

**PROTEIN COMPLEX PREDICTION BY
DATE HUB REMOVAL**

IANA PYROGOVA

(M.Math (Hons.), DNU O.Honchar, Ukraine)

Supervisor: Professor Limsoon Wong

Examiners:

Dr Ng See Kiong

Professor Sung Wing Kin

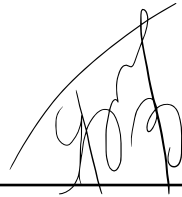
**A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE**

2017

DECLARATION

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

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

A handwritten signature in black ink, appearing to read 'Iana Pyrogova', is written over a solid horizontal line.

Iana Pyrogova

June 30, 2017

Acknowledgements

I would like to express my greatest appreciation to my supervisor, Professor Wong Limsoon, for his full support, invaluable guidance and mentorship throughout the course of this thesis.

Special thanks go also to the other members of the group, who in countless occasions were always available to spare some time and share their helpful suggestions.

I am also deeply grateful for all the understanding from my family. They give me their unconditional love and support during all my studies.

Contents

Summary	5
1 Introduction	11
1.1 Motivation	11
1.2 Thesis organization	13
2 Clustering Algorithms for Protein Complex Discovery	14
2.1 Introduction	14
2.2 Experimental techniques for inferring PPIs	15
2.3 A challenge in protein complex prediction	16
2.4 Protein complex prediction algorithms	17
2.5 Performance of current methods	21
2.5.1 Data sources	22
2.5.2 Evaluation methods	23
2.5.3 Performance on overlapping complexes	25
2.6 Results and discussion	25
2.7 Conclusion	29
3 Predicting Protein Complexes from PPI Network Decomposed by Known Date Hub Removal	31
3.1 Introduction	31
3.2 Background	32
3.2.1 Date and party hub proteins	32
3.2.2 Network decomposition by hub removal	33

3.3	Material and Method	34
3.3.1	Datasets	34
3.3.2	Approach to compare the quality of reference lists of date hubs	35
3.4	Results and discussion	36
3.4.1	Comparing the quality of reference lists of date hubs	36
3.4.2	Protein complex prediction	37
3.5	Conclusion	40
4	Predicting Protein Complexes from PPI Network Decomposed by Removing Predicted Date Hub Proteins	42
4.1	Introduction	42
4.2	Methods	43
4.2.1	Date Hub Prediction	43
4.3	Results and discussion	44
4.3.1	Network topology analysis	44
4.3.2	Impact of different thresholds on predicting date hubs	46
4.3.3	Protein complex prediction	51
4.3.4	Example complexes	54
4.4	Conclusion	55
5	A Strategy to Combine the Two Sets of Clusters Predicted by CMC Before and After PPI Network Decomposition	57
5.1	Introduction	57
5.2	Methods	58
5.2.1	Combining the two sets of predicted clusters	58
5.2.2	Quality of novel complexes	59
5.3	Results and discussion	60
5.3.1	Protein complex prediction	60
5.3.2	Quality of novel complexes	65
5.4	Conclusion	65

Bibliography	68
A Lists of Date Hubs	73

Summary

Proteins physically interact with each other and form protein complexes to perform their biological functions. Different approaches for the identification of protein complexes have been proposed. A common way to discover protein complexes is to use protein-protein interactions (PPI) network and search for clusters of highly-interconnected proteins. However, the prediction of protein complexes from PPI network is usually difficult when the complexes are overlapping with each other in a dense region of the network. It is hypothesized that many overlapping complexes are caused by the presence of date hubs, which are proteins that interact with many partners but at different time or locations. A possible solution is to remove date hubs (and associated edges) from a network before performing protein complex prediction. A recent method uses a hub-removal approach with some success. However, this method removes proteins with high degree instead of date hubs.

In this thesis, we propose a computational methodology to improve prediction of overlapping protein complexes. Particularly, we begin with collecting a gold-standard list of date hub proteins from previous literature. Then, we use these date hubs in a hub-removal approach. Next, we use some simple topological characteristics of gold-standard date hub proteins to predict date hubs from PPI network. Finally, we decompose the PPI network by removing a list of predicted date hub proteins and apply the CMC protein complex prediction algorithm. Overall, experimental results show that the CMC algorithm augmented with date hub-based PPI-network decomposition predicts more overlapping complexes than using CMC alone.

List of Tables

2.1	The four clustering algorithms and their parameters used for yeast and human complex discovery.	22
2.2	The results of various algorithms on yeast data, <i>match-thresh</i> = 0.75.	26
2.3	The results of various algorithms on human data, <i>match-threshold</i> = 0.5.	28
3.1	Different list of hub proteins used and the resulting number of proteins and PPIs discarded in the decomposed networks.	37
3.2	The results of CMC on yeast data, <i>match-threshold</i> = 0.75.	38
3.3	The results of CMC on human data, <i>match-thresh</i> = 0.5.	40
4.1	Different values of degree and transitivity used, and the resulting number of date hub proteins and PPIs discarded in the decomposed networks.	47
4.2	Performance statistics on yeast complex discovery, <i>match-thresh</i> = 0.75.	47
4.3	Different values of degree and transitivity used, and the resulting number of date hub proteins and PPIs discarded in the decomposed networks.	50
4.4	Performance statistics on human complex discovery, <i>match-threshold</i> = 0.5.	50
4.5	Performance statistics on yeast complex discovery, <i>match-thresh</i> = 0.75.	52
4.6	Performance statistics on human complex discovery, <i>match-thresh</i> = 0.5.	53
5.1	Performance statistics on yeast complex discovery, <i>match-thresh</i> = 0.75.	61

5.2	Performance statistics on human complex discovery, <i>match-thresh</i> = 0.5.	63
A.1	List of predicted date hubs for human: <i>degree</i> \geq 30 and <i>transitivity</i> \leq 0.05.	73
A.2	List of predicted date hubs for yeast <i>degree</i> \geq 23 and <i>transitivity</i> \leq 0.32	74

List of Figures

2.1	CMC, ClusterOne, COACH, and IPCA performance analysis on prediction of yeast complexes with <i>match-thresh</i> = 0.75.	27
2.2	CMC, ClusterOne, COACH, and IPCA the distribution of match score of real yeast complexes (a), overlapping real yeast complexes (b).	27
2.3	CMC, ClusterOne, COACH, and IPCA performance analysis on prediction of human complexes the precision, recall, F score, with <i>math - threshold</i> = 0.5 (a), and CMC, ClusterOne, COACH, and IPCA the distribution of match score (b).	28
2.4	CMC, ClusterOne, COACH, and IPCA performance analysis on human complexes, stratified by DENS. a: The match score of the best cluster for reference complexes in low DENS. b: The match score of the best cluster for reference complexes in medium DENS. c: The match score of the best cluster for reference complexes in high DENS.	29
3.1	a: Presence of date hubs in the intersection of yeast reference complex pairs, b: Presence of party hubs in the intersection of yeast reference complex pairs.	36
3.2	Performance of CMC, CMC with removing reference list of date hubs (<i>CMC_DH_reference</i>), and CMC with removing hubs with degree at least 50 (<i>CMC_HUB_50</i>), on prediction of yeast complexes with <i>match-thresh</i> = 0.75.	38
3.3	Match scores of the best clusters to yeast complexes. a: all real yeast complexes, b: overlapping real yeast complexes.	39
3.5	Match score of the best clusters to human complexes for CMC in low DENS (a), medium DENS (b), and high DENS (c).	39

3.4	Performance of CMC, CMC with removing reference list of date hubs (CMC_DH_reference), and CMC with removing hubs with degree at least 150 (CMC_HUB_150), on prediction of human complexes, <i>match-thresh</i> = 0.5.	40
4.1	Date and party hub analysis of topological features in the yeast high quality interaction network: degree (a), betweenness centrality (b), transitivity (c).	45
4.2	Date and party hub analysis of topological features in the human high-quality PPI network: degree (a), betweenness (b), transitivity (c).	45
4.3	Performance statistics for yeast complex discovery at <i>match-thresh</i> = 0.75 using different lists of date hubs.	48
4.4	Match scores of the best clusters for human complexes at different transitivity thresholds of 0.05, 0.1, 0.15 in the three DENS strata: low DENS, medium DENS, and high DENS. .	51
4.5	Match scores of the best clusters for human complexes at different transitivity thresholds of 0.05, 0.1, 0.15 in the three DENS strata: low DENS, medium DENS, and high DENS. .	52
4.6	Performance analysis on prediction of (a) yeast complexes with <i>match thresh</i> = 0.75 , (b) human complexes with <i>match thresh</i> = 0.5.	53
4.7	Reference yeast complexes Swr1p and Ino80p. (a) Swr1p and Ino80p complexes are overlapping with each other. Proteins within the intersection (green) were predicted as date hub proteins.(b) CMC included extraneous proteins (YLR399C and YOL12C) in its clusters and missed two proteins YLR085C and YLR385C.	55
5.1	Performance analysis on prediction of yeast complexes with <i>match thresh</i> = 0.75.	61
5.2	a: Match scores of the best clusters for yeast complexes. b: Match scores of the best clusters for overlapping yeast complexes.	62
5.3	Performance analysis on prediction of human complexes with <i>match thresh</i> = 0.5.	63
5.4	(a) Match scores of the best clusters to human complexes for (a) low DENS, (b) medium DENS , and (c) high DENS.	64
5.5	Number of novel predicted yeast complexes.	66
5.6	Coherence of predicted yeast complexes.	67

Chapter 1

Introduction

Proteins play a vital role in cellular processes. Generally, they do not act alone but form complexes with other proteins to carry out their biological functions. Protein complexes are formed by physical interaction among proteins at specific time and space. Detecting protein complexes is important for understanding the dynamics of biological processes within an organism. Therefore, a wide variety of computational approaches were developed specifically for the prediction of protein complexes. While protein complex prediction methods have evolved significantly in the last two decades, their performance still has room for improvement.

1.1 Motivation

Protein-protein interactions (PPI) are usually represented by PPI network, where nodes are proteins and edges are interactions between proteins. To discover protein complexes, earlier computational approaches commonly incorporate topological characteristics of the PPI network to find clusters of highly-interconnected proteins within the PPI network as protein complexes (Liu et al. 2009, Bader & Hogue 2003, Wu et al. 2009, Nepusz et al.

2012, Li et al. 2008). Nonetheless, the performance of these algorithms was not very satisfactory (Yong & Wong 2015*a*). A possible reason behind the poor performance is that one protein can participate in the formation of different complexes which perform distinct functions and occur at different time and location. Therefore, there are many complexes in the PPI network which overlap with each other. However, the PPI network does not contain any information about when, where and how a protein interacts with others. Overlapping complexes exist within a highly-connected region of the PPI network. Prediction of such complexes is challenging for existing complex discovery algorithms because the generated clusters may contain extra proteins which are outside a complex but connect to the proteins that participate in multiple other complexes (Yong & Wong 2015*a*). Therefore, predicted complexes cannot match true complexes. To overcome this issue, Liu et al. (2011) proposed a new technique to incorporate the dynamic nature of protein interactions by decomposing the PPI network into several smaller networks prior to clustering. This delimits the overlapping complexes more precisely. To decompose PPI network, Liu et al. (2011) removed hub proteins with large numbers of interaction partners, hypothesizing that these proteins may correspond to ‘date’ hub proteins, which bind their different partners at different time or location. However, proteins with high degree may represent ‘party’ hub proteins which interact with their partners simultaneously, and hence do not participate in the overlapping complexes. This motivates us to apply the network decomposition approach by removing a list of real date hub proteins, and investigate the impact of decomposition on the performance of protein complex prediction algorithms.

1.2 Thesis organization

The organization of this thesis is as follows. In Chapter 2, we give a background on popular methods for protein complex prediction, and specifically illustrate that CMC is a method that is generally effective. At the same time, we also highlight the challenges related to prediction of overlapping complexes. Chapter 3 discusses insights from our examination of the role of a gold-standard list of date hubs in overlapping complexes, and the impact of network decomposition on the performance of CMC. Chapter 4 describes our proposed methodology which predicts date hub proteins from PPI network using some simple topological features, incorporates the predicted list of date hubs for network decomposition by hub removal, and uses CMC to predict protein complexes from the resulting decomposed protein-protein interaction network. In Chapter 5, we also propose a double-barrel strategy to combine predicted clusters before and after we remove date hubs.

Chapter 2

Clustering Algorithms for Protein Complex Discovery

2.1 Introduction

Cellular processes are the result of the coordinated action of groups of interacting proteins. The most widely and successfully used methods for high-throughput screening of protein-protein interactions are the yeast two-hybrid system (Y2H) (Fields & Song 1989) and the tandem affinity purification with mass spectrometry (TAP-MS) (Rigaut et al. 1999) system. Once the PPI data are available from experimental detection methods, the PPI network may be used for further computational analysis to infer the protein complexes.

In this chapter, we first provide a brief description of experimental techniques for inferring PPIs. Then we describe challenges in protein complex prediction and review the computational methods to predict protein complexes. Finally, we evaluate some of the most well-known protein complex prediction methods applied to PPI network which could handle overlapping

protein complexes.

2.2 Experimental techniques for inferring PPIs

Yeast two-hybrid (Y2H)

A majority of published interactions have been detected using an Y2H screen. The classic Y2H system was developed by Fields & Song (1989). It involves a fragmented transcription factor to detect the interaction between the protein of interest X, called a bait, and the potential interacting protein Y, called a prey. Protein X is fused to the DNA binding domain of the transcription factor. At the same time, protein Y is fused to the activation domain (AD). The bait binds to the promoter region of a reporter gene, and the interaction between bait and prey leads to the transcription of the reporter gene. One major drawback of classic Y2H is that the interaction can occur only within the yeast nucleus. Therefore, Y2H does not detect interaction between two proteins if they do not localize into the nucleus after translation. Brückner et al. (2009) provides a comprehensive review on recent Y2H approaches which overcome this limitation. Another problem is that Y2H is limited to only direct-contact physical PPIs.

