, provide high credibility to this prediction. Indeed, the Uniprot descriptions for these proteins suggest that they may constitute subunits of a potassium channel complex [74].

3.4 Conclusion

In this chapter, we introduce a maximum-likelihood supervised approach for weighting composite protein networks for predicting protein complexes, called SWC (Supervised Weighting of Composite networks). First, we construct a composite protein network using three heterogeneous data sources: PPI, predicted functional association, and co-occurence in literature abstracts. Next, we weight each edge of the composite network based on its posterior probability of belonging to a protein complex, using a naive-Bayes maximum-likelihood model learned from a set of training complexes. The weighted composite network is then used by clustering algorithms to predict new complexes. We also propose a simple aggregative clustering strategy that combines clusters generated by multiple clustering algorithms, using simple voting.

We evaluate our weighting scheme using six clustering algorithms, as well our ag-

gregative clustering strategy, on the prediction of yeast and human complexes. We demonstrate that our proposed method outperforms a supervised data-integration approach using boosting, a predicted functional-association network from STRING, an unsupervised approach using a topological function to weight PPI networks, as well as a PPI network weighted with reliability estimation metric: our approach predicts more correct complexes at higher precision levels, and generates more high-confidence novel complexes with similar or better semantic coherence. We show that SWC gives the biggest improvement among complexes with many missing co-complex interactions. Using a few example complexes, we show that SWC increases the density of the complexes' subgraphs, and filters them to remove extraneous edges. Furthermore, our approach allows visualization of the evidence of predicted complexes, using learned likelihood parameters to express strengths of co-complex relationships of each data type. This aids human evaluation of the credibility of predicted complexes.

Finally, we present two novel predicted complexes: a four-protein yeast complex possibly involved in DNA metabolism and stress response, and a four-protein human complex possibly involved in transport processes. We show that these predictions appear credible from their evidences, being supported by diverse data sources with strong co-complexness. Indeed, a recent paper presents the predicted yeast complex as the Cul8-RING ubiquitin ligase complex, while the Uniprot database provides evidence that the predicted human complex may exist as a potassium channel complex.

SWC software package and data files are available at http://compbio.ddns.comp.nus.edu.sg/~cherny/SWC/.

Chapter 4

Decomposing PPI Networks for Complex Discovery

4.1 Introduction

Many algorithms have been developed to discover complexes from PPI networks, typically by searching for dense subgraphs [32, 35, 37, 39, 40, 42, 75, 76]. However, the performance of existing algorithms is not satisfactory, even in *Saccharomyces cerevisiae* (baker's yeast) where PPI data is fairly complete. One reason behind this is that the PPI network does not capture the dynamic nature of protein interactions and complexes: interactions do not all occur simultaneously, but rather may occur at different times with varying durations, and in different subcellular locations. In the cell, multiple copies of the same protein may exist bound in different complexes in different cellular locations; but in the PPI network, these copies of the protein are conflated into a single vertex, with all its temporally- and spatially-diverse interactions represented as undifferentiated edges connected to it. Existing complex-discovery algorithms do not take this into consideration. As a result, the clusters generated often contain extra proteins that preclude them from matching true complexes.

An ideal solution would be to decompose the PPI network into several smaller networks such that interactions within each smaller network are contextually coherent. In reality, it is very difficult to know which subset of interactions take place together. Here we choose to use cellular-component terms from Gene Ontology (GO [6]) to decompose PPI networks because a protein complex can be formed only if its proteins are localized within the same compartment of the cell. We use only localization GO terms that are relatively general for decomposition.

Hub proteins offer a second way to decompose the PPI network contextually. Hub proteins are proteins that have a lot of neighbors in the PPI network, and these neighbors often belong to multiple complexes [12]. Hubs make it difficult for complex-

discovery algorithms to correctly delimit the boundaries of complexes, and may cause complexes to be merged together as large clusters. To avoid this, we remove hub proteins from PPI networks prior to clustering. After the clusters are generated from the remaining PPI network, we then add the removed hub proteins back to the clusters.

We tested the above methods on the discovery of both yeast and human complexes. The results show that these methods can improve the performance of existing complex-discovery algorithms significantly. In the rest of the chapter, we first describe the two methods for decomposing PPI networks, and then show experiment results.

4.2 Methods

In this section, we first describe the two methods for decomposing PPI networks for complex discovery, and then briefly introduce the complex-discovery algorithms used in our experiments.

4.2.1 Decomposition by localization GO terms

A protein complex can only be formed if its proteins are localized within the same compartment of the cell. Hence we use cellular-component GO terms to decompose a given PPI network into several smaller PPI networks such that all proteins in each smaller network are annotated with the same localization GO term. We use only localization GO terms that are relatively general for decomposition. There are several reasons for this. First, it is relatively easy to obtain the rough localization of proteins, compared with obtaining precise and specific localization, so many proteins are already annotated with rough localizations in the public databases. Secondly, very specific GO terms are annotated to very few proteins. Using them to decompose PPI networks produces many small fragments, and lots of information may be lost due to the decomposition. Finally, some very specific cellular-component GO terms correspond to complexes, which we are trying to discover in the first place.

We use a threshold N_{GO} to select GO terms for decomposition, where N_{GO} should be large. The selected GO terms are annotated to at least N_{GO} proteins, and none of their descendant terms is annotated to at least N_{GO} proteins. If a GO term is selected, then none of its ancestor terms or descendant terms will be selected.

Given a selected GO term, we first remove all the proteins that are not annotated (explicitly, or implicitly via the true-path rule, i.e. via GO-ancestors of annotations) to the term from the given PPI network, and then apply a complex-discovery algorithm on the resultant network. This process is repeated for every selected GO term. The

final set of clusters is the union of the clusters discovered from every filtered network. Duplicated clusters are removed.

4.2.2 Hub removal

Hub proteins are those proteins that have many neighbors in the PPI network. We use a threshold N_{hub} to define hub proteins. We call a protein a *hub protein* if it has at least N_{hub} neighbors. A hub protein often connects proteins that belong to different complexes, which makes it hard to decide the boundary of the complexes and the membership of the hub proteins.

To alleviate the impact of the hub proteins, we first remove hub proteins from a given PPI network, and then use an existing complex-discovery algorithm to find clusters from the remaining network. After the clusters are generated, hub proteins are added back to the clusters. We add a hub protein u back to a cluster C based on the connectivity between u and C, which is defined as follows:

$$Connectivity(u, C) = \frac{\sum_{v \in C} w(u, v)}{|C|} \tag{4.1}$$

where $w(u, v)$ is the weight of edge (u, v), and it is calculated from the original PPI network using iterative AdjustCD [35] before removing hubs. If there is no edge between u and v, then $w(u, v)$=0. A hub protein u is added to a cluster C only if $Connectivity(u, C) \geq hub\_add\_thres$, where hub_add_thres is a number between 0 and 1.

4.2.3 Combining the two methods

We combine the two methods by first removing hub proteins from the given PPI network, and then decomposing the resultant PPI network using selected GO terms. The whole process is described below:

1. Let $\mathcal{C}$ be the set of clusters generated. Initially $\mathcal{C}$ is empty.

2. Remove hub proteins that have at least N_{hub} neighbors from the given PPI network G. Let G' be the resultant network.

3. Let $g_1, \cdots, g_m$ be the localization GO terms that are selected using threshold N_{GO}. For each g_i, do the following:

   - Remove proteins that are not annotated with g_i from G'. Let G'_i be the resultant network.

- Apply a complex-discovery algorithm on G'_i to find clusters. Let $\mathcal{C}_i$ be the set of clusters generated.
- $\mathcal{C}$=$\mathcal{C} \cup \mathcal{C}_i$;

4. Remove duplicated clusters from $\mathcal{C}$.

5. Add hub proteins back to clusters in $\mathcal{C}$.

4.2.4 Complex-discovery algorithms

We use the following complex-discovery algorithms in our study: MCL, RNSC, IPCA, CMC, ClusterONE, and COACH (described in Chapter 2.4). MCL and RNSC generate a partition of the PPI network, and they do not allow overlap among clusters. The other algorithms, CMC, ClusterOne, IPCA, and Coach, allow overlap among clusters.

4.3 Results and discussion

In this section, we first describe the datasets and the evaluation method used in our experiments, and then study the impact of the two decomposition methods on the performance of the four complex-discovery algorithms.

4.3.1 Experiment settings

We use PPI data as described in Chapter 2.5.1, obtained by combining physical interactions from multiple databases, then scored by reliability, and filtered to take the top $20,000$ edges.

We use precision-recall graphs to evaluate the predicted clusters. As described in Chapter 2.5.2, this is obtained by scoring predicted clusters by their weighted densities, then calculating the precision and recall at varying score thresholds. We also calculate the area under the curve (AUC) of the precision-recall graphs, and the F-measure.

We use manually-curated yeast and human complexes as reference complexes. For yeast, we use the CYC2008 [56] set which consists of 408 complexes. Only complexes of size greater than three proteins are used for testing; there are 149 such complexes in CYC2008. For human, we use the CORUM [57] set which consists of 1829 complexes, of which 714 are of size greater than three.

Parameter settings of the four complex-discovery algorithms

For each clustering algorithm, we first determined the parameters that gave the best performance (in terms of highest AUC) for complex discovery, when no decomposition

Clustering algorithm	Parameters for yeast	Parameters for human
CMC	overlap_thres=0.5, merge_thres=0.75	overlap_thres=0.5, merge_thres=0.75
ClusterONE	-s 4	-s 4
MCL	-I 3	-I 3
IPCA	-S4 -P2 -T0.4	-S4 -P2 -T0.6
RNSC	-e10 -D50 -d10 -t20 -T3	-e10 -D50 -d10 -t20 -T3
Coach	*default*	*default*

Table 4.1: The six clustering algorithms and their parameters used for yeast and human complex discovery.

	Yeast			Human		
N_{GO}	#GO terms	#prots discarded	#PPIs discarded	#GO terms	#prot discarded	#PPIs discarded
1,000	6	1,001	5,206	10	3,140	9,870
500	8	1,388	7,133	14	3,492	11,409
300	10	1,526	7,698	19	3,666	13,189
100	25	2,151	12,172	48	4,704	17,880
30	48	2,350	13,313	97	5,017	18,153

Table 4.2: Different values of N_{GO} used, and the resulting number of proteins and PPIs discarded in the decomposed networks.

methods are used. The parameters found are given in Table 4.1 (parameters not shown are set to their default values). We stick to the same parameters throughout all the experiments, including when the decomposition methods are used.

4.3.2 Decomposition by localization GO terms

The first experiment studies the impact of GO-term decomposition on the performance of the six algorithms. We use annotations in Gene Ontology to select GO terms for decomposition. If a protein is annotated to none of the selected GO terms, then the protein is discarded because it does not occur in any of the small PPI networks after decomposition. If the two proteins of an interaction do not share any common selected GO term, then the two proteins do not co-occur in any subnetwork after decomposition and the interaction between them is lost too. Table 4.1 shows the number of GO terms selected, and the number of proteins and interactions discarded, under different N_{GO} values. The numbers of discarded proteins and interactions are considerably large when N_{GO} is small.

Figure 4.1 shows the precision-recall graphs of the six clustering algorithms when different N_{GO} thresholds are used for selecting localization GO terms, for yeast complex prediction at $match\_thres = 0.5$. For clarity, we only show the graphs for $N_{GO} = 30, 100, 300$. When N_{GO} is small ($N_{GO} = 30, 100$), recall drops because too many interactions and proteins are discarded as shown in Table 4.2. When $N_{GO} = 300$, recall either improves considerably (for MCL and ClusterOne), or is maintained at the same level (for the remaining algorithms), while precision also improves considerably

	Yeast		Human	
N_{hub}	#hub prots removed	#PPIs removed	#hub prots removed	#PPIs removed
200	2	568	2	534
150	3	739	6	1,203
100	9	1,408	21	2,932
75	44	3,920	36	4,084
50	126	7,178	90	6,936
30	340	12,238	225	10,754

Table 4.3: Different values of N_{hub} used, and the resulting number of hub proteins and PPIs removed.

for almost all cases (except IPCA). Hence we should use GO terms that are relatively general to decompose PPI networks to avoid breaking the whole network into tiny fragments. Overall, the performance for all clustering algorithms improves, except for IPCA where performance remains similar.

We also compare the improvements with that of using random protein groups for decomposition. Random protein groups are generated by replacing proteins of the selected GO terms with randomly-picked proteins. We generated 100 sets of random protein groups and plot their mean precision-recall graphs in Figure 4.1. It is clear that using random protein groups to decompose the PPI network decrease the performance of all the algorithms greatly, where the random protein groups were generated from GO terms selected at a threshold of 300.

Figure 4.2 shows the corresponding precision-recall graphs for human complex prediction with GO decomposition, at $match\_thres = 0.5$. Here we only show the graphs for $N_{GO} = 300, 500, 1000$. As shown in Table 4.2, for a given N_{GO} value, much more interactions and proteins are discarded from decomposing the human PPI network, compared to yeast. Thus it is no surprise that even more general GO terms (using higher N_{GO}) are required to decompose of the human network. With $N_{GO} = 300, 500$, recall suffers for CMC, IPCA, and Coach, as too many interactions and proteins are discarded. With $N_{GO} = 1000$, both recall and precision improve substantially for all algorithms except Coach—in this case, the precision is poor in the low-recall range, but is better when all its clusters are considered (at the high-recall range).

Here again, using random protein groups to decompose the PPI network gives dismal performance, as seen in Figure 4.2.

4.3.3 Hub removal

The second experiment studies the impact of hub removal on the performance of the four algorithms. Table 4.3 shows the number of hub proteins and interactions removed under different N_{hub} values. The numbers indicate that a small number of hub proteins

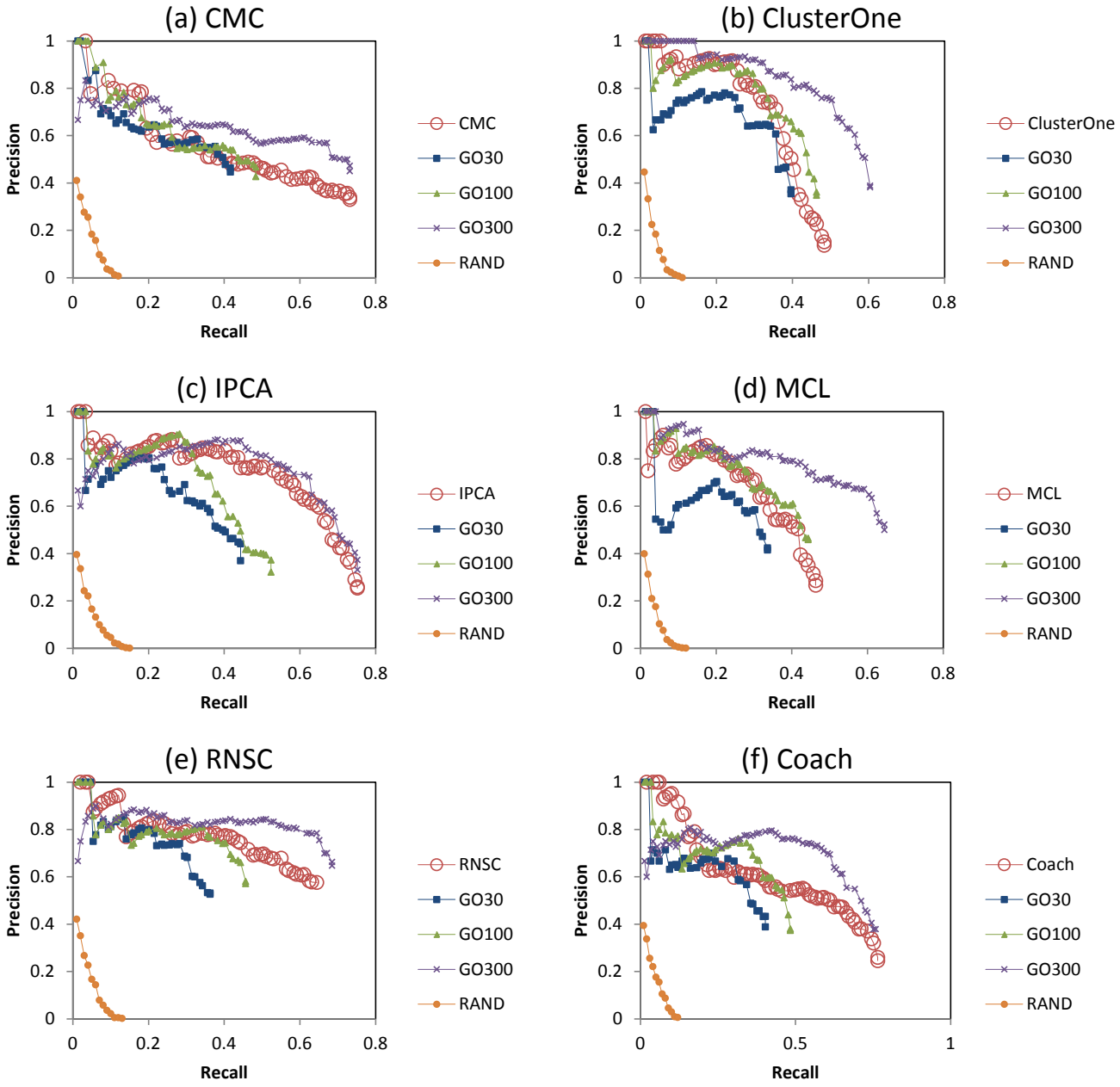

Figure 4.1: Precision-recall graphs for yeast complex prediction using GO decomposition at $N_{GO} = 30, 100, 300$, for the six clustering algorithms.

account for a large number of interactions. For example, in the human PPI network, the percentage of proteins with at least 30 neighbors is about 2%, while they account for about 24% of the interactions.

We use parameter *hub_add_thres* to determine when a hub can be added to a cluster. In our experiments, we found that the proper range for *hub_add_thres* is [0.2, 0.9]. In the rest of the experiments, we set *hub_add_thres* to 0.3.

Figure 4.3 shows the precision-recall graphs of the six complex-discovery algorithms when different N_{hub} thresholds are used for removing hub proteins, for prediction of

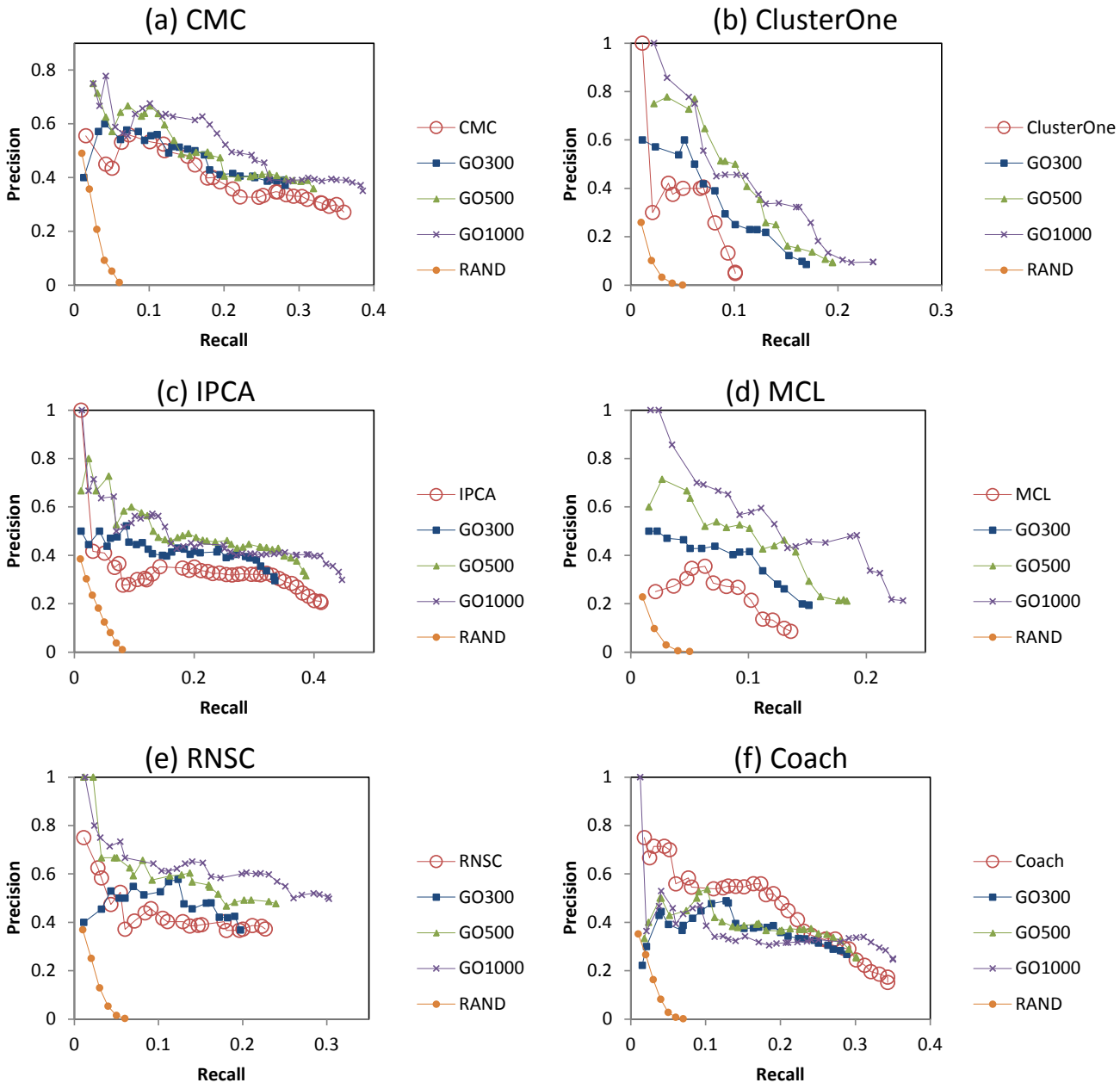

Figure 4.2: Precision-recall graphs for human complex prediction using GO decomposition at $N_{GO} = 300, 500, 1000$, for the six clustering algorithms.

yeast complexes. For clarity, we only show the results for $N_{hub} = 30, 50, 100$. The hub-removal strategy is helpful for CMC, ClusterOne, and Coach, giving mainly an improvement in precision (as well as recall for ClusterOne); however, hub removal does not change the performance much for IPCA and RNSC, and causes a slight decrease in precision for MCL.

