COMPUTATIONAL METHODS FOR IDENTIFYING CONSERVED PROTEIN COMPLEXES BETWEEN SPECIES FROM PROTEIN INTERACTION DATA

NGUYEN PHI VU
(B.Sc (Hons), Vietnam National University - HCMC)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE

2013

Acknowledgements

Firstly and most of all, I would like to extend my deep gratitude to my supervisor, Professor Leong Hon Wai. He taught me not only skills in doing scientific research but also the courage in pursuing the career of science. Many of his lessons are eye-opening and unforgettable to me. In particular, those are the habit of having evidences in any scientific claims, the positive attitude when listening to critiques, comments. My sincere thanks also go to Dr. Sriganesh Srihari for his co-authorship, suggestions and discussions during my works on this thesis. Without these supports from Professor Leong and Dr. Srihari, the thesis would not be possible.

The RAS Group at School of Computing – NUS has been a source of friendship as well as colleagueship. I have learnt so many things via discussions, coffee chats and activities from the group, especially from Nam Ninh Nguyen, Dr. Ket Fah Chong and Dr. Melvin Zhang.

I would be very grateful to the Computational Biology Group at SoC – NUS for all the seminars, lectures and activities which greatly enhanced my background knowledge in the area.

Finally, I would like to thank my parents for their unbounded love and belief in me during my oversea study.

Summary

Protein complexes *conserved* across species indicate processes that are *core* to cellular machinery. While numerous computational methods have been devised to identify complexes from the protein interaction (PPI) networks of individual species, these are severely limited by noise and errors (false positives) in currently available datasets. Our analysis using human and yeast PPI networks revealed that these methods missed several important complexes including those conserved between the two species.

In this thesis we first present a definition for the problem of identifying conserved protein complexes between species from protein interaction data. We then review the existing computational methods for this problem and its related issues. After that we propose a new and effective method for identifying conserved complexes by constructing *interolog networks* (IN). Our experiments were performed on human and yeast data. Here, we note that much of the functionalities of yeast complexes have been conserved in human complexes not only through sequence conservation of proteins but also of critical *functional domains*. Therefore, our method leverages the *functional conservation* of proteins between species through *domain conservation* in addition to sequence similarity. Our analysis revealed that the IN-construction removes several non-conserved interactions many of which are false positives, thereby improving the number of conserved protein complexes detected compared to direct complex prediction from the PPI networks. These additional complexes included the mismatch repair complex, MLH1-MSH2-PMS2-PCNA, and other important ones namely, RNA polymerase-II, EIF3 and MCM complexes, all of which constitute core cellular processes known to be conserved across the two species.

Our method based on integrating domain conservation and sequence similarity to construct interolog networks also helps to produce a better quality of interolog network between human and yeast compared to other local network alignment based methods. Therefore, integrating information of domain conservation might throw further light on conservation patterns between yeast and human complexes.

We observe from our experiments that protein complexes are not conserved from yeast to human in a straightforward way, that is, it is not the case that a yeast complex is a (proper) sub-set of a human complex with a few additional proteins present in the human complex. Instead complexes have evolved multifold with considerable re-organization of proteins and

re-distribution of their functions across complexes. This finding can have significant implications on attempts to extrapolate other kinds of relationships such as synthetic lethality from yeast to human, for example in the identification of novel cancer targets.

Content

List of Figures

List of Tables

Chapter 1 - Introduction

1.1. Background and Motivation

1.1.1. Protein-protein interaction networks

Protein interactions play a central role in most biological processes. In order to carry out biological functions as catalysts, signaling molecules, or building blocks in cells, proteins need to bind together via domain interfaces to make the corresponding chemical reactions happen. Thus, a critical step towards understanding the inner workings of cellular machinery is to build a complete map of protein-to-protein physical interactions, which is called the interactome.

Protein-protein interaction network (PPI network) is a mathematical model of the interactome in which nodes and edges of the network represent proteins and the physical interactions between them. There could be also edge weights which reflect the reliability of interactions. Figure 1.1b is a picture of the yeast PPI network [Jeong et al., 2001], one of the first eukaryotic interactomes that were studied.

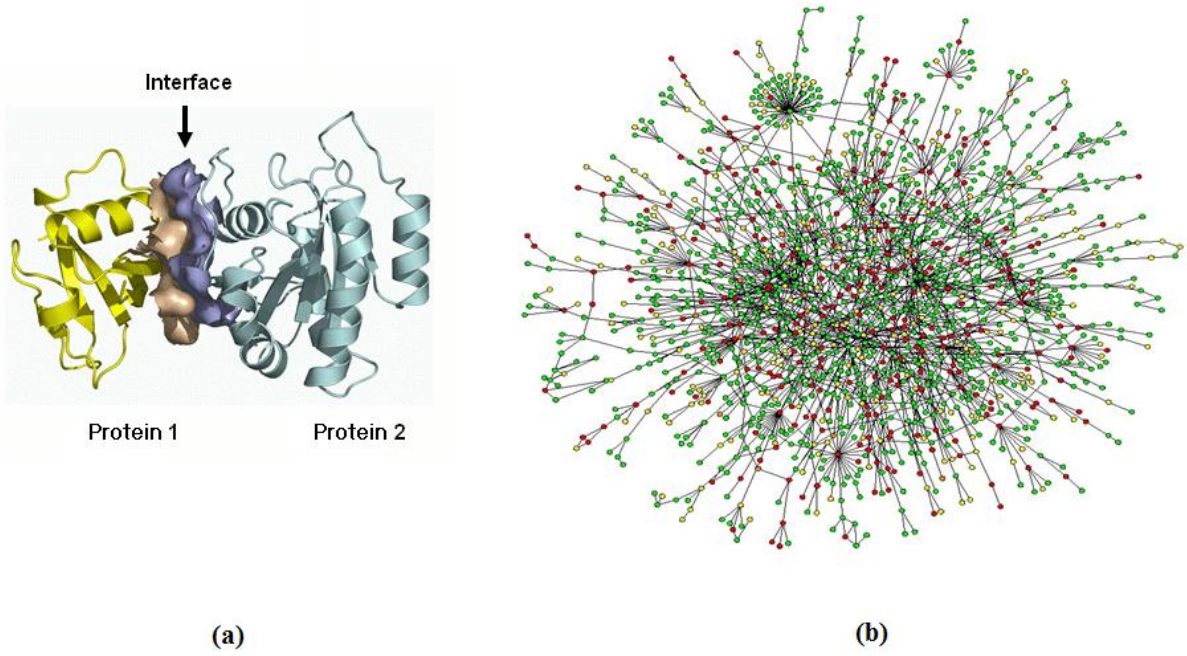


Figure 1.1 – (a) protein-protein interaction, (b) protein-protein interaction network.

As efforts to get a complete image of the interactome, many high-throughput techniques have been developed over the last decade to detect protein interactions on a genome-wide level not only in yeast, two typical techniques among them are: Yeast two hybrid (Y2H) [Uetz et al., 2000; Ito et al., 2001] and Tandem affinity purification combined with mass spectrometry (TAP-MS) [Gavin et al., 2006; Krogan et al., 2006] (See section for details 2.2.1).

1.1.2. Protein complex and predicting protein complexes from PPI networks.

Many proteins have to perform their functions together with other proteins to form protein complexes which are responsible for specific processes in a cell. Understanding how, why and when proteins associate into protein complexes is a critical part of understanding cellular life. Therefore, identifying protein complexes, along with protein pathways, which could be together referred to as cellular machinery, is known as one of the fundamental problems in molecular biology.

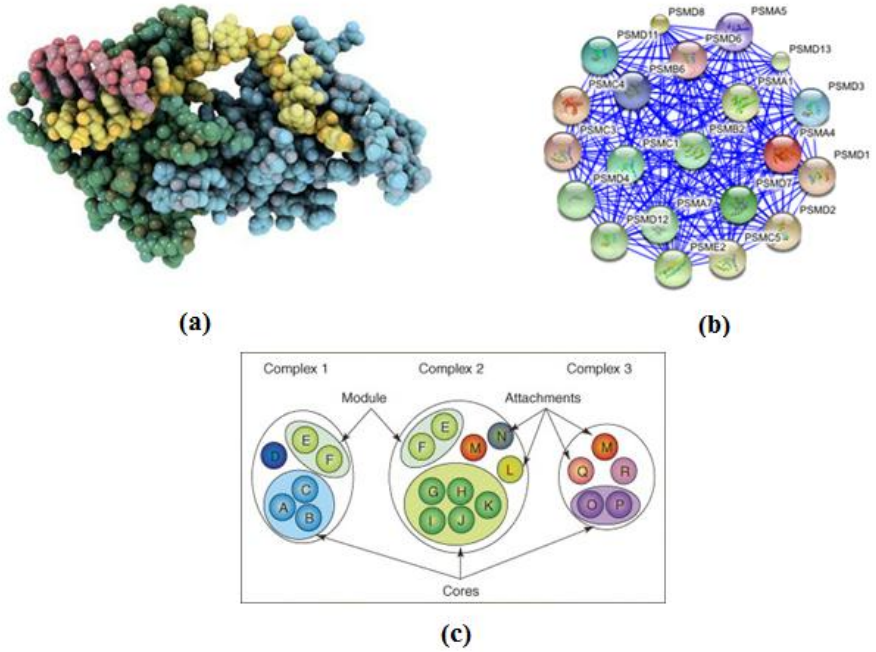


Figure 1.2 – (a) a picture of protein complex, (b) a graph representation of a protein complex.(c) core-attachment structure of protein complexes.

One of the biggest difficulties for computational methods to detect protein complexes from PPI networks is that there is no mathematical definition for protein complexes but the observation that proteins within a complex interact closely with each other (figure 1.2a).

Henceforth, computational biologists usually use an early accepted model of protein complexes as dense (or clique-like) subgraphs (figure 1.2b) and aims to seek for dense regions in the PPI networks as protein complex candidates. Typical complex detection methods that are based on graph clustering are: MCODE [Bader et al., 2003], MCL [van Dongen et al., 2000], CMC [Liu et al., 2009], HACO [Wang et al., 2009].

It is also known that protein complexes have a core-attachment structure [Gavin et al., 2006], in which cores are the stable parts of complexes, they keep recruiting attachment proteins to help perform specific functions. Among attachment proteins, there are instances where two or more proteins are always together, which are called 'modules' (figure 1.2c). Also, attachment proteins were seen to be shared between two or more complexes, thereby exemplifying the view that the same protein may participate in multiple complexes [Pu et al., 2007; Wang et al., 2009]. Typical complex detection methods incorporating core-attachment structure are CORE [Leung et al., 2009], COACH [Wu et al., 2009], MCL-CAw [Srihari et al., 2010]. For a complete literature survey on computational methods for predicting protein complexes from PPI networks, please refer to the recent papers [Li et al., 2010] and [Srihari et al., 2013].

Existing complex predicting methods have to face the difficulties in dealing with highly noisy interaction data (high false positive and false negative rates) and also low overlap between different data sources. Therefore, existing computational complex predicting methods still cannot have a complete coverage of known protein complexes. Shared proteins between multiple complexes in PPI networks also hinder graph-clustering based complex detection methods.

Current protein complex detection methods (all approaches) also rarely have 100% match for each detected complex, this hinders the comparisons between any two detected complexes from two species to identify the conserved pairs. Due to the above obstacles, protein complex detection from original PPI networks are still not an optimal approach for identifying conserved protein complexes among species.

1.1.3. Why do we need comparative interactomics and conserved protein complexes?

One of the most important reasons behind the searching for conserved biological entities between species is that: conservation implies functional significance. This accounts for the

birth of *comparative genomics* to identify proteins whose functions are conserved among species. While sequence-conserved proteins form the basis of *comparative genomics*, it is also very important to consider the conserved patterns of interactions between proteins themselves, which can be referred to as *comparative interactomics* [Kiemer et al., 2007]. The reason here is that comparing interactomes among different species helps to transfer biological knowledge and function annotation at a higher level than comparing only protein sequences.

Conserved protein complexes and functional modules is one of the main outcomes from solving comparative interactomics problems. Identifying conserved complexes between species is a fundamental step towards identification of conserved mechanisms from model organisms to higher level organisms, such as protein translation, DNA transcription, cell cycle, etc. These mechanisms, at the same time, are considered as back-bones for a unit living system as cell. Therefore, conserved protein complexes are highly related to core cellular processes and critical to be studied carefully.

Another advantage supporting the comparative interactomics approach is that despite the noises in data, comparative analysis helps us to use the cross-species conservation criteria to focus on the more reliable parts of protein interaction networks and infer likely functional components. Once the number of well-studied species increases, we can use this approach to guide the search for protein complexes in newly-sequenced species, thereby increase the precision of current computational protein complex predicting methods.

Identifying conserved protein complexes can also help to understand the evolutionary mechanisms of protein complexes and protein interaction networks between multiple species, such as deriving evolutionary rate and age measures for protein complexes [Yosef et al., 2009].

In summary, the generalization from finding orthologous proteins to *orthologous protein complexes* [Yosef et al., 2009] is a significant extension.

1.2. Research objectives

Due to the significance of detecting conserved protein complexes between species, and the fact that current protein complex detecting methods still cannot undertake this task, we now need an effective method for this purpose. There also exist methods specialized for

detecting conserved protein complexes, but most of them use only BLAST score for the whole protein sequence to decide which pairs of proteins between two species are considered to be conserved (see Chapter 3 for details). This can severely limit the number of protein pairs that are actually conserved in function. Identifying function-conserved proteins in this case is important because it serves as a corner-stone for predicting conserved protein complexes. For species that have far evolutionary distances, the above limitation causes a serious mistake because in these cases, their proteins have evolved many-fold in complexity, so simple BLAST scores for whole-sequence similarity may not be able to capture these complicated evolutionary processes. Henceforth, we also need an effective method in this aspect. Due to these research objective, the key contributions of this thesis are featured as follows.

1.3. Contributions of the thesis

1. *A survey on computational methods for identifying conserved protein complexes between species*: in this survey, computational methods for identifying conserved protein complexes are grouped into two classes, each uses a different approach. For each approach, a typical method is described in details, and the other methods are briefly described. Connections between methods and comparisons between the two approaches are also shown. Furthermore, a short summary on ortholog assignment methods is also presented due to its significance in the computational pipeline for identification of conserved protein complexes.

2. *A novel method for identifying conserved protein complexes by constructing interolog networks*: This method is novel in terms of: (i) employing an innovative and effective framework for detecting conserved protein complexes; (ii) hypothesizing an evolutionary mechanism among protein complexes that integrates protein domain information. Our experiments on yeast and human datasets revealed that our method can identify considerably more conserved complexes than plain clustering of the original PPI networks. Furthermore, we demonstrated that integrating domain information generates many-to-many ortholog relationships which significantly enhances the interolog network quality and throws further light on conservation of mechanisms between yeast and human.

3. *A gold standard dataset for conserved protein complexes between human and yeast*: By proposing a score to measure the conservation level between protein complexes, a collection of conserved complexes pairs between yeast and human is built and considered as a gold

standard dataset during this work. As currently there is no benchmark dataset for conserved protein complexes between human and yeast in the literature, the author hopes that this dataset could be useful for reference. Furthermore, this step also gives us a detailed examination on the conservation level between manually curated protein complexes of human and yeast.

1.4. Organization of the thesis

This chapter has briefly described the background and motivation, and outlined the research objectives of this work. The remainder of this thesis is organized as follows. Chapter 2 first gives the definition for the problem of identifying conserved protein complexes between species from protein interaction data, then presents the general computational pipeline to solve this problem. This pipeline includes the preparation for experimental data; a brief survey on ortholog assignment methods for defining conserved proteins; and protein complex detection from all the input data. Chapter 3 will survey existing methods specialized for detecting conserved protein complexes and functional modules from protein interaction data. The two main approaches presented are network alignment and network querying, which have interesting computational properties. Chapter 4 features the main contribution of this thesis, which designs a novel method for mining conserved protein complexes from the interolog network built from the two species' PPI networks. Chapter 5 concludes the work by figuring out the main contributions, limitations and recommendations for further research.

Chapter 2 - The problem of identifying conserved protein complexes from PPI data

2.1. Problem definition

The problem of identifying conserved protein complexes can be described as follows:

Given a PPI network and a collection of manually curated protein complexes of a well-studied species, a PPI network of a new species (the interaction data of this species might be far from complete, and both of the networks can contain many noisy interactions), and the homology information between the two species. How can we predict protein complexes in the new species that are conserved in the well-studied species? Conservation of protein interaction sub-networks is measured in terms of similarity in protein function (node similarity) and similarity in interaction patterns (network topology similarity).