Tandem affinity purification with mass spectrometry (TAP-MS)

The basic strategy underlying TAP-MS is tagging the protein of interest, which allows it to interact with its binding partners, and applying two affinity purifications to examine binding partners. Unlike Y2H, TAP-MS does not offer a list of pairwise binary interactions. Therefore, PPIs should be uncovered from the purified complexes. For this purpose two models may be used. The first is a matrix model where the bait protein and all the prey proteins are assumed to interact directly with each other. The second is the spoke model where only the bait is assumed to interact directly with all

the prey proteins. The first model may give a large number of false-positive interactions, while the spoke model leads to a few false positives and false negatives. TAP-MS is able to detect all the components of a larger complex, which do not necessarily interact directly with each other. However, Y2H has advantages over TAP-MS because it is an *in vivo* technique.

There are multiple drawbacks associated with the above two approaches. Both of them do not capture timing or localization information about the PPIs, and have limitations which lead to false-positive interactions.

Constructing PPI network

Today PPIs data are available for a range of organisms, from different interactions sources (curated PPIs, experimental PPIs, and predicted PPIs) and from different experimental detection methods. A PPI network is represented as an undirected graph where nodes are proteins and edges are the interactions between these proteins. Often, each protein pair has a weight which represents how likely the two proteins interact with each other.

2.3 A challenge in protein complex prediction

Using advances from high-throughput proteomic techniques, such as Y2H and TAP-MS, it has become possible to compile a large network of protein interactions. However, extracting useful knowledge from such networks is a non-trivial task. Therefore, a wide variety of sophisticated PPI network analysis algorithms to detect protein complexes have been proposed in the last two decades. They are often designed to detect sub-graphs with specific topological structures in a PPI network, such as cliques (Palla et al.

2005, Liu et al. 2009), dense sub-graphs (King et al. 2004, Nepusz et al. 2012), and core-attachment structures (Wu et al. 2009). Some algorithms incorporate topological features to assign weights to vertices or edges, such as the number of common neighbors (King et al. 2004, Li et al. 2008) and density (Bader & Hogue 2003). A comprehensive review is given by Srihari et al. (2015).

In reality proteins may perform different biological functions as members of different complexes. Therefore, protein complexes often overlap. Overlapping complexes exist within a highly-connected region of the PPI network. Prediction of such complexes is challenging for existing protein complex prediction algorithms and the generated clusters often contain multiple complexes merged into large clusters. Therefore, predicted complexes cannot match true complexes.

2.4 Protein complex prediction algorithms

In spite of the above challenge, previous works made considerable progress in protein complex prediction. In this subsection, we give a background on some popular protein complex prediction methods applied to PPI network, and specifically illustrate that CMC is a method that is particularly effective and consistent across multiple datasets.

Molecular Complex Detection (MCODE)

MCODE (Bader & Hogue 2003) is the earliest method to detect protein complexes based solely on the topology of PPIs network. It is a seed-and-grow algorithm and consists of three steps: vertex weighting, complex prediction and optional post-processing.

In the first step, for each vertex v , MCODE calculates the highest k -core

of the immediate neighborhood of v . Next, it computes a core-clustering coefficient of a vertex v as the density of the highest k -core. Then, MCODE assigns a weight to a vertex as the product of the vertex’s core-clustering coefficient and the highest k -core level of the immediate neighborhood of the vertex.

Next, for complex prediction, MCODE selects the highest weighted vertex as a complex seed and recursively includes adjacent vertices into the complex if their weight is above a given threshold. It stops once no more vertices can be included into the complex.

Clustering based on merging Maximal Cliques (CMC)

CMC (Liu et al. 2009) is a clique-based approach which uses maximal cliques to identify dense subgraphs from PPI networks. It first searches for all the maximal cliques in the PPI network. Then for each clique C , CMC calculates a score based on weighted density (Equation 2.1). Next, all clusters are ranked by weighted density. Many generated clusters may overlap. Therefore, highly overlapping clusters are merged if they have a high inter-cluster connectivity (Equation 2.2). On the other hand, if two highly overlapped clusters do not have a high inter-cluster connectivity, CMC removes the cluster with the lower density.

$$score(C) = \frac{\sum_{u \in C, v \in C} w(u, v)}{|C|(|C| - 1)} \quad (2.1)$$

$$inter_score(C_1, C_2) = \sqrt{\frac{\sum_{u \in (C_1 - C_2)} \sum_{v \in C_2} w(u, v)}{|C_1 - C_2||C_2|} \frac{\sum_{u \in (C_2 - C_1)} \sum_{v \in C_1} w(u, v)}{|C_2 - C_1||C_1|}} \quad (2.2)$$

where $w(u, v)$ is the weight of the interaction between proteins u and v .

CFinder

CFinder (Palla et al. 2005) is a clique-based approach that employs Clique Percolation Method (CPM) to predict protein complexes from a PPI network. The algorithm works by computing all k -cliques which are a complete subgraph of k vertices. A protein complex candidate is represented as a union of all adjacent k -cliques, where two k -cliques are adjacent if they share $k-1$ vertices. It should be noted that a larger value of k can lead to smaller subgraphs with higher density.

A core-attachment based method (COACH)

Another way to predict protein complexes from a PPI network is to look for clusters that have a core-attachment organization. COACH (Wu et al. 2009) works by considering the inherent structure of protein complexes. The basic idea is to generate protein-complex cores and then include attachments into the core. In general, a core is represented by a small and dense subgraph in a PPI network. To detect cores, COACH first constructs a neighborhood subgraph for each vertex. Next, vertices from a neighborhood graph are defined as core proteins if their degrees are higher than the average degree of that neighborhood graph. If the neighborhood subgraph is dense enough then COACH returns it as a preliminary core. Otherwise, COACH removes the core proteins from the it and forms several connected components. Next, COACH adds back the core proteins into each connected component and returns multiple preliminary cores.

Once a protein-complex core is detected, COACH adds the attachments to form a potential protein complex. Attachments are the neighbors of the vertices in the core which interact with at least half the core's members.

Clustering with Overlapping Neighborhood Expansion (ClusterONE)

ClusterONE (Nepusz et al. 2012) follows the general framework of the

seed-and-grow approach. The algorithm works by selecting the protein with the highest degree as a seed from weighted PPI networks. Then it grows the cluster by adding the neighboring vertices with high cohesiveness. A cohesiveness function of a cluster correlates with how likely it is for a group of proteins to form a protein complex, and is computed as the ratio of the sum of edge weights within the cluster versus the sum of edge weights within the cluster and outgoing edges from the cluster. The cluster extension process terminates once there are no more proteins to be added to the cluster. After all clusters are computed, highly overlapping pairs of clusters are merged into protein complex candidates. The overlap score of two protein sets A and B is calculated as follow:

$$\text{overlap_score}(A, B) = \frac{|A \cap B|^2}{|A||B|} \quad (2.3)$$

Finally, all protein complex candidates with less than three proteins are discarded.

IPCA

IPCA (Li et al. 2008) employs a seed-and-grow strategy similar to ClusterONE. IPCA first assigns the weight of an edge $[u, v]$ as a number of common neighbors for the vertices u and v . Then each vertex is assigned a weight which is a sum of the weights of its incident edges. The vertices with the highest weights are selected as the seeds. IPCA calculates two metrics — *the interaction probability* and *the cluster diameter* — to grow a cluster by adding highly-weighted neighboring vertices to it. The first metric is represented by (Equation 2.4), shows how strongly the neighboring vertex v is connected to a cluster C :

$$IN_{v \notin C} = \frac{m_{vC}}{n_C} \quad (2.4)$$

where m_{v_C} is the number of edges between the vertex v and the cluster C , and n_C is the number of all vertices in C .

The second metric represents the largest length of a shortest path between a pair of vertices in the cluster. The cluster is extended while neighboring vertices satisfy the thresholds for both metrics.

Restricted Neighborhood Search Clustering (RNSC)

RNSC (King et al. 2004) is a cost-based local-search clustering algorithm. It works by generating an initial set of random clusters, and then iteratively moves a vertex from one cluster to another to improve the value of the cost function. To move vertex v it calculates two scores. The first score is the sum between the number of neighbours that are not in the same cluster of v , and the number of node that are not neighbours of v but belong to the same cluster. The second score for vertex v is the ratio of its first score versus the sum of the number of nodes in cluster and the number of neighbours of vertex v . Next, to create a list of protein complex candidates, for each cluster RNSC computes the sum of these scores for each vertex in cluster. The process of moving a vertex is terminated once some move has been reached without decreasing the cost function. Finally, for each complex candidate a p -value is calculated as a probability that a set of proteins within a given cluster belongs to the same functional group by chance. All clusters with p -values above a given threshold are discarded. The main limitation of RNSC is that it does not support the detection of overlapping clusters.

2.5 Performance of current methods

In this section, we compare the performance of four protein complex prediction algorithms which allow overlapping complexes: CMC, ClusterOne,

Algorithm	Yeast	Human
CMC	overlap.thres=0.5 merge.thres=0.75	overlap.thres=0.5 merge.thres=0.75
ClusterOne	-s 4	-s 4
COACH	default	default
IPCA	-S 4 -P 2 -T 0.4	-S 4 -P 2 -T 0.6

Table 2.1: The four clustering algorithms and their parameters used for yeast and human complex discovery.

COACH, and IPCA. For all these methods, the optimal parameters were set as described in Yong & Wong (2015a) to maximize their F-measures (Table 2.1). For comprehensive comparison, we employed several evaluation measures, including recall, precision, F-measure, and the best match cluster score. Below, we explain the design of our experiments and present the results obtained from them.

2.5.1 Data sources

PPI datasets

In our experiments, we apply protein complex prediction methods on two PPI networks, yeast (*S. cerevisiae*) and human (*H. sapiens*). The PPIs were collected from (Yong & Wong 2015b), which are the union of physical PPIs from three databases (viz. BioGRID (Chatr-Aryamontri et al. 2013), IntAct (Orchard et al. 2013), MINT (Licata et al. 2012)) and the Consolidated PPI dataset (Collins et al. 2007). Collected PPIs have a reliability score for each interacting pair (a,b) which is estimated as:

$$reliability(a,b) = 1 - \prod_{e \in E_{a,b}} (1 - rel_e)^{n_{e,a,b}} \quad (2.5)$$

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

(a,b) . In our setup, we use top 20 000 edges with 3680 proteins in the yeast PPI network, and top 20 000 edges with 6352 proteins in the human PPI network.

Reference complexes for yeast and human

To evaluate the effectiveness of four algorithms for detecting protein complexes, we compare the predicted clusters produced by the algorithms with known protein complexes collected from CYC2008 (Pu et al. 2009) protein complex catalog for yeast, and from the CORUM (Ruepp et al. 2009) database for human. It is important to note that, reference complexes with less than four proteins were removed from the both datasets. Overall, there are 149 manually annotated complexes in the yeast set that each consists of four or more proteins. The human set consists of 651 protein complexes with size greater than three.

2.5.2 Evaluation methods

Best match cluster score

To investigate the performance of clustering algorithm, we compare a set of predicted clusters with a real reference protein complex set. The match between a predicted cluster and a reference protein complex may be often only partial, and a reference complex can match more than one predicted cluster and vice versa. In the present study, the best match cluster score evaluates a set of predicted protein complexes with respect to a set of reference complexes. To measure a match between a predicted cluster P and a reference complex C , we calculate the Jaccard similarity between the proteins contained in P and C .

Let V_C be the set of proteins in the reference complex C , and V_P be the set of proteins in the predicted cluster P . According to a definition provided

by Liu et al. (2011), a cluster P , created by a protein complex prediction algorithm, matches a reference complex C at given match threshold, $match-thresh$, only if $Jaccard(P, C) \geq match-thresh$, where $Jaccard(P, C)$ is the Jaccard similarity between the proteins contained in P and C :

$$Jaccard(P, C) = \frac{|V_P \cap V_C|}{|V_P \cup V_C|} \quad (2.6)$$

A threshold of $match-thresh = 0.75$ was used in matching yeast complexes, and $match-thresh = 0.5$ in matching human complexes. A more relaxed threshold is used for human complexes because the human protein interaction network is rather incomplete.

For each protein complex from reference dataset of protein complexes, we plot the distribution of the best-match cluster scores for complexes therein, to illustrate how well the predicted clusters represent the reference complexes, and to investigate which algorithm can predict the reference complexes with better matching clusters.

Recall and precision

For further performance assessment of each method, we report statistics such as the precision, the recall and F-measure of all predicted clusters. Suppose $P = P_1, P_2, \dots$ is a set of generated clusters, and $C = C_1, C_2, \dots$ is a set of real protein complexes from the reference database, then the precision, the recall and F-measure of the clusters are defined as follows:

$$\begin{aligned} precision &= \frac{|\{P_i \in P | \exists C_j \in C, P_i \text{ matches } C_j\}|}{|P|} \\ recall &= \frac{|\{C_i \in C | \exists P_j \in P, P_j \text{ matches } C_i\}|}{|C|} \\ F\text{-measure} &= \frac{2 \times precision \times recall}{precision + recall} \end{aligned}$$

2.5.3 Performance on overlapping complexes

To assess the performance of the methods on overlapping complexes, we, first, tested each possible pair of protein complexes for overlap. If two complexes have at least one common protein, these two complexes are overlapping. Next, for each overlapping complex, the best match cluster score was calculated.

Among large yeast complexes 66 pairs of complexes are overlapping, which includes 68 complexes overlapping with at least one complex. In human, 7446 pairs of complexes are overlapping, which corresponds to 625 large complexes overlapping with at least one complex. Because 96% of human reference complexes are overlapping, we do not investigate the performance on overlapping complexes for the human dataset.