Figure 4.4 shows the corresponding human precision-recall graphs of the six complex-discovery algorithms when different N_{hub} thresholds are used for removing hub proteins. Here we only show the results for $N_{hub} = 50, 100, 300$. Compared to

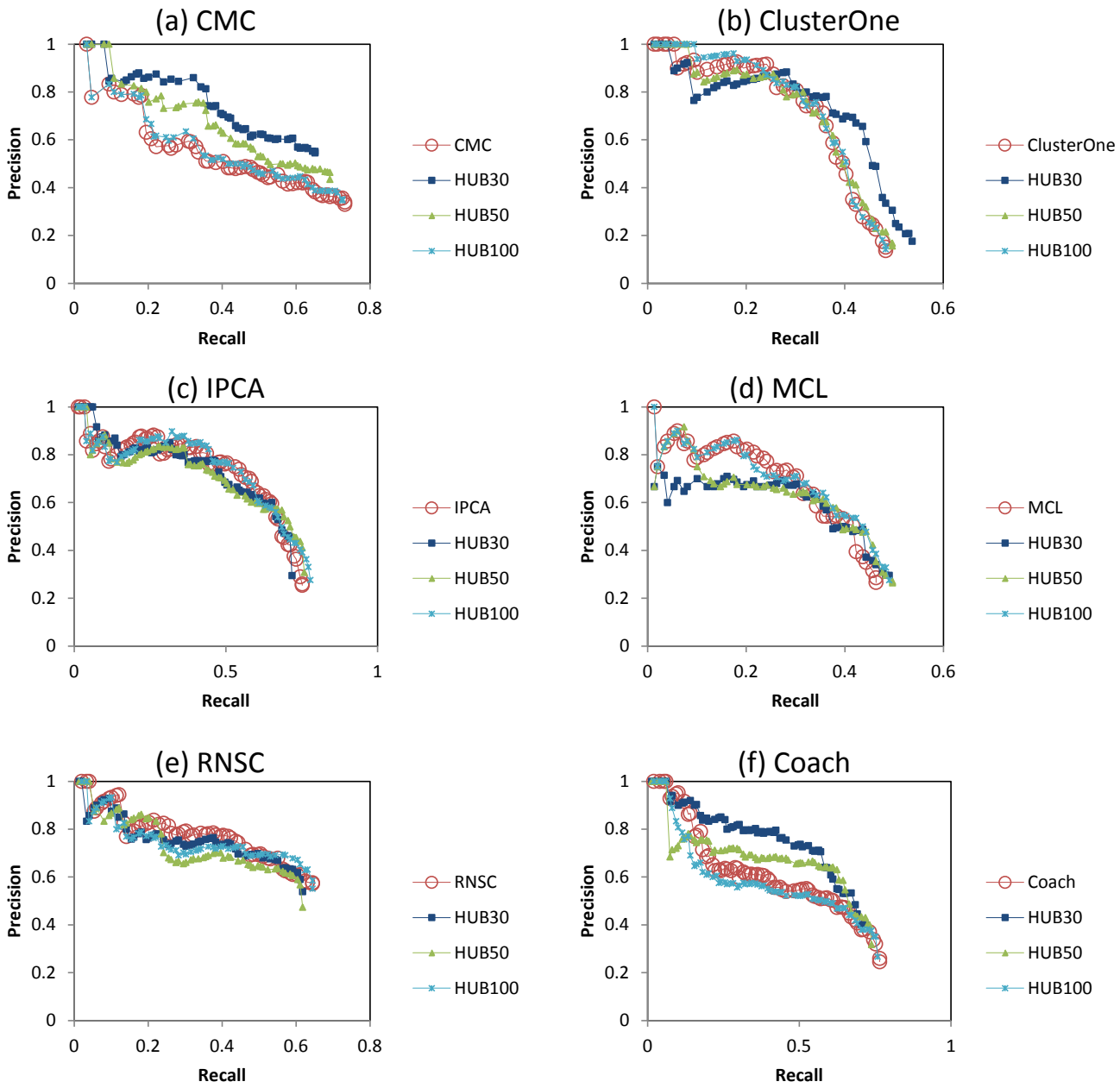

Figure 4.3: Precision-recall graphs for yeast complex prediction using hub removal at $N_{hub} = 30, 50, 100$, for the six clustering algorithms.

yeast, the benefits of hub removal are less clear in human: there is clear precision improvement for CMC, and only slight precision improvement for Coach; for ClusterOne the improvement is mainly in recall. Hub removal seems to have little effect for IPCA and RNSC, and causes a slight decrease in precision for MCL.

4.3.4 Combining the two methods

The last experiment is to examine the combined impact of the two decomposition methods. Figure 4.5 shows the precision-recall graphs for yeast complex prediction

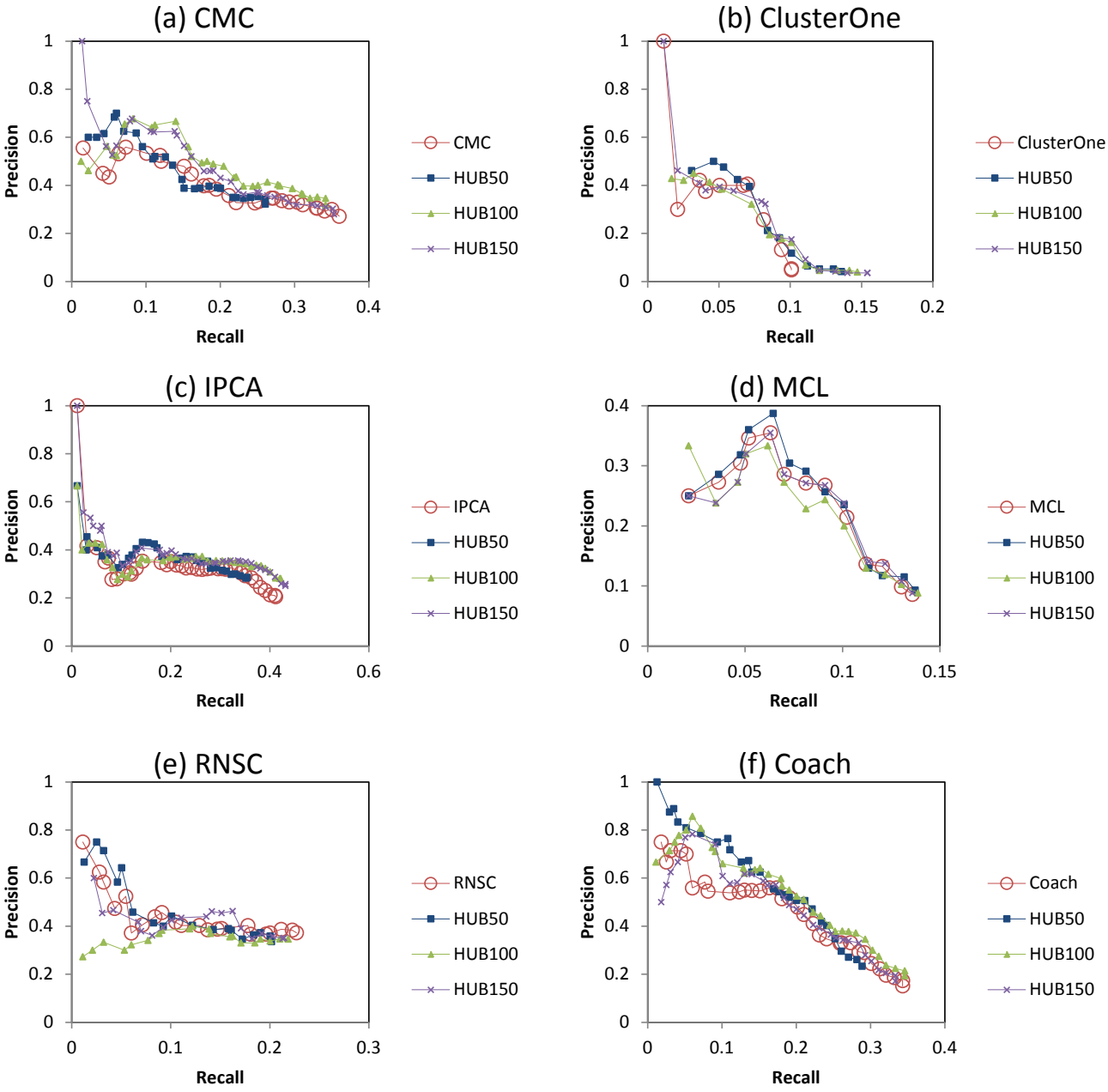

Figure 4.4: Precision-recall graphs for human complex prediction using hub removal at $N_{hub} = 50, 100, 150$, for the six clustering algorithms.

with both decomposition methods. IPCA, MCL, and RNSC do not benefit much from hub removal; so for these algorithms, combining the two decomposition methods yields little improvement compared with using GO decomposition alone. The performance of CMC, ClusterOne, and Coach improve when both methods are used.

Table 4.4 summarizes the performance of yeast complex discovery at $match\_thres = 0.5$ and 0.75, in terms of area under the curve of the precision-recall graph (AUC), and F-score, when using both GO decomposition and hub removal together, or each of them individually, or neither of them. In terms of AUC, only CMC, ClusterOne, and

	$Match\_thr$	F-Score				Prec-Rec AUC			
		Orig	HUB50 GO300	HUB50	GO300	Orig	HUB50 GO300	HUB50	GO300
CMC	.5	.455	**.615**	.533	.557	.417	**.508**	.479	.470
	.75	.275	**.391**	.330	.347	.204	**.278**	.243	.251
ClusterOne	.5	.213	**.483**	.238	.468	.361	**.531**	.362	.514
	.75	.105	**.270**	.107	.255	.209	**.323**	.194	.310
IPCA	.5	.380	**.531**	.438	.460	.564	.560	.549	**.572**
	.75	.143	**.240**	.160	.220	.308	.310	.276	**.323**
MCL	.5	.338	.553	.345	**.563**	.326	.496	.315	**.514**
	.75	.192	.328	.162	**.336**	.170	.255	.104	**.280**
RNSC	.5	.606	.636	.536	**.665**	.500	.560	.455	**.564**
	.75	.355	.377	.321	**.422**	.239	.284	.209	**.305**
Coach	.5	.372	**.573**	.444	.506	.477	**.564**	.505	.536
	.75	.182	**.312**	.223	.262	.218	**.302**	.220	.265

Table 4.4: Performance statistics for yeast complex discovery.

	$Match\_thr$	F-Score				Prec-Rec AUC			
		Orig	HUB150 GO1000	HUB150	GO1000	Orig	HUB150 GO1000	HUB150	GO1000
CMC	.5	.309	**.381**	.312	.367	.148	**.234**	.174	.204
	.75	.077	**.087**	.075	.092	.011	.016	.011	**.017**
ClusterOne	.5	.065	**.138**	.058	.136	.039	**.112**	.045	.104
	.75	.024	**.047**	.018	**.047**	.006	.031	.007	**.032**
IPCA	.5	.274	**.369**	.317	.358	.140	**.255**	.168	.215
	.75	.054	**.080**	.068	.073	.017	**.022**	.020	.019
MCL	.5	.105	**.222**	.107	**.222**	.032	.129	.032	**.131**
	.75	.035	**.078**	.006	**.078**	.006	**.023**	.006	**.023**
RNSC	.5	.282	**.378**	.266	.376	.101	.187	.092	**.191**
	.75	.090	.116	.085	**.118**	.014	**.022**	.016	**.022**
Coach	.5	.210	**.320**	.226	.289	.161	.163	**.165**	.131
	.75	.043	**.074**	.049	**.074**	.008	.011	**.013**	.011

Table 4.5: Performance statistics for human complex discovery.

Coach benefit from using both decomposition methods, while IPCA, MCL, and RNSC benefit most from using just GO decomposition. However, in terms of F-score, IPCA also benefits substantially from using both decomposition methods: this is because it attains much higher precision at the final recall point, when all its predicted clusters are considered.

Figure 4.6 shows the precision-recall graphs for human complex prediction with both decomposition methods. As in yeast, CMC and ClusterOne benefit from combining both decomposition methods; however, because Coach performs poorly using GO decomposition in human, it performs worse using both methods compared to just using hub removal. RNSC and MCL do not benefit much from hub removal, so combining the two decomposition methods gives no improvement compared with using GO decomposition alone (and actually gives poorer performance in RNSC). However, here IPCA obtains substantial improvement from combining both decomposition methods.

Table 4.5 summarizes the performance of human complex discovery at

$match\_thres = 0.5$ and 0.75, in terms of area under the curve of the precision-recall graph (AUC), and F-score, when using both GO decomposition and hub removal together, or each of them individually, or neither of them. In terms of AUC, only CMC, ClusterOne, and IPCA benefit from using both decomposition methods; MCL and RNSC benefit most from using just GO decomposition, while Coach benefits most from using just hub removal. However, in terms of F-score, combining both decomposition methods does better in most cases, and does no worse compared to using just GO decomposition or hub removal individually.

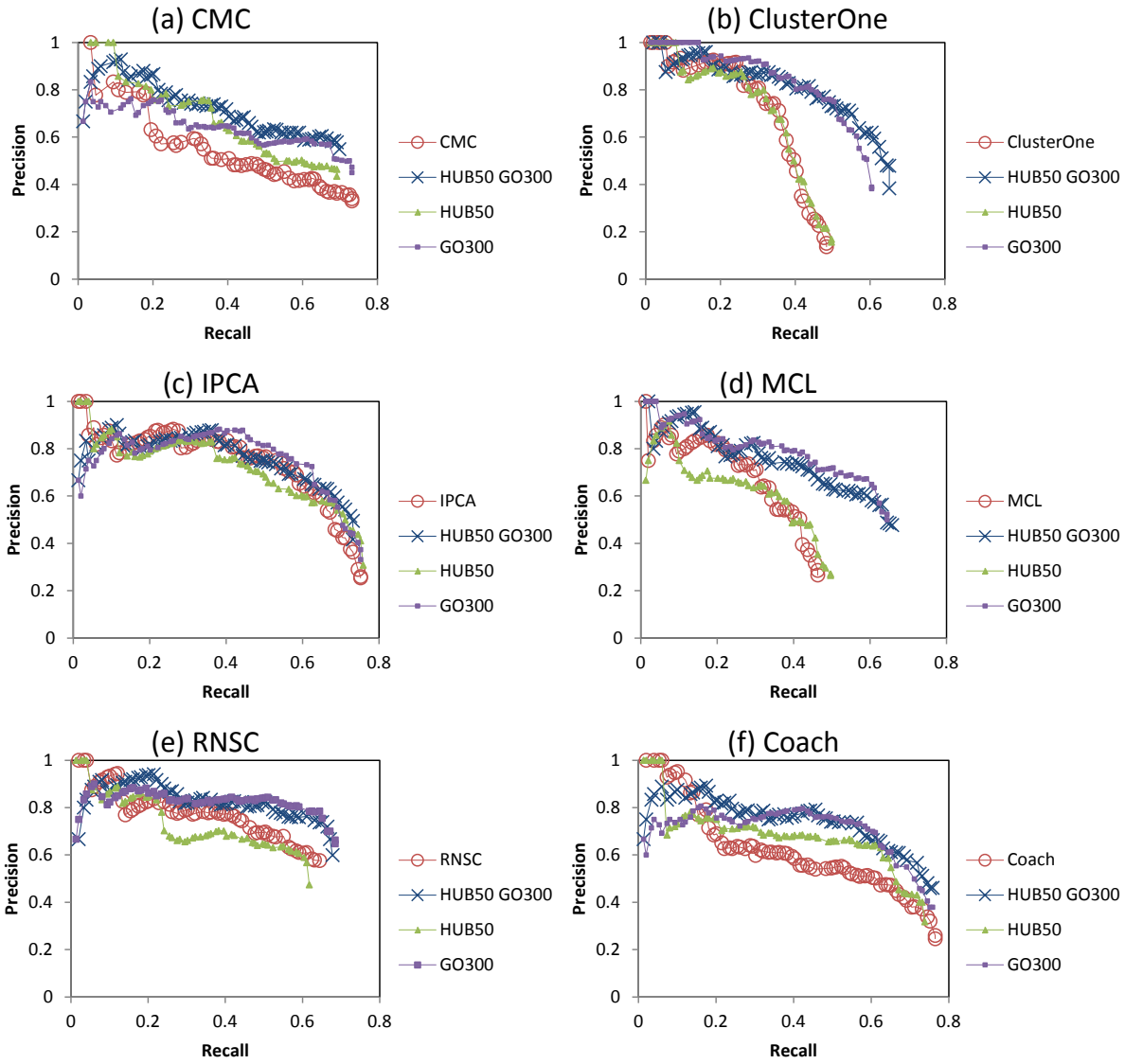

Figure 4.5: Precision-recall graphs for yeast complex prediction using both GO decomposition ($N_{GO} = 300$) and hub removal ($N_{hub} = 50$), for the six clustering algorithms.

