Mining protein complexes from protein interaction data

Wu, Min

2011

Wu, M. (2011). Mining protein complexes from protein interaction data. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/53200>

<https://doi.org/10.32657/10356/53200>

Mining Protein Complexes From Protein Interaction Data

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

by

WU MIN

SCHOOL OF COMPUTER ENGINEERING
NANYANG TECHNOLOGICAL UNIVERSITY

JANUARY 2011

Abstract

Protein complexes play key roles in various important cellular processes, such as transcription, translation and so on. However, experimental detection of protein complexes is very time-consuming and tedious. That is why currently we still have only a limited number of known protein complexes. Therefore, computational methods are highly desired to generate high-quality candidates of protein complexes for wetlab validation. Meanwhile, protein complexes are analyzed to have substructures, *i.e.*, a protein complex has several proteins to form a core and other proteins in periphery as attachments. Although there are already many computational methods for detecting protein complexes from PPI data, most of them fail to reveal this inherent organization of protein complexes.

This thesis introduces two novel algorithms for detecting protein complexes with core-attachment structures from protein interaction data. Both algorithms detect protein-complex cores first and then select attachments for cores to form protein complexes. The first algorithm, COACH, detects protein complexes from traditional PPI networks, while the second algorithm, CACHET, detects complexes from bipartite graphs which model protein interactions detected by Tandem Affinity Purification (TAP). This thesis also presents an approach to identify reliable interactions by integrating diverse biological and computational resources. We can control the quality of PPI networks by adding reliable false negatives and/or deleting false positives. Computational methods for protein complexes (not limited to COACH) can thus benefit from this study since more reliable and complete interactome are crucial for accurately detection of protein complexes.

Acknowledgments

My deepest gratitude goes to my supervisors, Professor Kwoh Chee-Keong and Professor Li Xiaoli, for their immense patience and invaluable advice they have provided me during this important part of my life. The work that I have done here would not have been possible without them.

I have furthermore to thank Professor Wong Limsoon, Dr. Ng See-Kiong, Professor Xie Zhipeng and Dr. Chua Hon-Nian for their effective comments, which have had a remarkable influence on my work during my Ph.D study. I warmly thank my colleagues and friends, Deng Dingxiong, Jiang Xing, Zhang Juyong, Zhuang Jinfeng, Zhao Peilin, Zhu Linhong and Liu Yong, for their valuable advice and friendly help. Thanks also go to my former supervisor, Dr. Jiang Daxin, for his kind help.

Especially, I would like to give my special thanks to my wife and parents. Their encouragement and understanding are my motive to finish my Ph.D study.

Contents

Abstract	i
Acknowledgments	ii
List of Figures	vii
List of Tables	x
List of Algorithms	xii
List of Notations, Abbreviations and Explanations	xiii
1 Introduction	1
1.1 Background	1
1.2 Challenges	2
1.3 Aims and Contributions	3
1.4 Scope	4
1.5 Organization	5
2 Literature Review	7
2.1 Protein-Protein Interactions	7
2.1.1 Detection of PPIs	7
2.1.2 Graph Models for Protein Interaction Data	9
2.2 Protein Complexes	9
2.2.1 Experimental Detection of Protein Complexes	11
2.3 Computational Methods for Protein Complexes	12
2.3.1 Terminologies	13
2.3.2 Topological Properties of Protein Complexes	14
2.3.3 Graph Clustering for Protein Complex Mining	15
2.3.4 Recent Emerging Techniques	23

- 2.3.5 Summary 25
 - 2.4 Evaluation of Protein Complexes 28
 - 2.4.1 Precision, Recall and F-measure 28
 - 2.4.2 Sensitivity, Positive Predictive Value and Accuracy 28
 - 2.4.3 P-values 29
- 3 COACH: Detecting Protein Complexes from PPI Networks 31
 - 3.1 Introduction 31
 - 3.1.1 Background 31
 - 3.1.2 Motivation 31
 - 3.2 Method 32
 - 3.2.1 Preliminaries 33
 - 3.2.2 Defining Preliminary Cores 33
 - 3.2.3 Mining Protein-Complex Cores 34
 - 3.2.4 Assembling Protein Complexes 37
 - 3.3 Experimental Results 38
 - 3.3.1 Experimental Data and Evaluation Metrics 38
 - 3.3.2 Comparative Evaluation 40
 - 3.3.3 Validation of Core-Attachment Structures 46
 - 3.3.4 Effect of the Parameter t 49
 - 3.3.5 Results Using Krogan *et al.*'s PPI data 51
 - 3.3.6 Performance Evaluation by the CYC2008 Benchmark 51
 - 3.3.7 Comparison with the CoreMethod 52
 - 3.3.8 Running Time Analysis 53
 - 3.4 Conclusions 54
- 4 Integrating Diverse Sources For Reliable Protein-Protein Interactions 56
 - 4.1 Introduction 56
 - 4.1.1 Background 56
 - 4.1.2 Motivation 58
 - 4.2 Method 58
 - 4.2.1 Individual Biological and Computational Evidences 58

- 4.2.2 InterBIN: Grouping Interactions by Their Original Scores 60
 - 4.2.3 InterCLASS: Integrating Evidences by Machine Learning Techniques 62
 - 4.3 Experimental Data and Evaluation Metrics 64
 - 4.3.1 Experimental Data 64
 - 4.3.2 Receiver Operating Characteristics 64
 - 4.3.3 Validation of False Positives/Negatives 65
 - 4.4 Experimental Results 65
 - 4.4.1 Comparison between Bin-Confidence Scores and Original Scores . 65
 - 4.4.2 Effect of the Parameter μ 67
 - 4.4.3 False Positive/Negative Detection 68
 - 4.4.4 Comparison with Patil’s Integration Method 71
 - 4.5 Expanding and Purifying Interactions for Protein Complexes 72
 - 4.5.1 From COACH to COACH* 73
 - 4.5.2 Expanding PPI Networks by Adding False Negatives 73
 - 4.5.3 Purifying PPI Networks by Deleting False Positives 74
 - 4.5.4 Concurrently Adding and Deleting Edges 75
 - 4.5.5 Summary 76
 - 4.6 Conclusions 76
- 5 CACHET: Detecting Protein Complexes Directly from TAP Data 78**
 - 5.1 Introduction 78
 - 5.1.1 Background 78
 - 5.1.2 Motivation 79
 - 5.2 Methods 80
 - 5.2.1 Preliminaries 80
 - 5.2.2 Mining Reliable Bicliques 81
 - 5.2.3 Mining Maximal Independent Set of Bicliques 82
 - 5.2.4 Detecting Protein-Complex Attachments 85
 - 5.3 Experimental Results 85
 - 5.3.1 Evaluation Metrics and Experimental Data 87
 - 5.3.2 Converting TAP Data to PPI Networks 88

5.3.3	Comparative Evaluation of Predicted Protein Complexes	89
5.3.4	Evaluation of Predicted Protein-Complex Cores	93
5.3.5	Effect of The Parameter r	94
5.3.6	Impact of Indexing Reliability and Filtering Redundancy	96
5.3.7	Running CACHET on Krogan <i>et al.</i> 's Data	96
5.3.8	Running CACHET on the Combined Data	97
5.3.9	Protein-Level Evaluations	99
5.4	Conclusions	101
6	Conclusions, Future Work and Discussions	102
6.1	Review of Main Contributions	102
6.2	Future Work	104
6.2.1	Neighborhood Counting	105
6.2.2	Random Walk with Restart (RWR)	106
6.2.3	CIPHER	106
6.2.4	Random Walk with Restart on Heterogeneous Networks (RWRH)	107
6.2.5	PRINCE	107
6.2.6	AlignPI	108
6.2.7	RegMOD	108
6.3	Discussions	109
6.3.1	Summary	109
6.3.2	Possible Solutions	110
	References	114
	Appendix	133
	List of Publications	135

List of Figures

2.1	The PPI network with 11000 yeast interactions among 2401 proteins from [Przu 04].	10
2.2	A sample TAP bipartite graph. $\{v_1, v_2, v_3\}$ is the set of baits, $\{v_1, v_4, v_5, v_6, v_7\}$ is the set of preys and the edges connecting baits and preys are interactions detected by TAP experiments.	10
2.3	RNA polymerase II complex in [Hort 06].	11
2.4	Nuclear pore complex [Fahr 03].	12
2.5	The procedure of Tandem Affinity Purification in [Hube 03].	13
2.6	The sample PPI graph consists of 9 nodes (proteins), which are labeled from 1 to 9.	17
2.7	CODEC on a sample TAP bipartite graph.	23
3.1	DNA repair complex, which repairs DNA damage by interacting with damaged DNA.	32
3.2	The Diagram of our Core-removal algorithm.	37
3.3	The procedure of our COACH algorithm.	39
3.4	The performance comparison for various algorithms on DIP data, using F-measure and Coverage Rate.	41
3.5	(A) The real SAGA complex in the benchmark; (B-E) The SAGA complex predicted by different methods. For each predicted complex, the proteins in red color are involved in the real SAGA complex and those in blue color are not.	42
3.6	Co-annotation, co-localization and co-expression scores of predicted complexes from different algorithms.	43
3.7	Examples of protein complexes predicted by COACH method.	46

3.8 Predicted complexes with interactions scored by gene-expression correlations. 47

3.9 Predicted complexes with interactions scored by functional similarities (using BP). 48

3.10 Predicted complexes with interactions scored by functional similarities (using CC). 49

3.11 The effect of t on F-measure and Coverage rate. 50

3.12 F-measure and Coverage Rate of various methods for Krogan’s data. . . 51

3.13 The average number of edges in the random graphs in our experiments. . 54

3.14 The running time of our COACH on the random graphs. 55

4.1 AUC for each method as the parameter varies on DIP data. 67

4.2 AUC for each method as the parameter varies on BioGrid data. 68

4.3 The average functional similarity of top-ranked interactions in DIP data. 69

4.4 The average functional similarity of top-ranked interactions in BioGrid data. 70

4.5 The average functional similarity of top-ranked false negative candidates generated from DIP data. 72

4.6 The performance of COACH* on the PPI network expanded by adding false negatives. 74

4.7 The performance of COACH* on the PPI network purified by deleting false positives. 75

4.8 The performance of COACH* on the PPI network modified by concurrently adding and deleting edges. 76

5.1 A example TAP bipartite graph. $\{v_1, v_2, v_3\}$ is the set of bait proteins and $\{v_1, v_4, v_5, v_6, v_7\}$ is the set of prey proteins. Please note that certain proteins (*e.g.*, v_1) can serve as baits and they can also be tagged as preys by other baits. 81

5.2 The diagram of our CACHET method working on a sample TAP graph. . 86

5.3 Comparison among various methods on Gavin *et al.*’s data using different reliability schemes. 91

5.4 F-measure of various methods when ω is varied from 0.2 to 0.5. 91

5.5 The proportion of significant complexes predicted by various methods. . . 92

5.6 The organization of our identified protein complexes. Each complex consists of four rows as shown in this figure. The two rows in the middle (the second and third rows with rectangles as proteins) form the protein-complex core and other two rows (with circles as proteins) are attachments. The second and third rows are baits and preys in the cores, respectively. Correspondingly, the first and fourth rows represent bait-attachments and prey-attachments, respectively. The predicted complexes are histone deacetylase complex in (A) and DNA-directed RNA polymerase II complex in (B). 95

5.7 Performance of our CACHET on Gavin *et al.*'s data under different values of r 95

5.8 Comparison among various methods on the combined data using different reliability schemes. 98

5.9 Performance of our CACHET on Gavin *et al.*'s data under different values of ω 100

5.10 Performance of our CACHET on Gavin *et al.*'s data under different values of ω 100

List of Tables

2.1	Details of various PPI databases.	14
2.2	Properties of CYC2008 complexes on each PPI dataset.	15
2.3	Key developments for computational detection of protein complexes. . . .	26
3.1	The results of various algorithms using DIP data.	41
3.2	The proportions of significant complexes predicted by different methods on DIP data.	44
3.3	Selected protein complexes predicted by COACH method on DIP data. .	45
3.4	Average similarity of interactions involved in protein-complex cores, pro- tein complexes and DIP data respectively.	47
3.5	Performance of various methods evaluated by the CYC2008 benchmark. .	52
3.6	Comparative analysis between CoreMethod and COACH.	53
4.1	AUC of each method using original raw scores and Bin-Confidences, re- spectively.	66
4.2	The comparison of AUC between Patil's method and ours.	73
5.1	Details of TAP datasets.	88
5.2	Contingency table from [Zhan 08] to evaluate the matching between pre- dicted interactions and known interactions.	89
5.3	PPI networks converted from TAP data	89
5.4	Performance for various <i>MWIS</i> heuristics using SA scores on Gavin <i>et al.</i> 's data ($\omega = 0.2$)	90
5.5	Performance for various methods on Gavin <i>et al.</i> 's data ($\omega = 0.2$ using SA scores)	92

5.6 Average functional similarity within protein-complex cores and protein complexes. 93

5.7 Inner-similarity of cores predicted by various methods on Gavin *et al.*'s data 94

5.8 Performance for CACHET on Gavin *et al.*'s data ($\omega = 0.2$) 96

5.9 Performance for various methods on Krogan-I network 97

5.10 Performance for various methods on Krogan-II network 97

5.11 CACHET on combined data using SA scores ($\omega = 0.2$). 99

5.12 Performance of various methods using SA scores on combined data ($\omega = 0.2$). 99

5.13 Protein-level evaluation for various methods 101

5.14 Protein level evaluation for various methods using SA scores on combined data. 101

List of Algorithms

1	Protein-complex Core Mining Algorithm	35
2	Core-removal(cg)// $cg = (V_{cg}, E_{cg})$	36
3	Redundancy-filtering(C, t, SC)	38
4	Mining Reliable Bicliques	83
5	Maximum Weighted Independent Set	84

List of Notations, Abbreviations and Explanations

Notations	Explanations
G	a general graph representing a pairwise PPI dataset
G_{TAP}	a biparite graph representing a TAP dataset
G'	a subgraph of either a general graph or a bipartite graph
G_v	the local neighborhood graph of vertex v
$density(G')$	the density of a subgraph G'
N_v	the set of neighbors of vertex v
$deg_{G'}(v)$	vertex v 's degree in G'
$Average(G')$	the average degree of a subgraph G'
$NA(G_1, G_2)$	neighborhood affinity score between G_1 and G_2
$A(G')$	available node set of a subgraph G'
$w(v)$	the weight or importance of a vertex v
$W(G')$	the weight or importance of a subgraph G'
GO	Gene Ontology
BP	'Biological Process' sub-ontology of GO
CC	'Cellular Component' sub-ontology of GO
MF	'Molecular Function' sub-ontology of GO
GE	Gene Expression

Chapter 1

Introduction

1.1 Background

Protein complexes are formed by multiple proteins with interactions among them. They are very important molecular entities within the cells because they are involved in many important cellular processes, such as transcription, translation, reproduction and so on. However, experimental detection of these biologically important protein complexes is very time-consuming and tedious. Meanwhile, the number of known protein complexes is currently quite limited. Fortunately, it is possible for us to detect protein complexes from the increasing amount of protein-protein interaction (PPI) data generated by high-throughput techniques.