2.6 Results and discussion

We now compare ClusterOne, CMC, IPCA and COACH comprehensively, as they all can predict overlapping complexes.

Yeast dataset

A good protein complex prediction algorithm should identify as many known complexes as possible. Table 2.2 provides the basic information of predictions by various methods. As can be seen from the table, the clusters generated by CMC match more real complexes than other methods for *match-thresh* = 0.75.

Figure 2.1 summarizes the performance of all methods on prediction of large yeast complexes in terms of recall, precision and F-measure. We observe

Algorithms	CMC	ClusterOne	COACH	IPCA
# predicted complexes	354	548	652	1136
# detected real complexes	68	45	65	57
# predicted complexes match real	67	43	84	158

Table 2.2: The results of various algorithms on yeast data, *match-thresh* = 0.75.

that CMC can achieve the highest recall and precision, which shows that it can predict protein complexes more accurately. IPCA yields the closest recall to CMC, but has much lower precision, suggesting that IPCA suffers from many false positives. At the same time ClusterOne and COACH perform very poorly (recall 30% and 44% respectively). It is clear that the prediction of the big yeast complexes is a difficult task.

Figure 2.2(a) illustrates the distribution of the best-match cluster scores for the reference protein complexes across various methods. The methods are in the *x-axis*, and the distribution of the best-match cluster scores are in the *y axis*. It is clear that CMC has the highest median score. Therefore, clusters predicted by CMC method are observed to match real protein complexes better than those predicted by other approaches. Figure 2.2(b) illustrates that all methods show lower median of best-match scores for overlapping complexes. Therefore, predicting overlapping complexes is difficult. The predicted clusters by CMC match real overlapping complexes with higher match score. The above results clearly indicate that CMC outperforms all other approaches on the yeast dataset.

Human dataset

Table 2.3 shows the basic information of predictions by various methods. In Table 2.3, CMC predicted less number of complexes (450) than other approaches, but more precisely (117 match 223 real complexes). Figure 2.3(a) shows the performance of the algorithms on the prediction of human complexes at a matching requirement of *match-thresh* = 0.5. As can be seen,

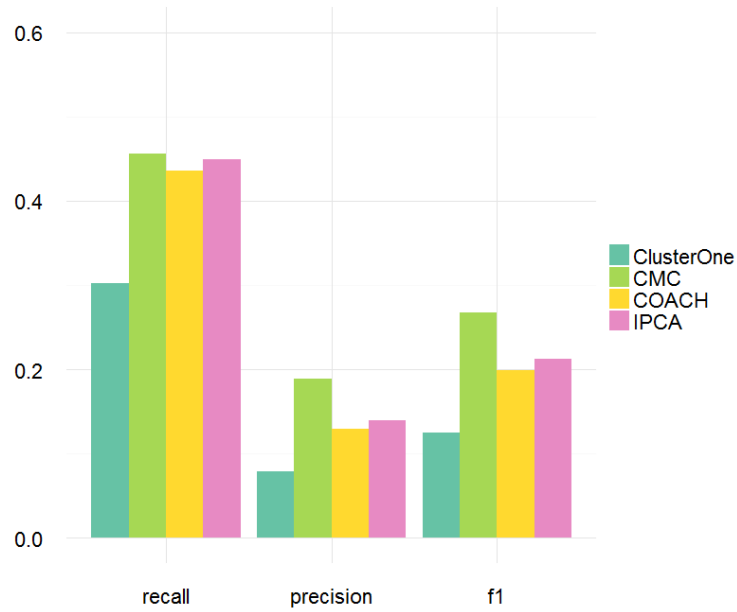


Figure 2.1: CMC, ClusterOne, COACH, and IPCA performance analysis on prediction of yeast complexes with $match-thresh = 0.75$.

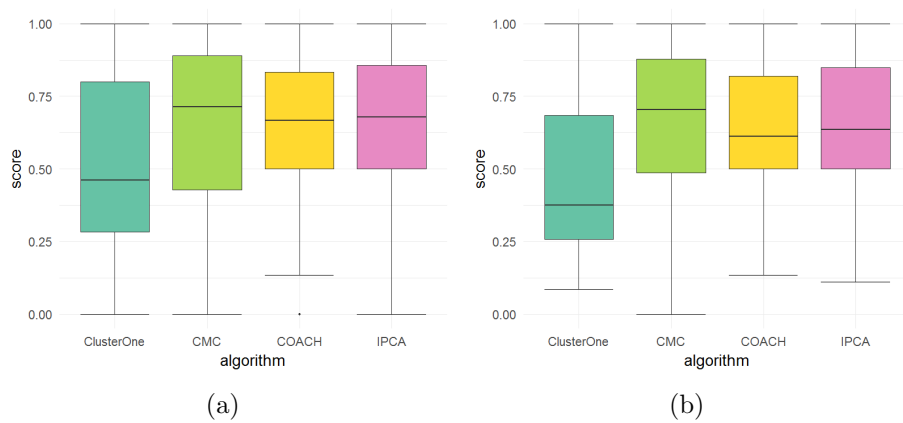


Figure 2.2: CMC, ClusterOne, COACH, and IPCA the distribution of match score of real yeast complexes (a), overlapping real yeast complexes (b).

Algorithms	CMC	ClusterOne	COACH	IPCA
# predicted complexes	450	1427	945	1886
# detected real complexes	223	86	205	250
# predicted complexes match real	117	53	138	369

Table 2.3: The results of various algorithms on human data, *match-threshold* = 0.5.

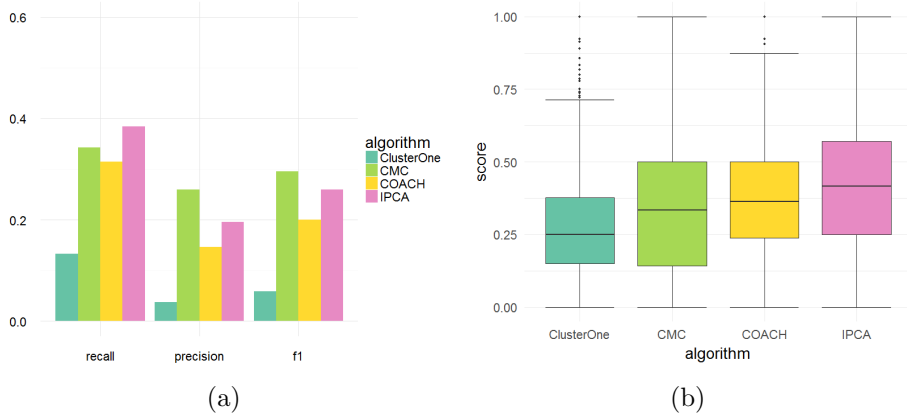


Figure 2.3: CMC, ClusterOne, COACH, and IPCA performance analysis on prediction of human complexes the precision, recall, F score, with *math - threshold* = 0.5 (a), and CMC, ClusterOne, COACH, and IPCA the distribution of match score (b).

more than half reference complexes cannot be predicted. It is clear that IPCA algorithm provides better recall, compared to other methods, but suffers from low precision. Again ClusterOne and COACH perform poorly. In terms of precision and F-measure, CMC outperforms other methods.

To investigate which reference complexes match predicted clusters with higher match score, we use the definitions from Yong & Wong (2015b) to stratify all reference complexes by density (DENS) into low, medium, and high DENS. The density for protein complex is defined as the ratio of the number of PPI edges in the complex versus the total number of all possible edges in the complex. Then complexes with low density correspond to DENS of [0,0.35], complexes with medium density correspond to DENS of (0.35,0.7], and complexes with high density correspond to DENS of (0.7,1].

We study the best-match score distribution for each subset reference com-

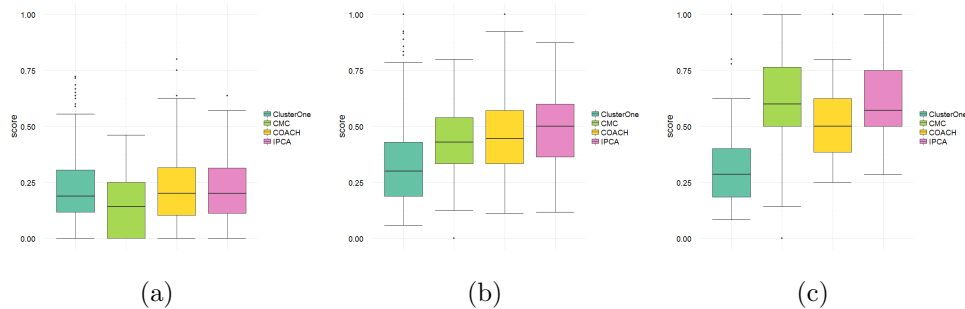


Figure 2.4: CMC, ClusterOne, COACH, and IPCA performance analysis on human complexes, stratified by DENS. a: The match score of the best cluster for reference complexes in low DENS. b: The match score of the best cluster for reference complexes in medium DENS. c: The match score of the best cluster for reference complexes in high DENS.

plexes among three DENS strata. In Figures 2.4(a) - 2.4(c), we observe that ClusterOne, COACH, IPCA perform very poorly on reference complexes with low DENS. CMC is also seen not performing very well on extremely sparsely-connected complexes. The possible reason is that complexes with low DENS do not form dense clusters that these algorithms can pick. IPCA has a higher median score for complexes with medium DENS. On the other hand, CMC leads the other methods on high DENS complexes, whereas ClusterOne performs the worst.

2.7 Conclusion

From our experiments above, we showed that the performance of existing protein complex prediction methods is not entirely satisfactory. We presented results of four algorithms for prediction of large complexes in yeast and human, and showed that in conjunction with higher recall, the predictions from CMC are more precise for the yeast dataset, and have the highest F-measure on the human dataset. Further analysis demonstrated that clusters predicted by CMC have the highest median match score. Therefore, in terms of all measures, CMC outperforms the other

approaches on identifying protein complexes. For this reason, we use only CMC in later comparisons in Chapters 3 to 5.

One possible reason why predicted clusters may not match real complexes is that the reference data are incomplete. Therefore, predicted complexes not matching any reference are not necessarily wrong and may correspond to novel complexes. In addition, many proteins participate in multiple complexes and form overlapping complexes in our reference datasets, leading to highly-connected regions, which make it difficult for protein complex prediction algorithms to identify the complexes' boundaries. To address the problem of predicting overlapping complexes, an approach like network decomposition (Liu et al. 2011) is promising. It decomposes a PPI network by e.g. removing proteins with high degree (hubs) which are thus more likely to participate in different complexes. This motivates us to examine a list of proteins, which bind their different partners at different time or at different location (viz. date hubs), manually collected from some recent study (Pritykin & Singh 2013), for network decomposition. We also propose a methodology to predict date hubs based on simple network features. Our approach and findings are discussed in the following chapters.

Chapter 3

Predicting Protein Complexes from PPI Network Decomposed by Known Date Hub Removal

3.1 Introduction

In this chapter, we first describe reference lists of date hubs. Then we investigate which list of date hubs is more reliable to use for network decomposition. Finally, we study network decomposition's impact on the performance of CMC by removing different lists of date hubs.

3.2 Background

3.2.1 Date and party hub proteins

To perform different cellular functions a protein may be recruited by more than one complex, and interacts with distinct sets of partners. A hub protein is a protein that has a lot of interacting partners in the PPI network. For a better understanding of protein interaction networks, Han et al. (2004) studied hub proteins with gene expression data, and first introduced the distinction between date and party hubs. First, they defined hubs as proteins with degree greater than 5. Then for each hub, they calculated the average of Pearson correlation coefficients (PCC) between the hub protein and each of its neighbors for mRNA expression. Their results suggest that party hubs are co-expressed with their interacting partners (have higher average PCC), while date hubs have significantly more diverse localization of partners. Therefore, party hubs are hubs that interact with their partners at the same time, whereas date hubs bind their different partners, which belong to multiple complexes, at different times or at different locations. Using an arbitrary average PCC threshold, the hub proteins with higher values of average PCC than the threshold were defined as party hubs, and all other proteins with the degree higher than 5 were indicated as date hubs.

A more recent study (Pritykin & Singh 2013) examined the correlation between the average co-expression of a hub protein with its partners and its different topological measures, such as betweenness centrality and clustering coefficient. While the first study used an arbitrary threshold to define hub proteins, this research considered hubs to be proteins in the top 10% by degree. It has been demonstrated that hubs with low betweenness or high clustering coefficient tend to have high average co-expression with

their partners. This suggests that the hub proteins can be partitioned into two classes based on their tendency to be co-expressed with their interacting partners, with significantly different network properties. Finally, it was confirmed that these simple topological and co-expression measures of hub proteins tend to be conserved across different organisms including *S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *A. thaliana*, and *E. coli* (Pritykin & Singh 2013).

3.2.2 Network decomposition by hub removal

Unfortunately, classic protein complex prediction algorithms are not able to delimit the overlapping complexes precisely. Therefore, Liu et al. (2011) proposed a possible solution via removing proteins with many interacting partners (hubs) and all their edges from a PPI network before clustering. The whole process of removing hub proteins from the given PPI network is described below:

1. Remove hub proteins from a given PPI network.
2. Apply a protein complex prediction algorithm (viz. CMC in this thesis) to find clusters from the remaining network.
3. Add a hub protein u back to a generated cluster C only if $Connectivity(u, C) \geq hub_add_thresh$:

$$Connectivity(u, C) = \frac{\sum_{v \in C} w(u, v)}{|C|} \quad (3.1)$$

where $w(u, v)$ is the weight of edge (u, v) , and $hub_add_thresh = 0.3$.