Figure 2.1 below illustrates a pair of conserved protein complex between a well-studied species as yeast and a newly sequenced species as human. For species that have a far evolutionary distance as human and yeast, many cellular mechanisms, though conserved in function, have in fact evolved many-fold in complexity. Consequently, the similarity in composition of the conserved protein complexes between these species is not expected to be

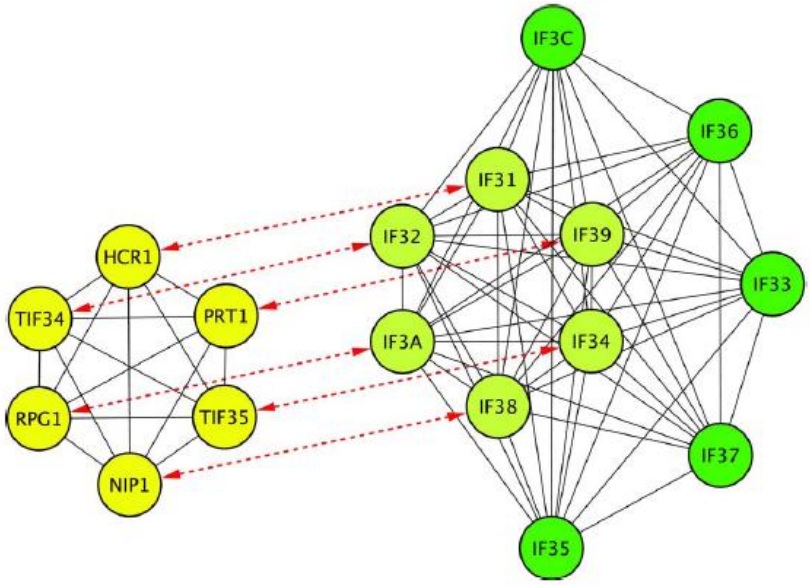


Figure 2.1 – An example about human (right) and yeast (left) Eukaryotic initiation factor (eIF3) complex.

very high, on the contrary, there might be a high portion of difference (in terms of insertions/deletions of proteins) in these pairs of protein complexes. Therefore, an efficient method for predicting conserved protein complexes from PPI networks needs to be able to recognize the evolutionary mechanisms responsible for the difference part of the two conserved protein complexes.

2.2. The computational pipeline

In order to carry on identifying conserved protein complexes between species from PPI data, we first need to gather physical protein interactions of the two species from various datasets and experiments to enhance the coverage of true positive interactions. Manually curated protein complexes (if available) of the well-studied species are also collected to aid predicting conserved complex in the other species. The second key step in this computational pipeline is to define the correspondence of function similarity between the two set of proteins, each from one species. This step is usually deemed to be identical to the task of ortholog assignment. And finally, when the input data is available, we need a method to detect conserved protein complexes from these data, followed by an evaluation for the resulting complexes.

2.2.1. Experimental data

Many high-throughput techniques have been developed over the last decade to detect protein interactions on a genome-wide level not only in yeast, the following are the two typical techniques among them:

Yeast two hybrid (Y2H) [Uetz et al., 2000; Ito et al., 2001]: is a screening technique for physical protein-protein and protein-DNA interactions which takes place in a living cell of yeast (*in vivo*). The two proteins of interest are injected into a genetically engineered strain of yeast. If they physically interact, a reporter is transcriptionally activated and we get a colour reaction on specific media. This technique is low-cost but can be degraded by a high number of false positive (as well as false negative) detections (about 70% false positive rate as in [Deane et al., 2002]) and a low overlap rate between the two experiments (only 20% as in [Shoemaker, 2007]).

Tandem affinity purification combined with **mass spectrometry (TAP-MS)** [Gavin et al., 2006; Krogan et al., 2006]: is an *in vitro* technique, which has two steps: in the TAP stage, the protein of interest is embedded in a cell lysate to act as a bait for its interact-able proteins (prey) to bind, then together they will be identified by mass spectrometry after washing out the contaminants. Although TAP-MS technique still has a large number of false positive interactions and miss a lot of known interactions as Y2H, it can report higher-order interactions as protein complexes while Y2H has an advantage of detecting transient interactions [Shoemaker et al., 2007].

As an inherent weakness of high-throughput techniques, protein interaction data generated by these techniques contains a large number of false positives. For this reason, PPI scoring methods are invented to assess the reliability of each interaction in the PPI network. Some typical PPI scoring methods are: FSweight [Chua et al., 2006], Iterative-CD [Liu et al., 2008], which use solely the PPI network topology to evaluate the reliability of PPIs and predict new interactions; TCSS [Jain et al., 2010] uses semantic similarity within gene ontology of proteins to score PPIs.

For manually curated protein complexes, the two famous databases providing wet-lab experiments and verification are: Wodak Lab CYC2008 [Pu et al., 2007, 2008], which is for yeast, and CORUM [Ruepp et al., 2008, 2009], which is for mammalian species. Other typical databases for manually curated protein complexes include: MIPS [Mewes et al., 2006], Aloy [Aloy et al., 2004] for yeast, and Emililab [Havugimana et al., 2012] for human.

2.2.2. Ortholog assignment

Ortholog assignment takes a key role in this work because it defines the correspondence of function similarity between the two set of proteins of the two species, which is the corner stone for identifying protein complexes with function similarity. Orthology prediction methods can be grouped into three main classes: “graph-based”, “phylogenetic tree-based” and “synteny based”. It would be a large topic to talk about ortholog identification methods. At the scope of this thesis, only a brief summary with very popular methods for orthology inferring, some of which were used throughout this work, are mentioned.

Graph-based methods perform pair-wise gene/protein sequence comparisons between whole genomes, typically using all-versus-all BLAST. A weighted graph is then constructed with genes as nodes and sequence similarity scores as weights. Finally, various graph

clustering techniques are used to identify homolog groups. COGs [Tatusov et al., 2003], Inparanoid [O'Brien et al., 2005], OrthoMCL [Li et al., 2003] belong to this class.

Phylogenetic tree-based methods have the first stage similar to graph based methods, in which homolog groups are identified. For each of these homolog groups, a gene tree are built from multiple sequence alignments of homologs. These gene trees are then analyzed and reconciled with a trusted species tree to localize speciation and duplication events, which is the basis for differentiating orthologs from paralogs. For these details in analysis, many studies have shown that phylogenetic methods have greater precision than graph-based methods [Chen et al., 2007]. Typical examples of phylogenetic methods are EnsemblCompara [Vilella et al., 2009], PHOG [Datta et al., 2009].

Synteny based methods use the information of synteny blocks. This is based on a property that an ortholog pair is usually surrounded by many others, or ortholog pairs tend to locate closely to each other on the two genomes to collaborate in specific conserved functions. This fact is reflected in typical examples as operons in prokaryotes and conserved gene clusters in eukaryotes. Some instances of methods in this class are MSOAR2 [Shi et al., 2009] and BBHLS [Zhang et al., 2012], in which sequence similarity is combined with gene context similarity.

In many existing methods for identifying conserved protein complexes, function similarity between proteins were measured by using BLAST score only ([Sharan et al., 2005], [Flannick et al., 2006], [Sharon et al., 2009]). This severely restricts the number of actual proteins whose functions are conserved. The following is one of the approaches that can overcome this weakness.

Orthology prediction considering protein domain similarity:

There are circumstances under which a domain-based phylogeny may be preferable to one that is based on whole-sequence similarity. First, the requirement that orthologs have to be aligned well over their entire lengths – neither much longer nor shorter – might be overly restrictive. This is because there are cases when species have far evolutionary distances, their othologs have evolved many-fold in complexity so that only their functional and structural domains – which are the parts that directly perform functions – are similar to each other. Secondly, existing methods for ortholog identification are usually based on BLAST, a local alignment protocol, which is not designed to distinguish between sequences sharing a

common domain architecture and those having only local matches. This may increase the potential for annotation errors.

For these reasons, there are some ortholog assigment methods consider protein domain similarity in the process of inferring functional similarity. Those include Ensembl orthology [Vilella et al., 2009] and PHOG [Datta et al., 2009].

2.2.3. Protein complex detection from PPI networks

Protein complex detection is the final stage in the computational pipeline for identifying conserved protein complexes, when all input data (PPI data of the two species, manual curated protein complexes, homology information) are ready. The recent literature surveys for computational methods for protein complex prediction are done in [Li et al., 2010] and [Srihari et al., 2013].

This part aims to focus on standard methods that are based on graph clustering for complex detection. While these methods proposed effective framework for mining protein complexes from protein interaction data, and some of which has reached the state-of-the-art performance compared to other approaches, the approach of modeling protein complexes as dense sub-graphs faces difficulty in having radical detection of complexes from original PPI networks due to the following facts. First, protein interaction datasets, especially for newly sequenced species as human, still contain substantial number of noisy interactions. This will break out the protein complex model. Secondly, in a PPI network, especially of multi-cellular species, each protein does not necessarily participate in all its known interactions simultaneously (as shown in [Liu et. al., 2011]). In other words, each protein can participate in many different complexes (shared attachment proteins is an example [Gavin et al., 2006]), so if using only the PPI network, it is difficult to know which subset of interactions take place together in a same complex. These factors can cause graph clustering based methods in missing many true complexes, many of which involve in core cellular processes that are conserved among species [Nguyen et al., 2013]. Some typical methods in this class are: MCODE [Bader et al., 2003], MCL [van Dongen et al., 2000], CMC [Liu et al., 2009], HACO [Wang et al., 2009].

Resulting complexes are subjected to a matching with manually curated protein complexes for evaluation. Current protein complex detection methods (all approaches) also rarely get 100% matched for each detected complex, this also hinders the comparisons

between any two detected complexes from two species to identify the conserved pairs. Due to the above obstacles, protein complex detection from original PPI networks are still not an optimal approach for identifying conserved protein complexes among species.

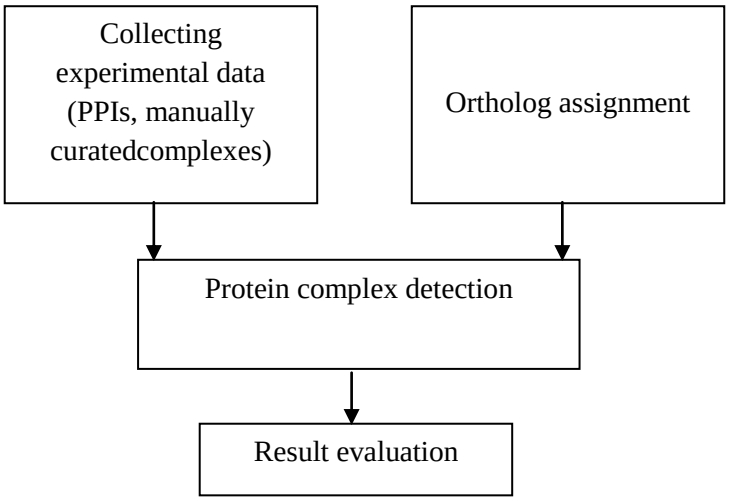


Figure 2.2 – The computational pipeline for identifying conserved protein complexes.

2.2.4. Result evaluation for conserved protein complexes

Detected conserved protein complexes need a benchmark dataset to be matched with. If there are no such datasets in the literature, we have to build one. Usually, for building a testing dataset for conserved protein complexes, we have to devise a model for protein complex conservation, or a score to measure the conservation level of two given protein complexes. We then apply this score to every pair of complexes that we need to check if they are conserved.

Chapter 3 – Computational methods for identifying conserved protein complexes

In general, there are two approaches for solving the conserved protein complexes from PPI networks, one compares the two whole PPI networks of the two corresponding species by aligning similar nodes and edges then searching for potential regions in the alignment network that could be conserved, which is called the local network alignment approach. Another approach uses information from the known protein complexes of a well-studied species then matches them to the PPI network of a new species to identify subnetworks that have similar shapes to the query complexes. Thus, the second approach is called network querying. Detailed descriptions for these two approaches are given in the following sections.

3.1. Local network alignment approach

Analogous to sequence alignment, network alignment is to measure the similarity between two networks by finding the best way to fit one network into the other. As for sequence alignment, there also exist local and global network alignments. Global network alignment searches for a unique alignment from every node in the smaller network to exactly one node in the larger network, even though this may lead to inoptimal matchings in some local regions. Because of this, global network alignment is aimed for discovering the common network topological properties that are preserved between the two networks. Several different formulations of the global network alignment problem have been proposed ([Flannick et al., 2008; Liao et al., 2009; Zaslavskiy et al., 2009]). On the other hand, local alignments look at small similar sub-networks between the two networks, thus aiming to identify pathways or protein complexes conserved in PPI networks of different species. By this, a node (or a sub-network) from one network can be mapped to many nodes (or many sub-networks) in another network. That is why this section is dedicated for local network alignment.

3.1.1. Problem definition and general solution framework

If a PPI network is represented by an undirected graph G(V, E), where V denotes the set of proteins, and (u, v) ∈ E denotes an interaction between proteins u, v ∈ V, then the local network alignment problem can be informally stated as follows:

Local network alignment problem: given k different PPI networks of k different species, how can we find conserved sub-networks between these networks?

In other words, a local network alignment is defined as a set of sub-networks chosen from the interaction networks of different species, together with a (label) mapping between corresponding (or aligned) proteins. To get an alignment uniquely specified, we require that the mapping is an mathematical equivalence relation. Consequently, the groups of aligned proteins are disjoint, and we refer to them as equivalence classes. Each of these classes can be called a protein family (or be usually referred to as a homology group), which represents a particular protein function. By this, a biological interpretation of an alignment is a collection of proten families whose interactions are conserved across a given set of species.

Generally, in order to find these conserved sub-networks, we have to build an *alignment graph* (or *orthology graph*), in which each of its nodes represents k sequence-similar (homologous) proteins (each protein belongs to a different species), and each edge represents a conserved interaction between k species.

When the number of species is 2 (k =2), this problem is called pair-wise network alignment. For the purpose of simplicity, henceforth, we will imply pair-wise network alignment when using the term network alignment. Figure 3.1 below gives a simple example of pair-wise network alignment.

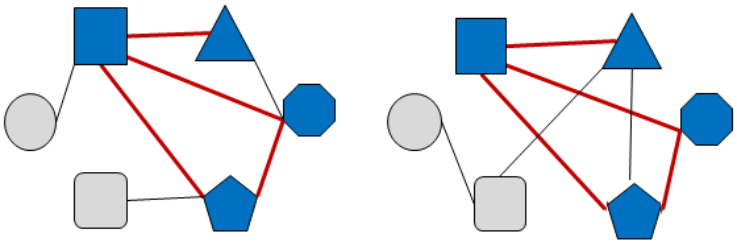


Figure 3.1 - A simple example for pair-wise network alignment, in which nodes having the same shape are considered as sequence-similar. Conserved sub-networks have thick edges.

With the purpose of applying network alignment to find conserved protein complexes from PPI networks, network alignment problem is extended to allow a limited number of

mismatches w.r.t. nodes and edges in the resulting subgraphs, some limited number of insertions/deletions of nodes.

General solution framework: a general framework for applying network alignment to identify conserved protein complexes can be illustrated in figure 3.2, where the first stage is defining a protein complex model in which every sub-network that satisfies this model will have a high chance being a true protein complex. The model accuracy is highly dependent on how good the knowledge (represented in terms of graphs) we use to define a protein complex. The second step is to devise a definition for protein complex conservation using the protein complex model of each species. This stage takes into account the homology information between the protein sets of the two corresponding species to build a so-called *alignment graph* (or *orthology graph*), which will be used for the searching stage afterwards.

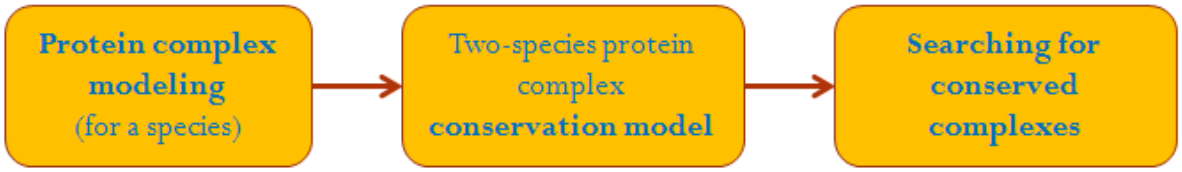


Figure 3.2 – A general solution framework for identifying conserved protein complexes using network alignment.

When the alignment graph is built, the problem of identifying conserved protein complexes will be equivalent to finding heavy subgraphs (in terms of node weight and edge weight) in the alignment graph. Moreover, the problem of searching for induced heavy subgraphs in a graph is NP-hard even when considering a single species where all edge weights are 1 or -1 and all vertex weights are 0 [Shamir et al., 2004]. Thus a heuristic is employed for searching the alignment graph for conserved protein complexes.

In this section, we will look at NetworkBLAST [Sharan et al., 2005a; Sharan et al., 2005b] as a typical method that bases on the above solution frame work for network alignment, other methods are usually variants of this.