Protein interaction data generally can be modeled as PPI graphs or networks in which nodes are proteins and edges are interactions among proteins. It is observed that densely connected subgraphs tend to be protein complexes [Spir 03, Tong 02]. Thus many computational methods have been proposed for protein complexes based on this observation and please refer to section 2.3 for a brief survey of some of these methods. In particular, protein interaction data from Tandem Affinity Purification (TAP) [Puig 01] or pull-down or even co-immunoprecipitation experiments (please refer to section 2.2 for details of these techniques) can be utilized in two different manners. On one hand, the relationships between baits and preys that are detected by above experiments can be converted to traditional PPI networks and protein complexes can then be detected from those PPI networks. On the other hand, bait-prey relationships can be modeled as the edges in a bipartite graph where two node sets are baits and preys, respectively. Similarly, dense bipartite subgraphs can be detected as protein complexes [Geva 11].

In addition, protein complexes are analyzed to have substructures, *i.e.*, a protein complex has several proteins to form a core and other proteins in periphery as attachments [Dezs 03, Gavi 06]. Proteins within cores tend to have higher functional similarities than attachments. Therefore, cores are “hearts” of protein complexes while attachments

are more like “secondary structures”. Although there are already many computational methods for detecting protein complexes from PPI data, most of them fail to approach, model and reveal this inherent organization of protein complexes.

1.2 Challenges

Currently, numerous computational methods have been proposed to detect protein complexes from PPI data. However, there are still several challenging issues to be addressed.

First, one challenge is originated from the definition and identification of core-attachment substructures. The study by Dezso *et al.* [Dezs 03] first mentioned that protein complexes like social communities, have some core members and some attachments based on the analysis of known protein complexes. Gavin *et al.* [Gavi 06] first identified core-attachment structures when predicting protein complexes. As we know, protein isoforms are different forms of the same protein. Here, many clusters are generated by the iterative clustering and “complex isoforms” are different clusters that refer to the same real protein complex. As such, core proteins of a complex are defined as those presented in most of the isoforms of this complex. In Zhang *et al.*’s study [Zhan 08], they detected protein complexes by merging overlapping maximal cliques. Given a complex, they first listed all the cliques whose merging resulted in this complex and then similarly defined those proteins that appeared in over 2/3 of all these cliques as core proteins. These two definitions of core proteins are based on redundant data sources and sound statistically reasonable because important proteins will appear more frequently in those isoforms or cliques [Navl 10]. However, a protein-complex core formed by individual significant proteins may not be significant as a whole in above methods. Meanwhile, local neighborhood search is the most popular strategy for protein complex detection [Chua 07a]. Then, how can we define and identify core and attachments during the local neighborhood search as well as taking the statistical significance of core proteins into consideration?

Second, current PPI data are still incomplete and of high noise, which may affect the performance of existing protein complex mining approaches. Each experimental method for detecting PPI has its own limitations. For example, yeast-two-hybrid suffers from high false positive rate. The reason for this high error rate lies in the principle of the technique: it investigates the interaction between (a) over-expressed (b) fusion proteins in the (c) yeast (d) nucleus. Each of these points (a-d) alone can give rise to false results [Zhan 10]. For example, over-expression can result in non-specific interactions (a). Moreover, a mammalian protein is sometimes not correctly modified in yeast (*e.g.*, missing phosphorylation), which can also lead to false results (b). Finally, some proteins might specifically interact when they are co-expressed in the yeast and nucleus, although in reality they are never present in the same cell at the same time (c, d). Currently,

CHAPTER 1. INTRODUCTION

the exact rate of false positive results is not known, but estimates are as high as 50% [Meri 02]. How can we address the reliability of interactions when detecting protein complexes?

Third, detecting protein complexes from TAP data is challenging due to the following reasons. Existing methods detecting protein complexes from TAP data usually consist of two steps. The first step is to convert the co-complex relationships in TAP data into binary interactions to form a PPI network. The second step is to detect densely connected subgraphs in the PPI network as protein complexes. In the first step, reference protein complex data are usually exploited for generating benchmark protein interactions and then for converting TAP data to a PPI network. In particular, this benchmark interaction set is generated by considering that all the connections within the same complexes are known positive interactions. As we know, it is not reasonable to assume all protein pairs within the same complexes are true interactions because two proteins in a large complex are far away from each other in 3-D spaces and thus may have no chance to bind. In addition, these methods are supervised and thus biased since current known protein complexes are still incomplete. Meanwhile, reference protein complexes will also be used to evaluate the predicted complexes and it is not unbiased because it is not a blind testing. Therefore, converting TAP data to accurate PPI data is still inconvenient and challenging for mining protein complexes. On the other hand, can we directly detect protein complexes from TAP data and avoid the step of converting the TAP data to address this issue?

1.3 Aims and Contributions

The purpose of this thesis is to investigate the characteristics of protein complexes, especially their topological properties in networks or graphs (traditional graphs or bipartite graphs) that represent protein interaction data, and to develop algorithms mining protein complexes from these graphs. Our detected complexes should have core-attachment structures, which would provide insights into their inherent functional organizations. We have addressed the above challenging issues associated with detecting protein complexes from protein interaction data and achieved the following contributions in this thesis.

- **Identifying protein complexes with core-attachment structures**

We detect protein complexes with core-attachment structures in an inside-out manners. Protein-complex cores are first located and then attachments are selected based on their closenesses/connectivities with respect to those given cores. Protein complexes are finally formed by combining cores and attachments. In

CHAPTER 1. INTRODUCTION

traditional PPI networks, we consider local dense neighborhood graphs as candidates of protein-complex cores and collect the final set of cores by selecting significant/promising candidates as well as filtering the redundancy among these candidates. In TAP bipartite graphs, we first compute all the maximal bicliques. All the edges in the bipartite graphs are indexed with reliability scores and the bicliques whose average edge reliability scores are no less than a pre-defined threshold are defined as reliable bicliques. We consider those non-redundant and reliable bicliques as protein-complexes cores and then form protein complexes similarly. This inside-out strategy guarantees that the proteins in our collected cores have higher functional similarity and are more likely to be co-localized than those attachments. Meanwhile, many of structured protein complexes detected by our methods are also validated by the original research in [Gavi 06].

- **Expanding and purifying protein interaction data for protein complexes**

The quality of predicted protein complexes is highly related to the reliability of protein-protein interactions in PPI networks. We present a method to identify reliable false negatives and filter false positives by integrating diverse biological and computational sources. We thus increase the coverage and overall confidence of PPI data by adding reliable novel interactions and removing un-reliable interactions. It is demonstrated that our methods for protein complexes gain higher accuracy from those PPI data with higher quality.

- **Detecting protein complexes directly from TAP data**

We model TAP data as a bipartite graph, where two node sets are baits and preys and edges represent bait-prey relationships. In traditional PPI networks, we detect protein complexes based on the observation that densely connected subgraphs are likely to be protein complexes. Here in TAP bipartite graphs, we similarly consider densely connected bipartite subgraphs as protein complexes. Then, we devise a biclique-based method to form protein complexes with core-attachment structures. Therefore, our method detects protein complexes directly from TAP data and avoids the challenging step that converts TAP datasets into traditional PPI networks.

1.4 Scope

Let us make a principled distinction between two biological concepts, namely, protein complexes and functional modules [Spir 03]. A protein complex is a physical aggregation of several proteins (and possibly other molecules) via molecular interaction (binding) with each other at the same location and time. A functional module also consists of a number of proteins (and other molecules) that interact with each other to control or

perform a particular cellular function. However, unlike protein complexes, these proteins do not necessarily interact at the same time and location.

In this thesis, we do not distinguish protein complexes from functional modules because the underlying protein interaction data that are utilized for protein complex detection do not provide temporal and spatial information. To distinguish between protein complexes and functional modules, it will be necessary to integrate additional biological resources that contain such information (*e.g.*, gene expression data)[Jin 09].

In addition, we only performed experiments over yeast PPI data and collected yeast complexes. The reason is that yeast as the model organism, has the most comprehensive PPI data for our experiments and also provides quite a number of benchmark protein complexes for our validation. In our future work, we plan to identify some human protein complexes from human PPI data, especially human disease-related complexes by considering gene-phenotype associations.

1.5 Organization

The rest of this thesis is organized as follows. In Chapter 2, protein-protein interactions, protein complexes and their experimental detection methods will first be reviewed in details. We conduct a survey on computational detection of protein complexes. Lastly in this chapter, we introduce evaluation measures for predicted complexes.

Chapter 3 presents a core-attachment based method called COACH to detect protein complexes in traditional PPI networks. COACH first detects dense neighborhood graphs as protein-complex cores and then selects attachments to form protein complexes.

In Chapter 4, we propose a strategy to identify false positives and predict reliable false negatives by integrating diverse biological and computational sources. The quality of PPI data will thus be enhanced after removing some false positive interactions or/and adding some reliable false negatives. We also show that the performance of our protein complex detection method will be improved after removing false positive interactions and adding reliable false negatives.

As we know, TAP experiments are designed to detect protein-protein interactions and protein complexes, although their generated results suffer from high false positives and false negatives. Therefore, accurate assessment and utilization of these co-complex relationships in TAP data may lead to more accurate identification of protein complexes. Chapter 5 presents an algorithm called CACHET to detect protein complexes directly from TAP data. Firstly, the TAP data are modeled as a bipartite graph and all maximal bicliques are computed. Secondly, all the bait-prey relationships are indexed with reliability scores and then only reliable bicliques are kept. Thirdly, we select non-redundant and reliable bicliques as protein-complex cores and then assemble attachments to form protein complexes.

CHAPTER 1. INTRODUCTION

Finally in Chapter 6, an important application of protein complex mining will be introduced — detecting human disease-related protein complexes. We will first introduce some background and related work in this domain. We will also provide some possible solutions for detecting disease-related protein complexes.

Chapter 2

Literature Review

When we talk about protein complexes, we can't ignore protein-protein interactions (PPIs) because protein complexes are formed by proteins via interactions among them. In this chapter, we will first briefly introduce PPIs, protein complexes and their experimental detection methods. After that, we will focus on the computational methods that detect protein complexes from protein interaction data, as well as the evaluation measures for these computational methods.

2.1 Protein-Protein Interactions

Proteins are large organic compounds made of amino acids arranged in a linear chain. Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of organisms and they participate in many processes within cells. Proteins, as the chief actors within the cell, carry out the duties specified by the information encoded in genes by interacting with other proteins or genes. For example, DNA-protein interactions are the key to understanding the initiation of transcriptions. More importantly, protein-protein interactions (PPI) have attracted the dramatic attentions of researchers in proteomics since they are important for understanding many biological functions. Understanding PPI is also important for us to improve our understanding of diseases and their corresponding therapies. For example, the interaction between proteins p53 and MDM2 can cause cancer and inhabiting this interaction will be attractive anticancer agents [Davi 05].

2.1.1 Detection of PPIs

The developments in high-throughput methods, such as *yeast-two-hybrid* and *protein chips*, have enabled biologists to experimentally detect PPIs for many organisms (*e.g.*, [Ito 01], [McCr 00], [Rain 01] and [Uetz 00]). Tandem affinity purification [Puig 01] can

CHAPTER 2. LITERATURE REVIEW

be also used to detect protein-protein interactions [Gavi 06, Krog 06]. There are also many computational methods for protein interaction prediction, such as gene-fusion [Marc 99] and phylogenetic profiles [Pell 99]. These predictions can provide the biologists some PPI candidates for experiments. Here, we briefly introduce three high-throughput methods, namely, *yeast-two-hybrid*, Phage Display and Affinity purification and Mass Spectrometry. A detailed review for detecting protein-protein interactions can be found in [Ng 03].

2.1.1.1 Yeast Two Hybrid

Yeast Two Hybrid (Y2H) is a molecular biology technique for detecting physical protein-protein interactions [Ito 01, Rain 01, Uetz 00].