However, some hub proteins may interact with their partners simultaneously (Han et al. 2004), and hence do not participate in overlapping complexes. This motivates us to inspect a gold-standard set of date hub pro-

teins collected from the literature, and evaluate the performance of CMC after applying the decomposition strategy using these date hub proteins.

3.3 Material and Method

3.3.1 Datasets

Reference lists of date and party hubs

We first manually collect the reference lists of date and party hubs for yeast from two studies (Han et al. 2004, Pritykin & Singh 2013). The first list consists of 91 date hubs and 108 party hubs. The second list has 358 date hubs and 178 party hubs. In both lists, the hub IDs are in the form of Uniprot IDs. We call the first dataset of date and party hubs *Han_2004*, and we call the second dataset *Pritykin_2013*.

The human reference lists of date and party hubs were collected from (Pritykin & Singh 2013), and have 294 date hubs and 146 party hubs. Human hub IDs are in the form of ENSEMBLE gene IDs, while the PPI network is in Uniprot id. Using Uniprot ID mapping tool we converted ENSEMBLE IDs to Uniprot IDs. ENSEMBLE gene IDs can map to multiple proteins' Uniprot IDs, but only one or two of these may actually be found in the PPI network. As a result, we have 271 date hubs and 141 party hubs for the human reference list.

List of hub proteins

Hub proteins were used by (Liu et al. 2011) to decompose PPI network. Different degree thresholds were tested in Liu et al. (2011) to define *hub proteins* and to optimize the performance of clustering algorithms. The best performance of yeast complex discovery by CMC in terms of F-measure

was when the degree threshold was set to 50. Therefore, for future analysis, all proteins in the yeast PPI network with degree not less than 50 form a list of 124 date hubs which we call *HUB_50*. According to Liu et al. (2011), for human dataset, the optimal degree threshold is 150, leading to the list *HUB_150* of 5 date hubs.

Thus, in total, four sets of input date hub proteins belonging to three different studies are used in our analysis.

3.3.2 Approach to compare the quality of reference lists of date hubs

According to the definition, date hubs are proteins which interact with many partners at different time and location. Hence, if two reference complexes of different complex families overlap then the proteins within their intersection should correspond to the date hubs. Therefore, we assume that the real date hubs are more likely to appear within the intersection of overlapping complexes than the party hubs. One way to evaluate the quality of reference lists of date hubs is to investigate how many date hub and party hub proteins are present within the overlap of real protein complexes.

As mentioned in Subsection 2.5.3, among large real yeast complexes 66 pairs of complexes are overlapping. Hence, for each reference list of date and party hubs, we calculate the number of overlapping real complexes which have at least one hub in the intersection.

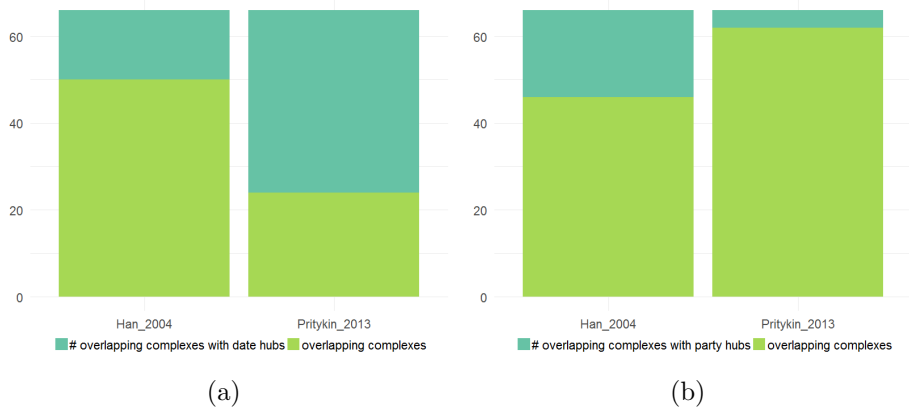


Figure 3.1: a: Presence of date hubs in the intersection of yeast reference complex pairs, b: Presence of party hubs in the intersection of yeast reference complex pairs.

3.4 Results and discussion

3.4.1 Comparing the quality of reference lists of date hubs

In this experiment we study the quality of date hubs collected from the literature. The experiment is performed only on the yeast dataset because the *Han_2004* date hubs are available only for yeast.

Figure 3.1(a) shows that 65% of overlapping complex pairs contain some date hubs from *Pritykin_2013* in their intersection, and only a few pairs contain party hubs (Fig. 3.1(b)). These findings confirm the difference in the role of date and party hubs across overlapping complexes. On the other hand, the proportion of overlapping complex pairs, which contain date hubs from *Han_2004* within intersection, are much lower compared to *Pritykin_2013* dataset. This shows that the date hubs from *Han_2004* are less relevant to overlapping complexes, and thus, for all further experiments we use the *Pritykin_2013* list of date hubs.

hubs list	# hubs	# PPIs discarded
HUB_50	124	6841
Pritykin_2013	358	7465

Table 3.1: Different list of hub proteins used and the resulting number of proteins and PPIs discarded in the decomposed networks.

3.4.2 Protein complex prediction

In the second experiment, we study the impact of date hub removal on the performance of CMC. We first apply the PPI network decomposition technique mentioned in subsection 3.2.2 to remove the reference list of date hubs in *Pritykin_2013*. We then run CMC on the remaining networks, and add back all date hubs to the generated clusters. For comprehensive comparison, we also run CMC on the decomposed network obtained by removing the hubs (for yeast: *HUB_50* list, for human: *HUB_150*), which is similar to the initial work by Liu et al. (2011).

Experiment settings

We use the same PPI datasets and the same sets of reference complexes as described in Chapter 2.5.1, considering only top 20 000 interactions. Only CMC is used for complex discovery. For comprehensive analysis, we employed the same evaluation measures, including F-measure and the distribution of the best match cluster score.

Observations on yeast dataset

Table 3.1 shows the number of interactions discarded by removing different sets of hubs. The *Pritykin_2013* set has almost three times more hubs than *HUB_50*, but the number of PPIs discarded from the network is not significantly different. The possible explanation is that the date hubs from the *Pritykin_2013* list are obtained from a network which is different from ours, leading to many proteins with a small degree in our PPI network.

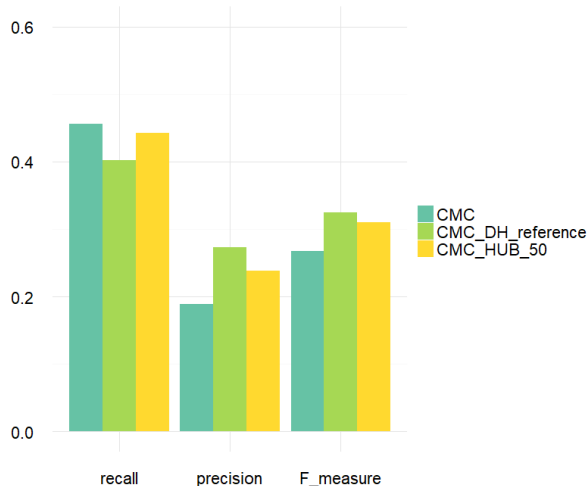


Figure 3.2: Performance of CMC, CMC with removing reference list of date hubs (*CMC_DH_reference*), and CMC with removing hubs with degree at least 50 (*CMC_HUB_50*), on prediction of yeast complexes with *match-thresh* = 0.75.

Algorithms	CMC	CMC_HUB_50	CMC_DH_reference
# date hubs	0	124	348
# predicted complexes	354	277	220
# detected real complexes	68	66	60
# predicted complexes match real	67	66	60

Table 3.2: The results of CMC on yeast data, *match-threshold* = 0.75.

From Figure 3.2, it is clear that the complexes predicted by CMC after network decomposition by date hub removal (*CMC_DH_reference*) are better in terms of precision and F-measure, and comparable, in terms of recall, to the predictions of its competitors.

In Figure 3.3, we observe that the network decomposition technique does not improve the performance in terms of best-match cluster. The conventional CMC approach gives the best matching scores compared to the predictions of *CMC_HUB_50* and *CMC_DH_reference*. This may happen because some reference complexes has only four proteins, and some of these proteins correspond to date hubs. Therefore, once we remove the date hub during network decomposition, the remaining complex has three proteins and CMC is not able to recover the complex anymore.

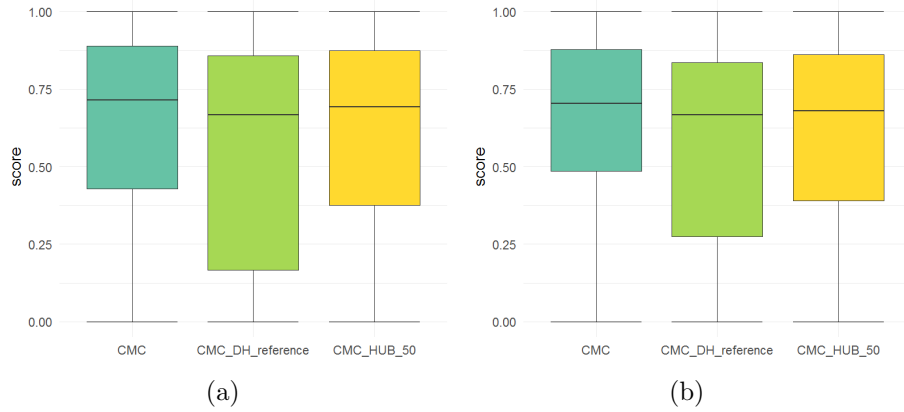


Figure 3.3: Match scores of the best clusters to yeast complexes. a: all real yeast complexes, b: overlapping real yeast complexes.

Observations: Human

For human dataset we also measure precision, recall, F-measure (Fig. 3.4), and inspect the best match cluster’s score distribution among three DENS strata (Fig. 3.5). We observe a similar trend that removing the reference list of date hubs gives mainly an improvement in precision. The recall drops significantly because many proteins and interactions were discarded. In addition, *CMC_DH_reference* shows the smallest median best match score across all DENS strata. Table 3.3 summarizes the basic information of predictions by CMC.

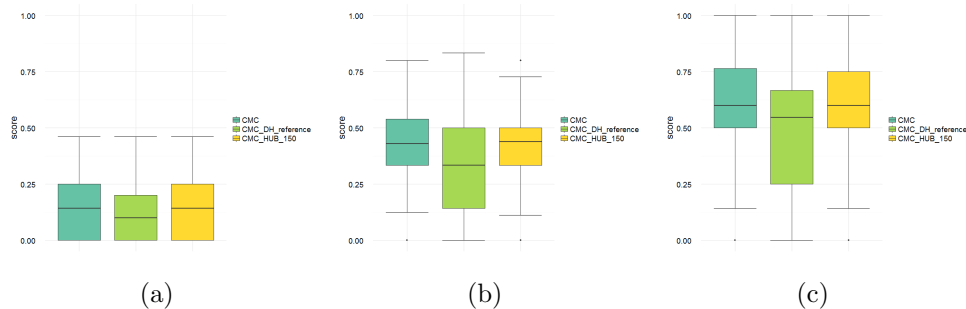


Figure 3.5: Match score of the best clusters to human complexes for CMC in low DENS (a), medium DENS (b), and high DENS (c).

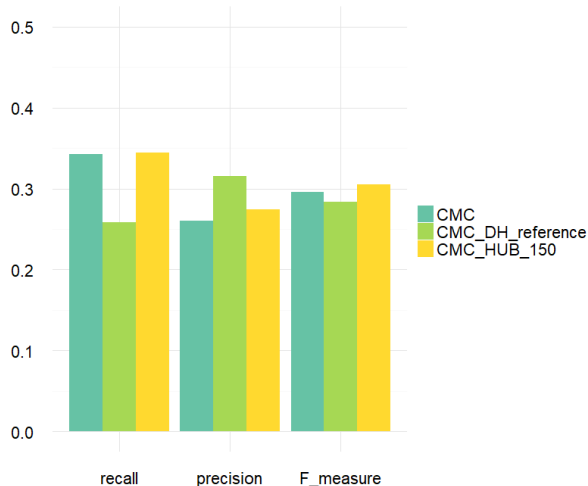


Figure 3.4: Performance of CMC, CMC with removing reference list of date hubs (CMC_DH_reference), and CMC with removing hubs with degree at least 150 (CMC_HUB_150), on prediction of human complexes, $match-thresh = 0.5$.

Algorithms	CMC	CMC_HUB_150	CMC_DH_reference
# date hubs	0	5	271
# predicted complexes	450	419	260
# detected real complexes	223	224	168
# predicted complexes match real	117	115	82

Table 3.3: The results of CMC on human data, $match-thresh = 0.5$.

3.5 Conclusion

The earlier study (Liu et al. 2011) showed reasonable performance improvement in CMC by removing hubs from the PPIs network before apply CMC. In this Chapter, we assume that a reliable list of date hubs may further improve the network decomposition technique. So, we manually collect a reference list of date hubs from the literature.

As a starting point, we find that the date hubs tend to occur within the intersection of real overlapping protein complexes. Moreover, we observe that *CMC_DH_reference* achieves the highest F-measure on yeast, with

the highest precision and comparable recall, which shows that it predicts protein complexes more accurately. However, the distribution of the best match cluster score has the lowest median score. This motivates us to create a reliable list of date hub proteins by inspecting the topological properties of the date hub reference list.