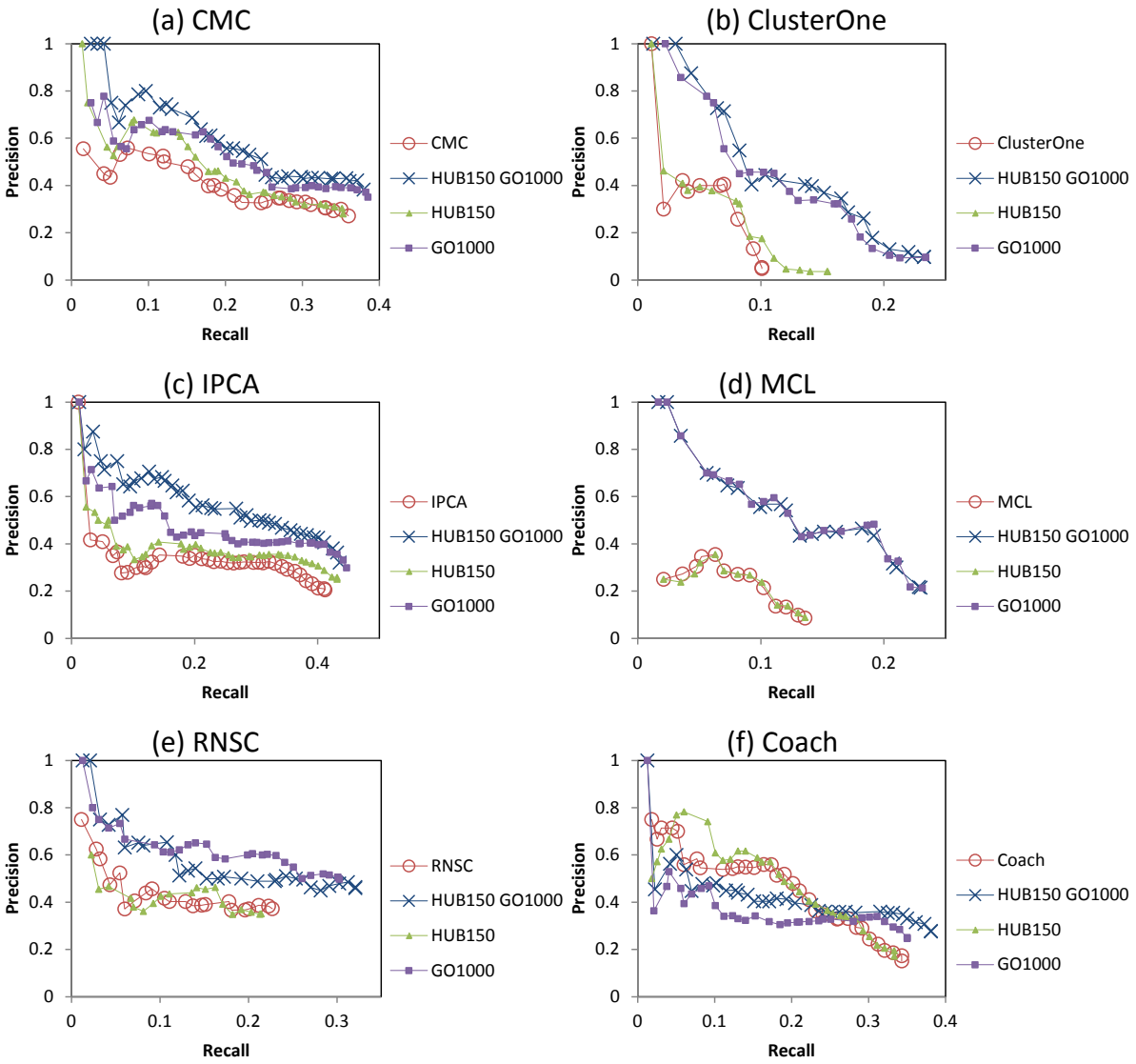

Figure 4.6: Precision-recall graphs for human complex prediction using both GO decomposition ($N_{GO} = 1000$) and hub removal ($N_{hub} = 150$), for the six clustering algorithms.

4.3.5 Performance among stratified complexes

To further investigate how PPI decomposition improves the performance of large-complex prediction, we study its effects on predicting large complexes with different degrees of extraneous and missing interactions.

As described in Chapter 2.5.1, we stratify the complexes by size, EXT (the number of external proteins that are highly connected to it), and DENS (density). Figures 2.3 and 2.4 show the distribution of the large complexes (containing four or more proteins) in terms of DENS, EXT, and our six analysis groups (stratified by DENS and EXT), for yeast and human.

In both yeast and human, around 40% of complexes have high EXT. We expect the prediction of these complexes to be extremely challenging, as it would be difficult to accurately delimit their borders from their highly-connected surroundings (the highly-connected external proteins are likely to be recruited into the predicted complexes). Most complexes in yeast have high density: only 10% of complexes have low DENS. On the other hand, in human about 35% of complexes are sparsely connected with low DENS. We expect these sparsely-connected complexes to also be difficult to predict, as they do not form dense clusters that are picked out by most clustering algorithms.

To investigate the benefits of PPI decomposition in predicting complexes from the different stratified groups of reference complexes, we compare how well the complexes from the different strata are matched by clusters generated from the decomposed network, versus without decomposition. Figure 4.7 shows the match scores of the best-matching clusters found for the yeast complexes in the different strata, using (a) PPI decomposition, and (b) without decomposition, while (c) shows the improvements in the score medians. Decomposition tends to give the bigger improvements (in terms of generating more well-matched clusters) among complexes with high EXT. As expected, decomposition helps with predicting complexes that are embedded within dense regions of the PPI networks, which frequently correspond to overlapping complexes: PPI decomposition splits such complexes (and their surrounding regions) into different temporal and spatial contexts, from which the boundaries of these complexes can be more accurately delimited. On the other hand, PPI decomposition does not work as well with complexes with low DENS, as it may remove edges from such complexes and make them even more difficult to find. Note that even though decomposition leads to worse matching scores for some low-density complexes, in yeast such complexes are in the minority, so this is offset by the improvements among higher-density complexes.

Figure 4.8 shows the corresponding charts for human complexes. Again, decom-

position gives improvements among complexes with high EXT, as it eliminates many extraneous edges when it splits the PPI network into separate spatial and temporal contexts. Decomposition can also give worse matching scores for some low-density complexes, as it may remove their edges and make them even more difficult to find. However, this is offset by the improvements among higher-density complexes to give an overall increase in prediction performance. For some clustering algorithms (ClusterOne, MCL, and Coach), decomposition can give even greater improvement among the less-challenging complexes (those with low EXT, high DENS). This is because decomposition still helps to remove their extraneous edges to generate better-matching clusters; moreover, decomposing the network into overlapping subnetworks allows MCL to find overlapping complexes, which is prevalent in human.

4.4 Conclusions

In this chapter, we proposed two methods to decompose PPI networks to account for spatial and temporal dynamics of PPIs, for the purpose of complex discovery. First, we used Gene Ontology localization terms to decompose the PPI network into spatially-coherent subnetworks; second, we removed hub proteins to break apart dense clusters that may correspond to distinct complexes that are fused together. We used six complex-discovery algorithms to experimentally study the effectiveness of the two decomposition methods for the prediction of yeast and human complexes.

The results show that network decomposition helps improve the performance of the six algorithms significantly. GO decomposition consistently improves the performance of all six algorithms, while hub removal appears less effective as it only benefits some of the algorithms. Nonetheless, combining both decomposition methods consistently gives better performance for all six algorithms for yeast and human complexes, compared to not performing decomposition at all.

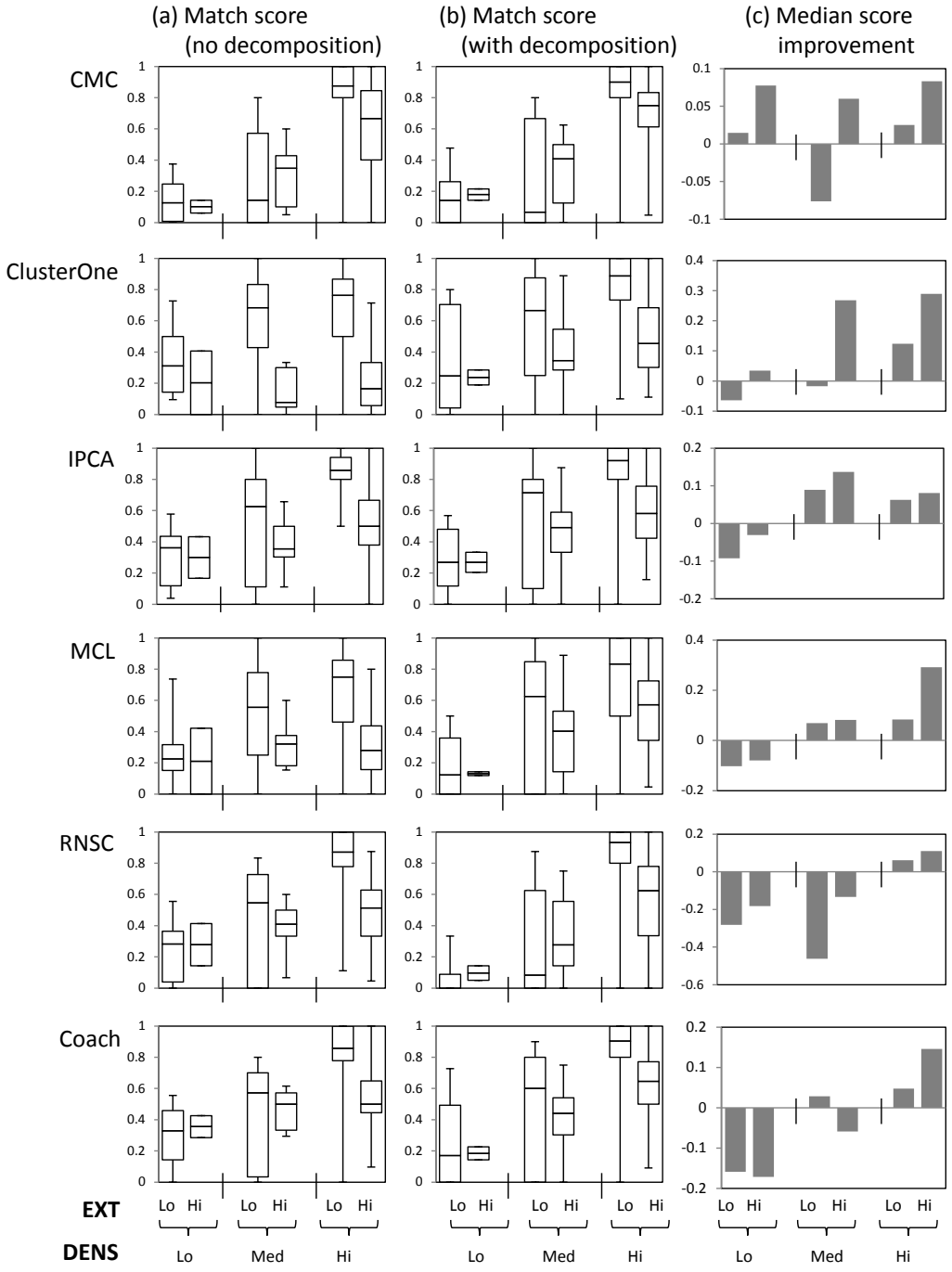

Figure 4.7: Match scores of the best clusters to yeast complexes in the six analysis strata, (a) without PPI decomposition, and (b) with PPI decomposition, generated by various clustering algorithms. (c) shows the improvements score medians.

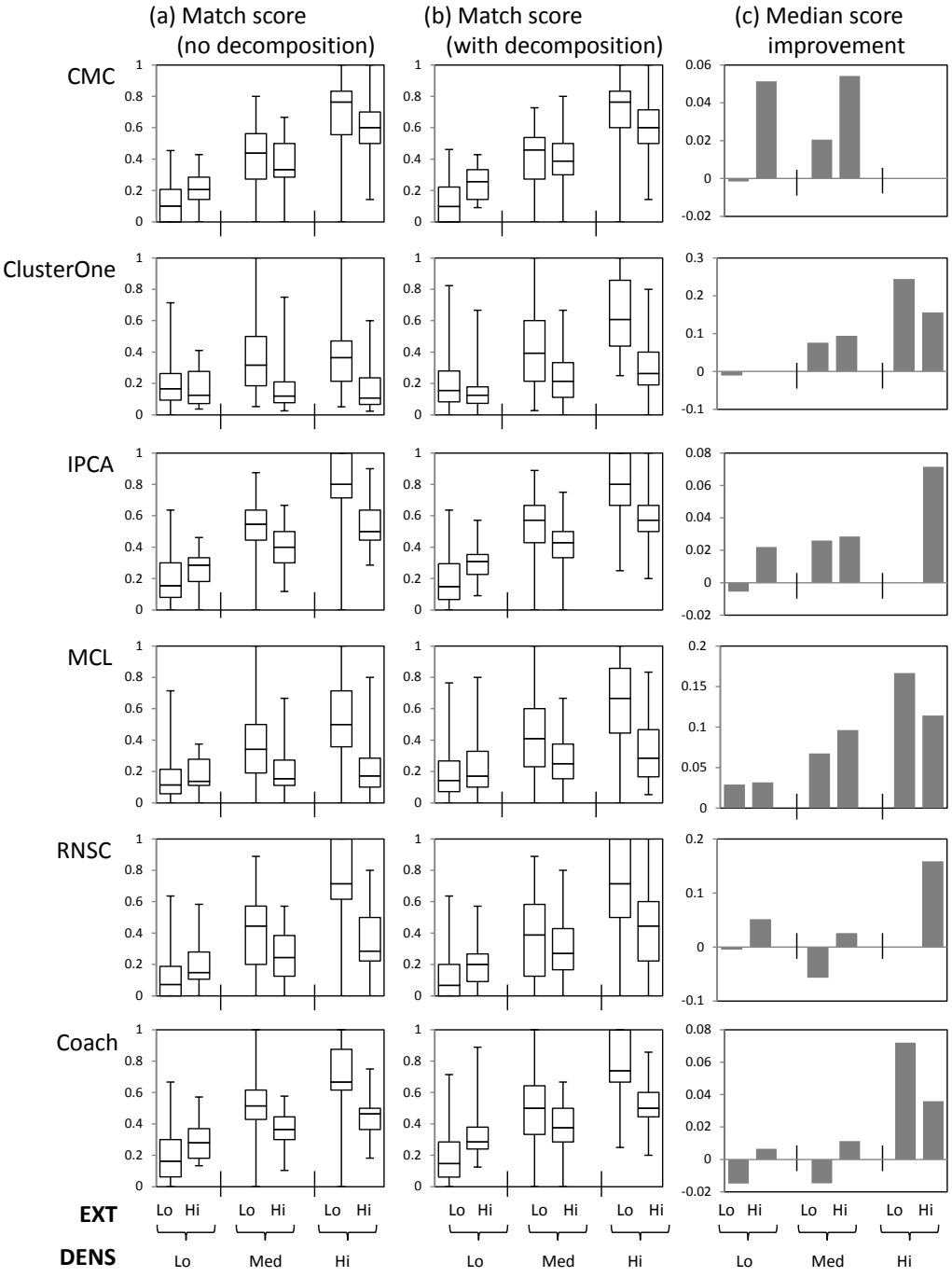

Figure 4.8: Match scores of the best clusters to human complexes in the six analysis strata, (a) without PPI decomposition, and (b) with PPI decomposition, generated by various clustering algorithms. (c) shows the improvements score medians.

Chapter 5

Discovery of Small Protein Complexes

5.1 Introduction

It has been noted that the distribution of complex sizes follows a power law distribution [31], meaning that a large majority of complexes are small. Thus the discovery of small complexes is an important subtask within complex discovery. An inherent difficulty is that the strategy of searching for dense clusters becomes problematic: fully-dense (i.e. cliques) size-2 and size-3 clusters correspond to edges and triangles respectively, and only a few among the abundant edges and triangles of the PPI network represent actual small complexes. Furthermore, high-throughput PPI data suffers from significant amounts of noise, in terms of false positives (spuriously detected interactions) as well as false negatives (missing interactions). This presents a challenge for complex discovery from PPI data, and is especially severe for the discovery of small complexes, which is more sensitive to extraneous or missing edges: for a size-2 complex, a missing co-complex interaction disconnects its two member proteins, while only two extraneous interactions are sufficient to embed it within a larger clique (a triangle).

Our proposed approach to address these challenges consists of two steps. First, we weight the edges of the PPI network with the probabilities of belonging to a complex, in a size-specific manner. Second, we extract the small complexes from this weighted network. In the first step, our weighting approach, called Size-Specific Supervised Weighting (SSS), integrates three different data sources (PPIs, functional associations, and literature co-occurrences) with their topological characteristics (degree, shared neighbours, and connectivity between neighbours), as well as an overall topological-isolatedness feature. SSS uses a supervised maximum-likelihood naive-Bayes model to weight each edge with two separate probabilities: that of belonging to a small complex, and of belonging to a large complex. In the second step, our complex-

extraction approach, called Extract, uses these weights to predict and score candidate small complexes, by weighting their densities with a cohesiveness function [41] that incorporates both small- and large-co-complex probabilities of edges within and around each cluster.

In our previous approach presented in Chapter 3, Supervised Weighting of Composite Networks (SWC [4]), we integrated diverse data sources (including topological characteristics) with a supervised approach to accurately score edges with co-complex probabilities, and attained good performance in predicting large complexes (of size greater than three) in yeast and human. However, SWC's performance in scoring edges from small complexes is unsatisfactory. This is because edges in small complexes have radically-different topological characteristics from edges in large complexes. And since there are a far greater number of edges from large complexes than from small complexes, the learned model reflects the features of the former rather than the latter. Thus, here we model both small complexes and large complexes separately, and use both models to weight the edges, which captures the characteristics of small-complex edges more accurately. Moreover, we incorporate additional topological features compared to SWC, to allow more discrimination between small and large complexes.

By integrating two additional data sources (functional associations and literature co-occurrences) with supervised learning, our approach reduces the amount of spurious interactions among the PPIs. Complexes tend to be characterized by certain topological characteristics in the PPI network (for example, they tend to be densely connected and bordered by a sparse region), but smaller groups of proteins are more likely to take on such characteristics by chance. Integrating topological features from multiple data sources reduces the discovery of false-positive complexes, as it is less likely that all data sources share such characteristics by chance in a random set of proteins.

An important topological characteristic of complexes, large and small, is that they tend to be topologically isolated, or bordered by a sparse region. Many complexes exhibit a core-attachment structure [18], where distinct complexes can share common subsets of proteins (called the core), with variations among the remaining proteins (attachments). Since distinct complexes can share proteins, they overlap in the PPI network, and thus are not expected to be completely isolated; nonetheless, proteins in small complexes with core-attachment structures are still more isolated than those in large complexes. Thus we incorporate an isolatedness feature derived from an initial posterior probability calculation, which contributes to discriminating between edges in small complexes, large complexes, or in no complex.

Predicted complexes are typically given some score indicative of confidence in the prediction. The weighted density of the predicted complex is frequently used for this purpose (for example in [4, 35]): assuming the edge weights represent co-complex estimates, the weighted density averages over the weights of all the edges within the predicted complex, giving an overall measurement of the prediction's reliability. However, for predicted small complexes the weighted density is derived from only one or three edges (corresponding to size-2 or size-3 clusters respectively), making it susceptible to noisy edge weights. Thus we incorporate a cohesiveness function in scoring predicted complexes, which includes both internal edges within the cluster, as well as outgoing edges around the cluster.

Some researchers have already noted the importance and difficulty of predicting small complexes, and proposed specialized approaches to address this challenge. For example, Ruan *et al.* proposed two methods for predicting size-two and size-three complexes separately [77, 78]. Both methods use weights of the interactions around putative small complexes as well as the number of domains in the constituent proteins to derive features for a kernel-based supervised approach. Our approach differs in several ways. We use a naive-Bayes model as it is transparent, so that learned parameters can be validated and used to understand predicted candidate complexes. Moreover, naive-Bayes models are known to be robust even when few training samples are available. We also incorporate data from other sources (functional associations and literature co-occurrence), as well as their topological characteristics, to aid in distinguishing small versus large complexes.

We test our approach on the prediction of small complexes in yeast and human, and obtain improved performance in both organisms. In the rest of the chapter, we first describe each of the two steps of our approach. Next we describe our experimental methodology, and finally present and discuss our results.

5.2 Methods

In this section, we describe our approach for predicting small protein complexes, which consists of two stages: first, Size-Specific Supervised Weighting (SSS) of the PPIs; second, extracting small complexes from this weighted PPI network.

5.2.1 Size-Specific Supervised Weighting (SSS) of the PPI network

SSS uses supervised learning to weight each edge of the reliable PPI network with two posterior probabilities, that of being a small-co-complex edge (i.e. of belonging to a

small complex), and that of being a large-co-complex edge, given the edge's features. These features consist of diverse data sources, their topological characteristics, and an isolatedness feature derived from an initial calculation of the posterior. We first describe the data sources and features we use, then describe our weighting approach.

Data sources and features

We use three different data sources (PPI, functional association, and literature co-occurrence) together with their topological characteristics as features. Each data source provides a list of scored protein pairs: for each pair of proteins (a, b) with score s, a is related to b with score s, according to that data source. For both yeast and human, the following data sources are used:

- *PPI*: PPI data obtained by combining physical interactions from multiple databases, then scored by reliability, as described in Chapter 2.5.1.,
- *STRING*: Predicted functional-association data obtained from the STRING database, as described in Chapter 3.2.1.
- *LIT*: Co-occurrence of proteins or genes in PubMed literature, as described in Chapter 3.2.1.

For each protein pair in each data source, we derive three topological features: degree (DEG), shared neighbors (SHARED), and neighborhood connectivity (NBC). For each data source, the edge weight used to calculate these topological features is the data-source score of the edge.

- *DEG*: The degree of the protein pair (a, b), or the sum of the scores of the outgoing edges from the pair:

$$DEG(a,b) = \sum_{x \in N_a \setminus \{b\}} w(a,x) + \sum_{x \in N_b \setminus \{a\}} w(b,x)$$

  where $w(x, y)$ is the data-source score of edge (x, y), N_a is the set of all neighbours of a, excluding a.