3.1.2. NetworkBLAST [Sharan et al., 2005a; Sharan et al., 2005b]

This method is to find conserved protein complexes by comparative analysis of two PPI networks, it assumes that proteins in a protein complex should be highly connected within themselves to help them act as a single organization. Thus a protein complex can be

represented in the form of a dense subgraph (clique-like). In order to evaluate how likely a subset of proteins can form a protein complex, and how statistically significant it is, a probabilistic model for protein complexes is devised as follows.

A probabilistic model for protein complexes:

At a top-down view, the complete protein complex model is a log likelihood ratio which is defined for each subset *U* of proteins to measure how likely they form a true complex (let us call it the *complex likelihood*):

$$L(U) = \log \frac{\Pr(O_U \mid M_c)}{\Pr(O_U \mid M_n)} \tag{3.1}$$

In this formula, O_U is the observation of all interactions within *U*; $\Pr(O_U \mid M_c)$ is a likelihood that measures how likely we can observe O_U given the complex model M_c (M_c represents for the fact that U is within a complex). The complex model M_c assumes that every two proteins in a complex interact with a high probability p (0.95 is used in this work). In terms of the graph, the assumption is that two vertices that belong to a same complex are connected by an edge with probability p, independently of all other pair-wise interactions and all other information.

In order to have a high chance becoming a true protein complex, a subset of proteins U with its observed interactions O_U need also to be statistically significant, and $\Pr(O_U \mid M_n)$ measures this quantity. In fact, this is the p-value for O_U in the null model M_n. The random model M_n assumes that each edge is present with the probability that one would expect if the edges of G (the graph that represents the PPI network) were randomly distributed but respected the degrees of the vertexes, which means edges incident to vertexes with higher degrees have higher probability. More precisely, let F^G represents the family of all graphs having the same vertex set as G and the same degree sequence. The probability of observing the edge (u, v) is defined to be the fraction of graphs in F^G that include this edge.

Given the assumption that all pair-wise interactions are independent, the log likelihood function in (3.1) can be decomposed into the log likelihood ratio for individual protein pairs as:

$$L(U) = \sum_{(u,v)\in U\times U} \log \frac{\Pr(O_{uv} \mid M_c)}{\Pr(O_{uv} \mid M_n)} \tag{3.2}$$

where $\Pr(O_{uv} \mid M_c) = \Pr(O_{uv}, T_{uv} \mid M_c) + \Pr(O_{uv}, F_{uv} \mid M_c)$ (law of total probability)

$$= \Pr(O_{uv} \mid T_{uv}, M_c) \Pr(T_{uv} \mid M_c) + \Pr(O_{uv} \mid F_{uv}, M_c) \Pr(F_{uv} \mid M_c)$$

$$= \beta \Pr(O_{uv} \mid T_{uv}) + (1-\beta) \Pr(O_{uv} \mid F_{uv}) \quad (3.3)$$

(O_{uv} and M_c are conditionally independent, $\beta = \Pr(T_{uv} \mid M_c)$)

T_{uv} (and F_{uv}) is the event that protein u truly interact (and not interact) with protein v; β is the probability that any two proteins u and v interact with each other in the complex model M_c.

Similarly, $\Pr(O_{uv} \mid M_n) = p_{uv} \Pr(O_{uv} \mid T_{uv}) + (1-p_{uv}) \Pr(O_{uv} \mid F_{uv})$ (3.4)

where here, as mentioned in the description of the null model M_n above, p_{uv}= $\Pr(T_{uv}|M_n)$ depends on the degrees of u and v. Hence, from (3.3) and (3.4), the log likelihood function in (3.2) can be rewritten as follows:

$$L(U) = \sum_{(u,v)\in U\times U} \log \frac{\beta \Pr(O_{uv} \mid T_{uv}) + (1-\beta) \Pr(O_{uv} \mid F_{uv})}{p_{uv} \Pr(O_{uv} \mid T_{uv}) + (1-p_{uv}) \Pr(O_{uv} \mid F_{uv})}$$

$$= \sum_{(u,v)\in U\times U} \log \frac{\beta \Pr(T_{uv} \mid O_{uv}) + (1-\Pr(T_{uv})) + (1-\beta)(1-\Pr(T_{uv} \mid O_{uv})) \Pr(T_{uv})}{p_{uv} \Pr(T_{uv} \mid O_{uv}) + (1-\Pr(T_{uv})) + (1-p_{uv})(1-\Pr(T_{uv} \mid O_{uv})) \Pr(T_{uv})} \quad (3.5)$$

(after applying Bayes's rule and cancelling common terms in the numerator and denominator)

So far, the log likelihood ratio can be calculated from: $\Pr(T_{uv} \mid M_c)$ or β, the probability of a truly interaction in the complex model, which is set manually in this work as 0.95; $\Pr(T_{uv} \mid M_n)$ or p_{uv}, the probability of an interaction if the edges are randomly distributed but respected the degree of vertexes, which can be estimated by Monte Carlo estimation; $\Pr(T_{uv} \mid O_{uv})$, the reliability of the interaction between u and v, estimated by using a PPI network scoring method; $\Pr(T_{uv})$, the prior probability that two random proteins interact.

Two-species protein complex conservation model:

Consider two subsets of proteins U^1 from species 1 and V^2 from species 2, and a many-to-many mapping $\theta : U^1 \to V^2$ between them. Then the likelihood score that measures how likely the 2 subsets of proteins are complexes can be computed as follows (let us call it the *concurrent complex likelihood*),

$$L(U^1,V^2)=\log\frac{\Pr(O_{U^1}\mid M_c^1)}{\Pr(O_{U^1}\mid M_n^1)}+\log\frac{\Pr(O_{U^2}\mid M_c^2)}{\Pr(O_{U^2}\mid M_n^2)} \tag{3.6}$$

which is the sum of the two corresponding complex likelihoods, each in one species. In order to get a conservation score of these two subsets of proteins, we have to take into account the sequence conservation among the pairs of proteins defined by θ, which assigns orthologous pairs between U^1 and V^2. Thus here, we need to define a so-called *homolog likelihood*, which measures how likely the two proteins u and v are homologs. This log likelihood ratio is also in the form of ratio between the likelihoods under the conserved complex model and the null model as follows:

$$H(u,v)=\log\frac{\Pr(E_{uv}\mid M_c)}{\Pr(E_{uv}\mid M_n)}$$

$\Pr(E_{uv}\mid M_c)=\Pr(E_{uv}\mid h_{uv})$: under the conserved complex model, u and v must be homologs;

$$\begin{aligned}\Pr(E_{uv}\mid M_n)&=\Pr(E_{uv},h_{uv}\mid M_n)+\Pr(E_{uv},\bar{h}_{uv}\mid M_n)\\&=\Pr(E_{uv}\mid h_{uv},M_n)\Pr(h_{uv})+\Pr(E_{uv}\mid \bar{h}_{uv},M_n)\Pr(\bar{h}_{uv})\\&=\Pr(E_{uv}\mid h_{uv})\Pr(h_{uv})+\Pr(E_{uv}\mid \bar{h}_{uv})\Pr(\bar{h}_{uv})\end{aligned}$$

(E_{uv} and M_n are conditionally independent.)

Using Bayes's rule, a simpler formula for the homolog likelihood can be derived as:

$$H(u,v)=\log\frac{\Pr(h_{uv}\mid E_{uv})}{\Pr(h)} \tag{3.7}$$

where E denotes the BLAST E-value between u and v; $\Pr(h_{uv}|E_{uv})$ is the probability that u and v are homologs given their BLAST E-value, this probability was calculated as in [Kelly et al., 2003]

Finally, the complete complex conservation score is formed as the sum of the concurrent complex likelihood $L(U^1, V^2)$ and the sum of homolog likelihood on all homolog pair between U and V. The first term measures how likely the two subsets of proteins U and V are true complexes in the two corresponding species while the second term measures how likely all homolog pairs assigned by θ are truly homologs.

$$S_\theta(U^1,V^2)=L(U^1,V^2)+\sum_{u\in U^1}\sum_{v\in\theta(u)}H(u,v) \tag{3.8}$$

Searching for conserved protein complexes:

After the complex model and complex conservation model are built, the problem of identifying conserved protein complexes reduces to the problem of identifying a subset of proteins in each species, and a correspondence between them, such that the complex conservation score S exceeds a threshold. In order to facilitate the search on all possible pairs of subsets U and V of proteins (each from one species) to test whether they are conserved complexes, a concept of orthology graph (or alignment graph) is introduced.

Let $G_1(E_1, V_1)$ and $G_2(E_2, V_2)$ be PPI networks of the two corresponding species, then the orthology graph $OG(E_{OG}, V_{OG})$ is built as follows:

Each node in V_{OG} is a pair (u, v) of proteins where $u \in V_1$ and $v \in V_2$.

Edges in OG connect all possible pairs of nodes. In other words, OG is a complete graph.

Each edge that connects two nodes (u_1, v_1) and (u_2, v_2) in OG has two weights: $w_1 = L_1(\{u_1, u_2\})$; $w_2 = L_2(\{v_1, v_2\})$, where L is the complex likelihood in (2), in this case, it measures how likely (u_1, u_2) and (v_1, v_2) form two co-complex relationships in the two corresponding species.

Each node (u, v) in OG has a weight that is the homolog likelihood between them, w(u, v) = H(u, v).

Figure 3.3 is an illustration of a node and an edge with two weights in the orthology graph. In this sense, if we can enumerate all possible subsets of nodes in OG, then those are all possible pairs of subsets U, V of nodes (each from one species).

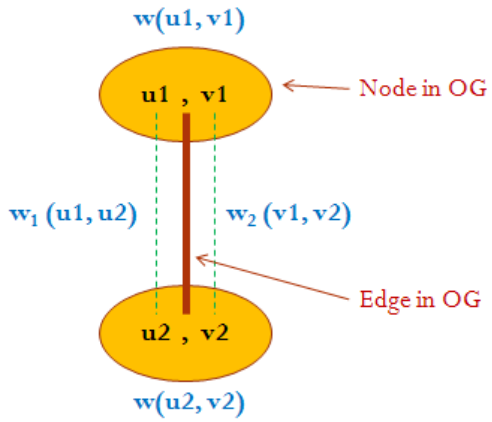


Figure 3.3 – An illustration of two nodes and their edge in the orthology graph.

Basing on the orthology graph, the problem of identifying a subset of protein in each species, and a correspondence between them, such that the complex conservation score is high, is equivalent to finding heavy subgraphs in the orthology graph. This is an NP-Hard problem, because it is reduced from the maximum clique problem. Thus a heuristic for searching was proposed as follows:

Compute a seed around each node v, which consists of v and all its neighbors u such that (u, v) is a strong edge.

If the size of this set is above a threshold (e.g. 10), iteratively remove from it the node whose contribution to the subgraph score is minimum, until we reach the desired size.

Enumerate all subsets of the seed that have size at least 3 and contain v. Each such subset is a refined seed on which a local search heuristic is applied.

Local search: Iteratively add a node, whose contribution to the current seed is maximum, or remove a node, whose contribution to the current seed is minimum, as long as this operation increases the overall score of the seed. Throughout the process, the original refined seed is preserved and nodes are not deleted from it.

For each node in the alignment graph, record up to k (e.g. 5) heaviest subgraphs that were discovered around that node.

Note that because the orthology graph is a complete graph, at any time, a constructed subgraph is also a clique. The resulting subgraphs may overlap considerably, thus a greedy algorithm is used to filter subgraphs whose percentage of intersection is above a threshold as follows:

Iterative find the highest weight subgraph.

Add that subgraph to the final output list.

Remove all other highly intersecting subgraphs.

Pruning the orthology graph:

In order to reduce the complexity of the graph and focus on potential conserved complexes, nodes with low homolog likelihood are removed from the graph. They are considered back only they satisfy the following condition: for every node $(p, y) \notin S$, we check whether there exist two nodes (p_1, y_1), $(p_2, y_2) \in S$ such that p interacts with p_1 and p_2,

and y interacts with y_1 and y_2. In this case, (p, y) serve as "bridges" in the orthology graph between protein pairs, whose members in each species are not known to directly interact.

Experimental results:

This method was experimented on yeast and bacterial data, it found 11 correct conserved protein complexes between these two species with the evaluation based on complex functional annotation. However, there was no benchmark data for estimating the sensitivity of the results.

3.1.3. Other local network alignment based methods

MaWIsh local network alignment method [Koyuturk et al., 2006] is based on the duplication/divergence models that focus on understanding the evolution of protein interactions. It constructs a weighted global alignment graph and tries to find a maximum induced sub-graph in it. Graemlin algorithm [Flannick et al., 2006] scores a possibly conserved module between different networks by computing the log-ratio of the probability that the module is subject to evolutionary constraints and the probability that it is under no constraints, taking into account the phylogenetic relationships of the species whose networks are being aligned. [Hirsh et al., 2007] also developed their own protein complex evolution model basing con protein interaction attachment/detachment and gene duplication events, then employed it to identify conserved protein complexes between yeast and fly. [Zhenping Li et al., 2007] formulate the local network alignment as an integer quadratic programming problem and then transform this into a quadratic programming problem, which almost always ensures an integer solution, thereby making the local network alignment problem tractable without any approximation.

3.2. Network querying approach

3.2.1. Problem definition

If we already have a list of known protein complexes, then it would be a natural thinking to match these complexes to a new species' PPI network for predicting conserved protein complexes, rather than aligning the whole two PPI networks and make no use of known

protein complex information in the well-studied species. The network querying problem can be stated as follows:

Network querying problem: given a query subnetwork G^Q and a target network G^T, how can we find subnetworks in G^T that are similar to G^Q? Similarity here is in terms of both node label and network topology.

Also, more general and suitable for identifying conserved protein complexes, insertion of proteins into the matched subnetwork, or deletion of vertices from the query subnetwork, as well as a limited number of mismatches, are allowed.

In this section, we will describe a typical method of network querying for identifying conserved protein complexes, Torque (TOpology-free netwoRk QUErying) [Bruckner et al., 2010].

3.2.2. Torque – Topology-free network querying [Bruckner et al., 2010]

"Topology-free" here means we only use the set of involved proteins of each query subnetwork and do not care about its topological information. The motivation of this work is that most of the protein complexes reported in the literature do not provide any information about their interaction patterns. Thus, Torque aims to find a connected component of proteins in the target network that matches the query set of proteins. This work first gives a formulation for the topology-free network querying and then devise three solutions to the problem those are: randomized dynamic programming, integer linear programming (ILP) solver (after formulating the network querying problem as an ILP problem), and a shortest-path based heuristic. In order to present the formulation for the problem, we firstly need to define a concept called *colorful*.

Let G= (V, E) be a PPI network where vertices represent proteins and edges correspond to PPIs. Given a set of color (1, 2, …, k), a *coloring constraint function* $\Gamma: V \rightarrow 2^C$ that assigns each vertex $v \in V$ a subset of colors of C (we can call this is the *color set* of v). For any subset S of C, we define a subset of vertices H of V as *S-colorful* if |H| = |S| and each vertex v in H can selected one color in its color set that is distinct from the selections of the other vertices in H.

Then the *topology-free network querying* problem can be formulated as a *C-colorful connected subgraph* basing on the colorful concept as follows.

C-colorful connected subgraph problem: Given a graph G = (V, E), a color set C, and a coloring constraint function Γ: $V \rightarrow 2^C$, is there a connected subgraph of G that is C-colorful?

This problem is corresponding to the topology-free network querying problem as follows: suppose we have a query complex with C proteins, if we assign each protein in this complex a distinct color (even if this protein has paralogs in this complex), then we have the color set C. If a protein in the target network G is orthologous with a protein in the complex, it will put the color of this protein complex into its color set. Thus, one protein in G can have multiple colors in its color set when it is orthologous with more than one protein complex. Therefore, if there is a connected subgraph of G that is C-colorful, then its node set will have the same set of protein families (or homolog groups), and each family has the same number of paralogs as the complex. And this subgraph is considered as a conserved protein complex of the query one.

We also can find another formulation for this problem that is somehow simpler to visualize as follows:

Let the query complex be a *multiset* M of colors in which each color represents a biological protein function. Thus, paralogs in this complex will have the same color. Then the problem is: does G have a connect subset of vertices whose multiset of colors equals M? (Note: two multisets are defined to be equal if they have the same multiplicity (number of occurrences) of each element).