To check whether two proteins interact, Y2H exploits the service of a third protein—a transcriptional activator that has the ability to switch on the reporter genes. The activator protein is separated into two functional fragments, a DNA-binding domain and a trans-activation domain. Two domains are then attached separately to each of the candidate interacting proteins. If the reporter genes are transcribed (switched on), these two proteins are considered to be interacting with each other because two domains of the activator are reunited when they interact. One inherent limitation of Y2H is that some proteins may have transcriptional properties and cause the reporter genes to be switched on by themselves. In this case, artifactual interactions are detected even though two proteins do not interact. In addition, many proteins need post-translation modification, but this is not accommodated by Y2H.

2.1.1.2 Phage Display

Phage display technology (PD), first reported by G. Smith in 1985, can also be applied for detecting protein-protein interactions.

A phage generally consists of a genetic material (either DNA or RNA) encapsulated by a protein coat. A key feature of phages is that their DNA can mix together with other foreign DNA, that is, a hybrid DNA will be formed by combining these two kinds of DNA. As the phage's DNA is replicated, the foreign DNA will also be replicated. As such, a hybrid protein coat will be synthesized and the foreign peptide will be “displayed” on the outer surface of the phages. Let target peptides pass through the “displayed” foreign peptide and all the captured targets are considered to interact with the foreign peptide. Compared with Y2H which detects interactions between two full length proteins, PD is more suitable for detecting the binding among short protein peptides.

CHAPTER 2. LITERATURE REVIEW

2.1.1.3 Affinity Purification and Mass Spectrometry

Affinity purification methods allow macro-molecules physically associated with a tagged bait to be retrieved and identified by mass spectrometry (*AP-MS*) [Gavi 06] and well suited for studying protein-protein interactions and protein complexes.

In this process, a bait protein is first immobilized by an affinity tag on a matrix or a column wall. Then, candidate proteins are allowed to pass through the column. Proteins binding to the bait will be captured and non-interacting proteins will be eluted away. This leads to the formation of a protein complex. Please refer to section 2.2.1 for more details about the whole process of a specific affinity purification—Tandem Affinity Purification (TAP). After this, mass spectrometry is applied to identify the content of the resulting protein complexes and the interactions involved in these complexes. Compared with Y2H only for pairwise interaction detection, AP-MS can detect many biological interactions (both direct physical interactions and indirect functional associations/interactions) that are involved in the identified protein complexes.

2.1.2 Graph Models for Protein Interaction Data

With a large amount of protein-protein interactions collected, we can represent them in the following two models. One is a general graph model, called PPI graph or network, in which vertices/nodes represent proteins and edges represent interactions between proteins. Figure 2.1 shows a real example of a yeast PPI network [Przu 04]. Another is bipartite graph model showing interactions from affinity purifications, where two sets of proteins are baits and preys respectively. Figure 2.2 shows a sample bipartite graph from TAP data.

Although high-throughput methods have provided a large amount of PPI data, recent studies have revealed that *Y2H* generates PPIs with high error rates [Meri 02, Spri 03]. The presence of false positives and false negatives in PPI data greatly diminishes the potential usefulness of the data. It would be better if the reliability of interactions is taken into consideration in research studies where protein interaction data are utilized, for example, detecting protein complexes from protein interaction data.

2.2 Protein Complexes

Protein complexes are molecular aggregations formed by the simultaneous interactions among multiple proteins. Many important biological functions, such as DNA transcription, mRNA translation, signal transduction, and cell cycle etc, involve protein complexes.

For example, the RNA polymerase II complex in Figure 2.3 transcribes genetic information into messages for ribosomes to produce proteins [Cram 00]. The 3-D structure of

CHAPTER 2. LITERATURE REVIEW

Figure 2.1: The PPI network with 11000 yeast interactions among 2401 proteins from [Przu 04].

Figure 2.2: A sample TAP bipartite graph. $\{v_1, v_2, v_3\}$ is the set of baits, $\{v_1, v_4, v_5, v_6, v_7\}$ is the set of preys and the edges connecting baits and preys are interactions detected by TAP experiments.

the RNA polymerase II complex in this figure shows that 10 member proteins are united together and form a compact structure by interacting with others.

Another example is the nuclear pore complex (NPC). NPC is responsible for the protected exchange of components between the nucleus and cytoplasm and for preventing the transport of material not destined to cross the nuclear membrane (nuclear envelop) [Fahr 03]. In other words, the nuclear pore complex is the gateway to the nucleus. The NPC complex is comprised of many distinct proteins, called nucleoporins—“nups” for short. Figure 2.4 shows the NPC complex is located at the nuclear membrane and then

Figure 2.3: RNA polymerase II complex in [Hort 06].

performs its functions.

2.2.1 Experimental Detection of Protein Complexes

Experimentally, protein complexes can be detected by affinity-purification with mass spectrometry [Gavi 02, Gavi 06, Krog 06]. Here, we introduce a specific affinity purification process—Tandem Affinity Purification.

The tandem-affinity-purification (TAP) tag consists of three components: a calmodulin-binding peptide, a tobacco etch virus (TEV) protease cleavage site and Protein A as an immunoglobulin G (IgG)-binding domain, as shown in Figure 2.5.

The target protein (so called “bait” protein), attached with the TAP tag, will pull down its interacting partners (“prey” proteins) as well as some other contaminations in the first affinity step. After washing, TEV protease is introduced to elute the bound material at the TEV protease cleavage site. This eluate is then incubated with calmodulin-coated beads in the presence of calcium. The second affinity step is required to remove the TEV protease as well as traces of contaminants remaining after the first affinity step. After washing, the eluate is then released with ethylene glycol tetraacetic acid (EGTA). The native elution can then be analyzed by mass spectrometry.

However, there are several limitations to this method [Tara 08]. First, its multiple washing and purification steps tend to eliminate transient low affinity protein complexes.

Figure 2.4: Nuclear pore complex [Fahr 03].

Second, the tag proteins used in the experiments may interfere the protein complex formation. Gavin et al. [Gavi 02] have shown that TAP-MS only captures limited known yeast protein complex subunits. Third, in TAP-MS the subcellular location of complexes is lost due to the *in vitro* purification of whole-cell lysates. This means that time-consuming preparation of subcellular fractionated lysates may be needed for a less-studied cellular process in order to employ subcellular localization information to validate the experimental results and detect false negatives or false positives. Therefore, TAP experiments provide us with additional but not conclusive sources of information. Due to these experimental limitations, computational detection of protein complexes is highly desired to complement to the experimental methods. Next, we will introduce computational methods for detecting protein complexes from protein interaction data.

2.3 Computational Methods for Protein Complexes

Let us now introduce some terminologies which are widely used in protein complex mining. Then, we will present traditional graph clustering techniques for complex mining followed by some new emerging techniques for this task.

CHAPTER 2. LITERATURE REVIEW

Figure 2.5: The procedure of Tandem Affinity Purification in [Hube 03].

2.3.1 Terminologies

A PPI network is often modeled as an undirected graph $G = (V, E)$, where V is the set of nodes (proteins) and $E = \{(u, v) | u, v \in V\}$ is the set of edges (protein interactions). A graph $G_1 = (V_1, E_1)$ is a subgraph of G if and only if $V_1 \subseteq V$ and $E_1 \subseteq E$. For a node $v \in V$, the set of v 's direct neighbors is denoted as N_v where $N_v = \{u | u \in V, (u, v) \in E\}$. v 's degree in G , $deg(v)$, is the cardinality of N_v , i.e., $|N_v|$.

Density. The density of the graph G , denoted as $density(G)$, is defined to quantify the richness of edges within G as shown in equation (2.1) [Cole 83]. Basically, $0 \leq density(G) \leq 1$. If $density(G) = 1$, then G is the fully connected graph or a clique, which has the maximum number of edges, i.e., every pair of distinct vertices is connected by an edge.

$$density(G) = \frac{2 \times |E|}{|V| \times (|V| - 1)}. \quad (2.1)$$

Clustering Coefficient. The clustering coefficient of a node v is the density of the subgraph formed by N_v and their corresponding edges, which quantifies how close v 's neighbors are to being a clique (complete graph) [Watt 98].

CHAPTER 2. LITERATURE REVIEW

Local Neighborhood. Given a node $u \in V$, its local neighborhood graph G_u is the subgraph formed by u and all its immediate neighbors with the corresponding interactions in G . It can be formally defined as $G_u = (V_u, E_u)$, where $V_u = \{u\} \cup N_u$, and $E_u = \{(v_j, v_k) | (v_j, v_k) \in E, v_j, v_k \in V_u\}$

2.3.1.1 Weighted PPI networks

If the edges in the PPI network are weighted (*e.g.*, weights represent the reliability of protein interactions [Yoon 06, Liu 09]), the definitions of vertex degree, density can be extended to their corresponding weighted versions as follows in equation (2.2):

$$\deg_w(v) = \sum_{e=(u,v) \in E} w(e) \text{ and } \text{density}_w(G) = \frac{2 \times \sum_{e \in E} w(e)}{|V| \times (|V| - 1)}, \quad (2.2)$$

where $w(e)$ is the weight of the edge e .

Similarly, the weighted clustering coefficient of the node v is the weighted density of the subgraph formed by N_v and their corresponding edges.

2.3.2 Topological Properties of Protein Complexes

Before introducing computational methods for protein complexes, we briefly show their topological properties in PPI networks.

We collected three PPI datasets, namely, DIP data [Salw 04], Krogan *et al.*'s data (Krogan data for short) [Krog 06] and BioGrid data [Star 06]¹. They have different characteristics, for example, coverage of proteins, number of interactions, the data sources (*e.g.*, the methods detecting these interactions) and so on. Table 2.1 shows the basic information of these datasets. DIP data consists of 4930 proteins and 17203 interactions among these proteins. The average degree for all proteins is 6.78, *i.e.*, each protein has 6.78 neighbors in DIP PPI network on average. Comparatively speaking, BioGrid data contains the most proteins and thus has the highest coverage of yeast proteins, while Krogan data has the least number of proteins. In addition, BioGrid data also has the highest density and DIP data has the lowest.

Table 2.1: Details of various PPI databases.

	# Proteins	# Interactions	Average degree	Density
DIP	4930	17203	6.98	1.42×10^{-3}
Krogan	3581	14077	7.86	2.20×10^{-3}
BioGrid	5163	69956	27.10	5.25×10^{-3}

¹Both DIP and BioGrid data were downloaded in August, 2009.

CHAPTER 2. LITERATURE REVIEW

Table 2.2 shows the properties of the complexes in the CYC2008 catalogue [Pu 09] on above three datasets. CYC2008 complexes downloaded from Wodak’s lab are the latest collection of yeast protein complexes. CYC2008 data have 408 protein complexes involving 1627 proteins. All these complexes have different numbers of proteins, ranging from 2 up to 81 (and the average size is 4.71). Given a PPI database, each complex is considered as a graph where two proteins have an edge if and only if there is an interaction between these two proteins in this PPI database. As such, we can obtain the average density of all these 408 complexes in the CYC2008 catalogue in each PPI database.

They achieve the highest average density on BioGrid data because BioGrid data cover the most proteins in CYC2008 catalogue (*i.e.*, 1616 out of 1627 proteins) as well as BioGrid data are the most enriched and comprehensive data with the largest number of interactions. They also have relatively high average density on DIP data and the lowest on Krogan data. Krogan data can only cover 1248 out of 1627 proteins in CYC2008 complexes, which is due to the fact that many proteins in complexes are isolated and have no interactions with other proteins. This is the main reason why CYC2008 complexes have the lowest average density on Krogan data. However, the average density of CYC2008 complexes is increased to 0.659 if we remove those isolated proteins that are not involved in Krogan data from CYC2008 complexes. That is, the parts of complexes that can be projected to Krogan PPI network are still quite dense. Existing methods detecting protein complexes are based on the assumption that dense subgraphs in PPI networks tend to be protein complexes or parts of protein complexes. Here, these number in Table 2.2 demonstrate that this assumption is reasonable because known protein complexes are indeed dense in PPI networks, compared with the overall density of these PPI networks as shown in 2.1.

Table 2.2: Properties of CYC2008 complexes on each PPI dataset.

	CYC on DIP	CYC on BioGrid	CYC on Krogan
# Proteins covered	1472	1616	1248
Average Density	0.615	0.902	0.462
Medium Density	0.667	1	0.377

2.3.3 Graph Clustering for Protein Complex Mining

We review the conventional graph clustering approaches for protein complex mining (especially, we will introduce methods in more details which are used for comparisons in Chapter 3 and Chapter 5). These methods mine for cliques or densely connected subgraphs in PPI networks which could correspond to protein complexes. While the methods mainly use the PPI networks for mining, additional information, such as gene expression data [Eise 98, Barr 09], functional information (*e.g.*, Gene Ontology data [Cons 06])

CHAPTER 2. LITERATURE REVIEW

as well as other biological information [Huh 03], may also be exploited to enhance the quality of predicted complexes.

2.3.3.1 Graph clustering based solely on PPI networks

In this section, let us describe the graph clustering approaches that use PPI networks as the sole underlying dataset for the mining task.