- *NBC*: The neighborhood connectivity of the protein pair (a, b), defined as the weighted density of all neighbors of the protein pair excluding the pair themselves:

$$NBC(a,b) = \frac{\sum_{x,y \in N_{a,b}} w(x,y)}{min(|N_{a,b}|, \lambda)(min(|N_{a,b}|, \lambda) - 1)}$$

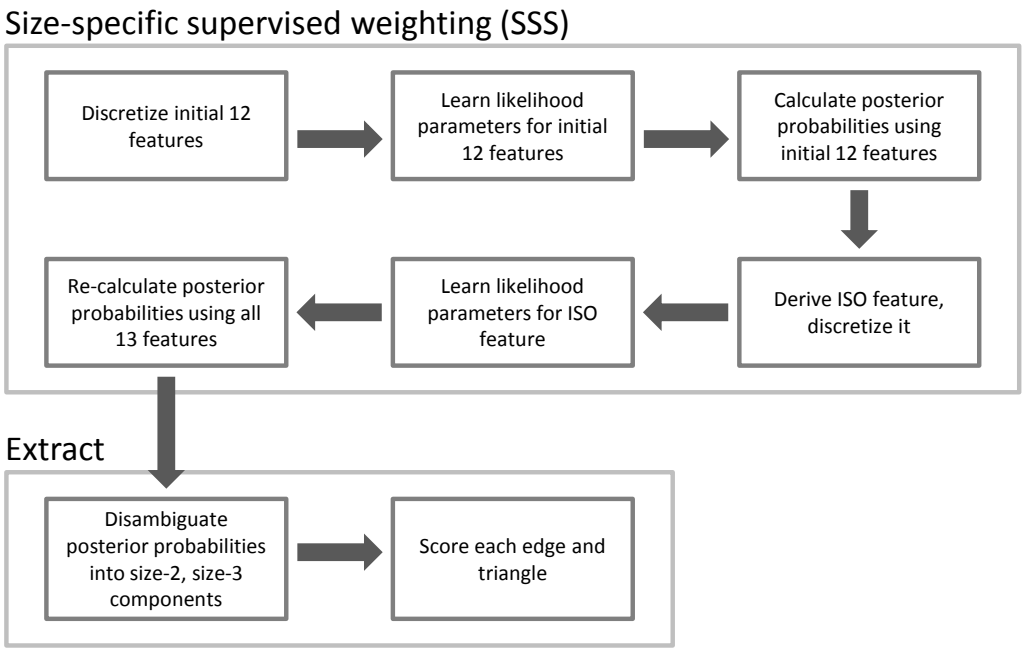

Figure 5.1: Flowchart of our approach, which consists of Size-Specific Supervised Weighting (SSS) and Extract.

where $w(x, y)$ is the data-source score of edge (x, y); $N_{a,b}$ is the set of all neighbours of a and b, excluding a and b themselves; λ is a dampening factor.

- *SHARED*: The extent of shared neighbors between the protein pair, derived using the Iterative AdjustCD function (with two iterations) [35], as described in Chapter 3.2.1.

This gives a total of twelve features: the three data sources PPI, $STRING$, and LIT, and nine topological features (three for each data source), DEG_{PPI}, DEG_{STRING}, DEG_{LIT}, $SHARED_{PPI}$, $SHARED_{STRING}$, $SHARED_{LIT}$, NBC_{PPI}, NBC_{STRING}, and NBC_{LIT}. In addition, a feature called isolatedness is incorporated after an initial calculation of the posterior probabilities, as described below.

Size-Specific Supervised Weighting (SSS)

In this step, we weight the edges of the PPI network with our Size-Specific Supervised Weighting (SSS) approach. We use a highly-reliable subset of the PPI network, by keeping only the top k edges with the highest PPI reliability scores. In our experiments we set $k = 10000$, but similar results are obtained for other values of k. SSS uses supervised learning to weight each edge with three scores: its posterior probability of being a small-co-complex edge (i.e. of belonging to a small complex), of being a large-co-complex edge, and of not being a co-complex edge, given the features of the edge.

These features consist of the twelve features described above (PPI, $STRING$, LIT, and nine topological features), as well as an isolatedness feature which is derived from an initial calculation of the posterior probabilities. We use a naive-Bayes maximum-likelihood model to derive the posterior probabilities. Each edge (a, b) is cast as a data instance, with its set of features $\mathbf{F}$. Using a reference set of protein complexes, each edge (a, b) in the training set is given a class label *lg-comp* if both a and b are in the same large complex; it is labelled *sm-comp* if both a and b are in the same small complex; otherwise it is labelled *non-comp*. Learning proceeds by the following steps (illustrated in Figure 5.1):

1. Minimum description length (MDL) supervised discretization [70] is performed to discretize the features (excluding the isolatedness feature). MDL discretization recursively partitions the range of each feature to minimize the information entropy of the classes. If a feature cannot be discretized, that means it is not possible to find a partition that reduces the information entropy, so the feature is removed. Thus this step also serves as simple feature selection.

2. The maximum-likelihood parameters are learned for the three classes *lg-comp*, *sm-comp*, and *non-comp*:

$$P(F = f|\textit{sm-comp}) = \frac{n_{sm,F=f}}{n_{sm}}$$

$$P(F = f|\textit{lg-comp}) = \frac{n_{lg,F=f}}{n_{lg}}$$

$$P(F = f|\textit{non-comp}) = \frac{n_{non,F=f}}{n_{non}}$$

   for each discretized value f of each feature F (excluding the isolatedness feature). n_{sm} is the number of edges with class label *sm-comp*, $n_{sm,F=f}$ is the number of edges with class label *sm-comp* and whose feature F has value f; n_{lg} is the number of edges with class label *lg-comp*, $n_{lg,F=f}$ is the number of edges with class label *lg-comp* and whose feature F has value f; n_{non} is the number of edges with class label *non-comp*, and $n_{non,F=f}$ is the number of edges with class label *non-comp* and whose feature F has value f.

3. Using the learned models, the class posterior probabilities are calculated for each edge (a, b) using the naive-Bayes formulation:

$$
\begin{aligned}
& P((a,b)\ is\ sm\text{-}comp | F_1 = f_1, F_2 = f_2, \ldots) \\
= \quad & \frac{\prod_i P(F_i = f_i | (a,b)\ is\ sm\text{-}comp) P(sm\text{-}comp)}{\sum_{class \in \{sm\text{-}comp, lg\text{-}comp, non\text{-}comp\}} \prod_i P(F_i = f_i | (a,b)\ is\ class) P(class)}
\end{aligned}
$$

The posterior probabilities are calculated in a similar fashion for the other two classes *lg-comp* and *non-comp*. We abbreviate the posterior probability of edge (a, b) being in each of the three classes as $P_{(a,b),sm}$, $P_{(a,b),lg}$, and $P_{(a,b),non}$.

4. A new feature ISO (isolatedness) is calculated for each edge (a, b), based on the probability that the edge is isolated (not adjacent to any other edges), or is part of an isolated triangle:

$$ISO(a,b) = ISO2(a,b) + ISO3(a,b)$$

$$ISO2(a,b) = P_{(a,b),sm} \prod_{x \in \{a,b\}, y \in N_{a,b}} P_{(x,y),non}$$

$$ISO3(a,b) = \sum_{c \in N_a \cap N_b} \left(P_{(a,b),sm} P_{(a,c),sm} P_{(b,c),sm} \prod_{x \in \{a,b,c\}, y \in N_{a,b,c}} P_{(x,y),non} \right)$$

where N_x denotes the neighbours of x, excluding x. The ISO feature is discretized with MDL.

5. The maximum-likelihood parameters for the ISO feature are learned for the three classes.

6. The posterior probabilities for the three classes, $P_{(a,b),sm}$, $P_{(a,b),lg}$, and $P_{(a,b),non}$, are recalculated for each edge (a, b), this time incorporating the new ISO feature.

5.2.2 Extracting small complexes

After using SSS to weight the PPI network, the small complexes are extracted. This stage, called Extract, consists of two steps (see Figure 5.1): first, the small-co-complex probability weight of each edge is disambiguated into size-2 and size-3 complex components; next, each candidate complex is scored by its cohesiveness-weighted density, which is based on both its internal and outgoing edges.

In the disambiguation step, the small-co-complex probability weight of each edge $(a, b) = P_{(a,b),sm}$, which denotes the probability of being in a small (either size-2 or

size-3) complex, is decomposed into two component scores (we use the term score instead of probability since its derivation is not probabilistic): $P'_{(a,b),sm2}$, which is the score of being in the size-2 complex composed of a and b; and $P'_{(a,b),sm3,abc}$, which is the score of being in the size-3 complex composed of a, b, and c. Intuitively, if an edge is contained within a triangle with high edge weights, then it is more likely to be a size-3 complex corresponding to the triangle rather than a size-2 complex; thus its size-2 component score should be reduced based on the weights of incident triangles:

$$P'_{(a,b),sm2} = P_{(a,b),sm} - \sum_{x \in N_a \cap N_b} P_{(a,b),sm} P_{(a,x),sm} P_{(b,x),sm}$$

Similarly, if an edge is contained within a triangle with high edge weights, and is also within another triangle with low edge weights, then it is more likely to form a size-3 complex with the former triangle rather than the latter; thus its size-3 component score corresponding to a specific triangle should be reduced based on the weights of its other incident triangles:

$$P'_{(a,b),sm3,abc} = P_{(a,b),sm} - \sum_{x \in N_a \cap N_b \setminus \{c\}} P_{(a,b),sm} P_{(a,x),sm} P_{(b,x),sm}$$

In the next step, each candidate complex is scored by weighting the density of the cluster with its cohesiveness, which is adapted from cluster cohesiveness as described in [41]. Here, we define cohesiveness of a cluster as the ratio of the sum of its internal edges' weights over its internal plus outgoing edges' weights, where the internal weights are the component scores as calculated above, and the external weights are the posterior probabilities of being either small or large co-complex edges. The cohesiveness of a size-2 cluster (a, b) and a size-3 cluster (a, b, c) respectively are:

$$Coh(a,b) = \frac{P'_{(a,b),sm2}}{P'_{(a,b),sm2} + \sum_{x \in \{a,b\}, y \in N_{a,b}} (P_{(x,y),sm} + P_{(x,y),lg})}$$

$$Coh(a,b,c) = \frac{P'_{(a,b),sm3,abc} + P'_{(a,c),sm3,abc} + P'_{(b,c),sm3,abc}}{P'_{(a,b),sm3,abc} + P'_{(a,c),sm3,abc} + P'_{(b,c),sm3,abc} + \sum_{x \in \{a,b,c\}, y \in N_{a,b,c}} (P_{(x,y),sm} + P_{(x,y),lg})}$$

We then define the score of a cluster as its cohesiveness-weighted density, or the product of its weighted density and its cohesiveness. The score of a size-2 cluster (a, b), and a size-3 cluster (a, b, c) respectively are:

Clustering algorithm	Parameters
CMC	overlap_thres=1, merge_thres=1
ClusterONE	*all default*
IPCA	-P1 -T0.4
MCL	-I 2
RNSC	-e10 -D50 -d10 -t20 -T3
PPSampler2	–f1DenominatorExponent 1 –f2

Table 5.1: The six clustering algorithms and their parameters used for small-complex discovery.

$$score(a,b) = Coh(a,b)P'_{(a,b),sm2}$$

$$score(a,b,c) = Coh(a,b,c)\frac{(P'_{(a,b),sm3,abc} + P'_{(a,c),sm3,abc} + P'_{(b,c),sm3,abc})}{3}$$

5.3 Results and discussion

5.3.1 Experimental setup

In our main experiments, we compare our two-stage approach (weighting with SSS, small-complex extraction with Extract) against using the original PPI reliability (PPIREL) weighted network with the following clustering approaches to derive small complexes: MCL, RNSC, IPCA, CMC, ClusterONE, and PPSampler (described in Chapter 2.4). Any predicted complex with size greater than three is discarded. We run these algorithms with a range of values for their respective parameters, and select the settings that give the optimal performance for predicting small complexes. The parameter settings used in our experiments are given in Table 5.1.

We also investigate the performance of using our SSS-weighted network with standard clustering approaches, and using the PPIREL network with our Extract approach.

We perform random sub-sampling cross-validation, repeated over ten rounds, using manually-curated complexes as reference complexes for training and testing. For yeast, we use the CYC2008 [56] set which consists of 408 complexes, of which 259 are small (composed of two or three proteins). For human, we use the CORUM [57] set (filtered to remove duplicates and small complexes that are subsets of large ones), which consists of 1352 complexes, of which 701 are small. In each cross-validation round, t% of the complexes (large and small) are selected for testing, while all the remaining complexes are used for training. Each edge (a, b) in the network is given a class label *lg-comp* if a and b are in the same large training complex; otherwise it is labeled *sm-comp* if a and b are in the same small training complex; otherwise its class label is *non-comp*. Learning in SSS is performed using these labels, and the edges of the network are weighted using

the learned models. Small complexes are then extracted from the weighted network. The predicted complexes are evaluated by matching them with only the small test complexes.

We design our experiments to simulate a real-use scenario of complex prediction in an organism where a few complexes might already be known, and novel complexes are to be predicted: in each round of cross-validation, the training complexes are those that are known and leveraged for learning to discover new complexes, while the test complexes are used to evaluate the performance of each approach at this task. Thus we use a large percentage of test complexes $t\% = 90\%$. In yeast, this gives about 233 small test complexes and 26 small training complexes per cross-validation iteration; in human, this gives about 631 small test complexes and 70 small training complexes.

5.3.2 Evaluation methods

We use precision-recall graphs to evaluate the predicted clusters, by matching the generated clusters with the reference test complexes, and calculating recall (sensitivity) and precision. We require a generated cluster to be identical to a complex to be considered a correct match. Each cluster P is ranked by its score, which is either the cohesiveness-weighted density (for Extract), or weighted density (for other clustering algorithms). To obtain a precision-recall graph, we calculate and plot the precision and recall of the predicted clusters in matching the test complexes, at various cluster-score thresholds (as described in Chapter 3.3.2). As a summarizing statistic of a precision-recall graph, we also calculate the area under the curve (AUC) of the precision-recall graph.

To measure the quality of a predicted complex, we derive the semantic coherence of its Gene Ontology (GO [6]) annotations across the three GO classes, biological process (BP), cellular compartment (CC), and molecular function (MF). First, we derive the BP semantic similarity between two proteins as the information content of their BP annotations' most informative common ancestor [72]. Then we define the BP semantic coherence of a predicted complex as the average BP semantic similarity between every pair of proteins in that complex (likewise for CC and MF).

5.3.3 Prediction of small complexes

In this section we compare the performance of small-complex prediction using our weighting approach (SSS) versus PPI reliability (PPIREL), and using our complex-extraction algorithm (Extract) versus other clustering algorithms (CMC, ClusterOne,

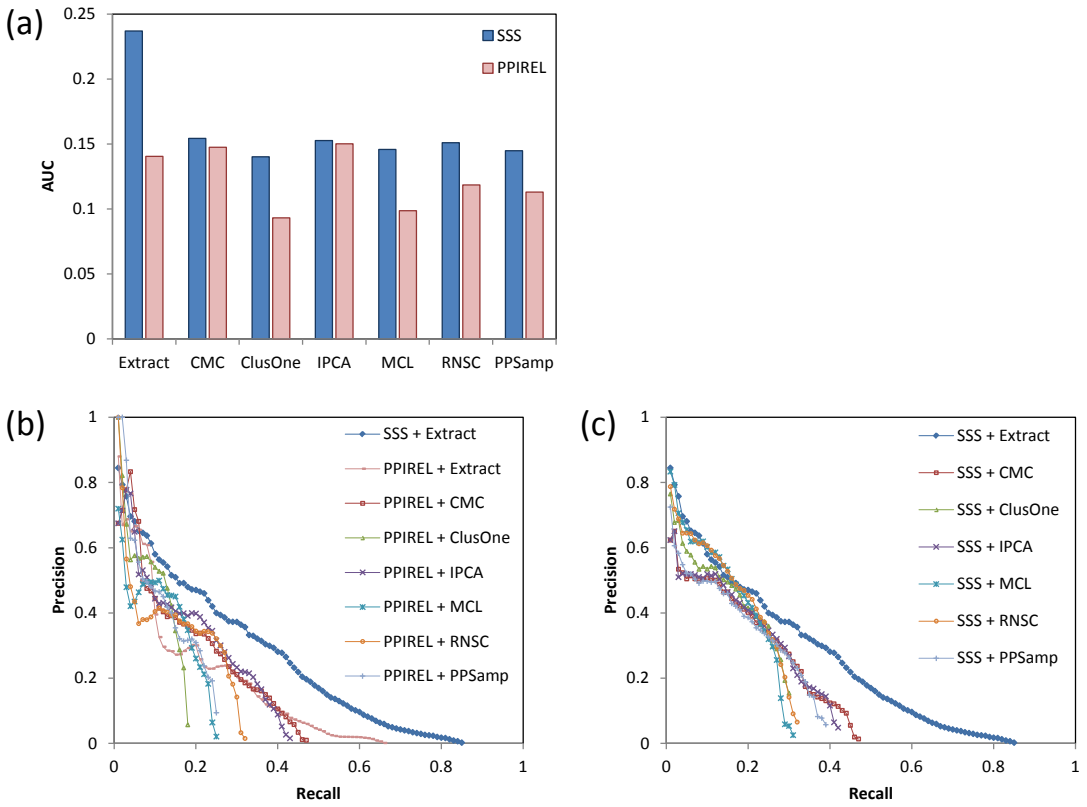

Figure 5.2: Performance of small-complex prediction in yeast, (a) precision-recall AUC, (b) and (c) precision-recall graphs.

IPCA, MCL, RNSC, PPSampler2). Figure 5.2a shows the performance of prediction of yeast small complexes, in terms of precision-recall AUC. Our 2-stage approach (SSS + Extract) outperforms all other approaches tested here, including using the PPIREL or SSS-weighted networks with standard clustering algorithms, or the PPIREL-weighted network with Extract. Furthermore, when using standard clustering algorithms to discover small complexes, weighting the network with SSS gives improved performance compared to using PPIREL (especially for ClusterOne, MCL, RNSC, and PPSampler2).

Figure 5.2b shows the precision-recall graphs comparing our approach (SSS + Extract) to the baselines of standard clustering algorithms applied on a PPIREL network. While our approach has lower precision among the initial top predictions (at recall less than 5%), beyond that we attain substantially greater precision: for example, at 40% recall, our approach attains more than three times the precision than the other clustering approaches (28% versus 9%). Furthermore, we attain substantially higher recall as well. Figure 5.2c shows the precision-recall graphs when the standard clustering algorithms are applied on the SSS-weighted network. Using the SSS-weighted network, most of the clustering algorithms achieve improved precision in the mid-recall ranges,

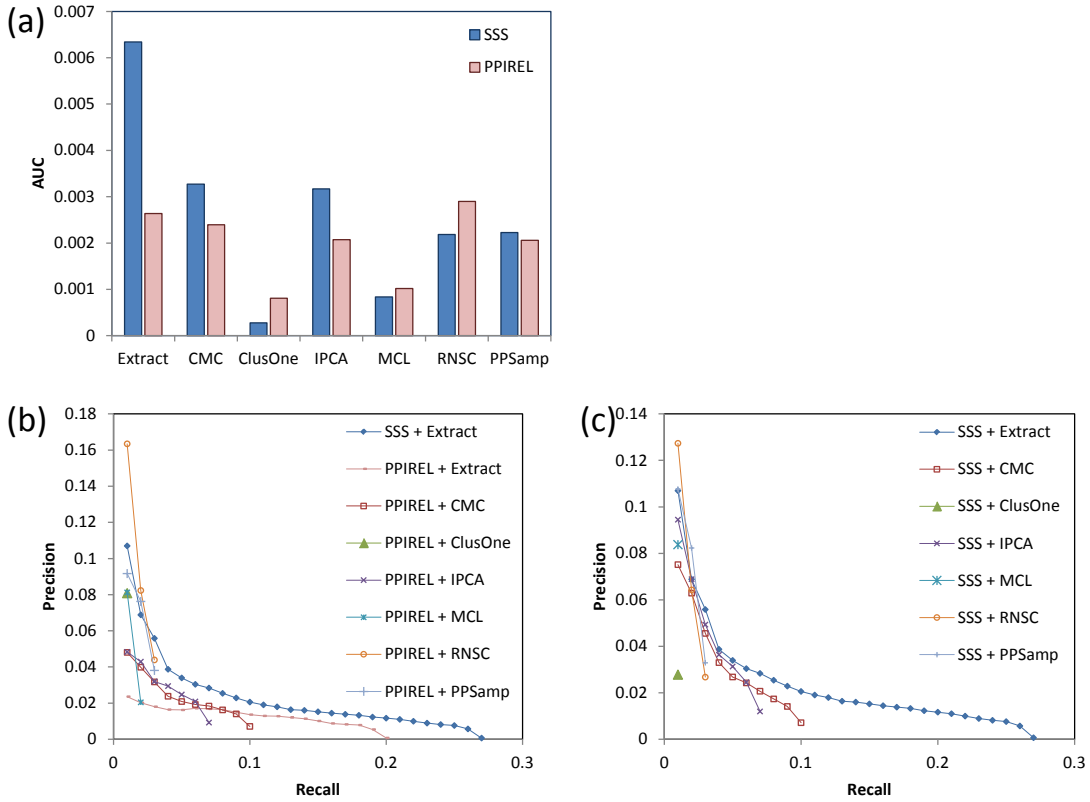

Figure 5.3: Performance of small-complex prediction in human, (a) precision-recall AUC, (b) and (c) precision-recall graphs.

as well as gains in recall. However, our approach (SSS + Extract) still maintains greater precision in most of the recall range.