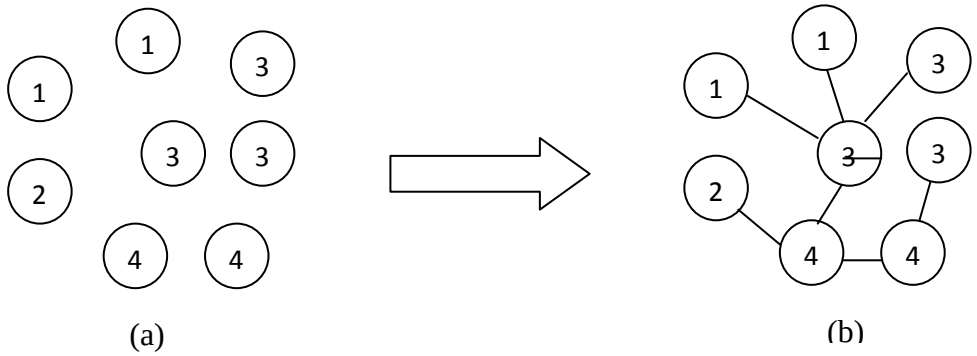


Figure 3.4 – An illustration for the query set of proteins (a) and its matched connected subgraph (b) in the target network, each number label represents a color. The multisets of colors, which represent multisets of biological protein function, in (a) and (b) are equal.

With the topological-free network querying problem defined above, Torque designs three approaches for solution:

Randomized dynamic programming approach:

This approach is used for firstly considering only coloring constraint functions that associates each vertex v ∈ V with a single color. Then the problem is to find a connected subgraph that has exactly one vertex of each color in the query protein complex. Since every subgraph has a spanning tree, this approach looks for colorful trees. A dynamic programming table B is constructed with rows corresponding to vertices and columns corresponding to subsets of colors. B(v, S) = true if there exists in G a subtree rooted at v that is S-colorful, and B(v, S) = false otherwise. As initialization, when S has a single color c and v∈ V we initialize B(v, c) = true iff the color set associated with v contains only c. Other entries of B can be computed using the following recurrence:

$$B(v,S) = \bigvee_{\substack{u \in N(v) \\ S_1 \cup S_2 = S \\ \Gamma(v) \in S_1, \Gamma(u) \in S_2}} B(v,S_1) \wedge B(u,S_2)$$

(N(v) is neighbor nodes of v)

This algorithm runs in $O(3^k m)$ time and can be generalize to the case of weithted graph by searching for heaviest colorful subtree rooted at each vertex and B(v, S) is a real number instead of a Boolean value. The weight of an optimum match is given by $\max_v$ B(v, C) and the recursion is modified as:

$$B(v,S) = \max_{\substack{u \in N(v) \\ S_1 \cup S_2 = S \\ \Gamma(v) \in S_1, \Gamma(u) \in S_2}} B(v,S_1) + B(u,S_2) + w(u,v)$$

After having the solution for the single-colored node case, this approach is extended for allowing a limited number of insertions and deletions in the resulting subgraphs by considering that: an S-colorful solution allowing j special insertions is a connected subgraph H ⊆ G, where ∃H' ⊆ H such that V(H') is S-colorful and all other vertices of H are non-colored, then finding a C-colorful connected subgraph with up to N_{ins} special insertions can be solved in $O(3^k m N_{ins})$ time. Deletions can be handled directly by the dynamic programming algorithm: if no C-colorful solution was found, then B(v, C) = false for all v. Allowing up to N_{del} deletions can be done by scanning the entries of B. If there exists $\hat{C} \subseteq C$ such that $|\hat{C}| \geq |C| - N_{del}$ and $B(v,\hat{C})$ = true, then a valid solution exists.

Finally, this approach is generalized to multiple color constraints, where a color constraint function can associate each vertex with a set of colors, not just a single color as above. This problem arises when a protein in the network is homologous to more than one protein in the query complex. The basic idea is to reduce the problem to the single color case by randomly choosing a single valid (distinct from other vertexes) color for every vertex. In order to do this, a coloring graph need to be defined as a bipartite graph B = (V, C, E) where V is the set of target network vertices, C is the set of colors and (v, c) $\in$ E iff vertex v has color c in its color set. Consider a possible match to the query, the probability for a subset of vertices of size k to become colorful in a random coloring is at least 1/(k!).

Integer linear programming:

An integer linear programming (ILP) formulation is also given to the C-colorful connexted subgraph problem, then ILP solvers can be employed. This method allows exactly N_{ins} arbitrarily insertions and exactly N_{del} arbitrarily deletions. Particularly, we are given edge weights $\omega: E \to Q$ and wish to find vertex subset K $\subseteq$ V of size t= k + N_{ins} – N_{del} that maximizes the total edge weight $\sum_{(v,w)\in E; v,w\in K} \omega_{vw}$. For expressing the connectivity of the C-colorful subgraph, it is formulated as finding a flow with t-1 selected vertices as sources of flow 1, and a selected sink r that drains a flow of t-1, while disallowing flow between non-selected vertices. For details of this formulation, please refer to [Bruckner et al., 2010].

Shortest-path based heuristic:

A heuristic based on a shortest-path algorithm is designed to obtain a fast solution for finding C-colorful subgraphs in the target network. This heuristic is suitable for the cases when the number of colored vertices is small and it does not allow insertions/deletions (indels) in the resulting subgraphs. This method is also used as a preliminary step, when it fails to return a solution or when indels are required, the dynamic programming or integer linear programming above will be run.

The heuristic aims to partition the initial vertex set V of the target network into two subsets: V_{in}, which is the final solution (the connected component that is C-colorful), and V_{out} for the remaining part. To get this final result, it has to maintain a partition of V into three sets , V_{in}, V_{out}, and V_{open}. Starting with V_{open}= V, vetices are then greedily moved from V_{open} either to V_{in}, meaning that they are part of the final solution, or to V_{out}, meaning that they are

rejected. Shortest-path is used in this heuristic as a criterion to move color nodes in V_{open} to V_{in}.

Experimental results:

Torque was applied to six collections of protein complexes from: yeast, fly, human and used complexes from one species as queries to query against the target PPI networks of the other species. The result comparison showed that it outdoes QNet (which was considered as a state-of-the-art method for finding conserved protein complexes and pathways at that time) in all the cases.

3.2.3. Other network querying based methods

QPath [Shlomi et al., 2006] is a technique for querying PPI networks with path-structured queries, QNet [Dost et al., 2008] is an extension of QPath for queries shaped as trees and graphs with bounded treewidth (though in its implementation, only tree-shaped queries are handled). Both QPath and QNet are based on the color coding technique [Alon et al., 1995], a randomized technique for finding simple paths and simple cycles of a specified length k within a graph (the basic idea is to randomly assign k colors to the vertices of the graph and then search for colorful paths in which each color is used exactly once). In both methods, the total number of node insertions and deletions in the potential solutions are bounded by two thresholds N_{ins} and N_{del}.

3.3. Comparison between the approaches

Local network alignment has a sound theoretical framework for complex conservation modeling and identifying conserved protein complexes, so that methods basing on this framework easily incorporate their own definitions of protein complex evolution into it [Sharan et al., 2005; Koyuturk et al., 2006; Flannick et al., 2006; Hirsh et al., 2007; Nguyen et al., 2013]. Because network alignment is based on the co-occurrences protein interactions between multiple species, it helps the complex detection focus on the more reliable parts of the PPI networks thereby increasing the precision of the task.

Network querying employs known protein complexes in well-studied species to query against PPI networks of other species. This can help to compensate for the incompleteness in PPI networks of some newly sequenced species. On the other hand, this approach is restricted

by the collections of known protein complexes and cannot be extended to detect novel complexes, which in turn highlights this advantage in network alignment approach. There are still not methods that combines the two approaches to exploit the best availability of information we have. Topology-free querying is flexible and robust to noises in protein interaction data but simultaneously, missing the important information of interaction pattern similarity. Table 3.1 below will summarize the comparisons between methods in local network alignment approach and network querying approach.

		Advantages	**Disadvantages**
Local network alignment approach		Sound theoretical framework and ease in incorporating protein complex evolution models. Releasing noises in data by focusing on co-occurring PPIs, which are more reliable PPIs. Can detect novel protein complexes.	Not using the information of known protein complexes.
	NetworkBLAST [Sharan et al, 2005a&b]	Using a simple probabilistic protein complex conservation model basing on dense subgraphs and protein sequence similarity.	Using only whole-sequence similarity (BLAST score) for aligning proteins.
	MaWIsh [Koyuturk et al., 2006]	Using the duplication/divergence models for protein interaction evolution.	Using only whole-sequence similarity (BLAST score) for aligning proteins.
	Graemlin [Flannick et al., 2006]	Combining phylogenetic relationships of proteins in different species and the evolutionary history of	Using only whole-sequence similarity (BLAST score) for aligning proteins.

		interactions.	
	[Hirsh et al., 2007]	Using protein complex evolution model basing on protein interaction attachment/detachment and gene duplication events.	Using only whole-sequence similarity (BLAST score) for aligning proteins.
	COCIN [Nguyen et al., 2013] **(our method)**	Considering protein domains in identifying functional conserved proteins.	
Network querying approach		Using the information of known protein complexes to compensate for incompleteness in the queried PPI networks, and as a good guide for searching for conserved complexes.	Not be able to detect novel protein complexes because it is restricted by the querying protein complexes.
	Topology-free querying [Bruckner et al., 2010]	Flexible and robust to noises in protein interaction data.	
	QPath [Dost et al., 2008]	Simple and fast	Only allows path-structured queries
	QNet [Shlomi et al., 2006]	Can allow both path-structured and tree-like queries.	

Chapter 4 – COCIN: Conserved protein complex detection from Interolog Networks

4.1. Overview

As mentioned in Chapter 1, in spite of the significant progress in computational identification of protein complexes from protein interaction (PPI) networks over the last few years (see the surveys [Srihari et al., 2013; Li et al., 2010]), computational methods are severely limited by noise (false positives) and lack of sufficient interactions (*e.g.* membrane-protein interactions) in currently available PPI datasets, particularly from human, to be able to completely reconstruct the complexosome [Srihari et al., 2013; Li et al., 2010]. For example, several complexes involved in core cellular processes such as cell cycle and DNA damage response (DDR) are not present in a recent (2012) compendium of human protein complexes (http://human.med.utoronto.ca/) assembled solely by computational identification of complexes from high-throughput PPIs [Havugimana et al., 2012]; a web-search (as of Feb 2013) in this compendium for BRCA1 does not yield any complexes even though BRCA1 is known to participate in three fundamental complexes in DDR *viz*. BRCA1-A, BRCA1-B and BRCA1-C complexes [Khanna et al., 2001; Xu et al., 2001; Wang et al., 2000]. A possible reason for missing these complexes is the lack of sufficient PPI data required for identifying them even using the best available algorithms. But, the authors of this compendium note that many human complexes appear to be ancient and slowly evolving – roughly a quarter of the predicted complexes overlapped with complexes from yeast and fly, with half of their subunits having clear orthologs [Havugimana et al., 2012]. Therefore, it is useful to devise effective computational methods that look for evidence from evolutionary conservation to complement PPI data to reconstruct the full set of complexes.

In the attempt to integrate evolutionary information with PPI networks, Kelley *et al.* [Kelly et al., 2003] and Sharan *et al.* [Sharan et al., 2005] devised methods to construct an *orthology graph* of conserved interactions from two species, which in their experiments were yeast (*S. cerevisae*) and bacteria (*H. pylori*), using a sequence homology-based (using BLAST E-score similarity) mapping of proteins between the species. Dense sub-graphs induced in this orthology graph represented putative complexes conserved between the two species. The complexes so-identified were involved in core cellular processes conserved

between the two species – e.g. those in protein translation, DDR and nuclear transport. Van Dam and Snel (2008) [Dam et al., 2008] studied rewiring of protein complexes between yeast and human using high-throughput PPI datasets mapped onto known yeast and human complexes. From their experiments, they concluded that a majority of co-complexed protein pairs retained their interactions from yeast to human indicating that the evolutionary dynamics of complexes was not due to extensive PPI network rewiring within complexes but instead due to gain or loss of protein subunits from yeast to human. Hirsh and Sharan [Hirsh et al., 2007] developed a protein evolution-based model and employed it to identify conserved protein complexes between yeast and fly, while Zhenping *et al.* [Zhenping et al., 2007] used integer quadratic programming to align and identify conserved regions in molecular networks. Marsh *et al.* [Marsh et al., 2011] integrated data on PPI and structure to understand mechanisms of protein conservation; they found that during evolution gene fusion events tend to optimize complex assembly by simplifying complex topologies, indicating genome-wide pathways of complex assembly.

Integrating domain conservation:

Inspired from these works, here we devise a novel computational method to identify *conserved complexes* and apply it to yeast and human datasets. A crucial point we note on the conservation from yeast to human is that many cellular mechanisms, though conserved, have in fact evolved many-fold in complexity – for example, cell cycle and DDR. Consequently, while several proteins in these mechanisms are conserved by sequence similarity (e.g. RAD9 and hRAD9), there are others that are unique (non-conserved) to human (e.g. BRCA1); see **Figure 4.1.** These non-conserved proteins perform similar functions (e.g. cell cycle and DDR) as their conserved counterparts, but do not show high sequence similarity to any of the yeast proteins. A deeper examination reveals that these proteins in fact contain *conserved functional domains* – for example, the BRCT domain which is present in yeast RAD9 and human hRAD9 is also present in the non-conserved human BRCA1 and 53BP1; all of these play crucial roles in DDR [Bork et al., 1997]. Similar structure can be seen in the case of RecQ helicases – several helicase domains are conserved from the yeast SGS1 to human BLM and WRN, but there are three helicases RECQ1,4,5 which are unique to human that also contain these helicase domains [Larsen et al., 2013]. Therefore, integrating information on *functional conservation*, mainly through *domain conservation*, can help to identify considerably more (functionally) conserved complexes than mere sequence similarity,

thereby throwing further light on the conservation patterns of complexes in particular and cellular processes in general.

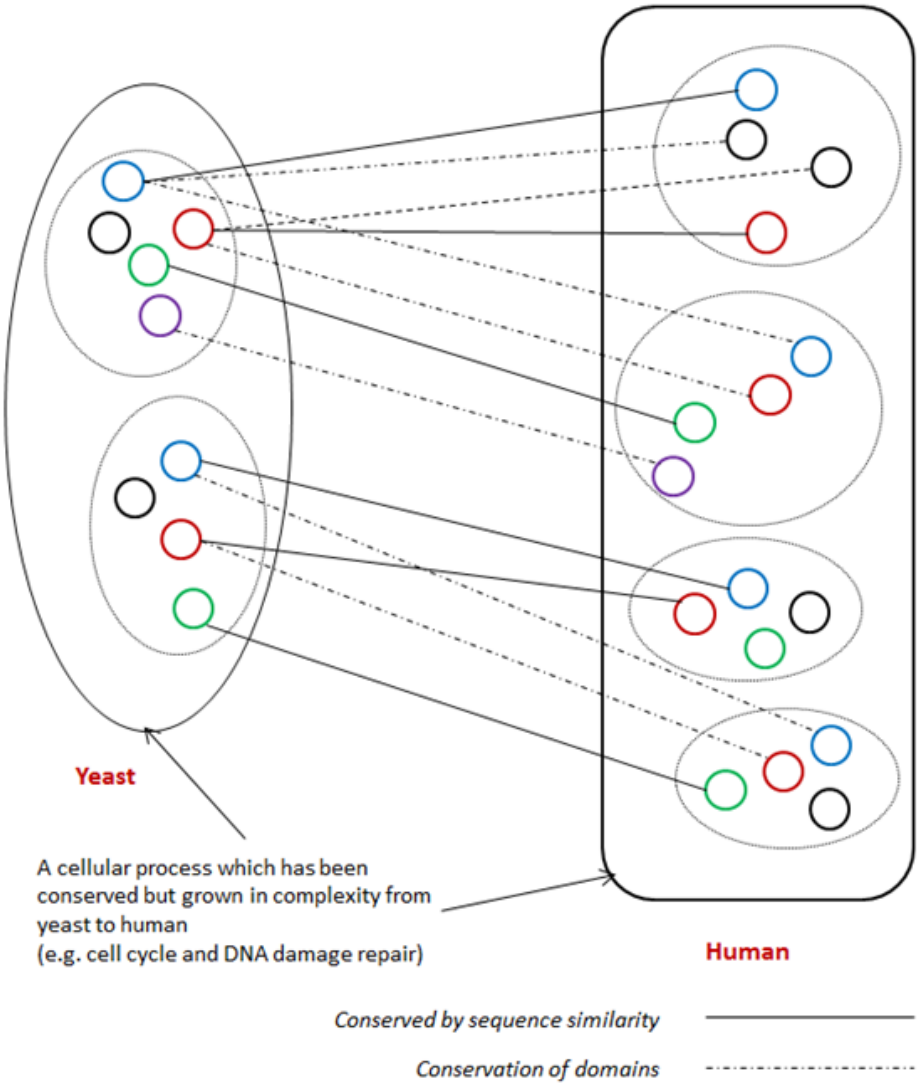


Figure 4.1 - Conservation of complexes between yeast and human

Many proteins in yeast have either 'split' into multiple proteins or fused into common proteins in human during evolution. This mechanism is a result of selecting optimal protein assemblies [Marsh et al., 2011] thereby resulting in multi-fold expansion of complexity in human. In order to capture these conservation mechanisms it is necessary to integrate domain along with PPI information.