MCODE. The MCODE algorithm proposed by Bader *et al.* [Bade 03a] is one of the first computational methods to detect protein complexes based on the proteins' connectivity values in the PPI network. MCODE first weighs every node based on their local neighborhood densities, and then selects seed nodes with high weights as initial clusters and augments these clusters by outward traversing from the seeds. In addition, MCODE has an optional post-processing step with operations such as filtering non-dense subgraphs and generating overlapping clusters. Figure 2.6 shows an example of how MCODE detects protein complexes from a small sample graph of protein-protein interactions. The sample PPI graph consists of 9 nodes (proteins), which are labeled from 1 to 9. MCODE first assigns each node a weight using its local neighborhood density. For example, the node set $\{4, 5, 6, 7\}$ is the highest k -core² in node 5's neighborhood graph with $k=2$ [Seid 83, Bata 03] and density $d = 5/6$ (5 interactions out of a total possible 6 interactions among the nodes $\{4, 5, 6, 7\}$). Thus, node 5 has an initial weight $w(5) = k \times d = 2 \times 5/6 = 1.67$. Next, the node with the highest weight (without loss of generality, node 1 is first selected here) is selected as an initial cluster. Node 2, as node 1's neighbor, satisfies the weight constraint to be included into the cluster because $w(2) = 3 \geq (1 - T_w) \times w(1)$. Here T_w is a threshold for cluster formation that is set as 0.2 by default. Similarly, nodes 3 and 8 are also added into this cluster and finally MCODE predicts $\{1, 2, 3, 8\}$ as a protein complex. Subsequently, $\{4, 5, 6, 7\}$ is detected as another putative protein complex from this sample PPI graph.

The experimental results of MCODE method showed that the number of predicted complexes is generally small and the size of many predicted complexes is often too large.

Clique. Spirin and Mirny [Spir 03] proposed three methods for protein complex prediction from PPI network. The first method is to exhaustively enumerate the full cliques (fully connected subgraphs) as protein complexes. However, the use of cliques was too constraining given the incompleteness in the PPI data. As such, they also applied the Super-Paramagnetic Clustering (SPC) and a Monte Carlo (MC) simulation for the same purpose. Their experiments show MC performed better than SPC for clusters that share common nodes and for high density graphs, whereas SPC has an advantage identifying clusters that have very few connections to the rest of the graph. The MC

² k -core is a subgraph where all its nodes with degree greater than or equal to k .

CHAPTER 2. LITERATURE REVIEW

current incompleteness limitation of protein interaction data. For each protein, LCMA efficiently locates a local clique in its neighborhood graph in the first step. In the second step, LCMA merges local cliques that share high similarity (with large overlaps) as protein complexes. Evaluation results show that LCMA was more efficient and effective in detecting complexes than the full clique method[Spir 03] described above.

DPCLUS. Amin *et al.* [Alta 06] proposed an algorithm (DPCLUS) to detect protein complexes by keeping track of the periphery of a detected cluster.

The seed is first selected as the one with the highest weight or the one with the highest degree if all the vertices have weights zero. Here, the weight of a node is the sum of the weights of the edges connecting to this node, and the weight of an edge is the number of the common neighbors between the two endpoints of this edge.

The initial cluster is a seed and then the cluster grows gradually by including vertices one-by-one from the neighbors of the current cluster. In each iteration, only the priority vertex, which has both a high weight and a high cluster property score with respect to this cluster, will be added into the cluster. Given a cluster c_k and a vertex v_n outside c_k , the cluster property score of v_n with respect to c_k , cp_{nk} , represents the cohesiveness between c_k and v_n .

$$cp_{nk} = \frac{|E_{nk}|}{d_k \times |N_k|}, \quad (2.4)$$

where d_k is the density of c_k , $|N_k|$ is the number of nodes in c_k and $|E_{nk}|$ is the number of nodes in c_k connecting to v_n .

Once a cluster is generated, it is removed from the PPI network. Thus, DPCLUS can only detect non-overlapping clusters if there is no further process. To generate overlapping clusters, DPCLUS is extended to allow a cluster to include vertices that are closely associated with it but appear in other clusters.

PCP Chua *et al.* [Chua 07a] proposed an algorithm, ProteinComplexPrediction (PCP), for complex prediction. PCP first applies FS-weights [Chua 06] to evaluate the reliability of protein interactions and then modifies the PPI network by removing interactions with low FS-weights and including novel indirect interactions with high FS-weights. In the modified PPI network, PCP detects and merges dense subgraphs as protein complexes, using an efficient clique-finding algorithm from [Tomi 06] and a partial clique merging.

Hub Duplication. Ucar *et al.* [Ucar 06] developed a refinement method to detect protein complexes in scale-free PPI networks. Hub proteins (with degree greater than 25 in [Ucar 06]) are first selected and their neighborhood graphs are subsequently constructed. A hub-duplication strategy is then applied to detect dense subgraphs in these neighborhood graphs with multi-functional hub proteins assigned to multiple clusters.

CFinder. Adamcsek *et al.* [Adam 06] provided a software called CFinder to find functional modules in PPI networks. CFinder detects the k -clique percolation clusters as

CHAPTER 2. LITERATURE REVIEW

functional modules using a *Clique Percolation Method* [Pall 05]. In particular, a k -clique is a clique with k nodes and two k -cliques are adjacent if they share $(k - 1)$ common nodes. A k -clique percolation cluster is then constructed by linking all the adjacent k -cliques as a bigger subgraph.

SCAN. Mete *et al.* [Mete 08] proposed a new methodology called SCAN to detect functional modules in PPI networks. SCAN is extended from a well-known density-based clustering called DBSCAN [Este 96]. SCAN first defines the structural similarity between two proteins based on their common neighbors in Equation (2.5). Two proteins are structure-reachable if their structural similarity is greater than a threshold and a protein is a core node if it has several structure-reachable neighbors. SCAN augments a cluster from a core node by iteratively including structure-reachable neighbors. Additionally, SCAN can also identify the hubs and outliers in the PPI networks.

$$\sigma(u, v) = \frac{|V_u \cap V_v|}{\sqrt{|V_u| \times |V_v|}}, \quad (2.5)$$

where $\sigma(u, v)$ is the structural similarity between two proteins u and v , V_u is $\{u\} \cup N_u$ and N_u is the set of neighbors of protein u .

GS. Navlakha *et al.* [Navl 09] applied a graph summarization (GS) technique [Navl 08] to cluster a PPI graph into functional modules. GS compresses the input PPI graph into a summary graph which shows a high-level structure of the input graph. In the summary graph, the nodes correspond to non-overlapping sets of proteins which share similar interacting neighbors in the PPI network and thus are predicted as functional modules.

CMC. Liu *et al.* [Liu 09] recently proposed a Clustering method based on Maximal Cliques (CMC) to detect protein complexes. CMC first obtains all the maximal cliques by applying a maximal clique mining algorithm [Tomi 06]. CMC then assigns each interaction a score based on the reliability measure in [Liu 08]. Therefore, each clique can be scored with its weighted density. Last, CMC removes or merges highly overlapping cliques to generate protein complexes. In particular, if two cliques are highly overlapping, CMC either merges these two cliques as a bigger one or simply removes the one with a lower score (weighted density) depending on their inter-connectivity.

2.3.3.2 Incorporating gene expression data

Proteins which interact with each other can be expected to be activated and repressed under the same conditions. In other words, interacting proteins are likely to exhibit similar gene-expression profiles. In fact, gene expression data has been widely exploited to annotate protein functions (guilt by association) and predict novel protein-protein interactions [Jans 03, Bhar 05, Li 06a]. We describe below some methods for incorporating gene-expression data to help identify protein complexes in PPI networks.

CHAPTER 2. LITERATURE REVIEW

GFA. Feng *et al.* [Feng 08] proposed a graph fragmentation algorithm (GFA) to detect protein complexes using protein interaction graphs weighted with microarray data. For a PPI graph $G = (V, E)$, two different density definitions as shown in equation (2.6) are used:

$$density(G) = \frac{|E|}{|V|} \text{ and } density_w(G) = \frac{|E|}{\sum_{v \in V} w(v)}, \quad (2.6)$$

where $w(v)$ is the weight of the protein v weighted by $e^{-expression(v)}$, and $expression(v)$ is the log fold change of v 's gene-expression profile. GFA first applies DSA (Densest Subgraph Algorithm) [Gall 89] to find the densest subgraphs by maximizing the ratios in equation (2.6). It then removes redundant subgraphs occurring in different samples of the microarray data. The resulting clusters are genes that are highly differentially co-expressed and hence likely to be protein complexes.

DMSP. Maraziotis *et al.* [Mara 07] developed an algorithm called DMSP (Detect Module from Seed Protein). DMSP operates in three phases. First, proteins are clustered based on the gene-expression data using fuzzy c-means algorithm. Given two proteins from different clusters, their similarity is calculated by the distance between two cluster centroids and the distance from each protein to its corresponding centroid. Second, the extensions of weighted degree and density are obtained for seed selection and dense subgraph formation. Third, given a seed protein s , its neighbors and even its indirect neighbors are iteratively included based on different criteria to form the module.

MATISSE. Ulitsky *et al.* [Uliti 07] also proposed a method called MATISSE (Module Analysis via Topology of Interactions and Similarity SETs) to grow a functional module from a set of seed proteins. First, the edge weights are mainly determined by the gene expression correlation between interacting proteins and the node weights are their weighted degree. Second, given a protein with the highest weight, k neighbors with the highest weights are picked to form a set of $(k + 1)$ seeds. Last, after selecting the seeds, Jointly Active Connected Subnetworks (JACS) are obtained by several operations (*e.g.*, adding proteins into the sets of seeds). Two small JACSs are merged to form a new one if they are closely connected.

2.3.3.3 Incorporating functional information

Functional information can also be incorporated to accurately detect protein complexes. Since proteins within the same protein complex are generally aggregated to perform a common function, various measurements for the functional enrichment of a cluster can be used to indicate its tendency to be a real complex. For example, we can have weighted density for clusters in Equation (2.2) where edge weights are evaluated by the consistency of functional similarity between two proteins.

CHAPTER 2. LITERATURE REVIEW

RNSC. King *et al.* [King 04] proposed a “Restricted Neighborhoods Search Clustering” (RNSC) algorithm to detect protein complexes based on both graph-theoretical and gene-ontological properties. RNSC starts with an initial random clustering and then searches for a better clustering with the minimum costs by vertex-moving. RNSC discards unpromising clusters based on their size, density and functional homogeneity. The functional homogeneity of a cluster is defined as the smallest p-value over all the functional groups. Relatively few complexes are predicted by RNSC and its results depended heavily on the quality of the initial clustering which is random or user-defined.

DECAFF. Li *et al.* [Li 07b] proposed an algorithm called DECAFF (Dense-neighborhood Extraction using Connectivity and conFidence Features) to incorporate functional information to detect dense and reliable subgraphs as protein complexes. Firstly, a Hub-removal algorithm is developed to mine multiple possibly overlapping dense subgraphs in a neighborhood graph. Secondly, the dense subgraphs detected by the Hub-removal algorithm and the local cliques by LCMA [Li 05] (in section 2.3.3.1, page 16) are processed by a merging operation if two subgraphs have a large overlap. Thirdly, DECAFF filters out possible false protein complexes with low reliability, ensuring that the proteins in the predicted protein complexes are connected by high confidence protein interactions in the underlying network. Here, the reliability of a subgraph is the average reliability of the edges within a complex, which is estimated by using a probabilistic model with the functional information of interacting proteins.

SWEMODE. Lubovac *et al.* [Lubo 06] presented an algorithm called SWEMODE (Semantic WEights for MODule Elucidation) to detect functional modules in PPI networks. SWEMODE assigns the weights to the nodes in a different manner. First, each edge is weighted by the semantic similarity of GO terms annotating its two proteins [Resn 95, Lord 03, Wang 07, Cout 07]. Second, the nodes have two different weighting schemes based on their weighted clustering coefficients. Finally, SWEMODE selects seeds and augments the clusters from the seeds in a similar way as MCODE [Bade 03a].

STM. Cho *et al.* [Cho 07] extended flow-based modularization approach called STM [Hwan 06] to identify functional modules and protein complexes by considering the functional information. In this work, two novel measures are developed to index the reliability of interactions based on GO terms. The weights of proteins are their corresponding weighted degree. Informative proteins (those with large weights) are then selected and the flow simulation from each informative protein will decompose the weighted PPI network into preliminary clusters. A post-processing step is also devised to merge similar preliminary clusters as protein complexes.

2.3.3.4 Using TAP datasets for mining

The techniques discussed above have used pairwise physical interactions detected by high-throughput experiments such as Y2H as the PPI dataset for detecting protein complexes.

CHAPTER 2. LITERATURE REVIEW

More recently, there are some researchers who attempt to detect protein complexes from interaction data obtained solely from TAP experiments. Unlike Y2H method which detects direct physical interactions, using TAP data requires careful weighing of the detected links as TAP also detects indirect interactions in protein complexes.

Krogan *et al.* [Krog 06] were one of the first to use high-throughput purification data to predict protein complexes. In their constructed PPI network, the edge weights are learned by machine learning techniques from both the purification records and the mass spectrometry scores. Markov Clustering (MCL) [Van 00, Pere 04] is then applied to generate non-overlapping clusters as protein complexes.

Caroline *et al.* [Frie 08] proposed a novel method to detect protein complexes by bootstrap sampling. First, several bootstrap samples (1000 in this work) are selected from the purification records with replacement. Then, MCL [Van 00, Pere 04] is applied to generate preliminary complexes from each sample with edge weights using socio-affinity indices [Gavi 06]. Finally, protein complexes are detected by MCL in the bootstrap network, where two proteins have an edge if they are clustered together in at least one sample. The reliability of an edge in the bootstrap network is inferred by the number of samples for which two proteins are in the same preliminary complex.

Zhang *et al.* [Zhan 08] proposed a dice coefficient to measure the reliability of protein interactions based on purification records. An unweighted PPI network is constructed by removing unreliable edges with weights less than a pre-defined threshold. Finally, maximal cliques are detected by the maximal clique finding algorithm in [Zhan 05] and some highly overlapping cliques are merged to form larger dense subgraphs. Predicted complexes in this work are with core-attachment structures. In particular, they defined the core proteins in a detected complex as those present in at least $\frac{2}{3}$ of the original cliques which are merged into this complex and the rest are all attachment proteins.