Figure 5.3 shows the performance of prediction of human small complexes. The prediction of complexes in human is much more challenging than in yeast, so the AUCs achieved here are correspondingly lower. Nonetheless, our approach (SSS + Extract) still outperforms all the other approaches, including using the PPIREL or SSS-weighted networks with standard clustering algorithms, or the PPIREL-weighted network with Extract. When using standard clustering algorithms to discover small complexes, weighting the network with SSS gives improved performance only for CMC and IPCA, while performance remains the same or decreases for the other clustering algorithms.

Figure 5.3b and c show the corresponding precision-recall graphs. As in yeast, our approach (SSS + Extract) outperforms the standard clustering algorithms applied on the PPIREL-weighted network by achieving substantially higher recall, as well as greater precision in almost the whole recall range (Figure 5.3b). Using the SSS- instead of the PPIREL-weighted network, CMC and IPCA achieve higher precision, while the other clustering algorithms suffer from lower precision or recall (Figure 5.3c).

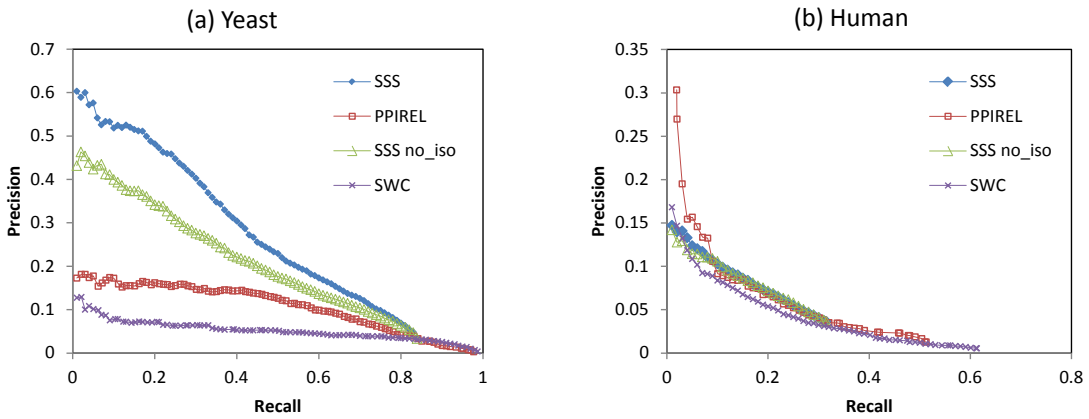

Figure 5.4: Performance of classification of small-complex edges, in (a) yeast, (b) human.

In the following section we investigate how the various techniques incorporated in SSS and Extract improve the performance of small complex prediction.

5.3.4 How do SSS and Extract improve performance?

Figures 5.2 and 5.3 showed that weighting the network with SSS improves yeast small-complex prediction in four of six clustering algorithms, while it only improves human complex prediction in two clustering algorithms. To investigate the benefits of SSS weighting, we compare the performance of the weighting approaches in *classifying* edges as belonging to small complexes. Each weighting approach is used to weight the edges of the network, and the precision-recall graph is obtained by varying a threshold on the edge weights. Figure 5.4a shows the precision-recall graph for classification of yeast small-complex edges. SSS achieves much higher precision than classifying by PPIREL, as the SSS weights more accurately reflect membership in small complexes. This leads to improved performance by clustering algorithms when applied to the SSS-weighted network to predict small yeast complexes. On the other hand, when classifying edges in small human complexes, Figure 5.4b shows that SSS has lower precision than PPIREL at the lower recall range, with only similar or marginally better precision at higher recall ranges. Thus, only two clustering algorithms obtain improved performance from clustering the SSS-weighted network.

Figure 5.4 also shows the poor performance of the previously-proposed supervised weighting approach SWC [4], which learns a model for all co-complex edges in general, as opposed to distinct models for small and large complexes. As the number of edges in a complex grows quadratically with its number of proteins, the edges from large complexes far outnumber those from small complexes, so SWC's learned model reflects the characteristics of large complexes. Thus, SWC suffers from poor performance in

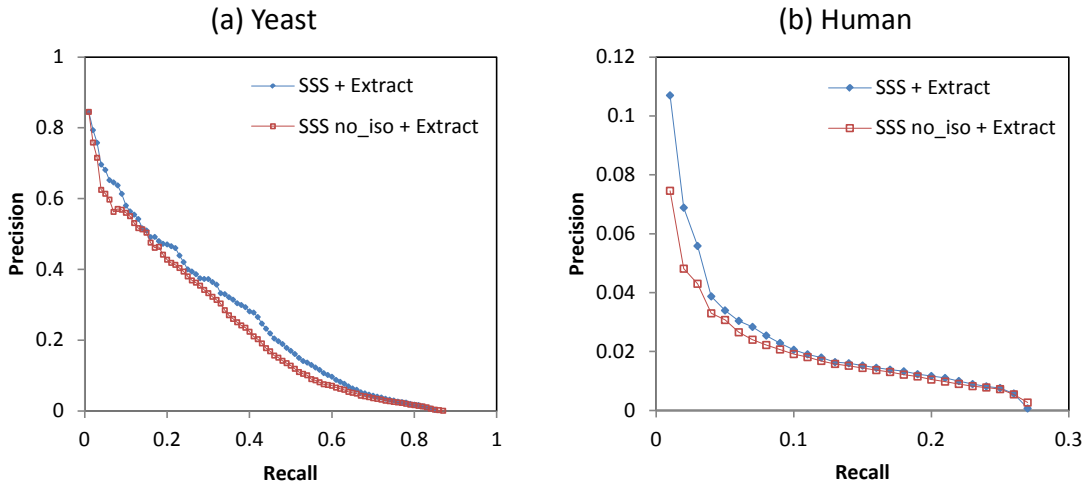

Figure 5.5: Performance of small-complex prediction with and without isolatedness feature in SSS, in (a) yeast, (b) human.

classifying edges from small complexes, demonstrating the importance of the size-specific modeling of SSS.

The *SSSno_iso* graph in Figure 5.4 shows that if the isolatedness feature is not incorporated into SSS (in other words, steps 4 to 6 of SSS are skipped), precision drops in yeast, showing the utility of the isolatedness function in predicting small complex edges. However, in human, incorporating the isolatedness feature gives only marginal improvement in precision. Figure 5.5 shows the performance of small-complex prediction, when SSS is used with and without the isolatedness feature, with the complexes derived by Extract. Incorporating isolatedness gives a noticeable boost to precision in both yeast and human, demonstrating that isolatedness benefits the prediction of small complexes by improving the SSS weighting of edges.

Next, we investigate the effect of cohesiveness weighting in Extract, applied on the SSS network versus the PPIREL network. Figure 5.6a shows the performance of the clustering algorithms applied on the SSS network, with and without scoring by cohesiveness weighting, for predicting yeast small complexes. For Extract (where cohesiveness weighting is used by default), scoring without cohesiveness weighting means a cluster's score is its weighted density. For the other clustering algorithms (where weighted density is used by default), scoring with cohesiveness weighting means a cluster's score is the product of its weighted density and its cohesiveness (ratio of sum of internal edges over internal and outgoing edges). With the SSS network, scoring by cohesiveness weighting improves performance across all clustering algorithms. On the other hand, Figure 5.6b shows that, with the PPIREL network, scoring by cohesiveness weighting decreases performance across most clustering algorithms. Thus,

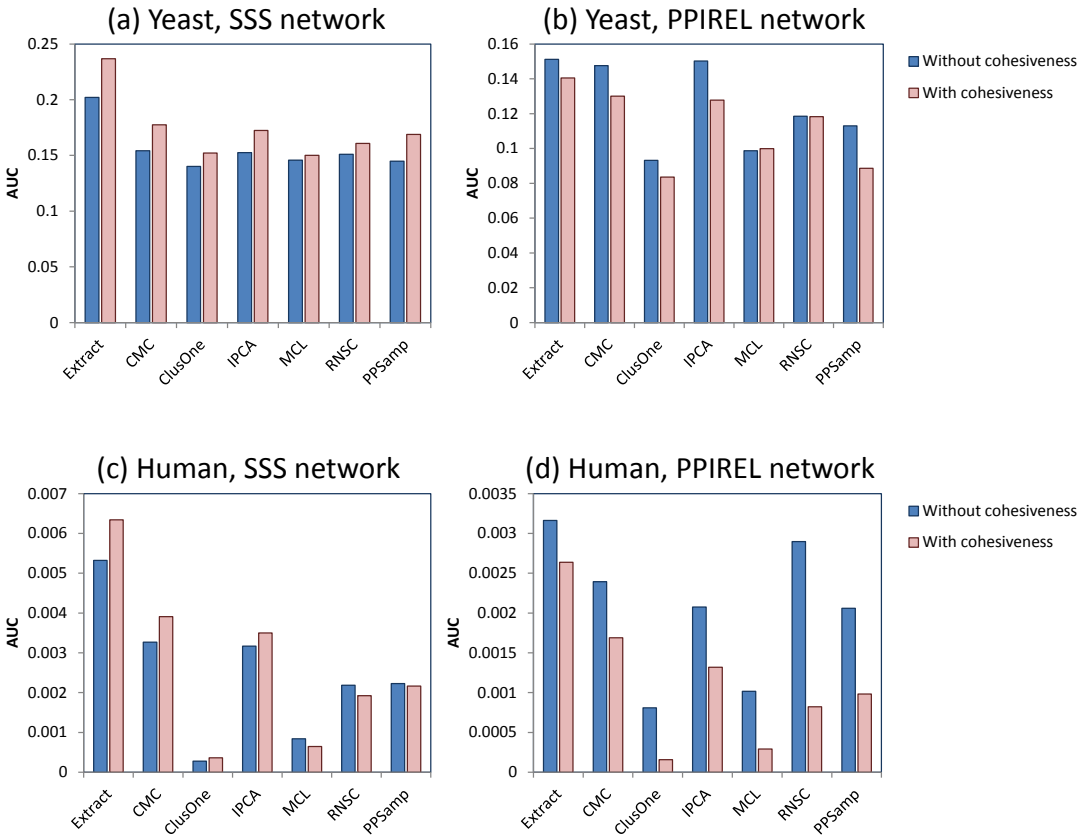

Figure 5.6: Performance of small-complex prediction with and without cohesiveness weighting for scoring clusters, for (a) SSS network in yeast, (b) PPIREL network in yeast, (c) SSS network in human, (d) PPIREL network in human.

cohesiveness weighting appears useful only when edges are weighted using SSS.

Figure 5.6c and d show the corresponding charts for human complexes, with and without cohesiveness weighting. With the SSS network, cohesiveness weighting improves performance in four of seven clustering algorithms; whereas with the PPIREL network, cohesiveness weighting decreases performance in all clustering algorithms. Thus, in human complexes as well, cohesiveness weighting appears useful only when edges are weighted using SSS.

5.3.5 Example complexes

In this section we present some example complexes that are difficult to predict using the PPIREL network with any standard clustering algorithm, but can be predicted with our approach (SSS + Extract). Since the various clustering approaches output different numbers of predictions, we consider only the top-scoring predicted clusters with a cross-validation precision level greater than some threshold. For yeast we use a precision threshold of 10%, but for human we use a lower precision threshold of 2%, since fewer human complexes are predicted with high precision.

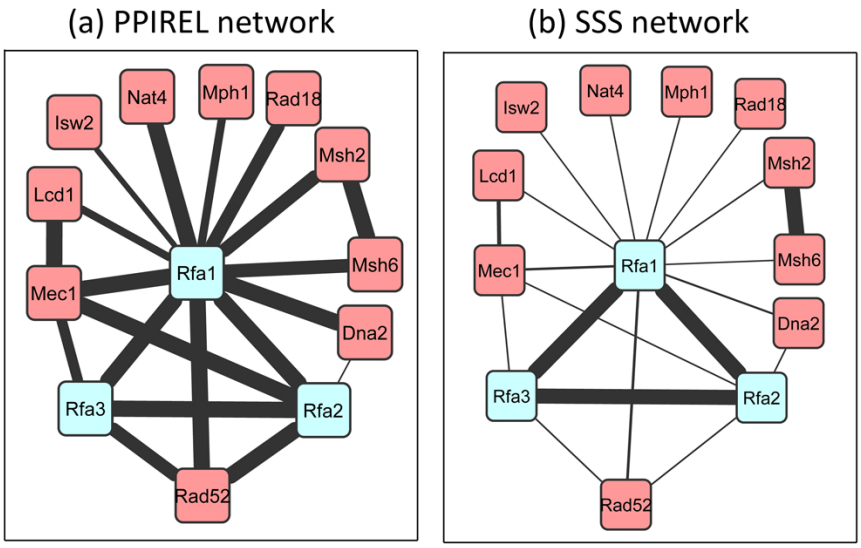

Figure 5.7: DNA replication factor A complex in yeast, in (a) PPIREL network, (b) SSS network.

The DNA replication factor A complex in yeast consists of three proteins, Rfa1p, Rfa2p, and Rfa3p. Figure 5.7a shows the PPIREL network around this complex, with edge widths scaled to PPI reliability scores. The complex is embedded within two size-4 cliques (with Rad52p, and Mec1p), with high PPIREL weights. Moreover, Rfa1p is also connected via high PPIREL weights to many external proteins, some of which form size-3 cliques as well. As a result, none of the standard clustering algorithms applied on the PPIREL network predicted this complex, in any cross-validation round. Figure 5.7b shows the SSS network, with edge widths scaled to the small co-complex posterior probability scores. The three proteins in the complex remain interconnected with high edge weights, while the extraneous edges' weights are now markedly lowered. Thus, our Extract algorithm is able to retrieve this complex from the SSS network consistently across all cross-validation rounds where it is tested.

Figure 5.8 shows two yeast complexes, with an overlapping protein (Sir2p), involved in transcriptional silencing: the chromatin silencing complex, consisting of Sir2p, Sir3p, and Sir4p, and the RENT complex, consisting of Sir2p, Cdc14p, and Net1p. In the PPIREL network (Figure 5.8a), each of the two complexes are connected via highly-weighted extraneous edges to many external proteins. Once again, none of the standard clustering algorithms applied on the PPIREL network could predict either of these complexes, in any cross-validation round. In the SSS network (Figure 5.8b), the chromatin silencing complex remains connected with high edge weights, with a marked reduction in the weights of the extraneous edges. Thus our Extract algorithm retrieves this complex from the SSS network consistently across all cross-validation

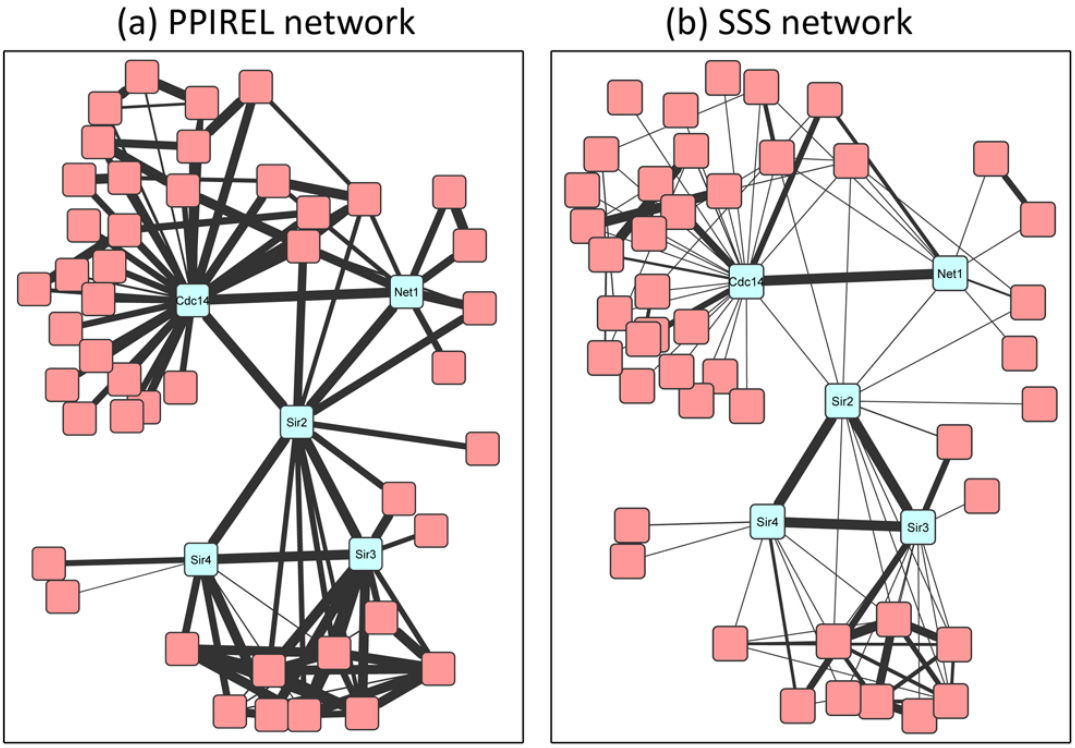

Figure 5.8: Chromatin silencing complex and RENT complex in yeast, in (a) PPIREL network, (b) SSS network.

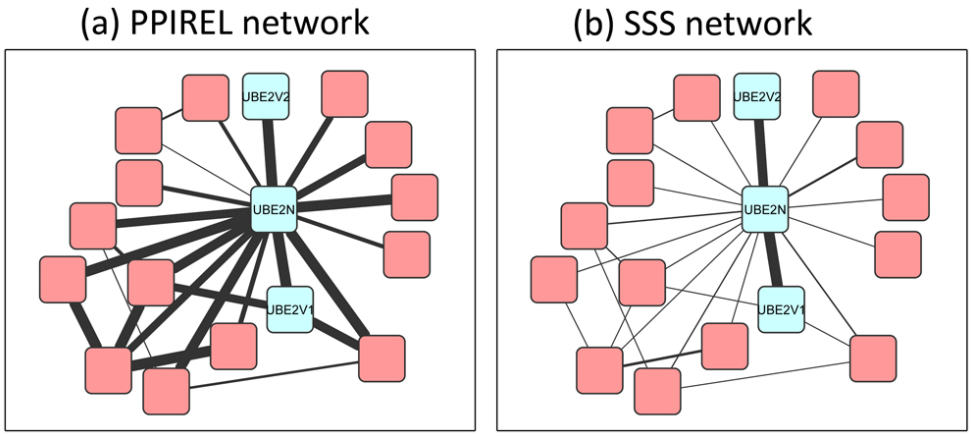

Figure 5.9: Two human ubiquitin ligase complexes, in (a) PPIREL network, (b) SSS network.

rounds where it is tested. On the other hand, in the RENT complex, the weights of two edges (from Sir2p to the other two proteins) are now even lower than some of its extraneous edges. As a result, our Extract algorithm retrieves this complex only 33% of the time. Nonetheless, this is still an improvement over using the PPIREL network with standard clustering algorithms.

Figure 5.9 shows two human ubiquitin ligase heterodimer complexes with an overlapping protein: the UBE2V1-UBE2N and UBE2V2-UBE2N complexes. In the PPIREL network (Figure 5.9a), UBE2N is connected via highly-weighted edges to

many other external proteins, forming a number of size-3 cliques with them. The UBE2V1-UBE2N complex is embedded within two size-3 cliques, making it difficult to discover: none of the standard clustering algorithms predicted this complex in any cross-validation round. On the other hand, the UBE2V2-UBE2N complex is relatively isolated as UBE2V2 is not connected to any other external protein, allowing CMC and IPCA to predict this complex consistently (none of the clustering algorithms could do so). In our SSS network (Figure 5.9b), all extraneous edges' weights have been dramatically lowered, leaving the co-complex edges with high weights. Thus our Extract algorithm retrieved UBE2V1-UBE2N 78% of the time, and UBE2V2-UBE2N 100% of the time.