Chapter 4

Predicting Protein Complexes from PPI Network

Decomposed by Removing Predicted Date Hub Proteins

4.1 Introduction

Earlier works on hub proteins in PPI networks show that they can be classified into date and party hubs on the basis of their partners' expression profiles. Previous research has established that properties of date and party hubs are significantly distinct. In the next several sections, we inspect three properties of date and party hubs on our PPI networks: degree, betweenness centrality, and transitivity (also called the clustering coefficient) to confirm whether date hubs can be distinguished from party hubs in terms of these properties. Degree is used as it has been shown from previous experiments that PPI network is sensitive to removal of proteins with high degree, also known as hubs. Protein betweenness centrality is calculated as

a fraction of shortest paths passing through a protein in the network (Equation 4.1). Proteins with higher betweenness usually lie between complexes, potentially indicating they are date hubs.

$$\textit{betweenness}(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (4.1)$$

where σ_{st} denote the total number of shortest paths from protein s to protein t , and $\sigma_{st}(v)$ is the number of those paths that pass through v .

Finally, transitivity of a vertex v measures the probability that its immediate neighbors are connected, and calculated as a ratio of the number of edges between all immediate neighbors for vertex v to the number of edges that could possibly exist between them. All these measures are purely topological and do not use any information other than interaction data.

4.2 Methods

4.2.1 Date Hub Prediction

Previous research has established that network properties of date and party hubs are distinct (Pritykin & Singh 2013). It should thus be possible to improve the network decomposition approach by predicting date hubs based on some simple topological features of PPI networks. First, we compute the hub protein lists based on the degree of a vertex. Across all hub proteins only some of them may correspond to the date hubs, the rest are party hubs. Here, three topological measures (degree, betweenness, and transitivity) are employed to characterize proteins and investigate whether date hubs can be distinguished from party hub proteins in terms of these properties. For each property, we discuss different thresholds to get a list of

date hub candidates. Then, we incorporate the predicted list of date hubs into network decomposition, run CMC on the decomposed network, and evaluate the performance by calculating precision, recall and F-measure.

4.3 Results and discussion

In this section, we analyze the top 20 000 edges of PPI networks for human and yeast, with 6352 and 3680 proteins respectively. First, three topological measures are employed to characterize proteins and investigate whether date hubs can be distinguished from party hub proteins in terms of these properties. Next, for each property, we discuss different thresholds to get a list of date hub candidates.

4.3.1 Network topology analysis

Experiment settings

Degree, betweenness, and transitivity of proteins were calculated by using the iGraph library (Csardi & Nepusz 2006) in the statistical computing environment R.

Observations on yeast dataset

For the yeast network, we confirm the clear differences between date and party hubs in terms of transitivity and degree. From Figure 4.1, we observe that party hubs have higher transitivity and higher average degrees, leading to denser neighborhoods. Therefore, it seems interesting to investigate how to effectively combine two measures to predict a list of date hubs. We also note that date hubs tend to have a higher betweenness, suggesting that they are more globally central in the network.

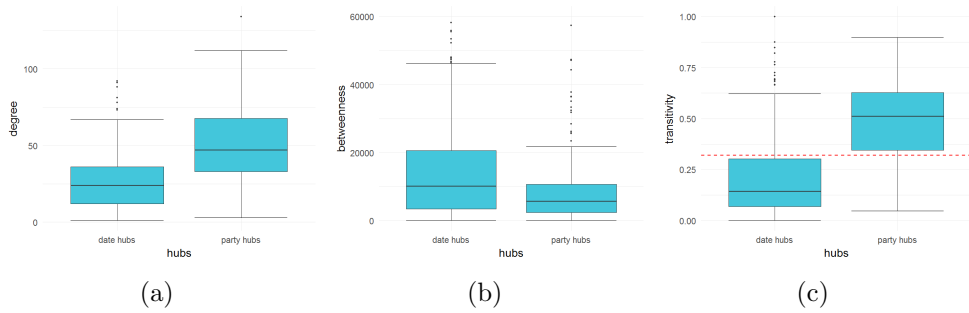


Figure 4.1: Date and party hub analysis of topological features in the yeast high quality interaction network: degree (a), betweenness centrality (b), transitivity (c).

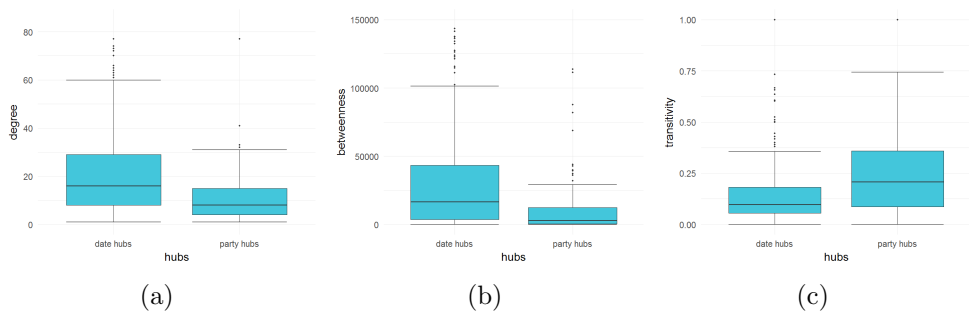


Figure 4.2: Date and party hub analysis of topological features in the human high-quality PPI network: degree (a), betweenness (b), transitivity (c).

Observations on human dataset

Next, we inspect the topological properties on the human PPI network. Figure 4.2 shows the distribution of degree, betweenness, and transitivity. We note that there are less obvious differences across all three measures. A possible explanation is that the current human PPI network is vastly incomplete (Venkatesan et al. 2009). Although the network is incomplete, the date hubs tend to have a higher betweenness and lower transitivity than party hubs.

4.3.2 Impact of different thresholds on predicting date hubs

Next we study the impact of different transitivity and degree thresholds on the selection of date hub proteins for network decomposition.

Different degree thresholds on yeast dataset

Analysis of topological properties showed significant differences in protein transitivity and degree for the yeast network, with date hubs having consistently lower values. Less consistent results are found for betweenness. Therefore, we examined only transitivity and degree as features to make predictions of date hub proteins by setting different thresholds.

The goal at this step is to compute the hub protein lists before predicting a list of date hubs. Liu et al. (2011) has already defined a protein as a hub protein if it has at least N_{hub} neighbors, and investigated different thresholds. They found that CMC achieves the best performance when $N_{hub} = 50$. On the other hand, Pritykin & Singh (2013) defined hubs as all proteins in the top 10% in the PPI network by the number of interactions, leading to the degree threshold of $N_{hub} = 23$. We include both thresholds into our experiments, and create two list of hub proteins:

1. *HUB_50* contains 124 proteins with at least 50 neighbors.
2. *HUB_23* contains 496 proteins with at least 23 neighbors.

Different transitivity thresholds on yeast dataset

As suggested by Figure 4.1(c), there is a clear separation between date and party hubs in terms of transitivity. Based on results from Figure 4.1, we set a transitivity threshold of 0.32 as the optimal point to separate date and

Name	# of date hubs	# of PPIs removed	Degree	Transitivity
HUB_23	496	13896	23	–
HUB_23_0.32	181	7694	23	0.32
HUB_50	124	6941	50	–
HUB_50_0.32	51	3837	50	0.32

Table 4.1: Different values of degree and transitivity used, and the resulting number of date hub proteins and PPIs discarded in the decomposed networks.

Algorithms	HUB_50	HUB_50_0.32	HUB_23	HUB_23_0.32
# predicted complexes	277	304	163	220
# detected real complexes	66	67	62	67
# predicted complexes match real	66	66	59	66

Table 4.2: Performance statistics on yeast complex discovery, $match-thresh = 0.75$.

party hubs. A hub protein with transitivity value lower than the threshold is predicted to be a date hub protein.

To support our choice of threshold, we examine the effect of setting transitivity threshold of the hub proteins from *HUB_50* and *HUB_23*. Table 4.1 shows the number of date hub proteins selected, and the number of interactions discarded, under different degree and transitivity values. The smaller the degree, the larger the number of interactions is discarded. To be more precise, from *HUB_50_0.32* and *HUB_23_0.32* we remove all hubs which were reported by Pritykin & Singh (2013) as party hubs. Finally, we remove the remaining date hubs from the given PPI network, and apply CMC on the resultant networks. After the clusters are generated, date hub proteins are added back to the clusters. Table 4.4 shows the numbers of known complexes matched to the clusters generated by CMC when different degree and transitivity thresholds are used for selecting date hub proteins, for yeast complex prediction at $match-thresh = 0.75$.

We subsequently examined the precision, recall and F-measure to further

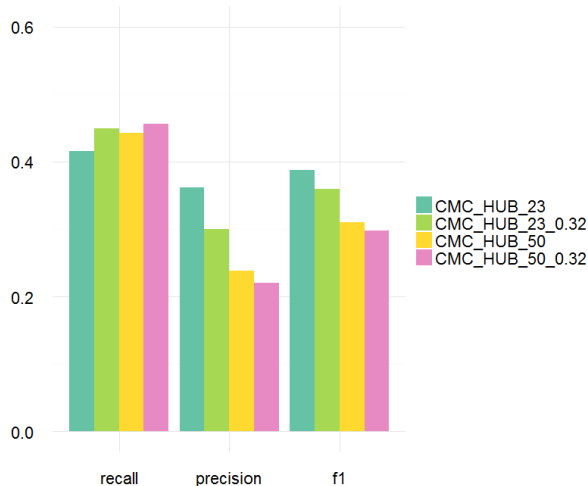


Figure 4.3: Performance statistics for yeast complex discovery at $match-thresh = 0.75$ using different lists of date hubs.

illustrate the effect of transitivity threshold on yeast complex prediction by CMC, at $match-thresh = 0.75$. From Figure 4.3, we observed that the precision and F-measure of different groups vary dramatically. It is clear that when we remove only the hub proteins from *HUB-50* the performance of predicted clusters is not perfect. Once we apply the transitivity threshold to *HUB-50*, the number of potential date hub proteins drops significantly, and we can recall more complexes with comparable precision. In addition, when we decrease the degree threshold and remove the date hub proteins (*HUB-23.0.32*), precision improves considerably, while recall remains similar to *HUB-50*. However, when we remove only the hub proteins (*HUB-23*), recall, as expected, drops substantially because too many proteins and interactions are discarded, and the precision is improved because many false-positive clusters are removed. Overall, the highest F-measure was obtained when we decompose PPI network by removing *HUB-23.0.32* date hub proteins. Therefore, we make predictions of date hubs by setting the degree threshold to 23 and the transitivity threshold to 0.32. The list of predicted date hubs can be found in Appendix A.

Different degree thresholds on human dataset

Next we inspect the degree and transitivity features for the human dataset. It has been reported that the hub-removal strategy gives the best performance for CMC when the hub proteins have at least 150 neighbors (Liu et al. 2011). However, we note that only five proteins in the high-confident human PPI network have degree more than 150. Therefore, we assume that this list of hub proteins for human dataset is incomplete, and potentially more date hub proteins can be found. According to (Pritykin & Singh 2013) the degree threshold for human hub proteins is 39. Therefore, in order to obtain the list of hub proteins, we tested three values for degree – 30, 40, and 50. A protein with degree value higher than the threshold is considered to be a hub protein. Hence, we examine three lists of hubs proteins – *HUB_30*, *HUB_40*, and *HUB_50*. For each list of hub proteins we apply different transitivity thresholds to shortlist the date hub proteins. Then the network is decomposed by removing date hub proteins, the clusters are predicted by CMC, and the date hubs are added back to the clusters.

Different transitivity thresholds on human dataset

Unfortunately, as discussed earlier, the results obtained from the preliminary analysis of topological features do not show a clear difference between the date and party hubs in terms of transitivity (Fig. 4.2 (c)) for human dataset. However, we note that the date hubs tend to have smaller transitivity. As the human PPI network is highly incomplete, the probability that a hub protein’s immediate neighbors have an edge between them is low. Thus, we assume that the transitivity value for date hub proteins is extremely small. Therefore, different transitivity thresholds are used for indicating hub proteins, for decomposing PPI network. We use the distribution of the match score of the best cluster for reference complexes stratified by DENS to study the effectiveness of combinations of our features in discriminating date hub proteins (Fig. 4.4). For clarity, we only

Name	# date hubs	# PPIs removed	Degree	Transitivity
HUB_30	78	5481	30	0.05
HUB_30	168	9060	30	0.1
HUB_30	197	9794	30	0.15
HUB_40	59	4897	40	0.05
HUB_40	119	7765	40	0.1
HUB_40	126	7982	40	0.15
HUB_50	46	4382	50	0.05
HUB_50	86	6626	50	0.1
HUB_50	88	6691	50	0.15

Table 4.3: Different values of degree and transitivity used, and the resulting number of date hub proteins and PPIs discarded in the decomposed networks.

Algorithms	HUB_30_0.05	HUB_40_0.05	HUB_50_0.05
# predicted complexes	296	302	308
# detected real complexes	191	192	191
# predicted complexes match real	98	99	98

Table 4.4: Performance statistics on human complex discovery, *match-threshold* = 0.5.

show the results for *transitivity* = 0.05, 0.1, and 0.15.