Pu *et al.* [Pu 07] also applied MCL to detect protein complexes from the purification data. In their work, the reliability of interactions are inferred from a scoring function, which combines the evidence in each purification for bait-prey and prey-prey relationships [Coll 07]. Similarly, Hart *et al.* [Hart 07] used alternative scoring scheme together with MCL to detect protein complexes.

Recently, Geva *et al.* [Geva 11] proposed a new approach called CODEC to detect protein complexes from TAP data. Unlike above methods that convert TAP data to PPI networks, CODEC models the TAP data as a bipartite graph $G = (U, V, E)$ where U and V represent the sets of baits and preys respectively and E describes the bait-prey relationships detected in the experiments as shown in Figure 2.7. CODEC defines a likelihood ratio score for a candidate bipartite subgraph to measure its density versus the chance that it is randomly generated. CODEC first identifies candidate complexes from the neighborhood of each prey protein and then modifies them by adding or deleting vertices to maximize their likelihood ratio scores. Subsequently, CODEC filters the redundant candidates and obtains the final list of protein complexes.

In Figure 2.7, $\{1, 2, 3, 4\}$ is the set of baits and $\{5, 2, 6, 7, 8, 9\}$ is the set of preys in this sample bipartite graph. CODEC finally predicts three protein complexes from this graph, namely $\{2, 6, 7, 8\}$, $\{2, 3, 7, 8\}$ and $\{3, 4, 9\}$ with likelihood ratio scores 2.39, 2.50 and 1.79, respectively.

Figure 2.7: CODEC on a sample TAP bipartite graph.

2.3.4 Recent Emerging Techniques

In this section, we review a number of emerging techniques for protein complex detection that are different from the application of traditional graph clustering described in the previous section.

2.3.4.1 Complex detection with supervised graph clustering

The previous graph clustering methods described above are unsupervised and are more or less based on the basic assumption that dense subgraphs in PPI networks are likely to be protein complexes. The protein complexes detected by many of these methods must be either cliques or defective cliques [Spir 03, Zhan 08] or are dense subgraphs that have density values above a pre-selected threshold (e.g., 0.7) [King 04, Li 05, Alta 06, Li 07b]. The flow-based clustering methods [Van 00, Cho 07] are also biased towards dense subgraphs. It is possible that some protein complexes may not have a dense structure in the underlying PPI network and they will thus not be detected by these methods.

Qi *et al.* [Qi 08] proposed a supervised graph clustering framework to predict protein complexes without such prior assumptions on the topological properties of protein complexes. By using a supervised approach, important topological and biological properties of known protein complexes will be learned as guideline to detect new complexes in the PPI networks.

To obtain the training data, they collect available known protein complexes and also generate some random subgraphs as non-complexes. Topological and biological

CHAPTER 2. LITERATURE REVIEW

properties of these training graphs are then summarized as features. A probabilistic Bayesian network (BN) is then applied to integrate all these features and the parameters of this BN model are learned from the training data.

Given a graph candidate and its corresponding features, a log likelihood ratio score can thus be calculated by the BN model to show whether it is qualified to form a complex. A simulated annealing search [Idek 02] is further employed to modify the candidate graph if possible. For example, a new protein can be included into a candidate protein complex if the resulting augmented cluster has a higher log likelihood ratio score.

Experimental results showed that the extracted features are capable of distinguishing complexes versus non-complexes and the supervised manner can provide more accurate identification of protein complexes. However, the knowledge learned from the limited training data could be biased and affect the complex formation during the clustering.

2.3.4.2 Complex detection by considering exclusive or cooperative interactions

Two adjacent interactions (those with a common protein) may be mutually exclusive [Kim 06, Alex 09] due to the overlapping binding interfaces on the common protein. A dense cluster detected by the traditional methods by blindly treating all protein-protein interactions in the PPI network to be able to occur simultaneously may be a false prediction as it may contain several mutually-exclusive interactions.

Jung *et al.* [Jung 08] recently proposed a method to extract Simultaneous Protein Interaction Clusters (SPIC) by considering this issue. They used both MCODE [Bade 03a] and LCMA [Li 05] to generate initial clusters. They then collect the SPICs by excluding the conflicting interactions so that it is possible for all the interactions within a SPIC to occur at the same time. Unfortunately, there are still many practical problems that need to be addressed, such as collecting more conflicting interactions based on 3D structure data and developing more efficient methods for SPIC.

Jin *et al.* [Jin 09] exploited the time-series of gene expression profiles to determine whether two adjacent interactions are exclusive or cooperative. Each protein has a gene expression time-series such that each interaction is associated with a time-range pair by aligning the corresponding two gene expression time-series of the proteins. Two adjacent interactions are considered to be cooperative and can thus occur simultaneously if their common partner protein has overlapping time-ranges for these two interactions.

2.3.4.3 Complex detection using evolutionary information

With the increasing availability of PPI data for most species (such as yeast, fly, worm and so on), it has become feasible to use cross-species analysis to derive insights into the evolution of the PPI networks for complex detection.

CHAPTER 2. LITERATURE REVIEW

Sharan *et al.* recently proposed a series of methods for comparative analysis in two or more species. They used these methods for conserved pathway detection [Kell 03], protein function analysis [Shar 05] and conserved protein complex detection [Shar 04, Hirs 06, Dost 07]. Basically, to detect the conserved protein complexes in two species, an orthology graph (also called network alignment graph) is constructed, in which each node represents a pair of sequence-similar proteins (homologous or orthology proteins) and each edge represents a conserved interaction between the corresponding protein pairs in each species.

In [Shar 04], each node (u, v) in the orthology graph is weighted by the sequence similarity between the protein pair u and v . An edge $((u_1, v_1), (u_2, v_2))$ is associated with a pair of weights $(w(u_1, u_2), w(v_1, v_2))$, where $w(u_1, u_2)$ is the weight of the interaction (u_1, u_2) . Two models, the protein-complex model and null model, are proposed to learn the weights of interactions and detect protein complexes in each species.

In [Hirs 06], the researchers detect conserved protein complexes across yeast and fly based on a probabilistic model which considers the evolution of PPI networks through link dynamics and gene duplications [Berg 04]. A subgraph in the orthology graph can have a likelihood ratio score from the protein-complex model and null model as above. Candidate subgraphs with highest scores are detected as conserved complexes, using a similar heuristic as in [Shar 04].

In [Dost 07], a tool called *QNet* was developed for queries in PPI networks. The similarity between two graphs is defined based on the node and edge similarity and the penalty scores for node deletion and insertion. *QNet* then performs tree queries and bounded-treewidth graph queries by the color coding algorithm [Alon 95]. Conserved protein complexes are obtained through querying known yeast complexes against the PPI networks of other species.

Another group of researchers, Dutkowski *et al.* [Dutk 07] also proposed an evolution-based framework to detect conserved protein complexes across multiple species. First, all the proteins from different species are clustered by MCL with BLAST E-scores as pairwise similarity and the proteins from each cluster are homologous and thus believed to have a common ancestral protein. Then, the interactions between the ancestral proteins are assigned weights under the duplication and speciation model of the PPI network evolution. Finally, in the conserved ancestral PPI network, the connected components after removing the edges with weights lower than an appropriate threshold are predicted as conserved protein complexes.

2.3.5 Summary

Identifying protein complexes is important for biological knowledge discovery since many important biological processes in the cell are carried out through the formation of protein

CHAPTER 2. LITERATURE REVIEW

Table 2.3: Key developments for computational detection of protein complexes.

Methods	Main contributions
MCODE [Bade 03a]	MCODE pioneered the computational detection of protein complexes from PPI networks.
MCL [Van 00, Pere 04]	MCL is a widely used method [Krog 06, Pu 07, Hart 07, Frie 08] with good convergence and robustness [Broh 06].
Krogan et al. [Krog 06]	They provided and exploited a comprehensive TAP dataset for computational protein complex detection.
Gavin et al. [Gavi 06]	In addition to providing a TAP dataset widely used for protein complex detection, they described the inherent organization of protein complexes with core-attachment structures.
Sharan et al. [Shar 04, Hirs 06, Dost 07]	They systematically conducted cross-species analysis of PPI networks to derive evolutionary information for protein complex detection.

complexes. However, there is currently a wide gap between data on protein complexes and (pairwise) protein-protein interactions. High throughput technologies for detecting pairwise protein-protein interactions *en masse* have already become routine in the laboratories for generating large datasets of protein interaction data, while the high-throughput technologies for detecting protein complexes remained relatively immature. Hence computational approaches for detecting protein complexes are needed to help fill up the relatively empty map for the protein “complexome”.

In this chapter, we have reviewed current computational approaches that have been proposed to exploit the abundant protein interaction data to bridge the data gap for protein complexes. Protein interaction graph mining algorithms that identify graphical subcomponents in the protein-protein interaction networks can be used for predicting protein complexes. We have surveyed the state-of-the-art algorithms by describing the traditional graph clustering methods as well as the recent emerging techniques for computational detection of protein complexes from PPI and other data sources. Table 2.3 lists five methods that, in our opinion, represent the key developments for computational protein complex detection so far.

On the other hand, we can still improve these existing methods. In the following discussions, we describe three key challenges for further improvements [Li 09].

- *PPI data is noisy with high false positive and false negative rates* [Bade 04, Hart 06, Gent 07]. The computational methods are highly dependent on the quality of the underlying interaction data. Unfortunately, despite the abundance of PPI data, the data quality of these data still leaves much to be desired. For example, the

CHAPTER 2. LITERATURE REVIEW

experimental conditions in which the PPI detection methods are carried out may cause a bias towards detecting interactions that do not occur under physiological conditions, resulting in false positive detection rates that could be alarmingly high [Meri 02]. At the same time, the high-throughput methods may also fail to detect various types of interactions, *e.g.*, loss of weak transient interactions, loss of post-translational modification and bias against soluble or membrane proteins [Tara 08]. These result in high false negative rates and low experimental coverage of the interactomes. The computational methods for complex mining are clearly limited by the poor quality of the underlying PPI data. Further improvements for complex mining can be obtained by improving the quality of the PPI data with new and more powerful experimental detection technologies, or with computational approaches for validating the existing protein interactions (to address the false positive interaction issue) and predicting novel protein interactions (to address the false negative interaction issue).

- *Graph mining of PPI networks is intrinsically computationally challenging*

PPI networks are very large graphs with thousands of vertices and tens of thousands of edges, even for a simple organism such as yeast. For the more complicated species such as the human being, the scale and complexity of the PPI networks are clearly overwhelming. The fact that many graph-based problems, such as subgraph isomorphisms [Ullm 76, Abdu 98] and enumeration [Shas 02, Groc 07], are NP-hard [Gare 79; Ausi 99], indicates that mining the PPI networks is intrinsically computationally challenging. Although mining objects such as complexes from protein interaction networks are computationally challenging problems, it may be possible to reduce the search space and time complexity and obtain better mining results by exploiting biological knowledge coupled with the development of novel graph mining techniques that are specialized on such networks [Bara 04].

- *Integrate various biological evidences into the mining process.*

Through integrating various independently obtained biological evidences, we can assess/weight the protein interactions by using appropriate confidence measures. For example, we can employ metrics from biological evidences such as reproducibility of the interactions from multiple experimental methods, support from such other non-interaction data as co-expression, co-localization and shared functions, as well as the conservation of the protein interactions across other genomes, etc to address the limitations in the current quality of PPI data. Computational methods, such as Bayesian network models [Jans 03] and kernel methods [Ben 05], have been proposed to integrate different biological resources (*e.g.* integrate weighted genomic features in [Jans 03] and mapping different features into high dimensional vector space in [Ben 05]).

2.4 Evaluation of Protein Complexes

Overall, there are three types of evaluation metrics used to evaluate the quality of the predicted complexes and compute the overall precision of the prediction methods.