5.3.6 Quality of novel complexes

In this section we compare the number and quality of high-confidence novel complexes predicted by our approach (SSS with Extract), against using standard clustering algorithms on the PPI reliability network. When weighing the network with SSS, the entire set of reference complexes is used for training. We filter the predicted complexes to remove those that match any reference complex, and to keep only high-confidence predictions: the score of each predicted complex is mapped to a precision value, using the cross-validation results, and only predicted complexes with estimated precision greater than a confidence threshold are kept. For yeast, this confidence threshold is 0.5; for human, a lower threshold of 0.1 is used, since much fewer complexes are predicted with high precision.

Figure 5.10a shows the number of high-confidence novel complexes predicted in yeast, and their average BP, CC, and MF semantic coherence, using the different approaches. Compared to the other approaches, SSS with Extract generates more than twice as many high-confidence novel predictions, with equal or greater quality: our predicted complexes have greater coherence than ClusterOne, MCL, or PPSampler ($p < .05$ in at least one of BP, CC, or MF), and similar coherence with the other approaches. The CYC2008 reference complexes have much higher BP and CC coherence, but lower MF coherence.

Figure 5.10b shows the corresponding charts for human predictions. Again, our approach generates more high-confidence novel predictions than the other approaches, with equal or greater quality: our predicted complexes have greater coherence than ClusterOne, MCL, RNSC, or PPSampler ($p < .05$ in at least one of BP, CC, or MF), and similar coherence with the other approaches. Our predicted complexes have similar

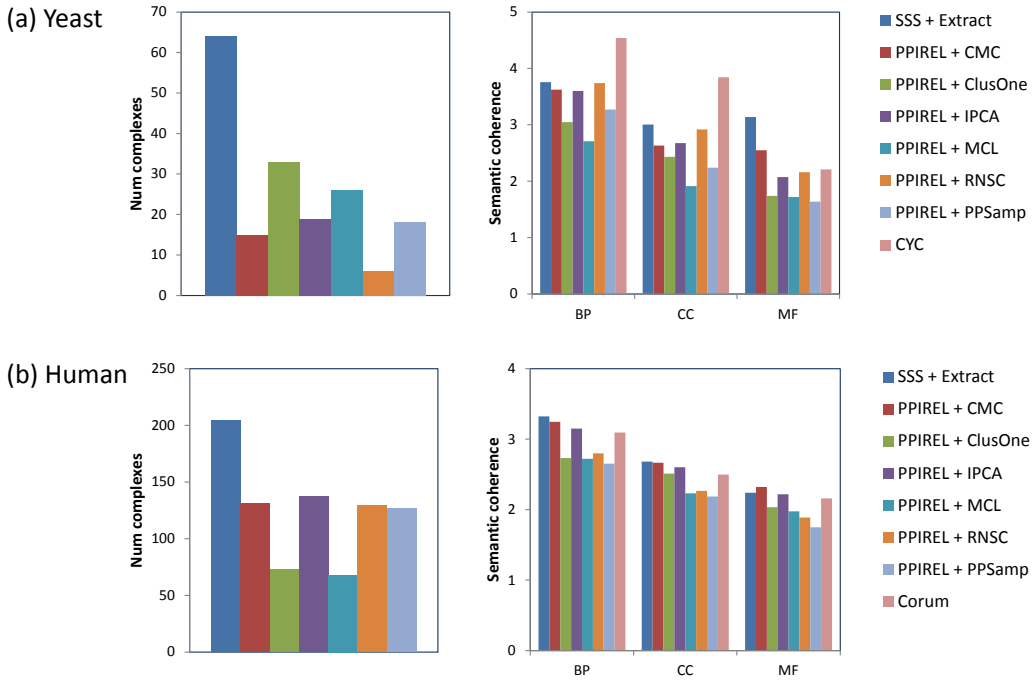

Figure 5.10: Number of high-confidence novel predictions, and their semantic coherences, in (a) yeast, (b) human.

semantic coherence compared to the Corum reference complexes.

Finally, we briefly mention two novel complexes, predicted by our approach, that we have validated via a literature scan. Our approach predicts a high-scoring yeast cluster consisting of Cap1p and Cap2p, which is not found in our reference database of complexes. However, a literature scan revealed this to be the capping protein heterodimer, which binds to actin filaments to control filament growth [79]. Our approach also predicts a novel high-scoring human cluster consisting of PKD1 and PKD2. A literature scan revealed that these two proteins, which are involved in autosomal polycystic kidney disease, have been found to form a PKD1-PKD2 heterodimer [80].

5.4 Conclusion

The size of protein complexes has been noted to follow a power distribution, meaning that a large majority of complexes are small (consisting of two or three distinct proteins). Thus the discovery of small complexes is an important subtask in protein-complex prediction. Predicting small complexes from PPI networks is inherently challenging. Small groups of proteins are more likely to take on topological characteristics of real complexes by chance: for example, fully-dense groups of two or three proteins correspond to edges or triangles respectively, but only a few of these actually correspond to small complexes. Furthermore, the prediction of small complexes is especially

susceptible to noise (missing or spurious interactions) in the PPI network, as these can easily disconnect a small complex, or embed it within a larger clique.

We propose a two-stage approach, SSS and Extract, for discovering small complexes. First, the PPI network is weighted by Size-Specific Supervised Weighting (SSS), which integrates heterogeneous data and their topological features with an overall topological-isolatedness feature, and uses a naive-Bayes maximum-likelihood model to weight the edges with their posterior probabilities of being in a small complex, and in a large complex. Integrating other data sources into the PPI network can help reduce noise, while incorporating the topological features across multiple data sources makes it less likely that random protein groups take on topological characteristics of complexes by chance.

In our second stage, Extract, the SSS-weighted network is analyzed to extract putative small complexes and score them by cohesiveness-weighted density, which incorporates both small-co-complex and large-co-complex weights of internal and outgoing edges. This reduces the impact of noisy edge weights in deriving reliable scores for predictions, as more edge weights around the candidate complex are utilized.

While a few previous approaches have used supervised learning to weight PPI edges, none of them have done so in a complex-size-specific manner, or incorporated isolatedness as a feature in this way. Our adaptation of cohesiveness to address the problem of the small number of edge weights available in scoring small complexes is also novel.

We test our approach on the prediction of yeast and human small complexes, and demonstrate that our approach outperforms some commonly-used clustering algorithms applied on a PPI reliability network, attaining higher precision and recall. Furthermore, our approach generates a greater number of novel predictions with higher quality in terms of Gene Ontology semantic coherence.

Nonetheless, there is still room for further work to improve the prediction of small complexes, as its performance still lags behind that of predicting large complexes, especially for human complexes. A possible future direction is to adapt other techniques that have proved useful for large-complex prediction, such as GO term decomposition and hub removal [5], which might further improve the performance of small-complex prediction.

Chapter 6

Integration of three approaches

6.1 Introduction

In the previous chapters we described three challenges in complex prediction that arise from, or are exacerbated by, a static view of PPIs and protein complexes which are in fact dynamic in nature. First, many complexes are sparsely connected in the PPI network, and cannot be picked out by clustering algorithms which search for dense subgraphs. Second, many complexes are embedded within highly-connected regions of the PPI network with many extraneous edges connecting them to external proteins, so that clustering algorithms cannot properly delimit their boundaries. Third, many complexes are small (that is, composed of two or three proteins), making measures of important topological features, such as density, ineffectual. We proposed three approaches that can help to address these problems.

First, Supervised Weighting of Composite Networks (SWC [4]), described in Chapter 3, addresses the problem of sparse complexes. SWC integrates PPI data with two additional data sources, functional associations and co-occurrence in literature, and uses a supervised approach to weight edges with their posterior probabilities of belonging to a complex. SWC fills in the missing edges in many sparse complexes through data integration, and reduces the amount of spurious non-co-complex edges through supervised weighting. Using this approach, improvements are obtained in both precision and recall for yeast and human complex discovery, especially among the sparse complexes.

Second, decomposing PPI network into spatially- and temporally-coherent subnetworks (abbreviated as DECOMP here [5]), described in Chapter 4, addresses the problem of complexes in highly-connected regions with many extraneous edges. DECOMP removes hub proteins with large numbers of interaction partners, as they tend to correspond to date hubs with non-simultaneous interactions. Next, it decomposes

the PPI network into spatially-coherent subnetworks using cellular-location Gene Ontology terms [6]. By splitting dense regions of the PPI network into less-dense but coherent subnetworks, complex-discovery performance is improved, with the biggest improvements among complexes in highly-connected regions.

Third, Size-Specific Supervised Weighting (SSS [7]) addresses the problem of predicting small complexes. SSS integrates PPI data with two additional data sources, functional associations and co-occurrence in literature, along with their topological features, and uses a supervised approach to weight edges with their posterior probabilities of belonging to small complexes versus large complexes. SSS then extracts small complexes from the weighted network, and scores them using the probabilistic weights of edges within, as well as surrounding, the complexes. This approach achieves significant improvements in precision and recall in discovering small complexes.

Although SWC and DECOMP both improve the prediction of large complexes in general, they have been shown to give the largest improvements among the complexes that they are designed for: sparse complexes for SWC, and complexes embedded in dense regions for PPI decomposition. The third technique, SSS, targets another separate group of complexes, the small complexes. Thus, we combine these three techniques into a single system that targets all three groups of challenging complexes, as this is likely to give a performance boost in complex discovery over using any single one of these techniques. In the integrated system, we also further modify DECOMP to incorporate the strategy of combining clusters derived from multiple clustering algorithms, using a simple voting scheme. This technique was used in SWC and found to improve complex-discovery performance (see Chapter 3), so it is likely to be beneficial in DECOMP as well.

6.2 Methods

In this section we describe how we integrate our three techniques, Supervised Weighting of Composite Networks (SWC), PPI decomposition (DECOMP), and Size-Specific Supervised Weighting (SSS), into a single system. We first describe the data sources and clustering algorithms used, then describe the integrated system.

6.2.1 Data sources and features

Table 6.1 lists the data features used in each of our three approaches. These features are derived from three different data sources (PPI, functional association, and literature co-occurrence), and their topological characteristics. Each data source provides a list

Feature	Description	SWC	DECOMP	SSS
PPI	PPI reliability	✓	✓	✓
$STRING$	Functional association	✓		✓
LIT	Literature co-occurrence	✓		✓
DEG_{PPI}	PPI topological degree			✓
DEG_{STRING}	STRING topological degree			✓
DEG_{LIT}	LIT topological degree			✓
$SHARED_{PPI}$	PPI shared neighbours	✓		✓
$SHARED_{STRING}$	STRING shared neighbours			✓
$SHARED_{LIT}$	LIT shared neighbours			✓
NBC_{PPI}	PPI neighbourhood connectivity			✓
NBC_{STRING}	STRING neighbourhood connectivity			✓
NBC_{LIT}	LIT neighbourhood connectivity			✓
ISO	Isolatedness			✓

Table 6.1: Data used for our three approaches.

of scored protein pairs: for each pair of proteins (a, b) with score s, a is related to b with score s, according to that data source. For both yeast and human, the following data sources are used:

- *PPI*: PPI data obtained by combining physical interactions from multiple databases, then scored by reliability, as described in Chapter 2.5.1.
- *STRING*: Predicted functional-association data obtained from the STRING database, as described in Chapter 3.2.1.
- *LIT*: Co-occurrence of proteins or genes in PubMed literature, as described in Chapter 3.2.1.

For each protein pair in each data source, we derive three topological features—degree (DEG), shared neighbors (SHARED), and neighborhood connectivity (NBC)—as described in Chapter 5.2.1. The final topological feature, isolatedness (*ISO(a,b)*), represents the probability that the protein pair (a, b) is in a size-2 or size-3 clique which is isolated from the rest of the network, as described in Chapter 5.2.1.

6.2.2 Clustering algorithms

We use the following clustering algorithms in our approach: MCL, RNSC, IPCA, CMC, ClusterONE, and COACH (described in Chapter 2.4).

In Chapter 3, SWC used a simple voting-based aggregative strategy, called COMBINE, to take the union of the clusters produced by the clustering algorithms. Here we use the COMBINE strategy for DECOMP as well. If two or more clusters are found to be similar to each other, then only the cluster with the highest weighted density is kept, and its score is defined as its weighted density multiplied by the number of algorithms that produced the group of similar clusters; otherwise its score is its weighted density

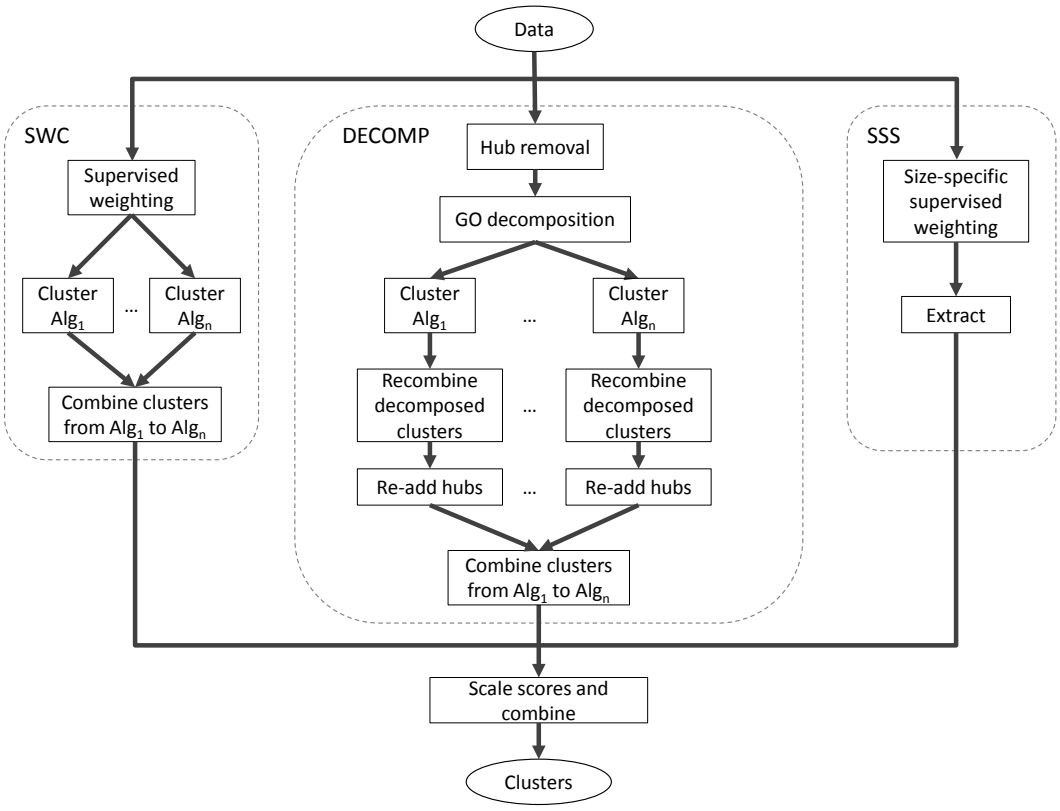

Figure 6.1: Flowchart of our integrated system consisting of Supervised Weighting of Composite Networks (SWC), PPI decomposition, and Size-Specific Supervised Weighting (SSS).

as usual. We define two clusters C and D to be similar if $Jaccard(C, D) \geq 0.75$, where $Jaccard(C, D)$ is the Jaccard similarity between the proteins contained in C and D.

6.2.3 Integrated complex-prediction system

Figure 6.1 shows a flowchart of our integrated system consisting of SWC, DECOMP, and SSS. Each of these approaches is run independently on the input data, and the resulting clusters are combined at the end. Here we give only a brief description of each approach, as they are described in greater detail in the respective chapters (Chapters 3 to 5).

First, SWC performs supervised weighting using its input data, to weight each edge with its posterior probability of being a co-complex edge. Then it runs the various clustering algorithms, and combines the resulting cluster sets with majority voting to produce its set of clusters. Each final cluster is scored by its weighted density (weights being the SWC posterior probabilities), multiplied by the number of clustering algorithms that produced it, and normalized to 1. We keep only clusters of size four or larger.

Next, DECOMP performs hub removal on its input PPI data (using $N_{hub} = 50$

for yeast, $N_{hub} = 150$ for human, as described in Chapter 4). Then it performs GO decomposition to split the PPI network into spatially-coherent subnetworks (using $N_{GO} = 300$ for yeast, $N_{GO} = 1000$ for human, as described in Chapter 4). For each of the clustering algorithms, the algorithm is run on the subnetworks, the clusters from the subnetworks are re-combined, and hubs are re-added to those clusters they are highly-connected to. Finally, the resulting clusters from the various clustering algorithms are combined with majority voting to produce a set of clusters. Each cluster is scored by its weighted density (weights being the PPI reliabilities), multiplied by the number of clustering algorithms that produced it, and normalized to 1. We keep only clusters of size four or larger.

Next, SSS performs size-specific supervised weighting using its input data, to weight each edge with its posterior probabilities of being small-co-complex, large-co-complex, and non-complex. Then the small complexes (size-2 and -3 complexes) are extracted and scored using Extract. Each final cluster is scored by its cohesiveness-weighted density, which takes into account the weights of both internal and surrounding edges.

Finally, the clusters produced by the three approaches are combined, also using the voting-based aggregative strategy. However, since each approach scores its clusters in a different manner, we first scale their scores to make them comparable. The clusters generated by DECOMP are scaled by a factor d, while those generated by SSS are scaled by a factor s. In our experiments we used $d = 0.6, s = 1$ for yeast, and $d = 0.6, s = 0.3$ for human. These factors were obtained by observing the relationship between scores and precision levels in the cross-validation results for each approach (e.g. a cluster predicted by DECOMP with a score of 0.6 obtained roughly the same precision as a cluster predicted by SWC with a score of 1.0). Then we take the union of the clusters produced by the three approaches. If a cluster from two or more approaches are found to be similar to each other (Jaccard similarity ≥ 0.75), we sum its scores from the different approaches.

6.3 Results

6.3.1 Experimental setup

We compare the performance of the following approaches:

1. SWC+DECOMP+SSS: integrated approach consisting of SWC, DECOMP, and SSS

2. SWC: Supervised Weighting of Composite network, using six clustering algo-

rithms combined with majority voting

3. DECOMP: decomposition of PPI network, using six clustering algorithms combined with majority voting

4. SSS: Size-Specific Supervised Weighting

5. PPI+COMBINE: PPI network weighted by reliability, using six clustering algorithms combined with majority voting

6. PPI+clustering algorithm: PPI network weighted by reliability, using a single clustering algorithm

We perform random sub-sampling cross-validation, repeated over ten rounds, using manually-curated complexes as reference complexes for training and testing. For yeast, we use the CYC2008 [56] set which consists of 408 complexes. For human, we use the CORUM [57] set which consists of 1829 complexes. In each cross-validation round, $t\%$ of the complexes are selected for testing, while all the remaining complexes are used for training. Thus we use a large percentage of test complexes $t\% = 90\%$, giving 41 training complexes in yeast, and 183 training complexes in human. Each edge (u, v) in the network is given a class label *co-complex* if u and v are in the same training complex, otherwise its class label is *non-co-complex*. For the supervised approaches, learning is performed using these labels, and the edges of the entire network are then weighted using the learned models. The top-weighted k edges from the network are then used by the clustering algorithms to predict complexes. In our experiments we use $k = 20000$ for SWC and DECOMP, and $k = 10000$ for SSS (as described in their respective chapters).

We use precision-recall graphs to evaluate how well the predicted clusters match the test complexes. Each cluster P is ranked by its score. To obtain a precision-recall graph, we calculate and plot the precision and recall of the predicted clusters at various cluster-score thresholds. The calculation of precision and recall differ slightly from those in Chapters 2 or 3, as here we define different matching thresholds for large and small complexes. Given a set of predicted clusters $\mathbf{P} = \{P_1, P_2, \ldots\}$, a set of test reference complexes $\mathbf{C} = \{C_1, C_2, \ldots\}$, and a set of training reference complexes $\mathbf{T} = \{T_1, T_2, \ldots\}$, the recall and precision at score threshold s are defined as follows:

$$Recall_s = \frac{|\{C_i | C_i \in \mathbf{C} \wedge \exists P_j \in \mathbf{P}, score(P_j) \geq s, P_j \; matches \; C_i\}|}{|\mathbf{C}|}$$

$$Precision_s = \frac{|\{P_j | P_j \in \mathbf{P}, score(P_j) \geq s \wedge \exists C_i \in \mathbf{C}, C_i \; matches \; P_j\}|}{|\{P_k | P_k \in \mathbf{P}, score(P_k) \geq s \wedge (\nexists T_i \in \mathbf{T}, T_i \; matches \; P_k \vee \exists C_i \in \mathbf{C}, C_i \; matches \; P_k)\}|}$$

$$C \; matches \; P = \begin{cases} \text{true} & \text{if } size(C) > 3 \wedge size(P) > 3 \wedge Jaccard(P,C) \geq lg\_match \\ & \text{or } size(C) \leq 3 \wedge size(P) \leq 3 \wedge Jaccard(P,C) \geq sm\_match \\ \text{false} & \text{otherwise} \end{cases}$$

The precision of clusters is calculated only among those clusters that do not match a training complex, to eliminate the bias of the supervised approaches for predicting training complexes well. We require small complexes to be matched perfectly, as a mismatch of just one protein in a small complex may render the prediction less useful; on the other hand we allow a slight tolerance for mismatch for large complexes. Thus we require that small complexes must be matched by small clusters with a match threshold of sm_match, and large complexes must be matched by large clusters with a different threshold of lg_match. We define $lg\_match = 0.75$ for large yeast complexes, $lg\_match = 0.5$ for large human complexes (since they are more challenging to predict), and $sm\_match = 1$ for small complexes in both yeast and human.