In order to achieve this, simple BLAST-based scores as used in earlier works [Kelly et al., 2003; Sharan et al., 2005; Dam et al., 2008; Hirsh et al., 2007; Zhenping et al., 2007] to measure homology between yeast and human proteins do not suffice. Here, we integrate

multiple databases including Ensembl [Flicek et al., 2012] and OrthoMCL [Li et al., 2003] to build homology relationships among proteins; these databases use a variety of information to construct *orthologous groups* among proteins including checking for *conserved domains*. The integration of these databases generates *many-to-many* correspondence between yeast and human proteins instead of the predominantly one-to-one correspondence obtained by from BLAST-based similarity.

We devise a novel computational method to construct an *interolog network* using domain information along with PPI conservation between human and yeast. Next, we identify dense clusters within the interolog network using current 'state-of-the-art' PPI-clustering methods (as against traditional clustering methods used in [Kelly et al., 2003; Sharan et al., 2005]). These clusters when mapped back to the PPI networks reveal conserved dense regions, many of which correspond to conserved complexes.

Our experiments in this work reveal that,

(i) integrating domain information generates many valuable interactions from the many-to-many ortholog relationships in the interolog network, thereby enhancing its quality;

(ii) interolog network also reduces false-positive interactions by accounting for conserved PPIs;

(iii) our interolog network construction aids clustering algorithms to identify far more conserved complexes than direct clustering of the individual PPI networks; and

(iv) many of these conserved complexes are involved in core cellular processes such as cell cycle and DDR throwing further light to the conservation of these cellular processes.

We call our method **COCIN** (COnserved Complexes from Interolog Networks).

4.2. Method

4.2.1. Constructing the interolog network

Given two PPI networks from two species S_1 and S_2, and the homology information between proteins of the two networks, we construct an *interolog network* G_I as follows. The two PPI networks are represented as $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, and the homology relationship between the proteins is governed by a *many-to-many correspondence* $\theta: V_1 \rightarrow V_2$. The interolog network is defined as $G_I(V_I, E_I)$, where $V_I = \{v_I = \{p, q\} \mid p \in V_1, q \in V_2, \text{ and } (p, q) \in \theta\}$, and $E_I = \{(v_I, v'_I) \mid v_I = \{p,q\}; v'_I = \{r,s\}; (p, r) \in E_1 \text{ and } (q,s) \in E_2\}$.

Each node in the interolog network represents a *pair of homologous proteins*, one from each species. Each edge in the interolog network represents an interaction that is *conserved* in both species (interolog). However, if a protein $p \in V_1$ can be orthologous to multiple proteins $x \in V_2$ and $x \in V_2$, then we add two vertices to G_I namely $\{p, x\}$ and $\{p, y\}$, and add an edge between two vertices. Doing so integrates the many-to-many relationships obtained due to domain conservation into the interolog network. **Figure 4.2** below gives a simple example of this network-construction.

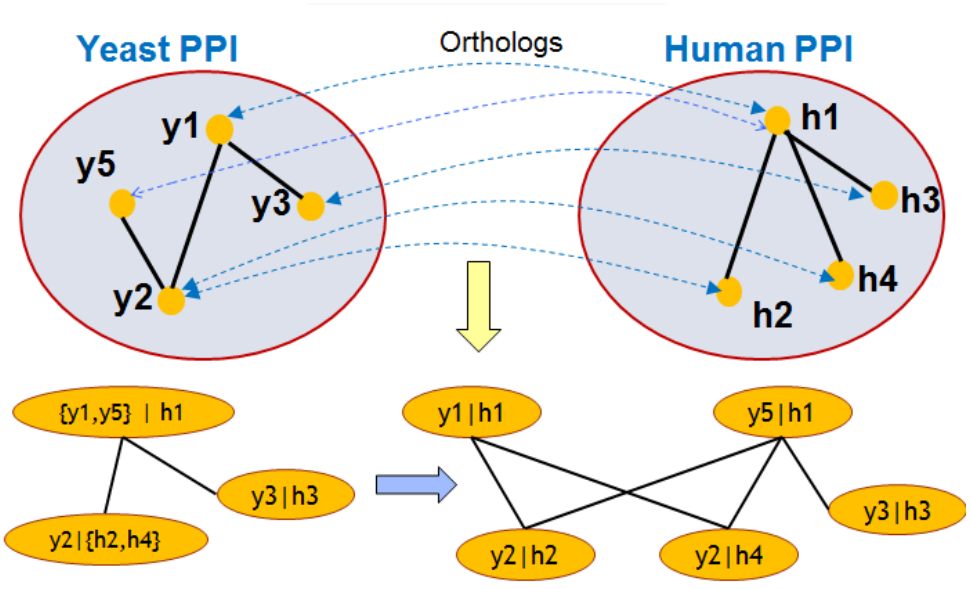


Figure 4.2 - Construction of the interolog network – a simplified example

Our interolog network constructing integrates PPI and domain conservation information to generate a network that is conducive for clustering algorithms to identify considerably

more conserved complexes compared to direct clustering of the original PPI networks from species.

Any connected sub-network in this interolog network can be mapped back to conserved sub-networks in the two PPI networks, and this is similar to the orthology graph method introduced by Kelley *et al.* [Kelly et al., 2003] and Sharan *et al.* [Sharan et al., 2005]. However, one unique advantage of our interolog network offers is that we can infer a *collection* of homologous complexes between the species. This property is highly relevant for identifying conserved complexes between yeast and human (revisit **Figure 4.1**).

In order to achieve this, we integrate multiple databases including Ensembl [Flicek et al., 2012] and OrthoMCL [Li et al., 2003] to build our homology relationships among proteins; these databases use a variety of information to construct orthologous groups among proteins including checking for conserved domains.

4.2.2. Clustering the interolog network and detection of conserved complexes

We identify dense clusters in the interolog network to detect conserved complexes between the two species. To do this, we tested a variety 'state-of-the-art' PPI network-clustering methods, and found the following three to perform the best – CMC (Clustering by merging Maximal Cliques) by Liu *et al.* [Liu et al., 2009], MCL (Markov Clustering) by van Dongen [Dongen et al., 2000] and HACO (Hierarchical Clustering with Overlaps) by Wang *et al.* [Wang et al., 2009]. The comparative assessment of these methods has been confirmed with earlier works [Srihari et al., 2013; Li et al., 2010; Srihari et al., 2010;2012a;2012b].

CMC operates by first enumerating all maximal cliques in network, and ranks them in descending order of the weighted interaction density. It then iteratively merges highly overlapping cliques to identify dense clusters in the network. MCL simulates a series of random paths (called a flow) and iteratively decomposes the network into a number of dense clusters. HACO performs hierarchical clustering by repeatedly identifying smaller dense clusters and merging these into larger clusters. HACO has an advantage over the traditional hierarchical clustering because it allows for overlaps (protein-sharing) among the clusters.

Upon finding each dense cluster in the interolog network, because one-to-many homology relationships may exist between human and yeast proteins (see **Table 4.10** and revisit **Figure 4.2**), we map back these clusters to sub-networks within the two PPI networks to eliminate

duplicated nodes in one species, thereby identifying the exact protein complex that is conserved.

4.2.3. Building a benchmark dataset for conserved protein complexes

Due to lack of benchmark datasets of conserved protein complexes between human and yeast in the literature, we built our own "gold standard" conserved dataset as follows. Using currently available datasets of manually curated protein complexes of human and yeast, we selected pairs of complexes that shared significant fraction of (homologous) proteins.

For measuring the conservation level of a given complex pair $\{C_1, C_2\}$, where C_1 belongs to species S_1 and C_2 belongs to species S_2, we adopted the following *Multi-set Jaccard score*:

Multi-set Jaccard score: Let G_{C1} and G_{C2} be the collections of ortholog groups in complexes C_1 and C_2, respectively. For any group $g_i \in Gc_i$ (i = 1, 2), let I_{Ci} represent the multiplicity of the group g_i in complex C_i,, which essentially is the number of paralogs within the group. Multi-set Jaccard score is given as:

$$MSJ(C_1, C_2) = \frac{\sum_{g_i \in (G_{C1} \cup G_{C2})} \min(\mathrm{I}_{C_1}(g_i), \mathrm{I}_{C_2}(g_i))}{\sum_{g_i \in (G_{C1} \cup G_{C2})} \max(\mathrm{I}_{C_1}(g_i), \mathrm{I}_{C_2}(g_i))},$$

There are often duplication of genes (paralogs) within complexes and clusters. Therefore, MSJ takes into account the multiplicity of the groups and does a more conservative and accurate estimation of the conservation between C_1 and C_2. See **Figure 4.3** for an illustration.

We selected pairs of complexes that show MSJ $\geq$ 50% (see result section for details).

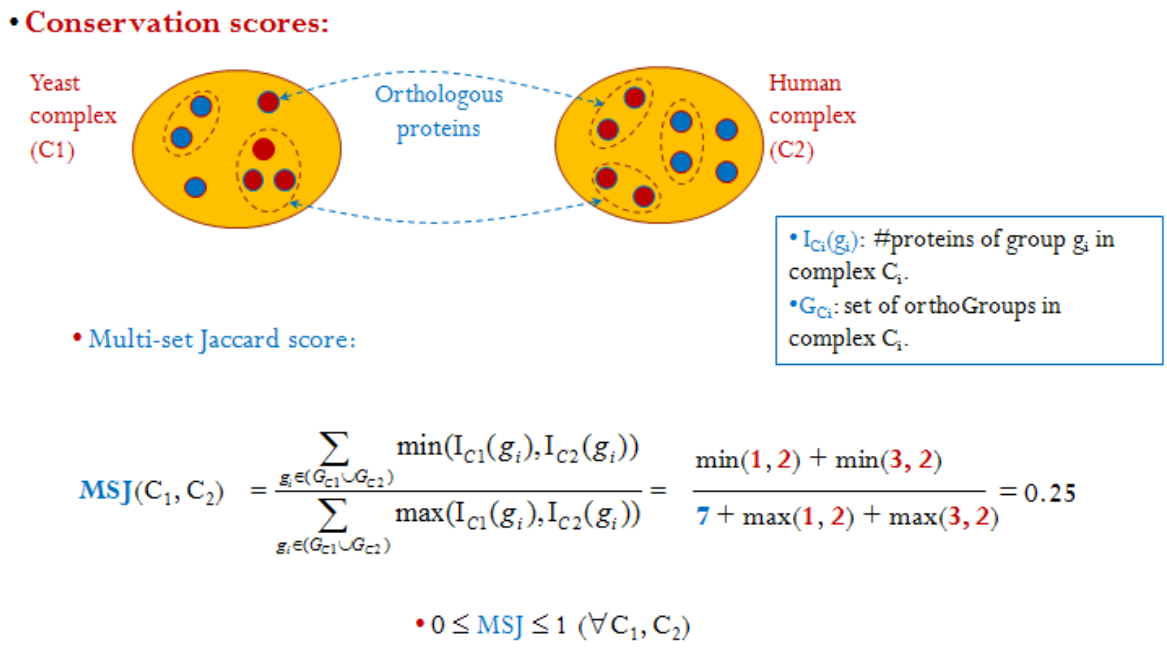


Figure 4.3 - Conservation scores for building benchmark complex datasets

We generate a "gold standard" conserved complexes dataset to test our method. We use two scores here – the Jaccard score for orthologous groups and multi-set Jaccard score.

4.3. Results

4.3.1. Preparation of experimental data

We combined multiple PPI datasets to enhance the coverage of our interactome. We collected PPIs from IntAct [Kerrien et al., 2007] (version November 13, 2012) and Biogrid [Stark et al., 2011] (versions 3.2.95 and 3.2.89) databases for yeast; and from Biogrid and HPRD [Keshava et al., 2009] (Release 9, 2010) for human. **Table 4.1** and **4.2** summarise these datasets.

Yeast curated complexes were gathered from Wodak database (CYC2008) [Pu et al., 2009] and human curated complexes from CORUM (version 09/2009) [Ruepp et al., 2008]; these form our *benchmark* complex datasets (details in **Table 4.3**). We used Ensembl [Flicek et al., 2012] and OrthoMCL [Li et al., 2003] for the homology mapping between human and yeast proteins.

Table 4.1 – Properties of yeast physical PPI datasets

Database	# proteins	# (non self and duplicated) interactions
IntAct (version Nov 13, 2012)	5276	18834
Biogrid (version 3.2.95, Nov 30, 2012)	5886	73923
IntAct ∪Biogrid	6332	83777
IntAct∩Biogrid	4620	8930
ICDScore(IntAct ∪ Biogrid)	5239	71636

Table 4.2 - Properties of human physical PPI datasets

Database	# proteins	#interactions
HPRD (Release 9, 2010)	9617	39184
Biogrid (April 25, 2012)	12515	59027
HPRD ∪Biogrid	13624	76719
HPRD∩Biogrid	8615	21491
ICDScore(HPRD ∪Biogrid)	8521(EntrezID)	61868
ICDEnrich(HPRD ∪Biogrid)	9764 (EntrezID)	192053 (EntrezID)

Table 4.3 - Properties of manually curated protein complex datasets

Databases	# complexes
Wodak [28] yeast complexes (CYC 2008)	**149** with size>3 (36.5%)
	Total: 408
CORUM [29] human complexes (September 2009)	**722** with size>3 (39.1%)
	Total: 1843

Criteria for evaluating predicted complexes:

For a predicted complex C_i of one species and a manually curated (benchmark) complex B_j, we used Jaccard score based on collections of complex proteins: $J(C_i, B_j) = \frac{|C_i \cap B_j|}{|C_i \cup B_j|}$, which considers C_i a correct prediction for B_j if $J(C_i, B_j) \geq t$, a *match threshold*. We chose t = 0.50 in our experiments as suggested by earlier works [Liu et al., 2009; Srihari et al., 2010]. C_i is then referred to as a *matched prediction* or *matched predicted complex*, and B_j is referred to as a *derived benchmark complex.*

Based on this, *precision* is computed as the fraction of predicted complexes matching benchmark complexes, and the *recall* is computed as the fraction of benchmark protein complexes covered by our predicted complexes. A correctly predicted complex is also checked against our "gold standard" testing dataset to see if it is a conserved complex, in which case the derived complex is a *derived conserved complex.*

4.3.2. Results of complex detection using interolog network (IN)

Table 4.4 summarizes the interolog network constructed from yeast and human PPIs. We map back each predicted cluster from the IN to the original PPI networks to predict conserved complexes between the two species.

Table 4.4 - Properties of the interolog network constructed from yeast and human PPIs

# Mapped nodes using orthology	2470
# Interologs	6133
Size of biggest connected component	2434 nodes, 6112 edges
#Other connected components	16 (size from 2-3)

Firstly, we compared the results of complex detection from COCIN with direct clustering of the original PPI networks using CMC, HACO and MCL as shown in **Tables 4.5** and **4.6**. Interestingly, we observed that COCIN, which employs CMC, HACO and MCL for clustering the interolog network, yielded a better recall than these methods on the original

PPI networks. Further, because IN capitalises on the existence of interactions in both PPI networks (that is, conservation of interactions), the number of noisy dense clusters in COCIN is considerably reduced thereby enhancing its precision.

Table 4.5 - Comparisons of different methods on yeast data

Predicted complexes: resulting network clusters

Matched predictions: resulting network clusters that match with benchmarks

Precision = #matched prediction / #predicted complexes

Recall = # detected conserved complexes / # gold standard conserved complexes

Method	#Predicted complexes	#Matched predictions	Precision	#Gold standard conserved complexes	# Detected conserved complexes	Recall (of conserved complexes)
COCIN	71	36	**50.7%**	42	32	**76.2%**
CMC	1202	145	12.1%	42	23	54.8%
HACO	1040	69	6.6 %	42	17	40.5%
MCL	387	37	9.6%	42	5	11.9%

Table 4.6 - Comparisons of different methods on human data

Predicted complexes: resulting network clusters

Matched predictions: resulting network clusters that match with benchmarks

Precision = #matched prediction / #predicted complexes

Recall = # detected conserved complexes / # gold standard conserved complexes

One predicted complex of COCIN can match with many benchmark complexes, this explains for #detected conserved complexes > #matched predictions (as illustrated in **Figures 5-8**)

Method	# Predicted complexes	# Matched predictions	Precision	#Gold standard conserved complexes	# Detected conserved complexes	Recall (of conserved complexes)
COCIN	71	36	**50.7%**	118	78	**66.1%**
CMC	1389	156	11.2%	118	66	55.9%
HACO	1290	80	6.2%	118	36	30.5%
MCL	631	45	7.1%	118	24	20.3%

Figure 4.4 compares a predicted complex Ci through COCIN with two predictions Cy and Ch from the original PPI networks; Cy and Ch form a pair of orthologous complexes, but by direct clustering of the original PPI networks and matching them and not using COCIN. We noticed that Cy and Ch contained several noisy proteins and interactions among them which were false positives. These false positives reduced the Jaccard accuracy of these complexes when matched to known benchmark complexes. We also note that when we computed the complex-derivability index called Component-Edge score (this index measures how much of chance a complex can be detected given the topology of a PPI network) proposed in [Srihari et al., 2012], Ci had a higher CE-score compared to Cy and Ch in the networks.