2.4.1 Precision, Recall and F-measure

Precision, recall and F-measure are commonly-used evaluation metrics in information retrieval and machine learning. For evaluating protein complex prediction, we need to define how well a predicted complex which consists of a set of protein members, matches an actual complex, which is another set of protein members. The neighborhood affinity score $NA(p, b)$ between a predicted complex $p = (V_p, E_p)$ and a real complex $b = (V_b, E_b)$ in the benchmark complex set, as defined in equation (2.7) below, can be used to determine whether they match with each other. If $NA(p, b) \geq \omega$, they are considered to be matching (ω is usually set as 0.20 or 0.25). Let P and B denote the sets of complexes predicted by a computational method and real ones in the benchmark, respectively. Let N_{cp} be the number of predicted complexes which match at least one real complex and N_{cb} be the number of real complexes that match at least one predicted complex. Precision and Recall are then defined as follows: [Chua 07a, Wu 09a, Geva 11] (Note that the definition of Recall here is different from the one in [Bade 03a, Li 05, Li 07b]):

$$NA(p, b) = \frac{|V_p \cap V_b|^2}{|V_p| \times |V_b|}, \quad (2.7)$$

$$N_{cp} = |\{p | p \in P, \exists b \in B, NA(p, b) \geq \omega\}|,$$

$$N_{cb} = |\{b | b \in B, \exists p \in P, NA(p, b) \geq \omega\}|,$$

$$Precision = \frac{N_{cp}}{|P|} \quad \text{and} \quad Recall = \frac{N_{cb}}{|B|}. \quad (2.8)$$

F-measure, or the harmonic mean of Precision and Recall, can then be used to evaluate the overall performance:

$$F = 2 \times Precision \times Recall / (Precision + Recall). \quad (2.9)$$

2.4.2 Sensitivity, Positive Predictive Value and Accuracy

Recently, sensitivity (Sn , also called as Coverage Rate in this thesis), positive predictive value (PPV) and accuracy (Acc) have also been proposed to evaluate the accuracy of the prediction methods [Broh 06, Frie 08]. Given n benchmark complexes and m predicted complexes, let T_{ij} denote the number of proteins in common between i^{th} benchmark complex and j^{th} predicted complex. Sn and PPV are then defined as follows:

CHAPTER 2. LITERATURE REVIEW

$$Sn = \frac{\sum_{i=1}^n \max_j \{T_{ij}\}}{\sum_{i=1}^n N_i} \quad \text{and} \quad PPV = \frac{\sum_{j=1}^m \max_i \{T_{ij}\}}{\sum_{j=1}^m T_j}. \quad (2.10)$$

Here N_i is the number of proteins in the i th benchmark complex, and $T_{.j} = \sum_{i=1}^n T_{ij}$. Generally, high Sn values indicate that the prediction has a good coverage of the proteins in the real complexes, while high PPV values indicate that the predicted complexes are likely to be true positives. As a summary metric, the accuracy of a prediction, Acc , can then be defined as the geometric average of sensitivity and positive predictive value,

$$Acc = \sqrt{Sn \times PPV}. \quad (2.11)$$

2.4.3 P-values

As we gained more and more biological knowledge about the proteins, we can associate a protein with (possibly multiple) functional annotations. Here, we exploit the p-value to show the statistical significance of the occurrence of a protein cluster (predicted protein complex) with respect to a given functional annotation, which follows the hypergeometric distribution in equation (2.12) [Bu 03, Przu 04]:

$$Pvalue = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i} \binom{|V| - |F|}{|C| - i}}{\binom{|V|}{|C|}} \quad (2.12)$$

where a predicted complex C contains k proteins in the functional group F and the whole PPI network contains $|V|$ proteins. The p-value of a predicted complex is the smallest p-value over all the possible functional groups. A predicted complex with a low p-value indicates a high functional homogeneity, *i.e.*, it is enriched by proteins from the same function group and it is thus likely to be true protein complex. By setting a common threshold which specifies the acceptable level of statistical significance, the numbers of predicted complexes with p-value under this threshold for the various methods can then be used for evaluating their respective overall performance.

It is important to realize that the evaluation metrics described above can only provide us some sense of how well the current graph mining techniques can be used to detect the protein complexes from protein interaction data. These metrics are by no means absolute measures — they all have their own limitations, especially for sensitivity (Sn), positive predictive value (PPV) and Accuracy (Acc). For sensitivity (Sn), if a method predicts a giant complex which covers many proteins in the known real complex set, then this method will get a very high Sn score. As for PPV value (PPV), it does not evaluate overlapping clusters properly. Here is a case: if the known gold standard MIPS complex

CHAPTER 2. LITERATURE REVIEW

set (with proteins that belong to multiple complexes) [Mewe 04] is taken to match with itself, then the resulting *PPV* value is 0.772 instead of 1 (indicating an imperfect match) while the Precision and Recall are both correctly 1. As such, the Accuracy (*Acc*) score, as the geometric average of *Sn* and *PPV*, will also not make good sense.

In addition, all the evaluation metrics described above assumed that a complete set of true protein complexes is available, while we are far from it in reality. If a method predicts an unknown but real protein complex (which is not similar with any of the known complexes), all of these evaluation metrics will regard it as a false positive. Furthermore, for P-values, since its calculation relies on the availability of the proteins' functional information, its applicability would be limited in the less studied genomes. As such, it has mainly been used in the model organism yeast for which rich molecular functional information is available.

Relatively speaking, the Precision, Recall, F-measure and P-values are thus more acceptable for evaluating the performance of current techniques. Still, we need to treat the current evaluation metrics with caution, as more research is needed to come up with a robust evaluation metric for the protein complex prediction task.

Chapter 3

COACH: Detecting Protein Complexes from PPI Networks

3.1 Introduction

3.1.1 Background

With the completion of many genome-sequencing projects, the focus in the post-genomic era has turned to proteomics. One important task in proteomics is to detect protein complexes based on the PPI data generated by various experimental technologies, *e.g.*, yeast-two-hybrid [Ito 01], affinity purification [Puig 01, Gavi 02, Ho 02] and others.

Pair-wise protein interactions can be modeled as a graph or network, where vertices are proteins and edges are protein-protein interactions (PPI). Such a network modeling provides a new perspective to understand the complicated biological systems [Li 09]. As introduced in our review chapter, numerous methods have already been proposed to detect protein complexes from this kind of PPI networks.

3.1.2 Motivation

Existing methods mainly focus on detecting highly connected subgraphs in PPI networks as protein complexes but ignore their inherent functional organization. However, recent analysis of experimentally detected protein complexes has revealed their inherent organization [Dezs 03]. A protein complex generally contains a core in which proteins are highly co-expressed and share high functional similarity. The protein-complex core is often surrounded by some attachments, which assist the core to perform subordinate functions. Gavin et al.'s work [Gavi 06] also demonstrates the similar architecture and modularity of protein complexes. Figure 3.1 shows an example of DNA repair complex [Gavi 06] with core-attachment architecture. In this figure, the DNA repair complex has 9 proteins, where the core consists of 4 red proteins (YAR007C, YDR097C, YML032C

and YIR002C) in the dashed circle. These four proteins form a fully connected subgraph (clique) in the PPI network. Proteins YBR114W, YER095W, YJL173C, YNL312W and YOL090W are the attachments of this complex. The interactions in this figure are from the DIP data.

Figure 3.1: DNA repair complex, which repairs DNA damage by interacting with damaged DNA.

In this chapter, we propose a new method, called COACH (**C**ore-**A**ttachment based method), to detect protein complexes in PPI networks by considering their inherent organizations. In particular, protein-complex cores, as the “hearts” of the protein complexes, are first detected from each vertex’s neighborhood graphs. We subsequently generate protein complexes by including attachments into the protein-complex cores. Experimental results using PPI data of *Saccharomyces cerevisiae* show that our COACH method does provide insights into the inherent modularity and organization of protein complexes. In addition, in terms of prediction accuracy, our COACH method also outperforms existing computational methods.

3.2 Method

A protein-protein interaction (PPI) network can be modeled as a simple graph $G = (V, E)$, in which a vertex in vertex set V represents a protein and an edge in edge set E represents an interaction between two distinct proteins. This graph structure modeling is helpful for global analysis of PPI data, such as graph clustering for protein complex detection. Our COACH method operates in two phases. COACH first detects protein-complex cores and then applies an outward growing strategy to generate protein

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

complexes by including attachments into the protein-complex cores. We will first briefly introduce some basic terminologies and then describe in detail our proposed method for protein complex detection.

3.2.1 Preliminaries

Given a PPI network $G = (V, E)$, the degree of a vertex $v \in V$ is the number of v 's neighbors in G , written as $\deg(v)$. The average degree of graph G is defined as the average of $\deg(u)$ for all $u \in V$, written as $Avdeg(G)$ in equation 3.1. The density of G , denoted as $den(G)$, is defined in the equation 2.1 in section 2.3.

$$Avdeg(G) = \sum_{u \in V} \frac{\deg(u)}{|V|}. \quad (3.1)$$

Given two graphs $A = (V_A, E_A)$ and $B = (V_B, E_B)$, their neighborhood affinity [Bade 03a], $NA(A, B)$, is defined as follows to measure the similarity between them (Note that this equation is the same as that defined in Equation (2.7) section 2.4),

$$NA(A, B) = \frac{|V_A \cap V_B|^2}{|V_A| \times |V_B|}. \quad (3.2)$$

For a vertex $v \in V$, the neighborhood graph of v consists of v , all its neighbors and the edges among them. It is defined as $G_v = (V', E')$, where $V' = \{v\} \cup \{u | u \in V, (u, v) \in E\}$, and $E' = \{(u_i, u_j) | (u_i, u_j) \in E, u_i, u_j \in V'\}$. In G_v , there will be some vertices with degree 1 (*i.e.*, only connect with the vertex v) and generally the interactions involving these proteins have low reliability with respect to the topological reliability measures in [Sait 02, Sait 03, Chua 06]. Therefore, all vertices with degree 1 will be removed from G_v . Since current PPI data are quite noisy [Meri 02], this preprocessing step can help us to filter out possible false positive interactions. The neighborhood graph of v , G_v , thereafter refers to the above remaining graph if it is not empty. As a result, every vertex in G_v has at least two neighbors and $Avdeg(G_v) \geq 2$.

3.2.2 Defining Preliminary Cores

A protein-complex core is a small group of proteins which show a high mRNA co-expression patterns and share high degree of functional similarity. It is the key functional unit of the complex and largely determines the cellular role and essentiality of the complex [Dezs 03, Gavi 06]. Protein-complex cores and their members often have specific topological properties in PPI networks. For example, a protein in a core often has many interacting partners and protein-complex cores often correspond to small, dense and reliable subgraphs in PPI networks [Gavi 06]. In addition, complex cores may have overlaps with each other.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

According to these properties of protein-complex cores, we first define their possible candidates in the neighborhood graphs, denoted as preliminary cores. A preliminary core in a neighborhood graph G_v is a dense subgraph where all its members should show higher significance and have heavier weights [Bade 03a, Alta 06] than those non-members. In particular, we first define a vertex $u \in G_v$ as a core vertex if u 's degree in G_v is larger than or equal to G_v 's average degree, *i.e.*, $\deg(u) \geq \text{Avdeg}(G_v)$. The core graph of G_v is defined as the subgraph formed by all the core vertices. A preliminary core in G_v defined in this chapter should satisfy the following three constraints: (1) it is a subgraph of the core graph, that is, all its vertices are core vertices, (2) it is dense (with density $\geq d$ and d is typically set as 0.7 in [Alta 06, Li 07b], which is also used in this chapter) and (3) it is maximal, that is, none of its supergraphs satisfy the first two constraints.

Above definition of preliminary cores is based on the definition of core vertices, which further relies on the degree distribution of the PPI network. As we know, PPI networks are considered to be scale-free ¹ [Yook 04]. Therefore, preliminary cores would mostly be formed around the proteins with relatively large degrees. This is reasonable to form preliminary cores because proteins with high degrees in PPI network serve important biological roles [Yook 04, Przu 04].

3.2.3 Mining Protein-Complex Cores

Based on the definition of preliminary cores, we are now ready to describe our proposed algorithm to detect protein-complex cores. In our algorithm, the preliminary cores are first detected from the neighborhood graph of each vertex in the PPI network. Specifically, given a neighborhood graph G_v , if its core graph CG is dense, CG is thus directly predicted as a preliminary core; otherwise, multiple possible preliminary cores would be detected from CG . Since some vertices have similar neighborhood graphs, the preliminary cores detected from their neighborhood graphs may have large overlaps, which will result in high redundancy. Hence, a Redundancy-filtering procedure is applied to process preliminary cores and finally generate protein-complex cores by eliminating such kind of redundancy.

Algorithm 1 illustrates the overall framework to detect protein-complex cores. For each vertex v in the PPI network $G = (V, E)$, we first construct its neighborhood graph G_v and G_v 's core graph CG in line 3. If CG is dense enough, our Core-removal algorithm in line 4 will return it as a preliminary core. If not, Core-removal algorithm may generate several subgraphs of CG , which will be further processed to be maximal dense as preliminary core in lines 5-14 (note that a subgraph sg obtained from the Core-removal

¹In scale-free networks, the degree distribution follows a power-law, *i.e.*, $P(k) \sim k^{-\alpha}$, where $P(k)$ is the probability of a nodes with a degree of k and $\alpha > 0$. That is, the majority of the nodes will have small degrees while a small part of nodes denoted as ‘‘hubs’’ will have very large degrees.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

algorithm may not be dense or maximal). If sg is not dense, we first iteratively remove vertices with the minimum degree until it is dense in lines 6-9. We may add some core vertices into sg , which are highly connected to the vertices in sg , to guarantee that sg is maximal and dense in lines 10-14. SC is the set of all currently detected preliminary cores. sg will then be discarded or preserved in SC by the Redundancy-filtering procedure in line 15. Note that after filtering the possible redundant preliminary cores, all the remaining preliminary cores in set SC will be regarded as protein-complex cores. The details of Core-removal algorithm, as well as Redundancy-filtering procedure, are described in Algorithm 2 and 3 respectively.

Algorithm 1 Protein-complex Core Mining Algorithm

Input: The PPI network $G = (V, E)$;
 Density threshold d ;
 Neighborhood affinity threshold to filter redundancy t .