Our experiments show that when the transitivity is high (*transitivity* = 0.15) and the degree is low (30), many reference complexes do not match any predicted cluster (Fig. 4.4: 0.15, *low/medium/high DENS*). This is because too many proteins and interactions were discarded, and CMC may not recover the reference complexes. But when we decrease the transitivity threshold, the median match score increases (Fig. 4.4: 0.1, *low/medium/high DENS*). Moreover, we observe that when the value for transitivity is extremely small (0.05) the distribution of the best match score is similar across different degree threshold, because the lists of predicted date hub proteins are almost the same for these thresholds settings. Table 4.4 shows the basic information of predictions by using different degree threshold. From Table 4.4 and Figure 4.5, we conclude that the CMC obtains the best performance when we decompose PPI network by removing proteins which have at least 30 neighbors and the transitivity value is not more

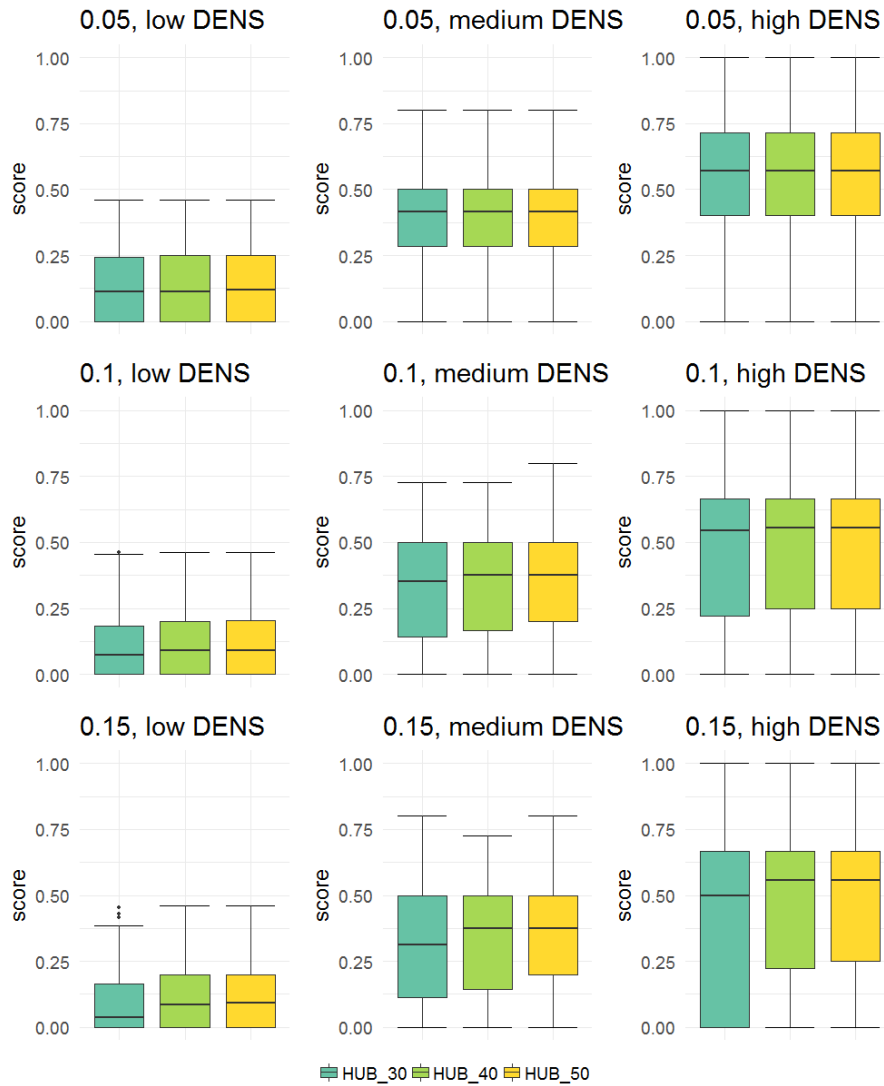


Figure 4.4: Match scores of the best clusters for human complexes at different transitivity thresholds of 0.05, 0.1, 0.15 in the three DENS strata: low DENS, medium DENS, and high DENS.

than 0.05 .

4.3.3 Protein complex prediction

Observations on yeast dataset

Figure 4.6a shows precision, recall and F-measure of CMC when predicted date hubs and date hubs from reference dataset are used for the network decomposition. Very clearly, CMC benefits much from removing predicted date hubs ($CMC_{DH_predicted}$) and shows significant improvement in re-

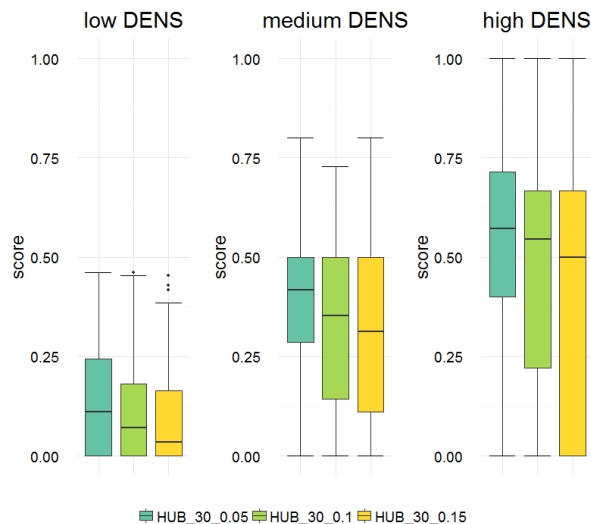


Figure 4.5: Match scores of the best clusters for human complexes at different transitivity thresholds of 0.05, 0.1, 0.15 in the three DENS strata: low DENS, medium DENS, and high DENS.

Algorithms	CMC	CMC_DH_predicted	CMC_ DH_reference
# predicted complexes	354	220	220
# detected real complexes	68	67	60
# predicted complexes match real	67	66	60

Table 4.5: Performance statistics on yeast complex discovery, $match-thresh = 0.75$.

call compared to $CMC_DH_reference$. $CMC_DH_predicted$ also attains the highest F-measure. In addition, we observe that conventional CMC has lower precision compared to CMC after we decompose network by removing the predicted list of date hubs. Table 4.5 summarizes the performance of different variants of CMC on yeast complex discovery at $match-thresh = 0.75$. Closer inspection of the table shows that $CMC+CMC_DH_predicted$ and $CMC+CMC_DH_reference$ generate the same number of potential protein complexes, but $CMC+CMC_DH_predicted$ attains a higher number of detected real complexes.

Observations on human dataset

The prediction of complexes in human is much more challenging than that

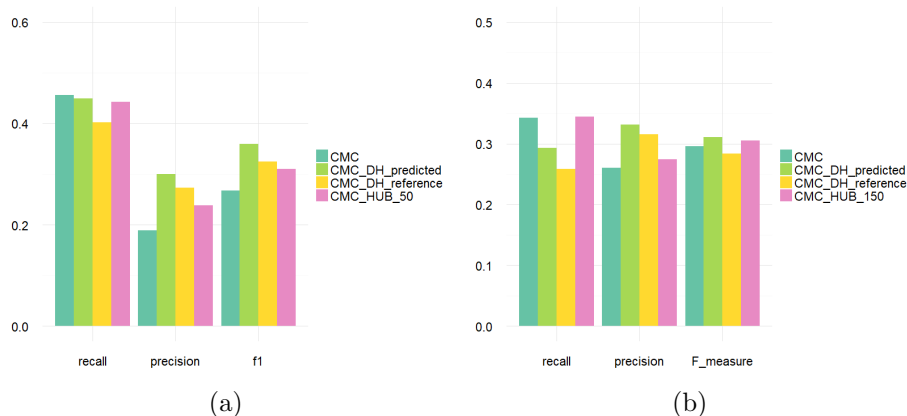


Figure 4.6: Performance analysis on prediction of (a) yeast complexes with $match\ thresh = 0.75$, (b) human complexes with $match\ thresh = 0.5$.

Algorithms	CMC	CMC_DH_predicted	CMC_DH_reference
# predicted complexes	450	296	260
# detected real complexes	223	191	168
# predicted complexes match real	117	98	82

Table 4.6: Performance statistics on human complex discovery, $match\ thresh = 0.5$.

in yeast. From Figure 4.6b, we observed that when we remove predicted date hubs before applying the CMC, *CMC_DH_predicted* achieves very good precision and the highest F-measure. Unfortunately, the recall drops significantly, because along with the date hubs we also remove a large number of interactions. In the human PPI network, the percentage of proteins with degree at least 30 is about 2%, while they correspond to about 24% of the interactions (Yong & Wong 2015b).

Table 4.6 presents an overview of predicted complexes by different variants of CMC on human dataset. Closer inspection of the table shows that *CMC_DH_predicted* attains a higher number of detected real complexes compared to *CMC_DH_reference*.

4.3.4 Example complexes

In this section we present a couple of real yeast complexes that are difficult to predict. Figure 4.7(a) shows two yeast complexes, with four overlapping proteins (YDR190C, YFL039C, YJL081C, YPL235W), involved in chromatin remodeling activity: the Ino80p complex, consisting of 12 proteins, and the Swr1 complex, consisting of 13 proteins. CMC applied on the yeast PPI network was able to predict the Ino80p complex, at *match-thresh* = 0.75. However, Swr1p complex was not predicted. Figure 4.7(b) shows the clusters generated by CMC. As can be seen, CMC found the cluster that match Swr1p complex, but with two extraneous protein (YOL012C and YLR399C) which are highly-connected to proteins in that complex. In addition, two proteins from Swr1p complex were missed (YLR085C, YLR385C). Complex Ino80p was predicted with one extraneous protein (YER092W), but one protein (YOR189W) was missed. Sw1p was not predicted by ClusterOne, COACH as well, at *match-thresh* = 0.75. It is possible to lower our *match-thresh* cutoff to include Swr1p complex.

On the other hand, when we decomposed the PPI network by removing the predicted list of date hub proteins before applying CMC, we were able to recover both complexes. Moreover, all four overlapping proteins were predicted as date hub proteins, and were incorporated into the network decomposition. In addition to Swr1 complex, after network decomposition, CMC was able to predict five new non-overlapping complexes which were not found by CMC earlier, at *match-threshold* = 0.75: Cytoplasmic exosome complex, DNA-directed RNA polymerase II complex, Ndc80p complex, nuclear cohesin complex, ubiquitin ligase ERAD-L complex. On the other hand, seven real protein complexes were no longer predicted after we removed the date hubs from network. However, we investigated that all seven complexes consist of four proteins, where at least one protein was

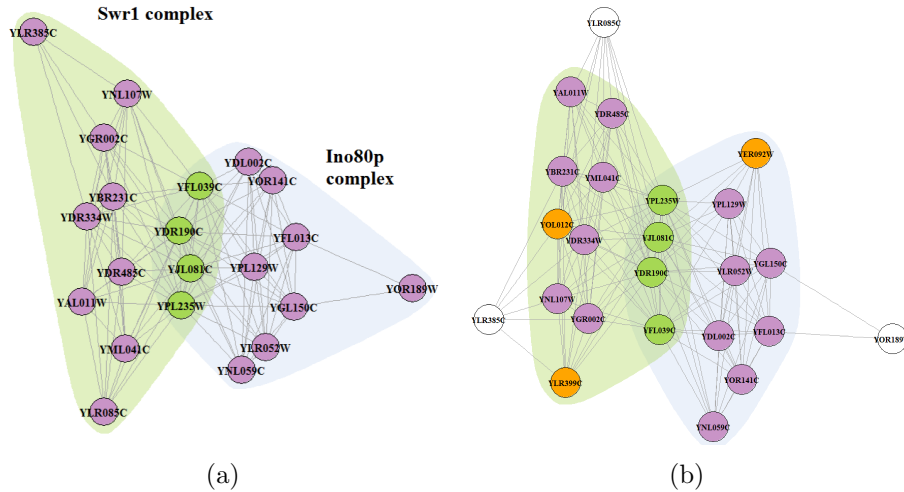


Figure 4.7: Reference yeast complexes Swr1p and Ino80p. (a) Swr1p and Ino80p complexes are overlapping with each other. Proteins within the intersection (green) were predicted as date hub proteins.(b) CMC included extraneous proteins (YLR399C and YOL12C) in its clusters and missed two proteins YLR085C and YLR385C.

predicted as a date hub. Therefore, once we decomposed the network, CMC is not able to generate clusters of size three to predict those complexes.

These results suggest that the combination of the two lists of generated clusters by CMC (before and after network decomposition) may improve the overall performance of CMC.

4.4 Conclusion

In this chapter, we have proposed a two-step approach to predict a list of date hubs based on the topological features of date and party hub proteins from the reference datasets. In step one, we computed the list of hub proteins by applying different degree thresholds. In step two, we analyzed different transitivity thresholds to eliminate party hub proteins from hub proteins. To test the predicted lists of date hubs, we examined their effect on the performance of CMC after applying the date hub removal strategy. Our results demonstrate that CMC benefits much from this, attaining

higher precision and recall across yeast and human datasets.

In addition, we analyzed protein complexes predicted by CMC before and after network decomposition. We noticed that there are missing complexes of size 4 after we decomposed a PPI network by removing predicted date hubs. Therefore, the focus of the next chapter is to overcome this limitation to further improve protein complex prediction performance.

Chapter 5

A Strategy to Combine the Two Sets of Clusters Predicted by CMC Before and After PPI Network Decomposition

5.1 Introduction

In the previous chapters, we verified that CMC after network decomposition by date hub removal has the ability to predict more overlapping complexes that were missed earlier. This illustrates that the date hub proteins are important in maintaining overlapping complexes. However, the research has also shown that the predicted complexes have lower median match score. An investigation of predicted clusters has shown that some real complexes of size 4 were not predicted after we removed the date hub proteins. Therefore, the combination of the two lists of generated clusters by CMC (before and after network decomposition) may further improve the overall performance of CMC.

5.2 Methods

5.2.1 Combining the two sets of predicted clusters

In this section, we propose a simple strategy to combine the two sets of predicted clusters. For this purpose, we modify the approach which was originally presented in the CMC algorithm (Liu et al. 2009) to merge highly overlapping cliques. The whole process to merge or remove highly overlapping clusters is described below. The modification proposed is the insertion of Step 3 and 8:

1. Let C' be the set of clusters generated by CMC without PPI network decomposition, and C'' be the set of clusters generated by CMC on the PPI network decomposed by removing predicted date hub proteins.
2. Let $C = C' \cup C''$ be the union of C' and C'' . $C = \{C_1, C_2, \dots, C_k\}$
3. Consider only clusters without date hub proteins.
4. Keep only clusters of size ≥ 4 .
5. Remove all duplicate clusters from C .
6. Sort all clusters in descending order of their weighted density.
7. For each cluster C_i :
 - Check whether there exists a cluster C_j such that C_j has a lower weighted density than C_i and $|C_i \cap C_j|/|C_j| \geq \textit{overlap_thresh}$, where *overlap_thresh* is a predefined threshold for overlapping;
 - If such C_j exists, calculate the interconnectivity score $\textit{inter_score}(C_i, C_j)$ between C_i and C_j using Formula 2.2;
 - If $\textit{inter_score}(C_i, C_j) \geq \textit{merge_thresh}$, then C_j is merged with C_i ; otherwise, C_j is removed.
8. Add the above processed clusters with date hub proteins back to clusters in C .