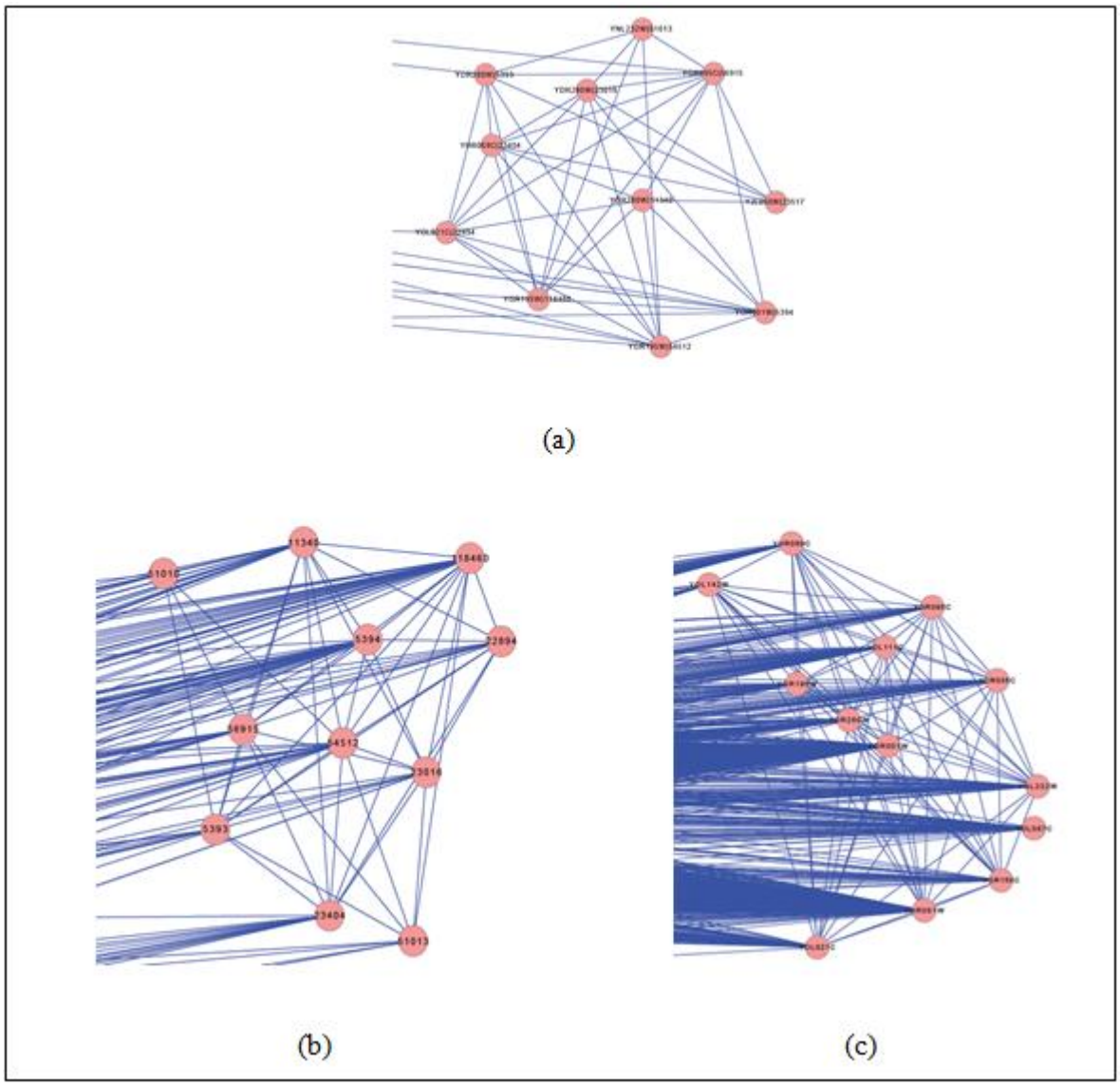


Figure 4.4 - An illustration on a predicted complexes from IN

(a) A predicted complex in the IN.

(b) The corresponding complex in the human PPI network.

(c) The corresponding complex in the yeast PPI network.

Figure 4.5 highlights the improvement of COCIN over CMC, that is, the additional protein complexes of human and yeast detected by COCIN. As many noisy interactions are removed in the IN, among the conserved complexes that are detected by both CMC and COCIN, COCIN on an average obtained higher Jaccard scores. Some important additional conserved complexes found using COCIN were: RNA Polymerase II, EIF3 complex, MSH2-MLH1-PMS2-PCNA DNA-repair initiation complex, MCM complex, MMR complex,

Ubiquitin E3 ligase, transcription factor TFIID, DNA replication factor C, 20S proteasomes (descriptions of these complexes are listed in **Tables 4.7** and **4.8**).

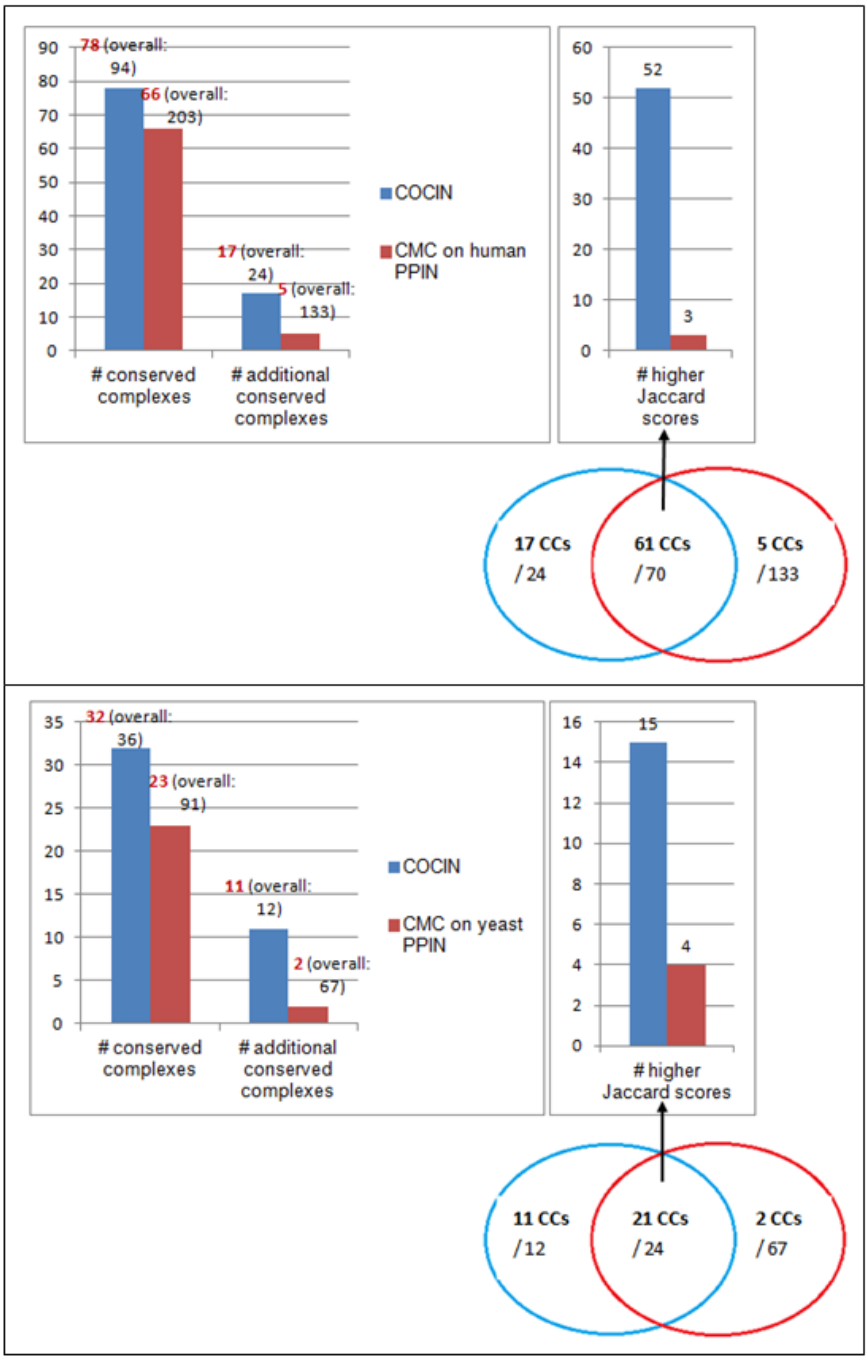


Figure 4.5 - COCIN compared to CMC

COCIN over the interolog network identifies significantly more conserved complexes compared to direct clustering of the original PPI networks using CMC [19].

Table 4.7 – Additional conserved complexes found in yeast

ID	Complex name	Size	Jaccard score	Functional category	Functional description
96	eIF3 complex	7	0.63	Translation	Eukaryotic translation initiation factor
247	Transcription factor TFIID complex	15	0.73	Transcription	mRNA synthesis
27	DNA-directed RNA polymerase II complex	12	0.69	Transcription	mRNA synthesis
45	DNA replication factor C complex (Rad24p)	5	0.67	DNA processing	DNA synthesis and replication
152	DNA replication factor C complex (Rcf1p)	5	0.67	DNA processing	DNA synthesis and replication
294	Mcm2-7 complex	6	0.6	DNA processing	Chromosome maintainance, DNA synthesis and replication
268	SF3b complex	6	0.57	RNA processing	mRNA splicing
65	U6 snRNP complex	8	0.5	RNA processing	This complex combines with other snRNPs, unmodified pre-mRNA, and various other proteins to assemble a spliceosome, a large RNA-protein molecular complex upon which splicing of pre-mRNA occurs.
375	AP-3 adaptor complex	4	0.67	Cellular transport, vesicular transport	This complex is responsible for protein trafficking to lysosomes and other related organelles.
25	20S proteasome	14	0.5	Cell cycle, protein fate	Proteasomal degradation (ubiquitin/proteasomal pathway), protein processing (proteolytic)
137	Chaperonin-	8	0.67	Protein fate	A multisubunit ring-shaped complex that mediates protein folding in the

	containing T-complex				cytosol without a cofactor.

Table 4.8 – Additional conserved complexes found in human

ID	Complex name	Size	Jaccard score	Functional category	Function description
4392	EIF3 complex (EIF3A, EIF3B, EIF3G, EIF3I, EIF3C)	5	0.57	Translation	Translation initiation
4403	EIF3 complex (EIF3A, EIF3B, EIF3G, EIF3I, EIF3J)	5	0.57	Translation	Translation initiation
104	RNA polymerase II core complex	12	0.69	Transcription	mRNA synthesis
2685	RNA polymerase II	17	0.59	Transcription	mRNA synthesis
2686	BRCA1-core RNA polymerase II complex	13	0.64	Transcription	mRNA synthesis
471	PCAF complex	10	0.6	Transcription, DNA processing	DNA conformation modification (e.g. chromatin), modification by acetylation, deacetylation, organization of chromosome structure.
2200	RFC2-5 subcomplex	4	0.5	DNA processing	DNA synthesis and replication
387	MCM complex	6	0.6	DNA processing	Chromosome maintainance, DNA synthesis and replication
369	MMR complex 2	4	0.67	DNA processing	DNA damage repair
290	MSH2-MLH1-PMS2-PCNA DNA-repair initiation complex	4	0.67	DNA processing	DNA damage repair initiation
1169	SNARE complex	4	0.6	Cellular transport, vesicular transport	Vesicle fusion, synaptic vesicle exocytosis
562	LSm2-8 complex	7	0.67	RNA processing	mRNA splicing
561	LSm1-7 complex	7	0.67	RNA processing	Control of mRNA stability during splicing
3036	Ubiquitin E3 ligase (SKP1A, SKP2, CUL1, CKS1B, RBX1)	5	0.5	Cell cycle, protein fate	Mitotic cell cycle and cell cycle control, modification by ubiquitination, deubiquitination
2188	Ubiquitin E3 ligase (CDC34, NEDD8, BTRC, CUL1, SKP1A,	5	0.5	Cell cycle, protein fate	Mitotic cell cycle and cell cycle control, modification by

	RBX1)				ubiquitination, deubiquitination
2189	Ubiquitin E3 ligase (SMAD3, BTRC, CUL1, SKP1A, RBX1)	5	0.5	Cell cycle, protein fate	Mitotic cell cycle and cell cycle control, modification by ubiquitination, deubiquitination

4.3.3. The result of complex detection in the conserved subnetworks

To further understand the advantage of the interolog network on leveraging conservation for better detection of complexes, we performed another experiment *alternative* to the interolog network as follows. We predicted complexes from the *subset of protein interactions of the first species that are conserved in the second* (we call this the *conserved subnetwork* in the first species). The advantage of conserved subnetworks is that is does not produce duplicated edges in cases of one-to-many and many-to-many homology relationships (revisit **Figure 4.2**). However, this can only find complexes of one species at a time, so we map these predicted complexes onto the PPI network of the other species to identify the corresponding conserved complexes. We employed CMC to do clustering on the conserved subnetworks.

Complex prediction from conserved subnetworks showed similar result as COCIN –16 additional conserved complexes in human and 9 additional conserved complexes in yeast are found. This supported the purpose of IN – to leverage conserved interactions for improving complex prediction.

Figure 4.6 shows two other examples that explain why additional conserved complexes are found by COCIN but missed by CMC. We see from this picture that the predicted human complex from IN (the leftmost figure) and the corresponding predicted complex from the conserved subnetwork (the center figure) were contained in a *larger* CMC-predicted complex (the rightmost figure) from the original PPI networks. This larger complex included several noisy proteins that reduce the accuracy of the complex, thereby causing the complex to be missed.

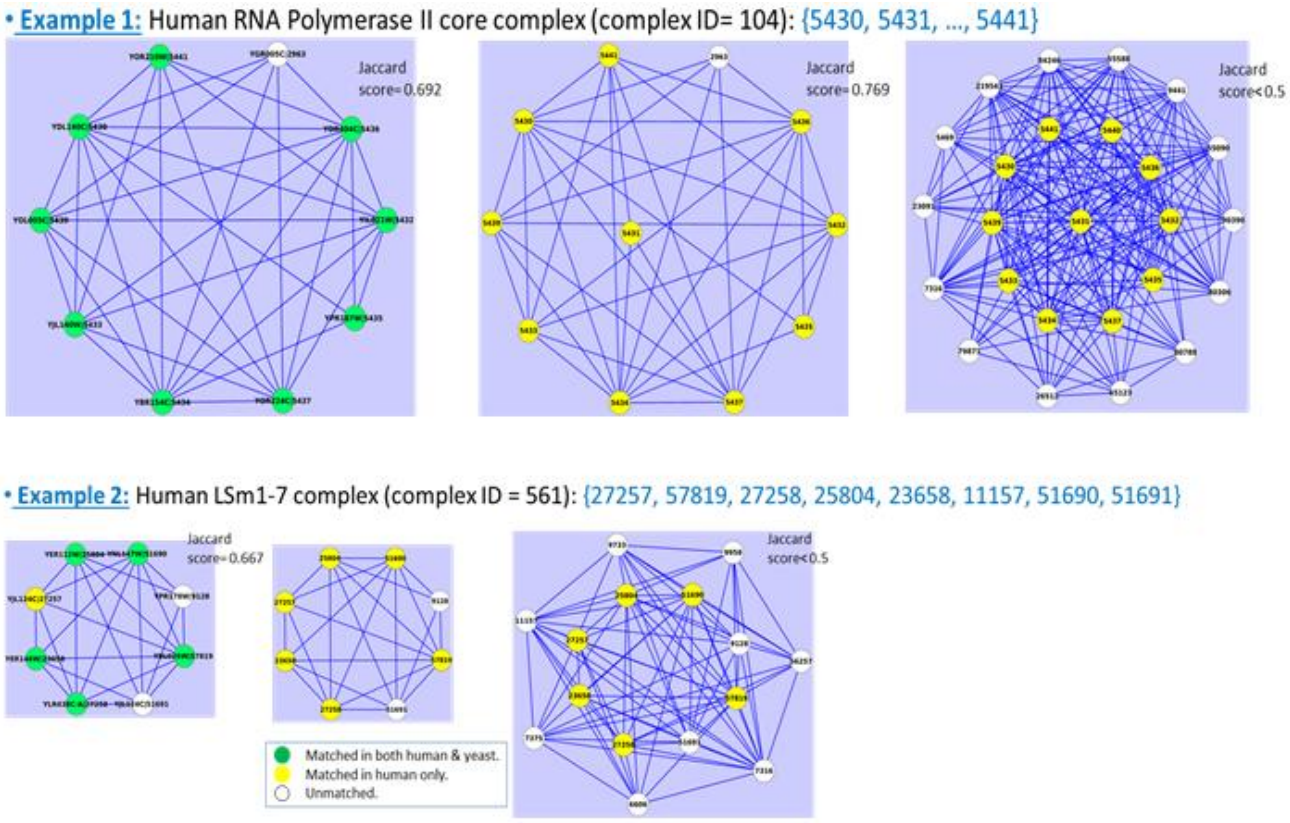


Figure 4.6 - Some examples of additional conserved complexes found in IN

The clusters detected from the original PPI networks include several noisy proteins and noisy interactions (false positives), thereby reducing their Jaccard accuracies.