6.3.2 Complex prediction

Figure 6.2 shows the precision-recall graphs for complex prediction in yeast. Figure 6.2a shows that SWC and DECOMP both attain higher precision than PPI+COMBINE, demonstrating the benefits of supervised weighting and PPI decomposition (note that all three of these approaches use the COMBINE strategy). As SSS' predictions are limited to small complexes, which is moreover a difficult challenge with a perfect matching requirement, it has lower precision levels compared to PPI+COMBINE. However, the integrated approach, SWC+DECOMP+SSS, is able to predict both large and small complexes, and achieves much higher recall as well as precision. Figure 6.2b shows that individual clustering algorithms (used with the PPIREL network) give lower precision and recall compared to PPI+COMBINE, showing the utility of combining the clusters from multiple clustering algorithms.

We noticed that the generated small clusters may depress the precision, as many of them are false positives. Figures 6.2c and d show the performance when these small clusters are removed. As expected, recall drops substantially, as the small complexes are now unable to be predicted: for example, for PPI+COMBINE, recall drops from over 40% to about 20%. However, precision is improved, as the many false-positive

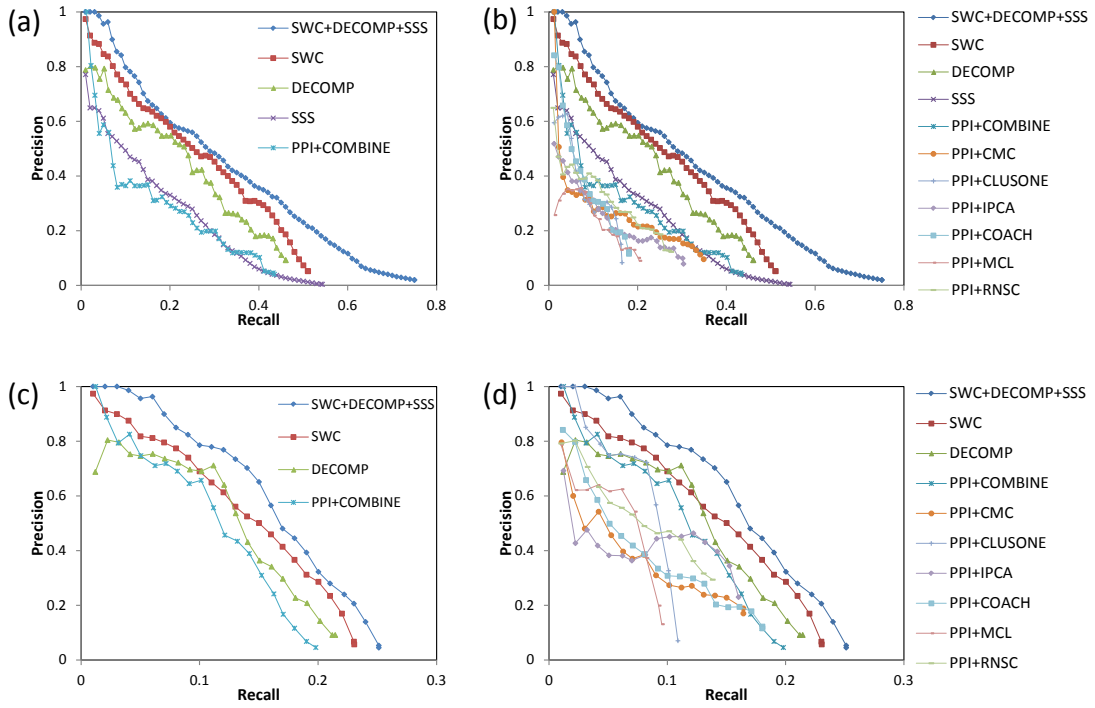

Figure 6.2: Precision-recall graphs for complex prediction in yeast. For clarity, (a) shows only the integrated approach (SWC+DECOMP+SSS), each of its constituent approaches, and the PPI+COMBINE approach, while (b) includes the individual clustering algorithms. (c) and (d) show the performance when the generated small clusters are removed, which lowers recall substantially but increases precision. In yeast we use a matching threshold of $lg\_match = 0.75$ for large complexes, and $sm\_match = 1$ for small complexes.

small clusters are removed. For our integrated approach (SWC+DECOMP+SSS), the removal of small clusters means removing those clusters generated by SSS. We still achieve higher precision and recall than the other approaches, showing that our integrated approach still outperforms other approaches when considering large complexes only. Moreover, *without* removing small clusters, our integrated approach maintains high precision as it uses a specialized approach, SSS, to predict small complexes.

Figure 6.3 shows the corresponding precision-recall graphs for complex prediction in human. Figure 6.3a shows that SWC and DECOMP both attain higher precision than PPI+COMBINE, showing the benefits of supervised weighting and PPI decomposition. SSS shows poor performance as it is limited to predicting small complexes, which is especially challenging in human. The integrated approach, SWC+DECOMP+SSS, is able to predict both large and small complexes, and achieves higher recall as well as precision. Figure 6.3b shows that most of the individual clustering algorithms (used with the PPIREL network) give lower precision and recall compared to PPI+COMBINE, showing the utility of combining the clusters from multiple clustering algorithms. The exception is Coach, which attains high precision as it does not generate small clusters by design, thereby cutting down on its false-positive predictions.

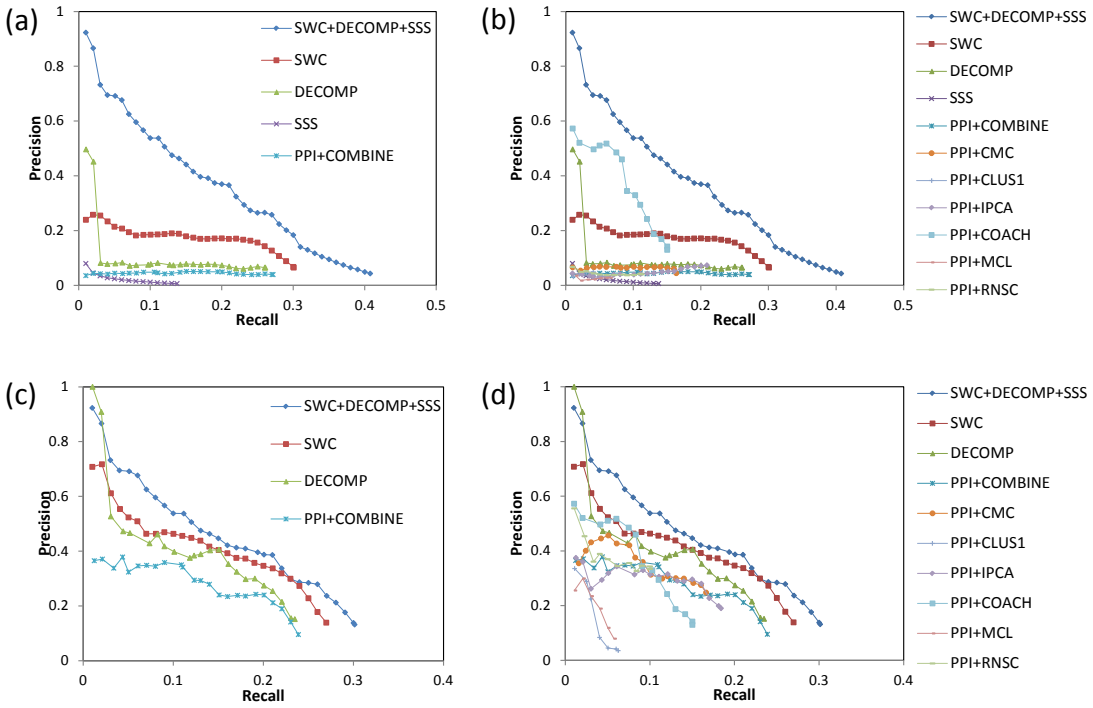

Figure 6.3: Precision-recall graphs for complex prediction in human. For clarity, (a) shows only the integrated approach (SWC+DECOMP+SSS), each of its constituent approaches, and the PPI+COMBINE approach, while (b) includes the individual clustering algorithms. (c) and (d) show the performance when the generated small clusters are removed, which lowers recall but increases precision. For human we use a matching threshold of $lg\_match = 0.5$ for large complexes, and $sm\_match = 1$ for small complexes.

Figures 6.3c and d show the performance when the generated small clusters are removed. Compared to yeast, here the recall does not drop as much: for example, for PPI+COMBINE, recall drops by about 5% only. However, the improvement in precision is substantial: for example, PPI+COMBINE sees more than fivefold increase in precision at many points in the graph. This reveals an issue in complex prediction which is more obvious in human but still apparent in yeast: predicting small complexes alongside large ones means accepting a drop in precision due to large numbers of false-positive small clusters; while improving precision by excluding small clusters means that no small complexes can be predicted. On the other hand, our integrated approach uses a specialized approach, SSS, to generate the small clusters separately from the large ones, which allows effective prediction of the small complexes while still maintaining high precision levels.

To investigate the performance of our integrated approach with respect to the three challenges that we highlighted, we stratify the reference complexes in terms of their size, EXT (the number of external proteins that are highly connected to it), and DENS (density), as described in Chapter 2.5.1. Figures 2.3 and 2.4 show the distribution of the complexes in terms of size, DENS, and EXT, for yeast and human.

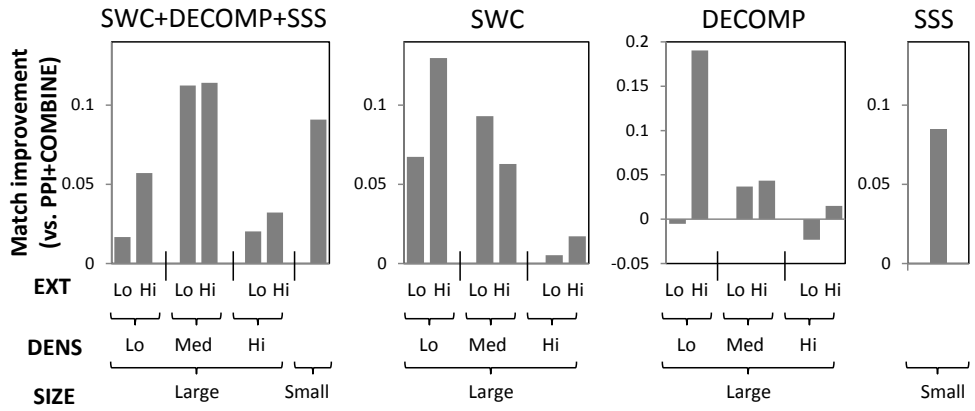

Figure 6.4: Match-score improvements among stratified yeast complexes.

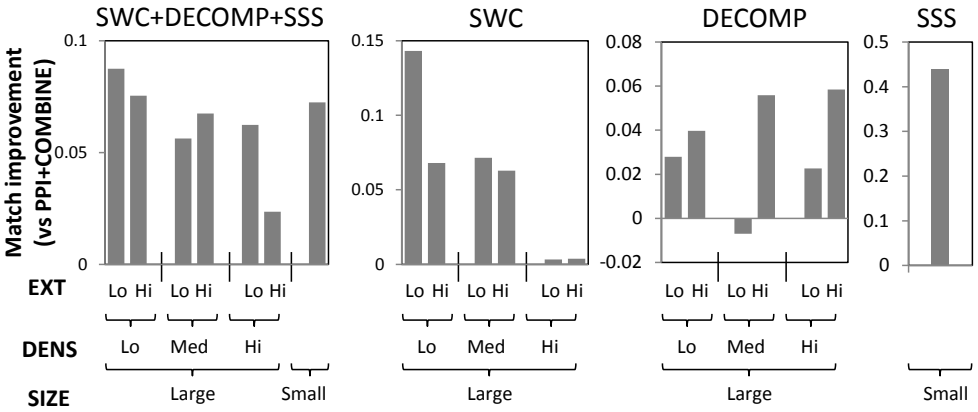

Figure 6.5: Match-score improvements among stratified human complexes.

We take the top 1000 clusters generated by each approach, and determine how well the reference complexes in the different strata are matched by these clusters. Figures 6.4 and 6.5 show the average improvements in matching scores among the stratified complexes for our approaches versus PPI+COMBINE, in yeast and human respectively.

Among yeast and human large complexes, SWC gives the biggest improvements among complexes with low to medium density: it uses data integration and supervised learning to fill in missing edges of sparse complexes to allow them to be predicted. Among sparse complexes, even those with high EXT see an improvement, showing that SWC's supervised weighting can effectively reduce the number of spurious edges in the PPI network. DECOMP gives the biggest improvements among complexes with high EXT, within each density stratum. This is because it decomposes the PPI network into spatially- and temporally-coherent subnetworks, in which complexes may become disconnected from their original densely-connected neighbourhoods, allowing their borders to be better delimited by clustering algorithms. As expected, SSS improves the performance among small complexes. Our integrated ap-

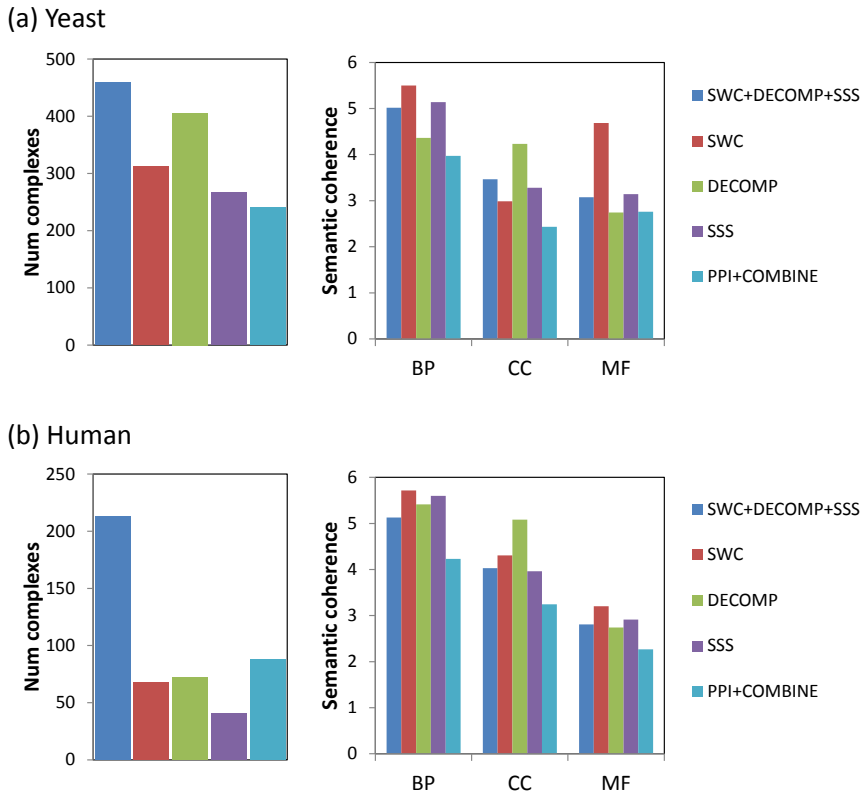

Figure 6.6: Number and quality of novel predictions in (a) yeast, (b) human.

proach (SWC+DECOMP+SSS) spreads out the improvements among the complexes in the different strata, showing that the different approaches complement each other to predict different types of challenging complexes.

6.3.3 Novel complexes

Here we investigate the number and quality of novel complexes predicted by our approaches. For the supervised approaches, we use the entire sets of reference complexes for training. We keep only predicted complexes that are novel, unique, and high-confidence. First, predicted complexes that are similar to each other are filtered to keep only the highest-scoring one. Next, we keep only the top-scoring predictions such that the precision of these predictions (i.e. proportion of predictions that match a reference complex) is greater than 0.4. Finally, we keep only novel predictions by removing those that match a reference complex. We measure the quality of these novel predictions by their semantic coherence in each of the three GO classes, as described in Chapter 3.3.2.

Figure 6.6a shows the number and quality of novel predictions in yeast. Each of our individual approaches (SWC, DECOMP, and SSS) predicts more novel complexes compared to the baseline (PPI+COMBINE), while the integrated approach generates

the highest number of novel complexes. The novel complexes from our individual approaches attain higher semantic coherence in one or more of the GO classes, compared to the baseline. The novel predictions from the integrated approach attain semantic coherence that is averaged out between its three constituent approaches, which gives it higher coherence than the baseline across all three GO classes.

Figure 6.6b shows the number and quality of novel predictions in human. As described above, PPI+COMBINE generates a great number of small clusters in human, most of which are false-positives; this gives it a greater number of novel predictions compared to each of our individual approaches. Nonetheless, our integrated approach still generates the greatest number of novel complexes. As in yeast, our individual approaches generate novel complexes with greater semantic coherence compared to PPI+COMBINE; the integrated approach achieves greater semantic coherence, in all three GO classes, in its predictions compared to the baseline. Thus, in both yeast and human, our integrated approach generates the greatest number of novel predictions, with higher quality compared to the baseline approach of combined clustering with a PPI network.

6.4 Conclusion

Three open problems remain within protein-complex prediction. First, many complexes are sparsely connected in the PPI network, and so do not form dense clusters that can be derived by clustering algorithms. Second, many complexes are embedded within highly-connected regions of the PPI network, which makes it difficult for clustering algorithms to accurately delimit their boundaries. Third, many complexes are small (composed of two or three distinct proteins), so that traditional topological markers such as density or sparse neighbourhoods are ineffective.

In previous chapters we proposed three approaches for addressing each of these challenges. In Chapter 3, we described Supervised Weighting of Composite Networks (SWC), which integrates diverse data sources with supervised learning to weight edges with their posterior probabilities of being co-complex. SWC was shown to improve the prediction of sparse complexes. In Chapter 4, we described PPI network decomposition using GO terms and hub removal (DECOMP), which was shown to improve the prediction of complexes embedded within highly-connected regions. In Chapter 5, we described Size-Specific Supervised Weighting (SSS), which integrates diverse data sources and topological features with supervised weighting to weight edges with their posterior probabilities of belonging to small complexes. SSS was shown to improve the

prediction of small complexes.

In this chapter we integrate these three approaches into a single system. SWC, DECOMP, and SSS are run independently on the input PPI data and other data sources, and the resulting clusters are weighted to standardize their scores, then combined using majority voting. We test the integrated approach on the prediction of yeast and human complexes, and show that it outperforms SWC, DECOMP, or SSS when run individually, achieving the highest recall, and the highest precision at all recall levels.

We also investigate which complexes benefit most from our individual approaches and the integrated approach, compared to a baseline of running a set of clustering algorithms on a reliability-weighted PPI network. In both yeast and human, we find that SWC improves the prediction of sparse complexes, DECOMP improves the prediction of embedded complexes, and SSS improves the prediction of small complexes. The integrated approach combines these improvements and distributes them among the different types of challenging complexes. Furthermore, we show that our integrated approach generates the greatest number of novel predictions with higher quality in terms of GO semantic coherence.

Although we have taken great strides in tackling the three challenges we highlight within complex prediction, and have obtained substantial improvements in prediction accuracy and recall as a result, there remains room for further improvement. Moreover, as increasing amounts of PPI data become available for other organisms, the techniques that we propose will be useful in enabling the discovery of novel complexes in those organisms.

Chapter 7

Conclusion

7.1 Summary

In the cell, many proteins interact physically to form stoichiometrically-stable multi-protein structures called protein complexes. Protein complexes participate in many biological processes, and perform a wide variety of molecular functions, so determining the set of existing complexes is important for understanding the mechanism, organization, and regulation of cellular processes.