In the remaining experiments, for the best performance, we always set *overlap_thresh* to 0.75, and set *merge_thresh* to 0.75.

5.2.2 Quality of novel complexes

To further investigate how PPI network decomposition by date hubs removal improves the performance of large complex prediction by CMC, we study its effects on predicting novel complexes. The incomplete reference dataset of real complexes can lead to artificial low precision of predicted clusters. I.e., when the predicted complexes do not match real complexes, this does not mean that the predicted complexes are necessarily wrong; they may be novel complexes.

Therefore, next we compare the number and quality of novel complexes predicted by our approach, against those predicted by other methods on the PPI network. First, we keep only predicted complexes which do not match any reference complex at *match_thresh* = 0.75. Next, we filter all duplicates by keeping only unique clusters (*match_thresh* = 0.75). In order to evaluate the quality of novel complexes, we examine the semantic coherence for each cluster. Using the constituent proteins' annotations to Gene Ontology (GO) terms, we calculate three measures of semantic coherence for each cluster: biological process (BP), cellular compartment (CC), and molecular function (MF), using the same procedure as presented by Yong & Wong (2015b).

5.3 Results and discussion

5.3.1 Protein complex prediction

In this section, we inspect whether combining the two sets of predicted clusters can improve the performance of CMC. First, we run CMC on the PPI network and collect all generated clusters. Then we apply the network decomposition strategy by removing all predicted date hubs. Next, CMC is used on the decomposed PPI network to generate clusters. Finally, we use the procedure described in subsection 5.2.1 to combine the two sets of predicted clusters, and then compare the performance of this "double-barrel" CMC with other protein complex prediction algorithms.

For each algorithm, we use the same parameters settings as given in Table 2.1. We use the same parameters for all CMC experiments, including when the decomposition method is used. Finally, we calculate recall, precision, and F-measure. We also plot the distribution of the best-match cluster score.

Observations on yeast dataset

Figure 5.1 shows that the double-barrel *CMC+CMC_DH_predicted* has higher recall and precision than CMC, revealing the strong benefit of this combined strategy. In fact, this combined strategy has the highest recall among all methods, as well as the fourth highest precision. Its precision is expectedly lower than *CMC_DH_predicted*, since the precision of the combined strategy has to be between that of CMC and *CMC_DH_predicted*. However, we should point out that, many predicted clusters may still correspond to novel complexes, because the set of reference complexes is incomplete. Table 5.1 summarizes the performance of different variants of CMC on yeast complex discovery at *match-thresh = 0.75*.

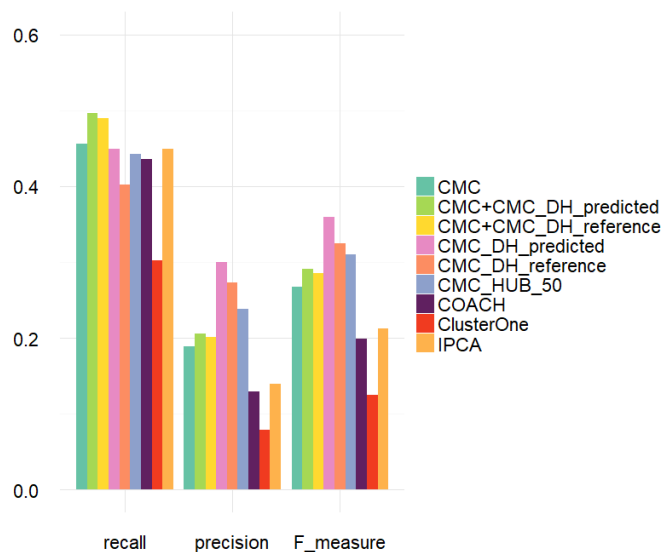


Figure 5.1: Performance analysis on prediction of yeast complexes with $match\ thresh = 0.75$.

Algorithms	CMC	CMC_DH_predicted	CMC+CMC_DH_predicted
# predicted complexes	354	220	413
# detected real complexes	68	67	74
# predicted complexes match real	67	66	85

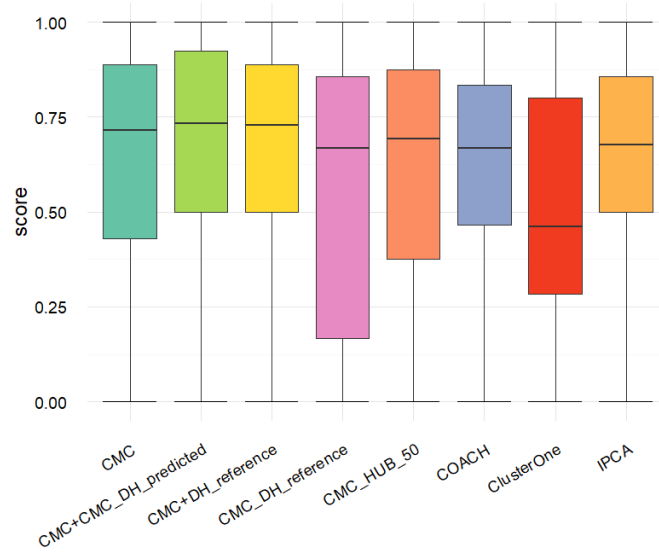
Table 5.1: Performance statistics on yeast complex discovery, $match-thresh = 0.75$.

From Figure 5.2, we observe a similar trend that *CMC+CMC_DH_predicted* generates potential protein complexes with larger value of median match score of the best clusters to yeast complexes than the other approaches.

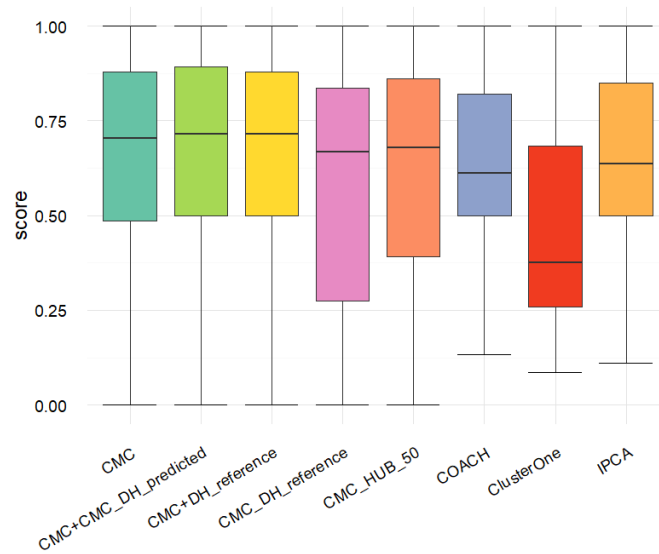
Observations on human dataset

The prediction of complexes in humans is much more challenging than that in yeast. From Figure 5.3, we observed that there is clear recall improvement for the double-barrel *CMC+CMC_DH_predicted*, with a small effect for precision.

Table 5.2 presents an overview of predicted complexes by different variants of CMC on the human dataset. Closer inspection of the table shows



(a)



(b)

Figure 5.2: a: Match scores of the best clusters for yeast complexes. b: Match scores of the best clusters for overlapping yeast complexes.

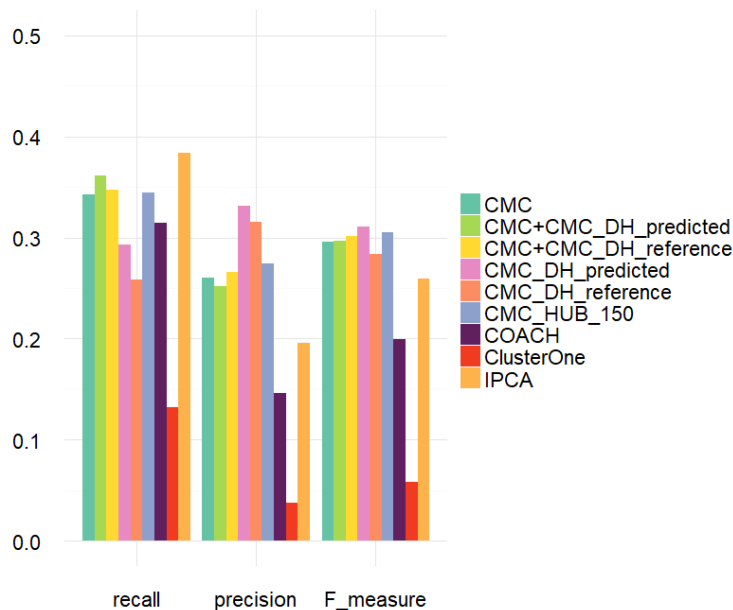


Figure 5.3: Performance analysis on prediction of human complexes with $match_thresh = 0.5$.

Algorithms	CMC	CMC_DH_predicted	CMC+CMC_DH_predicted
# predicted complexes	450	296	521
# detected real complexes	223	191	235
# predicted complexes match real	117	98	131

Table 5.2: Performance statistics on human complex discovery, $match_thresh = 0.5$.

that *CMC+CMC_DH_predicted* attains the highest number of detected real complexes.

Figure 5.4 shows the performance of CMC before and after network decomposition by removing date hubs, and the combination of the two in terms of the distribution of best match score compared to other clustering algorithms. The merged lists of generated clusters by the double-barrel *CMC+CMC_DH_predicted* has noticeably higher median score for the best match cluster to human complexes in medium and high density area. Furthermore, as in yeast, the combination of the two CMC outputs outperforms other algorithms in high density area.

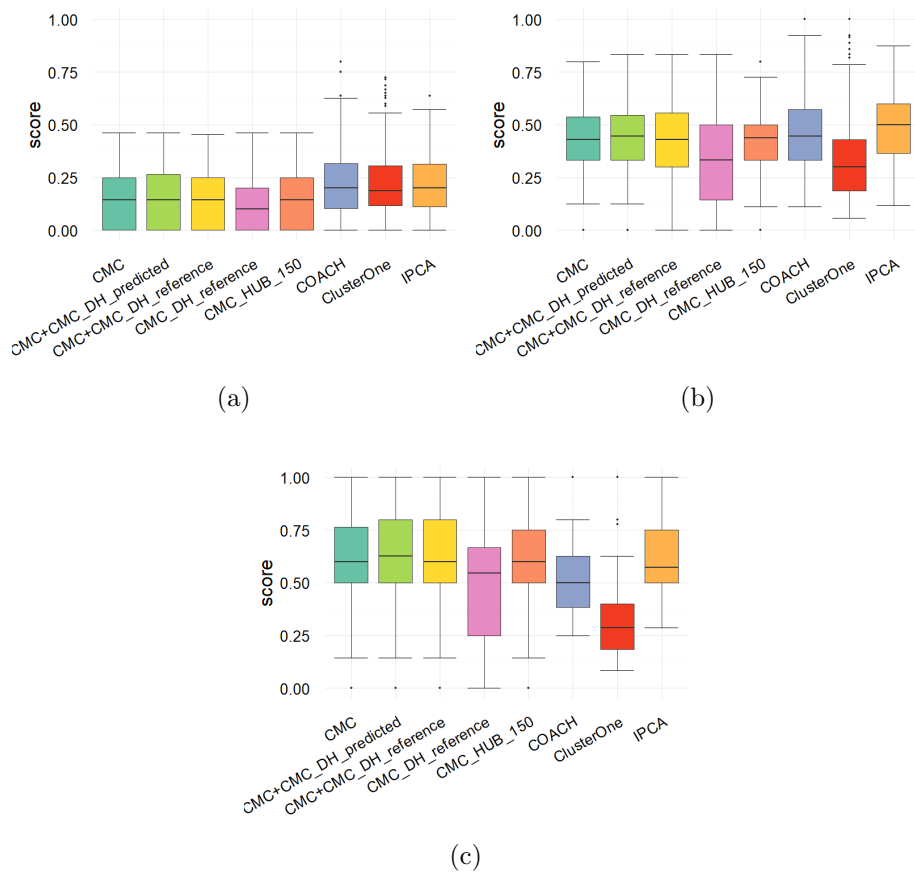


Figure 5.4: (a) Match scores of the best clusters to human complexes for (a) low DENS, (b) medium DENS , and (c) high DENS.

5.3.2 Quality of novel complexes

Figure 5.5 shows the number of novel complexes predicted in yeast by different algorithms. Compared to the other approaches, CMC generates fewer novel complexes. Once we combined the two CMC outputs, we get 293 novel complexes. Therefore, to evaluate the quality of novel predicted complexes we consider only the top 300 clusters generated by each algorithm and sorted by their density score.

The CYC2008 reference complexes are expected to demonstrate the highest values of the three measures of semantic coherence. From Figure 5.6, it is seen that predicted complexes by *CMC_DH_predicted* after network decomposition have greater BP, CC and MF coherence than other standard clustering algorithms. However, we observe that after we applied the combined strategy, *CMC+CMC_DH_predicted* has the highest MF coherence and competitive BP and CC coherence. This suggests that the *CMC+CMC_DH_predicted* generates a larger number of novel yeast complexes, but with similar semantic coherence compared to the conventional CMC. Moreover, the top 300 novel clusters generated by ClusterOne, COACH, and IPCA are of lower quality in terms of Gene Ontology semantic coherence compared to the *CMC+CMC_DH_predicted*. Furthermore, if we do not restrict to the top 300 clusters, ClusterOne, COACH, and IPCA have much lower coherence. Finally, we observe that the rest of the novel clusters (beyond top 300) generated by these methods have the lowest semantic coherence.