4.3.4. Comparisons with other complex detection methods in PPI networks

Similar results were obtained using the other two methods HACO and MCL as well, thereby supporting the effectiveness of COCIN in identifying conserved protein complexes. **Tables 4.5** and **4.6** present these comparisons in more details, while **Figures 4.7** and **4.8** highlight further substantiate these results.

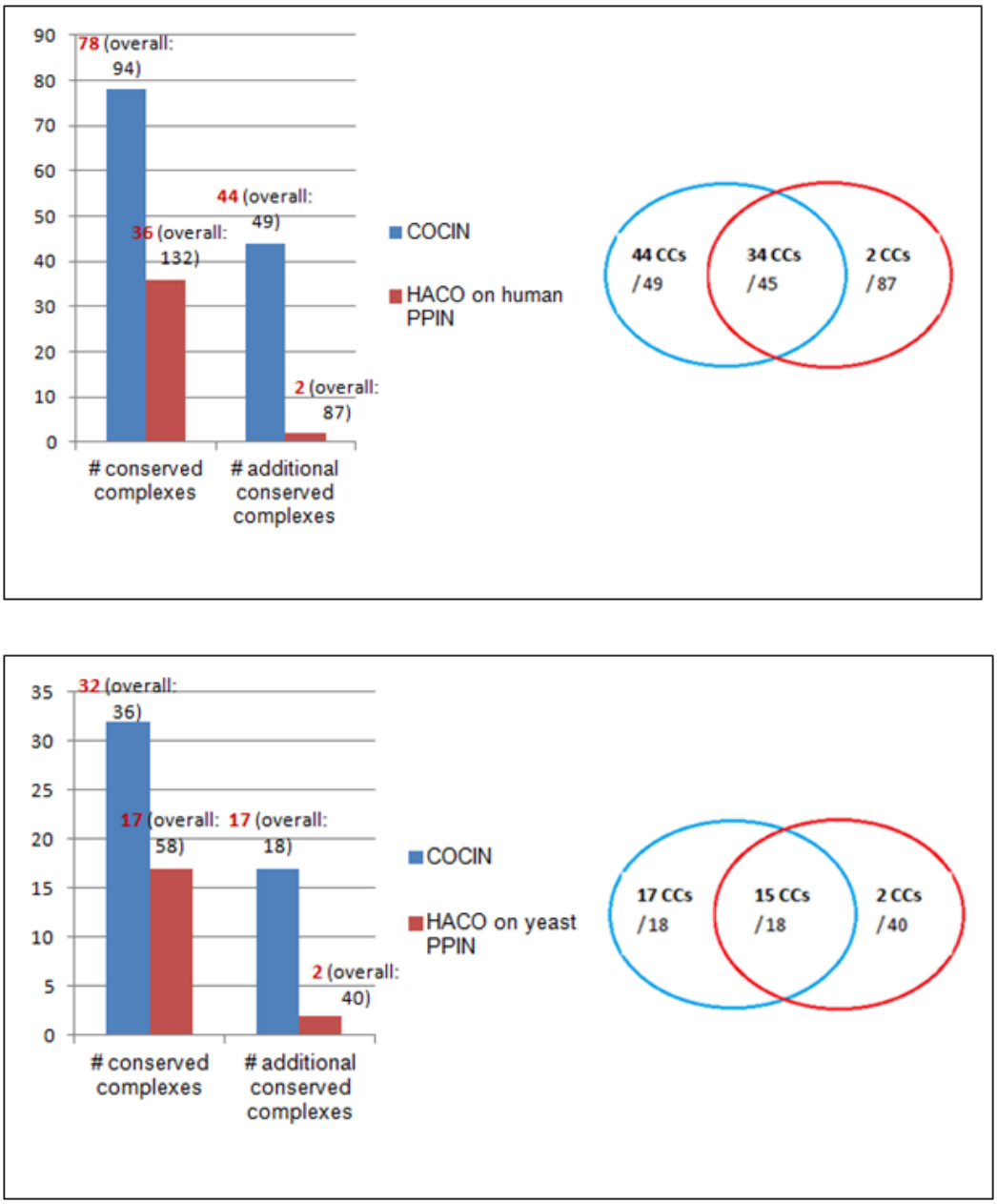


Figure 4.7 - COCIN compared to HACO

COCIN over the interolog network identifies significantly more conserved complexes compared to direct clustering of the original PPI networks using HACO [20].

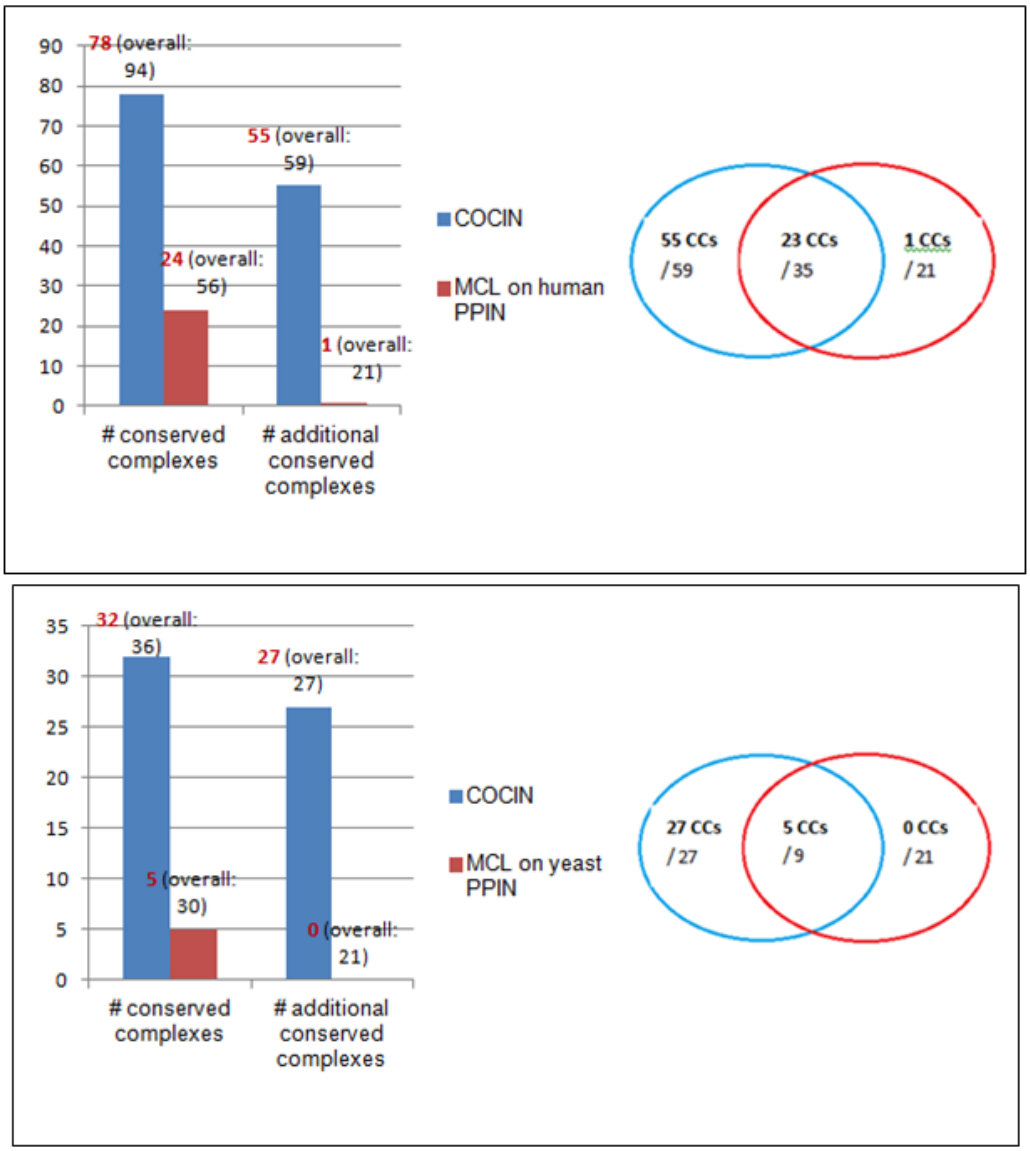


Figure 4.8 - COCIN compared to MCL

COCIN over the interolog network identifies significantly more conserved complexes compared to direct clustering of the original PPI networks using MCL [21].

4.3.5. Integrating domain information significantly enhances interolog construction

Finally**, Table 4.9** summarizes the quality of our testing dataset for conserved protein complexes between yeast and human. We compared the number of benchmark conserved complexes found in both human and yeast using mappings from Ensembl and OrthoMCL under multiple conservation score thresholds (**Figure 4.9**). *Note* that Ensembl contains homology information based on both sequence similarity as well as domain conservation,

while OrthoMCL is predominantly based on sequence similarity. We noticed that using Ensembl homology information can yield more conserved complexes at all conservation score thresholds. Further, **Figure 4.10** shows that there exist 1-to-many and many-to-many relationships of conservation between human and yeast complexes.

Table 4.9 – Details of gold standard testing dataset for conserved protein complexes between human and yeast

Score usage	MSJ ≥threshold
Threshold	50%
# conserved yeast complexes	**42**/149 with size>3 (28.1%)
	Total: 79/408 (19.3%)
# conserved human complexes	**118**/722 with size>3 (16.3%)
	Total: 219/1843 (11.9%)

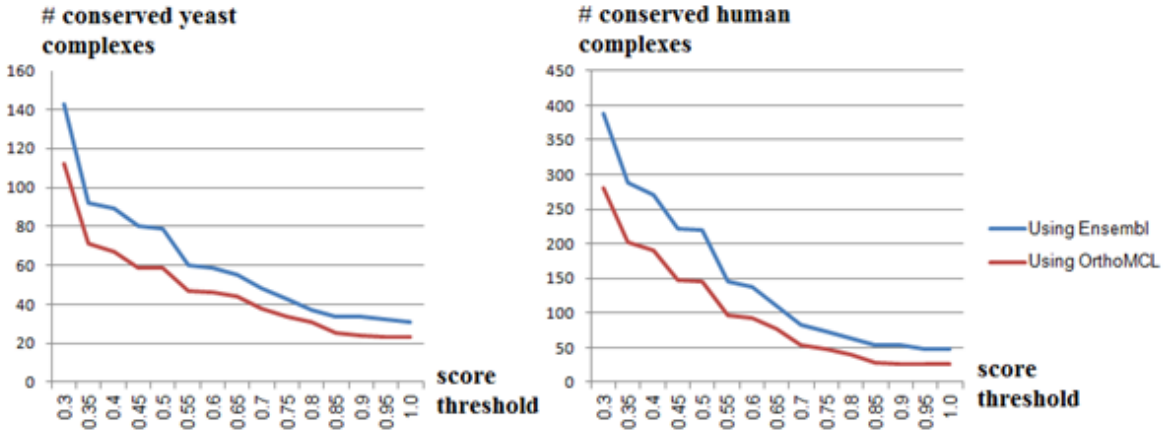
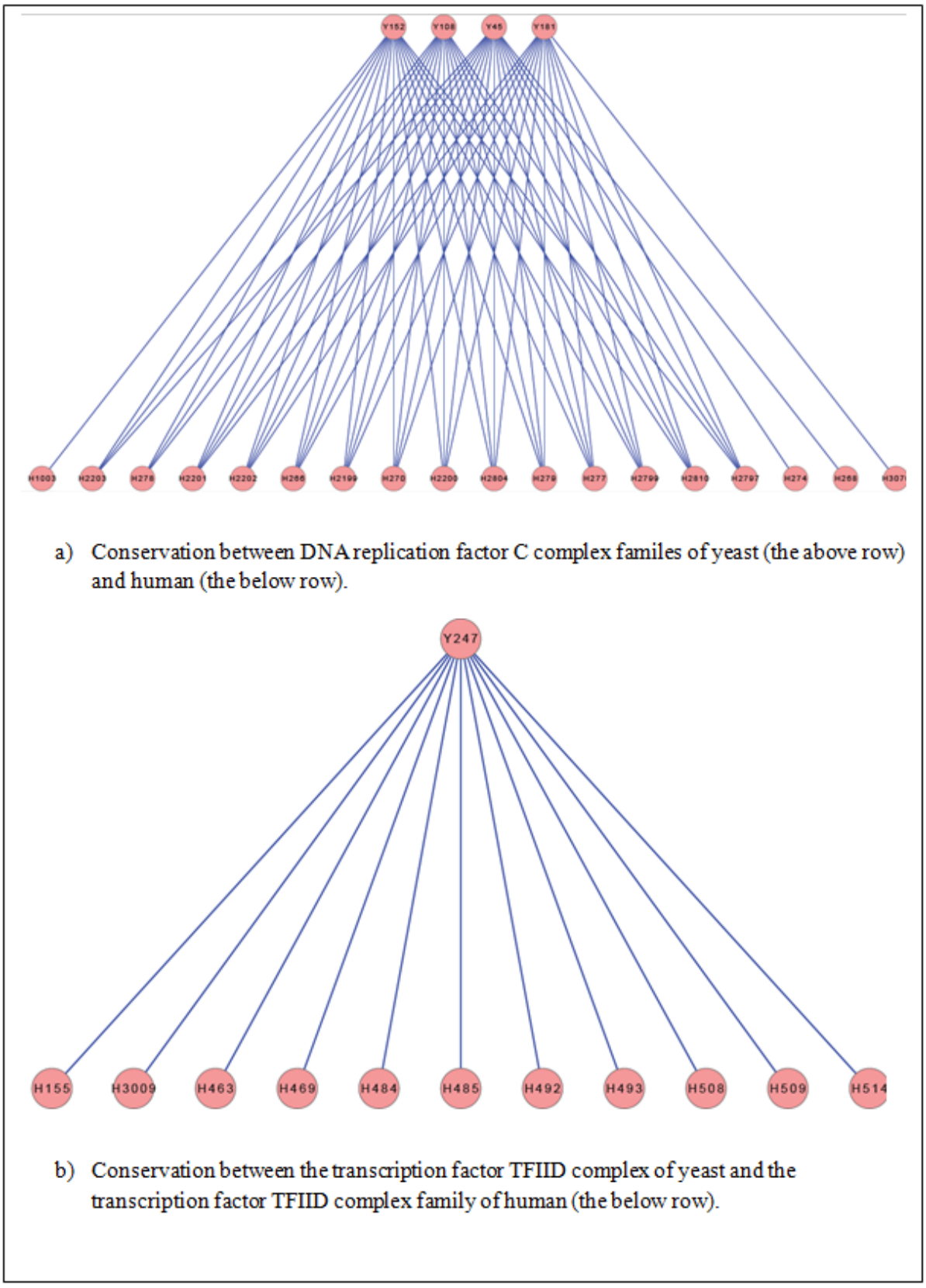


Figure 4.9 - Assessment of Ensembl and OrthoMCL based homology for IN construction and conserved-complex detection



a) Conservation between DNA replication factor C complex familes of yeast (the above row) and human (the below row).

b) Conservation between the transcription factor TFIID complex of yeast and the transcription factor TFIID complex family of human (the below row).

Figure 4.10 – Some examples of the one-to-many and many-to-many relationships of complex conservation between human and yeast

Existing local network alignment methods (NetworkBLAST [Sharan et al., 2005a], MaWIsh [Koyuturk et al., 2006], Graemlin [Flannick et al., 2006]) used whole-sequence BLAST score for identifying homologous proteins before constructing the aligned network, while COCIN uses homology that considers protein domain similarity. Because homologous proteins take the decisive role in identifying conserved protein complexes, the comparisons are made by comparing the aligned network (which is equivalent to an interolog network) produced by using whole-sequence BLAST score based homology (OrthoMCL homology) and the interolog network that uses homology with domain similarity (Ensembl homology). The result showed that the later produced a better-quality interolog network (with a higher number of aligned nodes and interologs) on human and yeast data, thereby improving the conserved complex detection (Section 4.3.5).