High-throughput experimental techniques have produced large amounts of protein-protein interaction (PPI) data, which makes it possible to discover protein complexes from PPI networks: since protein complexes are groups of proteins that interact with one another, they usually form dense subgraphs in PPI networks. Many algorithms have been developed to discover complexes from PPI networks based on this idea. However, the performance of these approaches still leaves room for improvement: for example, even in *Saccharomyces cerevisiae* (baker's yeast) where PPI data is fairly complete, accurate prediction of complexes at fine resolution remains difficult. One main stumbling block is that the representation of PPI data, and its analysis for complex discovery, do not take into account the dynamism of cellular PPIs and complexes.

In Chapter 2 we described how proteins interact in a dynamic fashion, with a variety of interaction timings, locations, and affinities. These are mediated by a wide range of factors including cellular state, cellular processes, and the interaction environment. Correspondingly, protein complexes exhibit dynamic behavior which are in fact important functional mechanisms, for example to allow complexes to be formed only at certain times, or to vary the composition of complexes to modulate or activate their functions. However, due to limitations in PPI-detection methodologies, it is difficult to interrogate the dynamics of PPIs (i.e. when, where, and how a protein interacts with others). Furthermore, this dynamism also precludes a faithful interrogation of PPIs in

the cell (e.g. condition-specific PPIs may be missed, or spurious PPIs may be detected in non-physiological experimental systems). Moreover, the representation of PPIs in the PPI network does not preserve any information about the dynamics of PPIs. Thus there exists a disparity between the dynamic nature of PPIs and protein complexes on the one hand, and the static representation and analysis of the PPI network on the other hand.

We identified three challenges in protein-complex discovery that arise from, or are exacerbated by, this static view of PPIs and protein complexes [8]. First, many complexes exist in sparse regions of the network, so that proteins within the complexes are not densely interconnected. This arises from undetected condition-specific, location-specific, or transient PPIs. Second, many complexes are embedded within highly-connected regions of the PPI network, with many extraneous edges connecting its member proteins to other proteins outside the complex. This arises from proteins that participate in multiple distinct complexes which correspond to dense overlapping regions in the PPI network, or from spuriously-detected interactions. Third, many complexes are small (that is, composed of two or three proteins), making measures of important topological features, such as density, ineffectual. This is further exacerbated by extraneous or missing interactions which can embed the small complex in a larger clique, or disconnect it entirely.

In this dissertation we proposed three approaches that can help to address these problems. In Chapter 3, we described an approach called Supervised Weighting of Composite Networks (SWC [4]) which can address the problem of sparse complexes. SWC integrates PPI data with additional data sources, and uses a supervised approach to weight edges with their posterior probability of belonging to a complex. By integrating diverse data sources that may support co-complex relationships between proteins, SWC fills in the missing edges in many sparse complexes, while reducing the amount of spurious non-co-complex edges. Using this approach, improvements are obtained in both precision and recall for yeast and human complex discovery, especially among the sparse complexes.

In Chapter 4, we described an approach to decompose the PPI network into spatially- and temporally-coherent subnetworks [5], which can address the problem of complexes in highly-connected regions with many extraneous edges. First, hub proteins with large numbers of interaction partners are removed before complex discovery, as they tend to correspond to date hubs with non-simultaneous interactions. Next, cellular-location Gene Ontology terms [6] are used to decompose the PPI network

into spatially-coherent subnetworks. By splitting dense regions of the PPI network into less-dense but coherent subnetworks, complex-discovery performance is improved, with the biggest improvements among complexes in highly-connected regions.

In Chapter 5, we described an approach called Size-Specific Supervised Weighting (SSS [7]) to address the problem of predicting small complexes. SSS integrates PPI data with additional data sources, along with their topological features, and uses a supervised approach to weight edges with their posterior probabilities of belonging to small complexes versus large complexes. SSS then extracts small complexes from the weighted network, and scores them using the probabilistic weights of edges within, as well as surrounding, the complexes. This approach achieves significant improvements in precision and recall in discovering small complexes.

In Chapter 6, we combined these three approaches into a single integrated system which addresses the three challenges of complex prediction: predicting sparse complexes, predicting complexes embedded within dense regions, and predicting small complexes. This integrated system obtains vast improvements compared to a baseline of using a set of clustering algorithms on a PPI-reliability-weighted network. For example, in yeast our integrated system doubles the recall (from 40% to 75%), while maintaining more-than-double the precision at most recall levels (for example, at 40% recall level, the precision is almost 40% compared to the baseline's 10%). In human, our integrated system increases the recall from 28% to 38%, while maintaining more-than-fivefold precision at most recall levels (for example, at 20% recall, the precision is 38% compared to the baseline's 5%). Furthermore, our integrated system also achieves greater performance in complex discovery over using any single one of the three proposed approaches.

7.2 Future work

7.2.1 Applications

A high-quality set of novel predicted protein complexes is not only an important resource for understanding cellular processes and functions. It can also support other bioinformatics analyses, of which we briefly discuss two here.

Gene-expression data has been analyzed to find genes that are differentially expressed between different phenotypes, in particular between diseased and normal samples. A challenge is that many diseases involve multiple genes that interact in complex ways, both physically and genetically. Thus various methods have been proposed for differential expression analysis among *gene sets* which correspond to higher-level

biological units, such as known pathways [81]. Of interest to us is differential expression analysis among novel predicted protein complexes, which can reveal novel disease mechanisms at the protein-complex level, as well as develop new biomarkers for disease subtype classification and diagnosis.

A different bioinformatics problem that can benefit from high-quality novel complexes is in the analysis of proteomics data. Traditional methods apply thresholds on mass-spectrometry proteomics data to select proteins that are present in the sample, which leads to large amounts of lost information as proteins present in low levels are discarded. Proteomics Signature Profiling (PSP [82]) instead analyzes this data at the level of protein complexes: by calculating the number of proteins present in each complex, it generates a Proteomics Signature Profile for each sample, which is successfully used to cluster moderate- and late-stage liver cancer patients. Given that the set of known biological complexes is far from complete, augmenting it with high-quality predicted complexes can help to expand the basis of such analyses.

7.2.2 Further improvements in complex prediction

Our proposed approaches achieve substantial improvements in the prediction of protein complexes in both yeast and human. However, there is still room for improvement especially for human complexes, where even at a rough matching requirement, less than 40% of the reference complexes can be predicted, at a 5% precision level.

A significant challenge for human complex prediction is insufficient PPI data. An estimate of the human interactome size is around $220,000$ PPIs [83]. Our human PPI data consists of around $140,000$ PPIs, and with an estimated false-positive rate of 50%, this means that our human PPI network represents only a third of the true human PPI network. In comparison, in yeast an estimate of the interactome size is around $50,000$ PPIs. Our yeast PPI data consists of around $120,000$ PPIs, so even with an estimated false-positive rate of 50%, our yeast PPI network can be believed to be a good representation of the actual yeast PPI network. The much poorer representation of the true human interactome partially explains the poorer performance of our approach on human complexes.

PPI coverage is even poorer for other model organisms. For example, other organisms with significant numbers of experimental PPI data are *Arabidopsis thaliana* (about 6000 experimental PPIs reported), *Drosophila melanogaster* (about 6000 PPIs), and *Caenorhabditis elegans* (about 2000 PPIs), all of which cover less than 10% of their interactomes (assuming the interactomes consist of at least $50,000$ PPIs, which is a

conservative estimate). Indeed, our preliminary experiments included predicting complexes from these organisms, which gave extremely poor results.

As more experimental PPI data from these organisms becomes available, prediction of their complexes will become more viable. Parallel to this effort, the integration of other data sources, as well as the development of new techniques to do this more accurately, can also help to boost interactome coverage. In our work we integrated PPI data with functional associations and literature co-occurences, but other data sources should also be explored, such as protein domains, gene expression, and interologs, as well as what is the best way to integrate them for complex discovery.

Aside from increasing interactome coverage, another important step to help the prediction and understanding of complexes is to directly interrogate the dynamism of PPIs and complexes. Recently, researchers have begun analyzing the composition of complexes under different perturbation states, using quantitative AP-MS approaches: affinity purification with selected reaction monitoring (AP-SRM [25]) was proposed to probe quantitative changes in interactions of the Grb2 protein after stimulation with various growth factors; while affinity purification combined with sequential window acquisition of all theoretical spectra (AP-SWATH [26]) was used to study changes in the 14-3-3β protein interactome following stimulation of the insulin-PI3K-AKT pathway. Both works represent key advances in methodologies that will allow dynamic and condition-specific views and analyses of interactomes in the near future; but for now, the range of the proteins and PPIs probed, as well as the conditions tested, remain limited. Moreover, as data about the dynamism of PPIs and complexes becomes available, more sophisticated representations of PPIs need to be developed that can capture such information, and that can enable its analysis to derive useful biological knowledge.

For now, the data and representation of PPIs are overwhelmingly static. The work described in this thesis shows that a consideration of the dynamism of PPIs and complexes can be very useful in the analysis of static PPI networks, giving improved performance in the discovery of protein complexes.

Bibliography

[1] Nooren IMA, Thornton JM: **Diversity of protein-protein interactions**. *EMBO J* 2003, **22**(14):3486–3492.

[2] Mendenhall A, Hodge A: **Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae**. *Microbiol Mol Biol R* 1998, **62**(4):1191–1243.

[3] Enserink JM, Kolodner RD: **An overview of Cdk1-controlled targets and processes**. *Cell Div* 2010, **5**(11).

[4] Yong CH, Liu G, Chua HN, Wong L: **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Syst Biol* 2012, **6**(Suppl 2):S13.

[5] Liu G, Yong CH, Chua HN, Wong L: **Decomposing PPI networks for complex discovery**. *Proteome Sci* 2011, **9**(Suppl 1):S15.

[6] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene Ontology: Tool for the unification of biology**. *Nature Genet* 2000, **25**:25–29.

[7] Yong CH, Maruyama O, Wong L: **Discovery of small protein complexes from PPI networks with size-specific supervised weighting**. *BMC Syst Biol* 2014, **8**(Suppl 5):S3.

[8] Yong CH, Wong L: **From the static interactome to dynamic protein complexes: Three challenges**. *J Bioinform Comput Biol* 2015, **13**(2):15710018.

[9] Li X, Wu M, Kwoh CK, Ng SK: **Computational approaches for detecting protein complexes from protein interaction networks: A survey**. *BMC Genomics* 2010, **11**(Suppl 1):S3.

[10] Srihari S, Leong HW: **A survey of computational methods for protein complex prediction from protein interaction networks**. *J Bioinform Comput Biol* 2013, **11**(2):1230002.

[11] Chen B, Fan W, Liu J, Wu FX: **Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks**. *Brief Bioinform* 2014, **15**(2):177–194.

[12] Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 2004, **430**:88–93.

[13] Batada NN, Hurst LD, Tyers M: **Evolutionary and physiological importance of hub proteins**. *PLoS Comput Biol* 2006, **2**(7):e88.

[14] Jones S, Thornton JM: **Principles of protein-protein interactions**. *Proc Natl Acad Sci USA* 1996, **93**:13–20.

[15] Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: **Transient protein-protein interactions: Structural, functional, and network properties**. *Structure* 2010, **18**(10):1233–1243.

[16] Deshaies RJ, Seol JH, McDonald WH, Cope G, Lyapina S, Shevchenko A, Shevchenko A, Verma R, Yates JR: **Charting the protein complexome in yeast by mass spectrometry**. *Mol Cell Proteomics* 2002, **1**:3–10.

[17] de Lichtenberg U, Jensen LJ, Brunak S, Bork P: **Dynamic complex formation during the yeast cell cycle**. *Science* 2005, **307**(5710):724–747.

[18] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**:631–636.

[19] Fields S, Song O: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**(6230):245–246.

[20] Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U: **Yeast two-hybrid, a powerful tool for systems biology**. *Int J Mol Sci* 2009, **10**(6):2763–2788.

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

[22] Gavin AC, Maeda K, Kühner S: **Recent advances in charting protein-protein interaction: Mass spectrometry-based approaches**. *Curr Opin Biotech* 2011, **22**:42–49.

[23] Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2007, **6**(3):439–450.

[24] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature* 2006, **440**:637–643.

[25] Bisson N, James DA, Ivosev G, Tate SA, Bonner R, Taylor L, Pawson T: **Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor**. *Nat Biotechnol* 2011, **29**(7):653–658.

[26] Collins BC, Gillet LC, Rosenberger G, Röst HL, Vichalkovski A, Gstaiger M, Aebersold R: **Quantifying protein interaction dynamics by SWATH mass spectrometry: Application to the 14-3-3 system**. *Nat Methods* 2013, **10**(12):1246–1253.

[27] Srihari S, Leong HW: **Temporal dynamics of protein complexes in PPI networks: A case study using yeast cell cycle dynamics**. *BMC Bioinformatics* 2005, **13**(Suppl 17):S16.

[28] Jung SH, Hyun B, Jang WH, Hur HY, Han DS: **Protein complex prediction based on simultaneous protein interaction network**. *Bioinformatics* 2010, **26**(3):385–391.

[29] 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.

[30] Ideker T, Krogan NJ: **Differential network biology**. *Mol Syst Biol* 2012, **8**:565.

[31] Tatsuke D, Maruyama O: **Sampling strategy for protein complex prediction using cluster size frequency**. *Gene* 2012, **518**:152–158.

[32] 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.

[33] Palla G, Derenyi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex networks in nature and society**. *Nature* 2005, **435**:814–818.

[34] Farkas I, Ábel D, Palla G, Vicsek T: **Weighted network modules**. *New J Phys* 2007, **9**(6):180.

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

[36] Li X, Tan S, Foo C, Ng S: **Interaction graph mining for protein complexes using local clique merging**. *Genome Informatics* 2005, **16**:260–269.

[37] Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.

[38] Rhrissorrakrai K, Gunsalus KC: **MINE: Module identification in networks**. *BMC Bioinformatics* 2011, **12**:192.

[39] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S: **Development and implementation of an algorithm for detection of protein complexes in large interaction networks**. *BMC Bioinformatics* 2006, **7**:207.

[40] Li M, Chen J, Wang J, Hu B, Chen G: **Modifying the DPClus algorithm for identifying protein complexes based on new topological structures**. *BMC Bioinformatics* 2008, **9**:398.

[41] Nepusz T, Yu H, Paccanaro A: **Detecting overlapping protein complexes in protein-protein interaction networks**. *Nat Methods* 2012, **9**:471–472.

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

[43] King AD, Przulj N, Jurisica I: **Protein complex prediction via cost-based clustering**. *Bioinformatics* 2004, **20**(17):3013–3020.

[44] Widita CK, Maruyama O: **PPSampler2: Predicting protein complexes more accurately and efficiently by sampling**. *BMC Syst Biol* 2013, **7**(Suppl 6):14.

[45] Blatt M, Wiseman S, Domany E: **Superparamagnetic clustering of data**. *Phys Rev Lett* 1996, **76**(18):3251–3254.

[46] Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ, Koller D: **A complex-based reconstruction of the Saccharomyces cerevisiae interactome**. *Mol Cell Proteomics* 2009, **8**(6):1361–1381.

[47] Girvan M, Newman MEJ: **Community structure in social and biological networks**. *Proc Natl Acad Sci USA* 2002, **99**(12):7821–7826.

[48] Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH: **Modular organization of protein interaction networks**. *Bioinformatics* 2007, **23**(2):207–214.

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

[50] Srihari S, Ning K, Leong HW: **MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure**. *BMC Bioinformatics* 2010, **11**:504.

[51] Rivas JDL, Fontanillo C: **Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks**. *PLoS Comput Biol* 2010, **6**(6):e1000807.

[52] Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M: **The BioGRID interaction database: 2013 update**. *Nucleic Acids Res* 2013, **41**(Database Issue):D816–D823.

[53] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H: **The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases**. *Nucleic Acids Res* 2014, **42**(Database Issue):D358–D363.

[54] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2012 update**. *Nucleic Acids Res* 2012, **40**(Database Issue):D857–D861.

[55] Chua HN, Sung WK, Wong L: **An efficient strategy for extensive integration of diverse biological data for protein function prediction**. *Bioinformatics* 2007, **23**(24):3364–3373.

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

[57] Ruepp A, Waegele B, Lechner M, Brauner B, I DK, Fobo G, Frishman G, Montrone C, Mewes H: **CORUM: The comprehensive resource of mammalian protein complexes–2009**. *Nucleic Acids Res* 2010, **38**:D497–D501.

[58] Bylund GO, Majka J, Burgers PMJ: **Overproduction and purification of RFC-related clamp loaders and PCNA-related clamps from Saccharomyces cerevisiae**. *Metholds Enzymol* 2006, **409**:1–11.

[59] Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z: **Protein complex identification by supervised graph local clustering**. *Bioinformatics* 2008, **24**(13):i250–i258.

[60] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498–2504.

[61] Edwards AM, Kus B, Jensen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: Assessing protein interaction data with known complexes**. *Trends Genet* 2002, **18**(10):529–536.

[62] Gilchrist M, Salter L, Wagner A: **A statistical framework for combining and interpreting proteomic datasets**. *Bioinformatics* 2004, **20**(5):689–700.

[63] Liu G, Li J, Wong L: **Assessing and predicting protein interactions using both local and global network topological metrics**. In *Proceedings of 19th International Conference on Genome Informatics* 2008:138–149.

[64] Han DS, Kim HS, Jang WH, Lee SD, Suh JK: **PreSPI: A domain combination based prediction system for protein-protein interaction**. *Nucleic Acids Res* 2004, **32**:6312–6320.

[65] Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques**. *Bioinformatics* 2006, **22**(7):823–829.

[66] Scott MS, Barton GJ: **Probabilistic prediction and ranking of human protein-protein interactions**. *BMC Bioinformatics* 2007, **8**:239.

[67] Chua HN, Hugo W, Liu G, Li X, Wong L, Ng SK: **A probabilistic graph-theoretic approach to integrate multiple predictions for the protein-protein subnetwork prediction challenge**. *Ann N Y Acad Sci* 2009, **1158**:224–233.

[68] Qiu J, Noble WS: **Predicting co-complexed protein pairs from heterogeneous data**. *PLoS Comput Biol* 2008, **4**(4):e1000054.

[69] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: Protein-protein interaction networks, with increased coverage and integration**. *Nucleic Acids Res* 2013, **41**(Database issue):D808–D815.

[70] Fayyad UM, Irani KB: **Multi-interval discretization of continuous valued attributes for classification learning**. In *Proceedings of the 13 Annual International Joint Conference on Articial Intelligence* 1993:1022–1027.

[71] Hand DJ, Yu K: **Idiot's Bayes not so stupid after all?** *Int Stat Rev* 2001, **69**(3):385–398.

[72] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM: **Semantic similarity in biomedical ontologies**. *PLoS Comput Biol* 2009, **5**(7):e1000443.

[73] Mimura S, Yamaguchi T, Ishii S, Noro E, Katsura T, Obuse C, Kamura T: **Cul8/Rtt101 forms a variety of protein complexes that regulate DNA damage response and transcriptional silencing**. *J Biol Chem* 2010, **285**:9858–9867.

[74] The Uniprot Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2012, **40**(Database Issue):D71–D75.

[75] Przulj N, Wigle DA: **Functional topology in a network of protein interactions**. *Bioinformatics* 2003, **20**(3):340–348.

[76] Chua HN, Ning K, Sung WK, Leong HW, Wong L: **Using indirect protein-protein interactions for protein complex predication**. *J Bioinform Comput Biol* 2008, **6**(3):435–466.

[77] P R, M H, O M, T A: **Prediction of heterodimeric protein complexes from weighted protein-protein interaction networks using novel features and kernel functions**. *PLoS ONE* 2013, **8**(6):e65265.

[78] P R, M H, O M, T A: **Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels**. *BMC Bioinformatics* 2014, **15**(Suppl 2):S6.

[79] Kim K, Yamashita A, Wear MA, Maéda Y, Cooper JA: **Capping protein binding to actin in yeast: Biochemical mechanism and physiological relevance**. *J Cell Biol* 204, **164**(4):567–580.

[80] Tsiokas L, Kim E, Arnould T, Sukhatme VP, Walz G: **Homo- and heterodimeric interactions between the gene products of PKD1 and PKD2**. *Proc Natl Acad Sci USA* 1997, **94**(13):6965–6970.

[81] Lim K, Wong L: **Finding consistent disease subnetworks using PFSNet**. *Bioinformatics* 2014, **30**(2):189–196.

[82] Goh WWB, Lee YH, Ramdzan ZM, Sergot MJ, Chung M, Wong L: **Proteomics Signature Profiling (PSP): A novel contextualization approach for cancer proteomics**. *J Proteome Res* 2012, **11**(3):1571–1581.

[83] Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120.