5.4 Conclusion

Identification of protein complexes is necessary to understand cellular organization and machinery. Many computational methods have been proposed

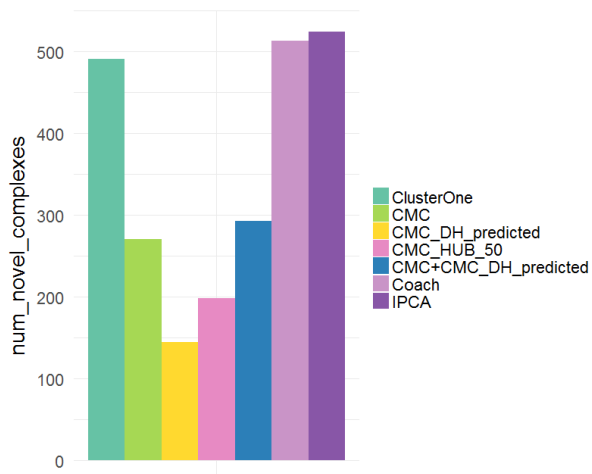
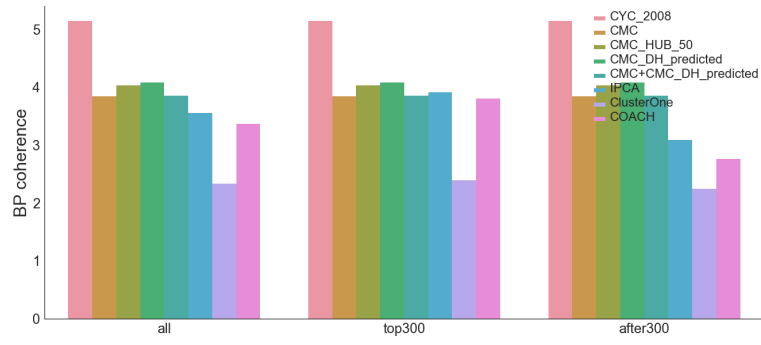


Figure 5.5: Number of novel predicted yeast complexes.

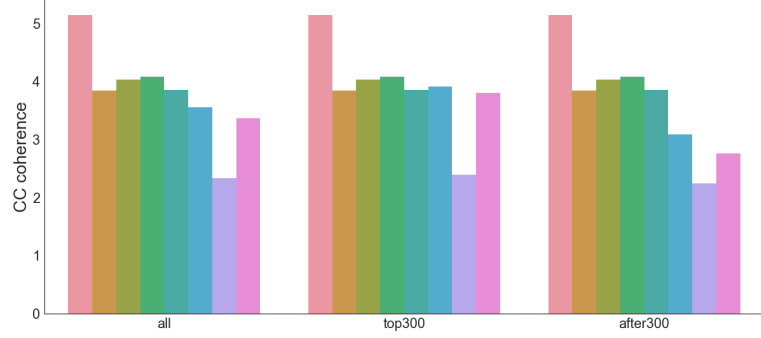
for predicting protein complexes of yeast and human. This thesis investigated the impact of PPI network decomposition by removing known and predicted date hub proteins on the performance of protein complex prediction.

Specifically, in Chapter 3, we evaluated the reliability of a reference set of date hub proteins through their participation in overlapping reference complexes. Then we inspected the potential benefits of removing real date hubs from PPI network before clustering, and demonstrated the performance advantages through the comparison to other methods.

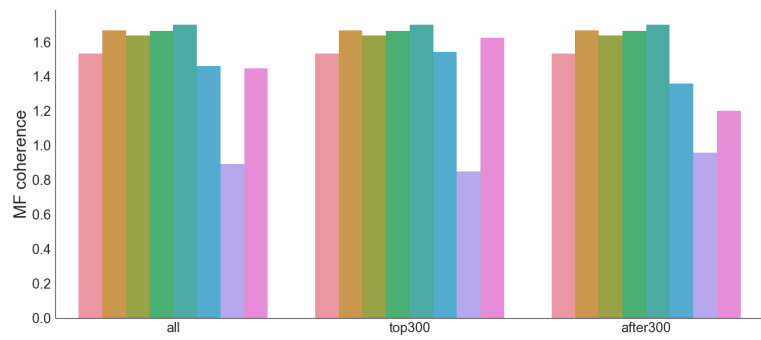
Chapter 4 was undertaken to design a reliable list of date hub proteins, and examined its effects on the performance of CMC after applying the date hub removal strategy. Following previous research, we first analyzed the topological features of date and party hub proteins on yeast and human PPI networks. Our findings confirmed significant differences in proteins degree and transitivity for the yeast network, with party hubs having consistently higher values. Although date hubs have a higher median betweenness and lower median transitivity than party hubs, less consistent results were found for the human network.



(a)



(b)



(c)

Figure 5.6: Coherence of predicted yeast complexes.

Then we demonstrated that degree and transitivity can be used to predict a reliable list of date hub proteins. The investigation of different threshold values has shown that proteins with degree above a certain threshold and with transitivity below a certain value are likely to be date hubs. Moreover, we successfully applied the network decomposition approach and confirmed that CMC benefits much from this, with improvements obtained in both precision and F-measure for yeast complex discovery.

In Chapter 5, we proposed a double-barrel strategy to combine the clusters predicted by CMC before and after we remove date hubs. We tested this strategy on the prediction of yeast and human complexes, and demonstrated that this strategy gave a performance boost in complex discovery over using a single run of CMC, and outperformed some commonly-used clustering algorithms applied on a PPI network. Moreover, the results suggested that taking the double-barrel run tends to give bigger improvements (in terms of generating more well-matched clusters) among overlapping complexes across different datasets (yeast and human). Furthermore, we also investigated that our approach generates novel predictions with higher quality in terms of Gene Ontology semantic coherence.

In summary, our observations provide a better understanding of the deconvolution of overlapping protein complexes from PPI networks. As more specific, high-quality PPI data become available, we believe our approaches to predict date hubs can reveal a further improvement on protein complex prediction.

Bibliography

- Bader, G. D. & Hogue, C. W. (2003), ‘An automated method for finding molecular complexes in large protein interaction networks’, *BMC bioinformatics* **4**(1), 2.
- Brückner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. (2009), ‘Yeast two-hybrid, a powerful tool for systems biology’, *International journal of molecular sciences* **10**(6), 2763–2788.
- Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L. et al. (2013), ‘The biogrid interaction database: 2013 update’, *Nucleic acids research* **41**(D1), D816–D823.
- Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S. & Krogan, N. J. (2007), ‘Toward a comprehensive atlas of the physical interactome of *saccharomyces cerevisiae*’, *Molecular & Cellular Proteomics* **6**(3), 439–450.
- Csardi, G. & Nepusz, T. (2006), ‘The igraph software package for complex network research’, *InterJournal, Complex Systems* **1695**(5), 1–9.
- Fields, S. & Song, O.-k. (1989), ‘A novel genetic system to detect protein protein interactions’, *Nature* **340**(6230), 245–246.
- Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. et al. (2004),

- ‘Evidence for dynamically organized modularity in the yeast protein–protein interaction network’, *Nature* **430**(6995), 88–93.
- King, A. D., Pržulj, N. & Jurisica, I. (2004), ‘Protein complex prediction via cost-based clustering’, *Bioinformatics* **20**(17), 3013–3020.
- Li, M., Chen, J.-e., Wang, J.-x., Hu, B. & Chen, G. (2008), ‘Modifying the dpclus algorithm for identifying protein complexes based on new topological structures’, *BMC bioinformatics* **9**(1), 398.
- Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardoza, A. P., Santonico, E. et al. (2012), ‘Mint, the molecular interaction database: 2012 update’, *Nucleic acids research* **40**(D1), D857–D861.
- Liu, G., Wong, L. & Chua, H. N. (2009), ‘Complex discovery from weighted ppi networks’, *Bioinformatics* **25**(15), 1891–1897.
- Liu, G., Yong, C. H., Chua, H. N. & Wong, L. (2011), ‘Decomposing ppi networks for complex discovery’, *Proteome science* **9**(1), S15.
- Nepusz, T., Yu, H. & Paccanaro, A. (2012), ‘Detecting overlapping protein complexes in protein-protein interaction networks’, *Nature methods* **9**(5), 471–472.
- Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N. et al. (2013), ‘The mintact projectintact as a common curation platform for 11 molecular interaction databases’, *Nucleic acids research* p. gkt1115.
- Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005), ‘Uncovering the overlapping community structure of complex networks in nature and society’, *Nature* **435**(7043), 814–818.

- Pritykin, Y. & Singh, M. (2013), ‘Simple topological features reflect dynamics and modularity in protein interaction networks’, *PLoS Comput Biol* **9**(10), e1003243.
- Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. (2009), ‘Up-to-date catalogues of yeast protein complexes’, *Nucleic acids research* **37**(3), 825–831.
- Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999), ‘A generic protein purification method for protein complex characterization and proteome exploration’, *Nature biotechnology* **17**(10), 1030–1032.
- Ruepp, A., Waagele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. & Mewes, H.-W. (2009), ‘Corum: the comprehensive resource of mammalian protein complexes2009’, *Nucleic acids research* **38**(suppl_1), D497–D501.
- Srihari, S., Yong, C. H., Patil, A. & Wong, L. (2015), ‘Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes’, *FEBS letters* **589**(19), 2590–2602.
- Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I. et al. (2009), ‘An empirical framework for binary interactome mapping’, *Nature methods* **6**(1), 83–90.
- Wu, M., Li, X., Kwok, C.-K. & Ng, S.-K. (2009), ‘A core-attachment based method to detect protein complexes in ppi networks’, *BMC bioinformatics* **10**(1), 169.
- Yong, C. H. & Wong, L. (2015*a*), ‘From the static interactome to dy-

namic protein complexes: Three challenges', *Journal of bioinformatics and computational biology* **13**(02), 1571001.

Yong, C. H. & Wong, L. (2015*b*), 'Prediction of problematic complexes from ppi networks: sparse, embedded, and small complexes', *Biology direct* **10**(1), 40.

Appendix A

Lists of Date Hubs

Proteins (H. sapiens)

P38398	P04637	Q13616	P12004	P00533	Q15428	P54253	Q99873	P07948	Q13573
Q09472	P06400	P62993	P45983	P63165	Q03135	Q9UBN7	P53350	P08238	P63244
P35222	Q13547	P40337	P84022	P10415	Q96J02	Q99683	O00716	Q5S007	Q92731
Q00987	P01106	Q9UNE7	P03372	P46937	O14744	Q16539	P13569	Q99459	P62136
Q9Y4K3	P07900	P49841	Q9Y297	P11940	Q9UKV8	P42858	P28482	P61981	P05067
P63000	P46934	O75381	P55072	Q13263	P31946	P0CG48	O60260	P62837	Q96EB6
P12931	Q13501	Q9H492	P31749	P60953	P61956	P62158	P51668	P29350	P27348
Q15843	Q13618	P06241	O15379	P63104	Q9UL18	O43678	O43463		

Table A.1: List of predicted date hubs for human: $degree \geq 30$ and $transitivity \leq 0.05$.

Proteins (S. cerevisiae)

YIL035C YOR061W YDR394W YGL190C YDR448W YGR252W YBR079C
YPL031C YOR039W YPL129W YLR293C YER148W YGR274C
YOR326W YMR047C YOR341W YPR110C YBR198C YBR081C
YPL203W YDL145C YGL137W YML010W YPL082C YOR181W
YDL042C YDR388W YDR477W YGL112C YDR328C YLR442C
YER171W YOR151C YJR022W YDL188C YPL153C YDL132W
YBR109C YOL012C YJL187C YCR088W YOL139C YGR262C
YML064C YER125W YBR009C YBR010W YPL240C YMR309C
YLL036C YER165W YFR028C YJL076W YCR086W YGR162W
YLL039C YGL019W YNL031C YDR224C YPR041W YDL140C
YDR225W YGL237C YPR010C YNL161W YNL189W YER133W
YBL007C YIL106W YDL160C YFL039C YOL004W YGR040W
YFR004W YDR190C YER095W YNL330CYDL029W YLR347C
YDR386W YGR240C YMR186W YLR113W YPL235W YBR114W
YBL023C YOR080W YDR381W YDR510W YBR160W YMR223W
YER110C YDL155W YER177W YMR139W YJR066W YJL098W
YKL196C YAL005C YDL047W YDL126C YLR026C YBL016W
YAR007C YGL049C YPL169C YDL185W YPL106C YJL081C
YGL206C YDL043C YDL059C YLR423C YMR109W YJL041W
YHR030C YPL140C YPL204W YIL061C YKR048C YGL207W
YML069W YDR172W YHR064C YJL115W YER151C YFR024C-A
YNL030W YJL095W YNL209W YDR247W YGR218W YBL002W
YLR096W YMR304W YGR052W YDL101C YDL028C YAL035W
YGL173C YLL024C YDR212W YJL014W YPR115W YKR001C
YMR116C YIL142W YDR170C YGR220C YMR059W YOR267C
YLR180W YKL104C YNL307C YLR249W YOR204W YBR084W
YMR012W YLR427W YCL037C YOL086C YLR342W YDR188W
YIL131C YMR308C YCL011C YGL195W YOR304W YMR106C
YOL054W YKL081W YBL003C YDR432W YDR356W YMR001C
YHR099W YDL229W

Table A.2: List of predicted date hubs for yeast $degree \geq 23$ and $transitivity \leq 0.32$