Output: The set of protein-complex cores, SC .

```

1:  $SC = \phi$ ; //initialization
2: for each vertex  $v \in V$  do
3:   construct the core graph of  $G_v$ ,  $CG = (V_{CG}, E_{CG})$ ;
   //  $V_{CG} = \{u | \deg(u) \geq Avdeg(G_v), u \in G_v\}$ 
4:    $S = \text{Core-removal}(CG)$ ;
5:   for each element  $sg \in S$  do
6:     while  $den(sg) < d$  do
7:        $w = \arg \min_{u \in V_{sg}} \deg(u)$ ; //  $\deg(u)$  is  $u$ 's degree in  $sg$ 
8:        $sg = sg - \{w\}$ ; // update  $sg$  by deleting  $w$  and its corresponding edges
9:     end while
10:     $w = \arg \max_{u \in V_{CG}, u \notin V_{sg}} \deg_1(u)$ ; //  $\deg_1(u)$  is  $u$ 's degree in  $sg + \{u\}$ 
11:    while  $w$  exists and  $den(sg + \{w\}) \geq d$  do
12:       $sg = sg + \{w\}$ ; // update  $sg$  by adding  $w$  and its corresponding edges
13:       $w = \arg \max_{u \in V_{CG}, u \notin V_{sg}} \deg_1(u)$ ;
14:    end while
15:    Redundancy-filtering( $sg, t, SC$ );
16:  end for
17: end for

```

3.2.3.1 Core-removal algorithm

Given a vertex v , if the core graph of G_v , CG , is dense enough, core-removal algorithm will return it directly; otherwise, there may be multiple preliminary cores in CG . Our core-removal algorithm works as follows. If CG is not dense enough, all the core vertices of CG

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

are first removed from CG , and the remaining graph may consist of a number of connected components. We recursively repeat this procedure to find highly-connected subgraphs in each of above connected components. The removed core vertices are added back into these subgraphs to form larger subgraphs of CG , which will be further processed to be maximal dense in Algorithm 1. Algorithm 2 shows the details of our core-removal algorithm and Figure 3.2 also provides an example to illustrate the process of our proposed core-remove algorithm. In this example, we assume that (A) shows the core graph of vertex 1's neighborhood graph, denoted as CG_1 . The density and average degree of CG_1 is 0.607 and 4.25 respectively. In (B), the core vertices of CG_1 , $\{1, 6\}$, are removed from CG_1 and two connected components are left in the remaining graph. In (C), $\{1, 6\}$ are added back into each connected component. Two subgraphs with vertices $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 6, 7, 8\}$ are obtained and finally returned.

Algorithm 2 Core-removal(cg) // $cg = (V_{cg}, E_{cg})$

```

1:  $result = \phi$ ;
2: if  $den(cg) \geq d$  do
3:   insert  $cg$  into  $result$ ;
4: else
5:    $cv = \{u | u \in V_{cg}, deg(u) \geq Avdeg(cg)\}$ ; //  $deg(u)$  is  $u$ 's degree in  $cg$ 
6:   remove all the vertices in  $cv$  from  $cg$  and obtain a set of connected components;
7:   for each connected component  $comp$  do
8:      $tresult = \text{Core-removal}(comp)$ ;
9:     for each element  $tc \in tresult$  do
10:      insert  $tc \cup cv$  into  $result$ ;
11:   end for
12: end for
13: return  $result$ ;

```

3.2.3.2 Redundancy-filtering procedure

Assume that SC is the set of all currently detected preliminary cores and $C = (V_C, E_C)$ is a newly detected preliminary core. We will first detect an element $B = (V_B, E_B)$ in SC , which has the highest similarity (NA score) with C .

In Algorithm 3, the procedure Redundancy-filtering(C) is used to check and decide whether to discard or preserve the newly detected preliminary core C . In particular, if B and C are not quite similar (with $NA(B, C) < t$), C will be inserted into SC in lines 2-3; otherwise, we prefer to preserve the preliminary cores that have larger sizes and densities in lines 4-8.

Figure 3.2: The Diagram of our Core-removal algorithm.

3.2.4 Assembling Protein Complexes

In previous subsections (the first phase of our COACH method), we have presented our techniques to detect the protein-complex cores. In the second phase, we will extract the peripheral information of each protein-complex core and select reliable attachments cooperating with it to form a protein complex. Given a PPI network $G = (V, E)$, the neighborhood of a complex core $C = (V_C, E_C)$ is defined as $N(C) = \{u | (u, v) \in E, v \in V_C, u \in V, u \notin V_C\}$. $N(C)$ consists of those direct neighbors of the vertices in the complex core C . For a vertex $v \in N(C)$, N_v is the set of all v 's neighbors. $|N_v \cap V_C|$ is the number of vertices in C connected with v . Thus, $\frac{|N_v \cap V_C|}{|V_C|}$ can be used to quantify the closeness between the vertex v and the core C , denoted as $\text{closeness}(v, C)$. Each vertex $v \in N(C)$ with $\text{closeness}(v, C) > 0.50$ will be selected as an attachment, indicating that selected attachments interact with more than half of the proteins in the core. In this way, the attachments are closely-associated with the complex core, showing that these attachments are in stable and reliable cooperation with the core.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Algorithm 3 Redundancy-filtering(C, t, SC)

```

1:  $B = \arg \max_{G' \in SC} NA(G', C); // B$  is  $C$ 's most similar subgraph in  $SC$ 
2: if  $NA(B, C) < t$  do
3:   insert  $C$  into  $SC$  (Inserting);
4: else
5:   if  $den(C) \times |V_C| \geq den(B) \times |V_B|$  do
6:     insert  $C$  into  $SC$  in place of  $B$  (Substituting);
7:   else
8:     discard  $C$  (Discarding).

```

In summary, our COACH method consists of the two stages above, *i.e.*, protein-complex core detection and complex formation (the available COACH system can be downloaded from <http://www1.i2r.a-star.edu.sg/~xlli/coach.zip>). An example in Figure 3.3 illustrates our proposed algorithm to detect protein complexes in PPI networks. For simplicity, the Redundancy-filtering procedure is not shown in this example. As we all know, detecting the complete set of preliminary cores or maximal dense subgraphs is a NP-complete problem (maximal clique finding is a special case when density threshold is 1). However, our heuristic, Core-removal algorithm, detects preliminary cores (not always the complete set of preliminary cores) from the core graphs, which are small-scale subgraphs within each vertex's neighborhood graphs, *e.g.*, the average size of core graphs is 4.30 in DIP data [Xena 02] and 5.45 in a denser Krogan *et al.*'s data [Krog 06], respectively. Therefore, our COACH method is very efficient to detect preliminary cores and protein complexes in PPI networks.

3.3 Experimental Results

In this section, we first briefly introduce the data and evaluation metrics. Then, we will focus on presenting the results on DIP data in detail. We perform both comprehensive comparisons among various existing computational methods and validation of our predicted core-attachment structures. In addition, the results on Krogan *et al.*'s data from [Krog 06] will also be briefly presented to demonstrate the effectiveness of our proposed technique.

3.3.1 Experimental Data and Evaluation Metrics

In our experiments, we performed our COACH method on two yeast PPI networks. One is from DIP (the Database of Interacting Proteins [Xena 02]), which consists of 17203 interactions among 4930 proteins. Krogan *et al.*'s PPI data [Krog 06] consists of

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Figure 3.3: The procedure of our COACH algorithm. In this example, COACH generates two complexes, where $\{0, 2, 3\}$ and $\{0, 8, 9\}$ in the dashed rectangles are their cores.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

14077 reliable interactions involving 3581 proteins (with a cut-off of 0.101 as shown in their supplementary Table S8). For evaluating our identified complexes, the set of real complexes from [Frie 08] was selected as benchmark. This benchmark set consists of 428 protein complexes, from three sources: (I) MIPS [Mewe 04], (II) Aloy *et al.* [Aloy 04] and (III) SGD database [Dwig 02] based on Gene Ontology (GO) annotations. In addition, the gene expression data was downloaded from Eisen’s lab [Genea] and GO data was downloaded from [Geneb]. As to evaluation metrics, we used Precision, Recall, F-measure and Coverage rate (or Sensitivity) defined in equations 2.8, 2.9 and 2.10 in section 2.4.

3.3.2 Comparative Evaluation

In this subsection, we compared the performance of our COACH method with other three competing algorithms, DPCLus [Alta 06], DECAFF [Li 07b] and MCL [Pere 04, Krog 06, Frie 08], using DIP data. For comprehensive comparisons, we employed several evaluation measures, including co-annotation, co-localization, functional enrichment of GO terms (p-values), F-measure and coverage rate. For all these methods, the optimal parameters were set to maximize their F-measures. For example, the *inflation* parameter in MCL was set as 1.9 when using DIP data [Wu 08a]. Note that for fair comparisons, we turned off the filtering step in DECAFF because it used the functional information to filter away possible false positive complexes while other techniques only used topological properties of PPI networks.

3.3.2.1 F-measure and coverage rate

Table 3.1 shows the basic information of predictions by various methods. In Table 3.1, MCL predicted 1116 complexes, of which 193 match 242 real complexes; DPCLus detected 1143 complexes, of which 193 match 274 real ones and DECAFF detected 2190 complexes, of which 605 match only 243 real ones. Our COACH method managed to predict 746 complexes, out of which 285 match 249 real complexes. In addition, MCODE [Bade 03a] predicted 182 complexes and correctly matched only 128 real complexes in the benchmark. Both the number of complexes predicted by MCODE and its N_{cb} (N_{cb} is the number of benchmark complexes that are correctly predicted, see section 2.4 for details) are far fewer than those of other algorithms. For this reason, MCODE is not included in the later comparisons.

Figure 3.4 shows the overall comparison in terms of F-measure and coverage rate. On DIP data, the F-measure of COACH is 46.1%, which is 19.6%, 19.4% and 8.9% higher than MCL, DPCLus and DECAFF respectively. Our COACH method can achieve the highest F-measure by providing the highest precision and comparable recall, which shows that our method can predict protein complexes very accurately. In Figure 3.4, our COACH method obtains the highest coverage rate of 34.9%, which is 1.7%, 4.5%

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Table 3.1: The results of various algorithms using DIP data.

Algorithms	MCODE	MCL	DPCLUS	DECAFF	COACH
# predicted complexes	182	1116	1143	2190	746
# covered proteins in DIP	1173	4930	2987	1832	1837
N_{cp}	93	193	193	605	285
N_{cb}	128	242	274	243	249

Figure 3.4: The performance comparison for various algorithms on DIP data, using F-measure and Coverage Rate.

and 8.3% higher than MCL, DPCLUS and DECAFF respectively. That is, our predicted complexes can cover the most proteins involved in the real complexes.

Figure 3.5 illustrates an example, in which our predicted SAGA complex [Gran 98] can cover more proteins in the real SAGA complex. In this example, the real SAGA complex in the benchmark consists of 20 proteins (Figure 3.5A). The complex predicted by our COACH method has 13 proteins and manages to cover 11 proteins (in red color). Meanwhile, MCL, DPCLUS and DECAFF cover only 6, 8 and 8 proteins of the real SAGA complex respectively.

3.3.2.2 Co-annotation, co-localization and co-expression

Since protein complexes are formed to perform a specific cellular function, proteins within the same complex tend to share common functions, be co-localized and co-expressed.

An interaction can have a similarity score which could be semantic similarity of the GO terms of its two interacting proteins or their gene expression correlation. As such, a

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Figure 3.5: (A) The real SAGA complex in the benchmark; (B-E) The SAGA complex predicted by different methods. For each predicted complex, the proteins in red color are involved in the real SAGA complex and those in blue color are not.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Figure 3.6: Co-annotation, co-localization and co-expression scores of predicted complexes from different algorithms.

complex can also have a similarity score which is the average similarity score for all interactions within this complex. Finally, co-annotation, co-localization and co-expression scores are defined as the average similarity scores of all complexes based on “Biological Process” terms, “Cellular Component” terms and gene expression profiles, respectively [Frie 08]. Generally, higher co-annotation, co-localization and co-expression scores show that proteins within the same protein complexes tend to share higher functional similarity, be more co-localized and more co-expressed, and hence they can be used to evaluate the overall quality of predicted protein complexes.

Figure 3.6 shows the co-annotation, co-localization and co-expression scores of complexes predicted by various methods. In terms of these three measures, the complexes predicted by our COACH method are observed to have comparable quality with those predicted by DECAFF, but much better than those predicted by MCL and DPCLus.

3.3.2.3 Statistical evaluation of biological significance of predicted protein complexes

To substantiate the biological significance of our predicted complexes, we calculate their p-values, which represent the probability of co-occurrence of proteins with common functions. Given that proteins in a protein complex are assembled to perform common biological functions, they are thus expected to share common functions. As such, low p-value of a predicted complex generally indicates that the collective occurrence of these proteins in the complex does not happen merely by chance and thus the complex has high

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Table 3.2: The proportions of significant complexes predicted by different methods on DIP data.

Algorithms	MCL	DPClus	DECAFF	COACH
# significant complexes	312	352	1653	622
# predicted complexes	913	1143	2190	746
Proportion (%)	34.2	30.8	75.5	83.4

statistical significance. In our experiments, the p-values (with Bonferroni correction) of complexes are calculated by the tool, SGD’s GO::TermFinder [Boyl 04].

Using DIP data, 622 out of 746 complexes predicted by COACH are considered to be significant, with corrected p-value ≤ 0.01 [Hu 05]. The proportion of significant complexes over all predicted ones can thus be used to evaluate the overall performance of various methods [Mara 07]. Table 3.2 shows the comparison results based on this measure. In Table 3.2, the majority of our predicted complexes (83.4%) are significant and our COACH method also predicts a higher proportion of significant complexes than other three algorithms.

Meanwhile, both MCL and DPClus predict many protein complexes with extremely small size (*e.g.*, with two proteins) and generally predicted complexes with small size tend to have large p-values [Mara 07] (in Table 3.2, we have discarded the predicted complexes with only one protein when calculating their p-values). Therefore, MCL and DPClus only predicted a small proportion of significant complexes. This result is also consistent with the results in Table 3.1 where both MCL and DPClus achieve very low precision scores.