Here, we used OrthoMCL as a substitute for the whole-sequence similarity due to technical difficulties of running BLAST for a large number of proteins. We compared the performance of using OrthoMCL against using Ensembl, which uses domain conservation along with sequence similarity to determine orthology. **Table 4.10** and **Figure 4.11** show that we obtain an overall improvement in terms of the number of mapped protein pairs, interologs, as well as conserved protein complexes in both human and yeast by incorporating domain information (through Ensembl). This substantiates the improved performance of COCIN over traditional sequence-similarity based methods.

Table 4.10 - Homology data: Ensembl and OrthoMCL

Ensembl [Flicek et al., 2012] contains protein orthologs based on sequence similarity as well as domain information, while OrthoMCL [Li et al., 2003] is predominantly based on sequence similarity. As we can see from the table, using domain information (through Ensembl) generates significantly more many-to-many ortholog mappings thereby enhancing our interolog construction.

		Ensembl database	OrthoMCL database
# Ortholog groups:	# 1-to-1 groups	1096	1153
	# 1-Yeast-to-many groups	756	434
	# 1-Human-to-many groups	116	116

	# many-to-many groups	197	167
	Total:	2165 (5503 pairs)	1870
# Human paralog groups:		2573	2435
# Yeast paralog groups:		426	393
Total # homolog groups:		5164	4698

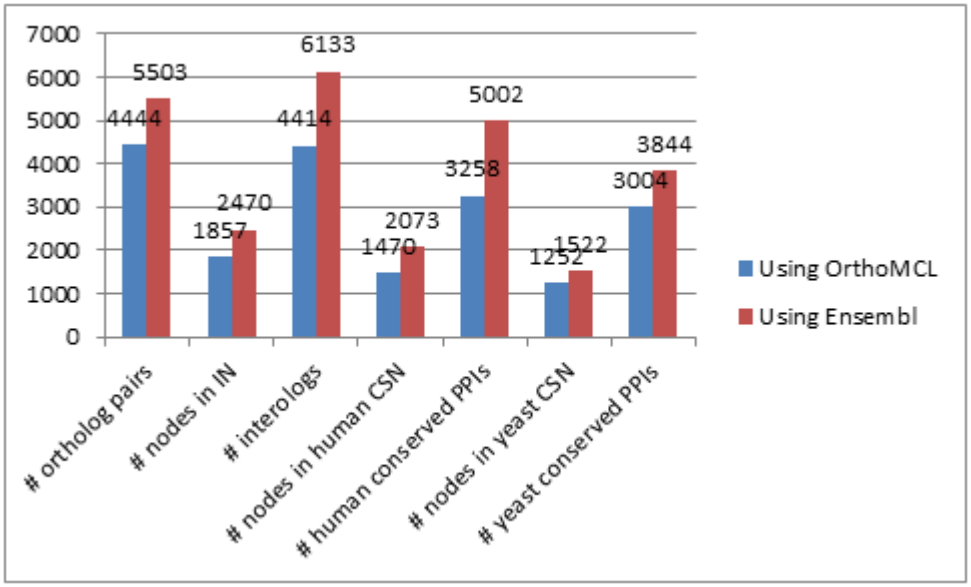


Figure 4.11 – Comparison between using Ensembl and OrthoMCL in constructing the interolog network

Ensembl [17] contains protein orthologs based on sequence similarity as well as domain information, while OrthoMCL [18] is predominantly based on sequence similarity. As we can see from the table, using domain information (through Ensembl) generates significantly more many-to-many ortholog mappings thereby enhancing our interolog construction.

Chapter 5 – Conclusion

5.1. Main contributions

Identifying conserved complexes between species is a fundamental step towards identification of conserved mechanisms from model organisms to higher level organisms. Current methods based on clustering PPI networks do not work well in identifying conserved complexes, and they are severely limited by lack of true interactions and presence of large amounts of false interactions in existing PPI datasets. Therefore, the main contributions of this thesis can be summarized as:

1. We first presented a detailed survey on computational methods for identifying conserved protein complexes between species, which classifies the existing methods into two approaches: local network alignment and network querying (Chapter 3). A brief overview on ortholog assignment methods are also given in Chapter 2.

2. We proposed a novel method, COCIN, which is based on building interolog networks from the PPI networks of species to identify conserved complexes. Our experiments on yeast and human datasets revealed that our method can identify considerably more conserved complexes that plain clustering of the original PPI networks. Further, we demonstrated that integrating domain information generates many-to-many ortholog relationships which significantly enhances the interolog network quality and throws further light on conservation of mechanisms between yeast and human.

3. We built a testing dataset for conserved protein complexes between human and yeast. By proposing a score to measure the conservation level between protein complexes, a collection of conserved complexes pairs between yeast and human is built and considered as a gold standard dataset during this work. As currently there is no benchmark dataset for conserved protein complexes between human and yeast in the literature, the author hopes that this dataset could be useful for reference. Furthermore, this step also gives us a detailed examination on the conservation level between manually curated protein complexes of human and yeast.

5.2. Limitations

The thesis is not without limitations. All the experiments and analyses about conserved protein complexes were performed on only one pair of species: human and yeast. This is because yeast is the most widely studied organism and its PPI network is more complete compared to other species, while human is the most important species we want to study and its PPI network is far from complete. Though this might be an ideal pair of species to study the protein complex conservation, this work need be also extended on many other pairs of species such as: human and mouse, human and fly. All of such studies will broaden our understanding about human protein complexes based on those that are well known in other species. Based on this we recommend the following directions for future research.

5.3. Recommendations for further research

1. Detection of conserved protein complexes between human and other well studied species: Mouse (Musculus) should be an important species to be compared to human in terms of protein complexes. Because mouse is mammalian, it is curious to know if the level of conservation in protein complexes between human and mouse is higher than human and yeast. The answer for this question can also help us in understanding more about protein complex evolution.

2. Protein complex evolution by protein domains needs more explorations. One of the things we can do is to union many existing homology datasets that considering protein domain conservation to increase the coverage of function-conserved proteins. We can also devise a ortholog assignment method by using protein domains queried from Pfam database, because we can infer whether two proteins having similar functional domains by querying Pfam.

Bibliography

[Alon et al., 1995] Alon N, Yuster R, Zwick U. **Color-coding**. *Journal of ACM 1995*, 42(4)**:**844-856..

[Aloy et al., 2004] Aloy P, Bottcher B, Ceulemans H, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB. **Structure-based assembly of protein complexes of yeast**. *Science 2004*, 303:2026-2029.

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

[Bork et al., 1997] Bork P, Hoffman K, Bucher P, Neuwald AF, Alstchul SF, Koonin EV. **A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins.** *FASEB Journal* 1997, 11(1):68—76.

[Bruckner et al., 2010] Bruckner S, Hüffner F, Karp RM, et al. **Topology-free querying of protein interaction networks.** *Journal of Computational Biology*, 17(3):237-252, 2010.

[Chen et al., 2007] Chen F, Mackey AJ, Vermunt JK, et al. **Assessing performance of orthology detection strategies applied to eukaryotic genomes.** *PLoS ONE*, 2:383, 2007.

[Dam et al., 2008] van Dam JP, Snel B. **Protein complex evolution does not involve extensive network rewiring.** *PLoS Computational Biology* 4(7):e1000132, 2008.

[Datta et al., 2009] Datta RS, Meacham C, Samad B, et al. **Berkeley PHOG: PhyloFacts orthology group prediction web server.** *Nucleic Acids Res*, 37:W84–9, 2009.

[Dongen et al., 2000] van Dongen SM. **Graph clustering by flow simulation.** *PhD thesis* 2000, University of Utrecht.

[Dost et al., 2008] Sharan, R., Dost, B., Shlomi, T., et al. **QNet: a tool for querying protein interaction networks.** *Journal of Computational Biology*, 15, 913–925., 2008.

[Flannick et al., 2006] Flannick J., Novak A., Srinivasan B. S., McAdams H. H., Batzoglou S. **Graemlin: General and robust alignment of multiple large interaction networks.** *Genome Research*, 16, 1169–118, 2006.

[Flicek et al., 2012] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM. **Ensembl 2012**, *Nucleic Acids Research* 2012, 40: D84—D90.

[Gavin et al., 2006] Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klien K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwin C, Heurtier MA, Copley RR, Edelmann A, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Sepharin B, Kuster B, Neubauer G, Furga GS. **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, 415:141-147.

[Gavin et al., 2006] Gavin A, 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*, 440(7084):631-636, 2006.

[Havugimana et al., 2012] Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boultz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlashblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ER, Paccanaro A, Marcotte EM, Emili A. **A consensus of human soluble protein complexes**. *Cell*, 150(5): 1068—1081, 2012.

[Hirsh et al., 2007] Hirsh E, Sharan R. **Identification of conserved protein complexes based on a model of protein network evolution.** *Bioinformatics* 23(2):e170–e176, 2007.

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

[Kelley et al., 2003] Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T. **Conserved pathways within bacteria and yeast as revealed by global protein**

network alignment. *Proceedings of the National Academy of Sciences* USA 2003, 100:11394—11399.

[Kerrien] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. **IntAct – open source resource for molecular interaction data.** *Nucleic Acids Research* 2007, 35:D561—D565.

[Keshava] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. **Human Protein Reference Database – 2009 Update.** *Nucleic Acids Research* 2009, 37:D767—D772.

[Khanna et al.,2001] Khanna KK, Jackson SP. **DNA double-strand breaks: signalling, repair and the cancer connection**. *Nature Genetics* 2001, 27:247—254.

[Kiemer et al., 2007] Kiemer L., Cesareni G. **Comparative interactomics: comparing apples and pears?** *Trends in Biotechnology* 2007, 25, 448-454.

[Koyuturk et al., 2006] Koyuturk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., and Grama, A. (2006). **Pairwise alignment of protein interaction networks.** *Journal of Computational Biology*, 13, 182–199, 2006.

[Krogan et al., 2006] Krogan N, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, *et al.* **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, 440(7084):637-643.

[Larsen et al., 2013] Larsen NB, Hickson ID. **RecQ helicases: conserved gaurdians of genomic integrity.** *DNA Helicases and DNA Motor Proteins : Advances in Experimental Medicine and Biology* (Springer New York) 2013, 973 :161—184.

[Leung et al., 2009] Leung, H., Xiang, Q., Yiu, S.M., Chin, F.Y., **Predicting protein complexes from PPI data: a core-attachment approach.** *Journal of Computational Biology* 16(2):133–144, 2009.

[Li et al., 2003] Li L, Stoeckert CJ, Ross DS. **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Research* 2003, 13:2178—2189.

[Li et al., 2010] Li X, Min Wu, Ng SK. **Computational approaches for detecting protein complexes from protein interaction networks: a survey.** *BMC Genomics* 2010, 11(Suppl 1): S3.

[Liao et al., 2009] Liao, C.-S., Lu, K., Baym, M., Singh, R., and Berger, B. (2009). **IsorankN: spectral methods for global alignment of multiple protein networks.** *Bioinformatics*, 25, 253–258, 2009.

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

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

[Marsh et al., 2011] Marsh JA, Hernandenz H, Hall Z, Ahnhert SE, Perica T, Robinson CV, Teichmann SA. **Protein complexes are under evolutionary selection to assemble via ordered pathways.** *Cell* 2011, 153(2) :461—470.

[Mewes et al., 2006] Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A. **MIPS: analysis and annotation of proteins from whole genomes.** *Nucleic Acids Res*, 34:D169-D172, 2006.

[Nguyen et al., 2013a] Nguyen PV, Srihari S, Leong HW. **Identifying conserved protein complexes between species by constructing interolog interaction networks**, (poster paper) *17th International Conference on Research in Computational Molecular Biology (RECOMB 2013)*, April 2013.

[Nguyen et al., 2013b] Nguyen PV, Srihari S, Leong HW. **Identifying conserved protein complexes between species by constructing interolog networks.** *BMC Bioinformatics*, 14 (S8), 2013.

[O'Brien et al., 2005] O'Brien KP, Remm M, Sonnhammer EL. **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res*, 33:D476–80, 2005.

[Pu et al., 2007] Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S.J. **Identifying functional modules in the physical interactome of Saccharomyces cerevisiae.** *Proteomics*, 7, 944-60 (2007).

[Pu et al., 2008] Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S.J. **Up-to-date catalogues of yeast protein complexes.** *Nucleic Acids Res*. 2008.

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

[Ruepp et al., 2008] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. **CORUM: the comprehensive resource of mammalian protein complexes**, *Nucleic Acids Research* 2008, 36(Database issue):D646—650.

[Ruepp et al., 2009] Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Mpflen VS, *et al*. **CORUM: the comprehensive resource of mammalian protein complexes – 2009**. *Nucleic Acids Research,* 36:D646-D650, 2009.

[Shamir et al., 2004] Shamir R., Sharan R., and Tsur D. **Cluster graph modification problems.** *Journal of Discrete Applied Mathematics,* 2004.

[Sharan et al., 2005a] Sharan R, Ideker T, Kelley B, Shamir R. **Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data.** *Journal of Computational Biology* 2005, 12(6): 835—846.

[Sharan et al., 2005b] Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci USA* 2005, 102:1974-1979.

[Shi et al., 2009] Guanqun Shi, Liqing Zhang, Tao Jiang. **MSOAR 2.0: Incorporating Tandem Duplications into Ortholog Assignment Based on Genome Rearrangement.** *Proc. of 8th LSS Computational Systems Bioinformatics Conference (CSB)*, Stanford, August, 2009, pp.12-24

[Shlomi et al., 2006] Shlomi T, Segal D, Ruppin E, and Sharan R. **QPath, a method for querying pathways in a protein-protein interaction network.** BMC Bioinformatics 2006, 7(199).

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

[Srihari et al., 2012] Srihari S, Leong HW. **Employing functional interactions for characterization and detection of sparse complexes from yeast PPI networks.**

International Journal of Bioinformatics Research and Applications, Vol. 8, Nos. ¾, pp. 286-304, September 2012.

[Srihari et al., 2012] Srihari S, Leong HW. **Temporal dynamics of protein complexes in PPI networks: a case study using yeast cell cycle dynamics.** *BMC Bioinformatics* 17(S16), 2012.

[Srihari et al., 2013] Srihari S, Leong HW. **A survey of computational methods for protein complex prediction from protein interaction networks.** *Journal of Bioinformatics and Computational Biology* 2013, 11(2): 1230002.

[Srihari et al., 2013] Srihari S, Ragan MA. **Systematic tracking of dysregulated modules identifies novel genes in cancer.** *Bioinformatics* 2013, 29(12):1553—1561.

[Stark et al., 2011] Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M. **The BioGrid Interaction Database: 2011 Update.** *Nucleic Acids Research* 2011, 39(Suppl. 1):D698—D704.

[Tatusov et al., 2003] Tatusov RL, Fedorova ND, Jackson JD, et al. **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics,* 4:41, 2003.

[Uetz et al., 2000] Uetz P., Giot L., Cagney G., Mansfield T. A., Judson R. S., Knight J. R., Lockshon D., Narayan V., Srinivasan M., Pochart P., et al. **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, 403, 623– 627.

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

[Vilella et al., 2009] Vilella AJ, Severin J, Ureta-Vidal A, et al. **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res*earch, 19:327–35, 2009.

[Wang et al., 2000] Wang Y, Cortez D, Yazdi P, Neff N, Elledge SJ, Qin J. **BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures.** *Genes and Development* 2000, 14:927—939.

[Wang et al., 2009] 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. *Molecular and Cellular Proteomics* 2009, 8(6):1361—1381.

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

[Xu et al., 2001] Xu B, Seong-tae K, Kastan MB. **Involvement of BRCA1 in S-phase and G2-phase checkpoints after ionizing irradiation.** *Molecular Cell Biology* 2001, 21: 3445—3450.

[Yosef et al., 2009] Yosef, N., Kupiec, M., Ruppin, E., et al. **A complex-centric view of protein network evolution.** *Nucleic Acds. Res.* 2009, 37, e88.

[Zaslavskiy et al., 2009] Zaslavskiy, M., Bach, F., and Vert, J. P. **Global alignment of protein-protein interaction networks by graph matching methods.** *Bioinformatics*, 25, i259–i267, 2009.

[Zhang et al., 2012] Melvin Zhang and Hon Wai Leong. **BBH-LS: An Algorithm for Computing Positional Homologs Using Sequence and Gene Context Similarities.** *BMC Systems Biology*, Vol. 6(Supp 1):S22, (11 pages), 2012.

[Zhenping Li et al., 2007] Zhenping L, Zhang S, Wang Y, Zhang XS, Chen L. **Alignment of molecular networks by integer quadratic programming.** *Bioinformatics* 2007, 23(13) :1631—1639.