In addition, Table 3.3 shows 10 protein complexes with very low p-values, predicted by our COACH method. The fifth column in Table 3.3 refers to the *NA* scores between our predicted complexes (in the third column) defined in Equation (3.2) and real complexes (in the fourth column). The last column shows the number of proteins in the real complexes correctly covered by our predicted complexes. In this table, proteins in bold form the protein-complex cores and the rest are attachments. Figure 3.7 gives three examples of complexes predicted by our COACH method. The first example in figure 3.7(A) is CCR4-NOT complex [Liu 01] (ID = 8). COACH managed to cover 9 out of 12 proteins in the real complex and also had two new proteins (in blue color). The predicted complex in figure 3.7(B) managed to cover 8 proteins in Oligosaccharyl transferase complex (OST complex) [Schw 05] and had two novel proteins (YBL105C and YPL076W) (ID = 9). The third example in figure 3.7(C) is our predicted HOPS complex, which succeeded to cover all 7 proteins in HOPS [Seal 00] (ID = 10). We find that many of our predicted complexes match well with the known complexes. Due to the incompleteness of the benchmark, our non-matched predicted complexes, especially for those with low p-values, may provide potential candidate complexes for biologists to validate.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Table 3.3: Selected protein complexes predicted by COACH method on DIP data.

ID	P-values	Predicted protein complexes	Real complexes	NA	# common proteins
1	9.85e-33	YBL084C YDL008W YDR118W YFR036W YGL240W YHR166C YKL022C YLR102C YLR127C YNL172W YOR249C	anaphase-promoting complex	0.688	11
2	5.85e-30	YBR154C YDL150W YDR045C YKL144C YKR025W YNL113W YNL151C YNR003C YOR116C YOR207C YOR224C YPR110C YPR190C	DNA-directed RNA polymerase III complex	0.765	13
3	7.0e-25	YCR035C YDL111C YDR280W YGR090W YGR095C YGR158C YGR195W YHR069C YHR081W YNL189W YNL232W YOL021C YOL142W YOR001W YOR076C	exosome (RNase complex)	0.805	13
4	5.00e-24	YBR081C YBR198C YBR253W YCL010C YDR167W YDR176W YDR216W YDR448W YGL112C YGR252W YHR099W YMR236W YPL254W	SAGA complex	0.452	11
5	9.90e-23	YBR081C YBR198C YDR167W YDR176W YDR392W YDR448W YEL009C YER148W YGL112C YGR274C YHR099W YMR236W YOL148C YPL254W YPR086W	SLIK (SAGA-like) complex	0.475	11
6	6.61e-23	YBL093C YBR193C YBR253W YCR081W YDL005C YDL140C YDR308C YER022W YGR104C YHR041C YHR058C YLR071C YNL236W YOL051W YOL135C YOR174W YPR070W	RNA polymerase II mediator complex	0.602	16
7	1.57e-20	YAR003W YBR175W YBR258C YDR469W YHR119W YKL018W YLR015W YPL138C	COMPASS complex	1.0	8
8	1.36e-19	YAL021C YCR093W YDL165W YDR443C YER068W YGR134W YGR274C YIL038C YNL288W YNR052C YPR072W	CCR4-NOT complex	0.614	9
9	1.43e-15	YBL105C YDL232W YEL002C YGL022W YGL226C-A YJL002C YMR149W YOR085W YOR103C YPL076W	OST complex	0.71	8
10	1.66e-13	YAL002W YDL077C ⁴⁵YDR080W YLR148W YLR396C YMR231W YPL045W	HOPS complex	1.0	7

Figure 3.7: Examples of protein complexes predicted by COACH method.

3.3.3 Validation of Core-Attachment Structures

We first analyzed the difference between the cores as consistent functional “hearts” and attachments as “secondary” units by using various biological evidences, such as the GO annotations and gene expression data. Then, we validated a few examples of predicted protein complexes using biological knowledge from literature.

3.3.3.1 Analysis of protein-complex cores

Proteins within the same complex core should have higher degree of functional similarities and tend to co-localize to the same subcellular compartment than those attachments [Dezs 03, Gavi 06]. Two interacting proteins (or an interaction) can have a similarity score based on their GO terms or gene expression profiles. In our experiments, functional similarity between two proteins is calculated based on the method in [Wang 07] and expressional correlation is measured by the Pearson correlation coefficient. The overall quality of all interactions involved in protein-complex cores is an aid to analyze those complex cores. Using DIP data, we managed to identify 746 complex cores, involving 3886 interactions among 1536 proteins. Table 3.4 shows the average similarity of all these interactions, using two sub-ontologies of GO (BP-“Biological Process” and CC-“Cellular Component”) and gene expression data respectively. We also obtained the average similarity of two other sets of interactions, all the interactions in DIP data and those involved in our inferred protein complexes. In Table 3.4, we can find that interactions within protein complexes have higher similarities than those in the whole DIP PPI data, while interactions within complex cores even have higher similarity than those in complexes, which indicates the cores’ biological meanings.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Table 3.4: Average similarity of interactions involved in protein-complex cores, protein complexes and DIP data respectively.

Interactions	Biological Process	Cellular Component	Gene Expression
In COACH complex cores	0.558	0.706	0.274
In COACH complexes	0.502	0.674	0.264
In DIP data	0.357	0.570	0.235

Figures 3.8-3.10 present some examples in which proteins within the protein-complex cores share much higher gene-expression correlations or functional similarities. Proteins in the dashed circles form the cores of those predicted complexes in these figures.

In Figure 3.8, each interaction within the predicted complexes is annotated with the gene-expression correlation between its two interacting proteins. The predicted complex in (A) consists of 6 proteins and matches the DASH complex [Mira 05]. In our predicted DASH complex, proteins YDR016C, YGR113W and YKR037C form the complex core. All 3 interactions in the core have a average gene-expression correlation of 0.78, which is much higher than that of interactions not in the core (0.45). The predicted complexes in (B) and (C) match the trehalose-6-phosphate synthase/phosphatase complex [Bell 98] and RNA polymerase II mediator complex [Balc 95], respectively. The average gene-expression correlations for interactions within and without the cores are 0.66 and 0.33 in (B) and 0.41 and 0.17 in (C), respectively.

Figure 3.8: Predicted complexes with interactions scored by gene-expression correlations.

In Figure 3.9, each interaction is annotated with the similarity of biological processes between its two interacting partners. The predicted complex in (A) matches the CCAAT-binding factor complex [Ganc 98]. In this predicted complex, the core consists of proteins YBL021C, YGL237C and YOR358W, which share exactly the same GO annotations, *e.g.*, all of them are involved in both the transcription (GO:0006350) and regulation of carbohydrate metabolic process (GO:0006109). Two predicted attachments, proteins

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

YDR277C and YOR047C, are involved in glucose transport (GO:0015758) and glucose metabolic process (GO:0006006), respectively. Obviously, the interactions not in the core have a much lower average functional similarity (0.28). The predicted complexes in (B) and (C) match Cbf5-Nop10 complex [Hamm 05] and nucleotide-excision repair factor 1 complex [Guzd 96], respectively. The average similarities of biological processes for interactions within and without the cores are 0.93 and 0.77 in (B) and 0.89 and 0.31 in (C), respectively.

Figure 3.9: Predicted complexes with interactions scored by functional similarities (using BP).

Similarly in Figure 3.10, each interaction is annotated with the similarity of cellular components between its two interacting partners. The predicted complexes in (A-C) match the GPI-anchor transamidase complex [Frae 01], CCAAT-binding factor complex (note that the predicted complex in (B) is the same one in Figure 3.9(A)) and FACT complex [Form 01], respectively. The average similarities of cellular components for interactions within and without the cores are 1.0 and 0.73 in (A), 1.0 and 0.46 in (B) and 0.94 and 0.61 in (C), respectively.

From Table 3.4 and Figures 3.8-3.10, it is clear that interactions within or outside the protein-complex cores have similarity scores at different-levels. This fact supports that proteins in complexes should be categorized into different organization levels, *i.e.*, core member level and attachment level. All above evidences also constitute proofs that our identified protein-complex cores are mostly biological hearts of protein complexes.

3.3.3.2 Validating examples for core-attachment structures

To illustrate the organization of our predicted complexes, we further analyzed the second and third predictions in Figure 3.7.

In the OST complex as shown in Figure 3.7(B), the core consists of 3 proteins (YEL002C, YGL022W and YJL002C). This is also reported in Gavin *et al.*'s analysis

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Figure 3.10: Predicted complexes with interactions scored by functional similarities (using CC).

[Gavi 06]. In addition, our method correctly identified 5 known attachments (YDL232W, YGL226C-A, YMR149W, YOR085W and YOR103C) and predicted two new attachments (YBL105C and YPL076W). Since the protein YML019W in the real OST complex interacts with only one protein (YJL002C) within the core in DIP data, our predicted complex did not identify it as an attachment. However, the interactions between YML019W and all three proteins in the core have been reported in [Gavi 02]. As more PPI data are accumulated, we can expect our COACH can work even better.

As to the HOPS complex in Figure 3.7(C), proteins YDR080W and YDL077C are its attachments and the core consists of proteins YLR148W, YLR396C, YMR231W, YPL045W and YAL002W. Biological experiments show the core have the function of vacuole protein sorting [Ried 97]. With the help of attachments YDR080W and YDL077C, this complex can perform the function of homotypic vacuole fusion [Seal 00]. This demonstrates that our computational discovery is consistent with the current biological knowledge, indicating that some novel knowledge could be discovered by our proposed method. Of course, biological experiments are necessary for further validating.

3.3.4 Effect of the Parameter t

Recall that COACH method employs a user-defined parameter t (see Algorithm 3 in section 3.2) to filter redundant preliminary cores. It is obvious that overlaps among protein-complex cores are allowed when $t > 0$ and are not allowed when $t = 0$. We now investigate how the variation of t affects the performance of our COACH method. Figure 3.11 shows the F-measure and coverage rate of our COACH method under different values of t , using DIP data.

Firstly, the number of predicted complexes increases with the increase of t , *i.e.*, COACH generates 268, 715 and 1040 complexes under $t = 0, 0.2$ and 0.4 respectively.

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

Figure 3.11: The effect of t on F-measure and Coverage rate.

This is because the bigger the value of t , the more overlaps among protein-complex cores are allowed, resulting in more predicted complexes. With more complexes predicted, it is reasonable that more proteins in the benchmark complexes are covered with increasing the value of t . However, the number of cases that multiple predicted complexes match the same real complex (also denoted as the redundancy in predicted complexes) is also increased with the increase of t . For example, 267 correct predictions match 247 real complexes when $t = 0.2$ (1.08 predicted complexes match one real complex on average), while 371 match only 253 when $t = 0.4$ (1.47 predicted complexes match one real complex on average).

Secondly, as we increase the values of t , the curve of the resulting F-measure in Figure 3.11 is observed to have 3 distinct and stable ranges for values of t , *i.e.*, $[0, 0.1]$, $[0.125, 0.225]$ and $[0.25, 0.4]$ respectively. Higher F-measure is achieved in the first range ($t \in [0, 0.1]$), which sacrifices the Coverage Rate, resulting the lowest Coverage Rate compared with those in other two ranges. Additionally in this range, COACH only generates a small number of complexes (*e.g.*, it predicts 308 complexes, 145 out of which match 211 real complexes when $t = 0.05$). On the contrary, higher Coverage Rates are achieved in the third range ($t \in [0.25, 0.4]$), which sacrifices the F-measure. However, as stated above, the redundancy in our predicted complexes would become more severe with the increase of t . To reduce the redundancy involved in predicted complexes, as well as to encourage one-to-one matching between predicted complexes and real ones, we recommend that the suitable setting of t would be in the second range *i.e.*, $t \in [0.125, 0.225]$. In fact, the performance of COACH does not change significantly

Figure 3.12: F-measure and Coverage Rate of various methods for Krogan’s data.

in this range. For the setting of $t = 0.225$ in our experiments, COACH can achieve a good balance of both F-measure and Coverage Rate.

3.3.5 Results Using Krogan *et al.*’s PPI data

We also performed our COACH on Krogan *et al.*’s PPI data (see Methods section). The F-measure and coverage rate of each method using this data ($t = 0.225$) are shown in Figure 3.12. The F-measure of our COACH is 44.2%, which is 18.7%, 10.3% and 4.6% higher than MCL, DPCLus and DECAFF respectively. From the perspective of Coverage Rate, our COACH still performs the best as shown in Figure 3.12.

In particular, we check both N_{cp} and N_{cb} (see Method section) of DECAFF and find that DECAFF generates protein complexes with some redundancy on both DIP and Krogan *et al.*’s PPI data. In other words, the complexes predicted by DECAFF overlap a lot with each other. For example, DECAFF predicts 2190 complexes using DIP data, 605 of which match 243 real complexes. Likewise, it predicts 2143 complexes using Krogan *et al.*’s data, 759 of which only match 192 real complexes (almost 4 correct predictions match a real one on average). The ratio N_{cp}/N_{cb} of DECAFF is much higher than those of other methods, which suggests that some post-processing (such as, clustering highly overlapping complexes) is needed for DECAFF.

3.3.6 Performance Evaluation by the CYC2008 Benchmark

Previously, we combined three sources and collected a benchmark with 428 protein complexes. However, this benchmark contains some overlapping protein complexes. Now,

Table 3.5: Performance of various methods evaluated by the CYC2008 benchmark.

Methods	MCODE	MCL	DPCLus	DECAFF	COACH
Precision	0.571	0.209	0.227	0.289	0.417
Recall	0.240	0.586	0.642	0.525	0.527
F-measure	0.338	0.308	0.336	0.372	0.466
Coverage	0.403	0.541	0.511	0.454	0.545

the CYC2008 catalogue provides us a new benchmark of complexes for evaluating existing prediction methods, consisting 408 protein complexes. It can be downloaded from Wodaklab [Pu 09]. We also evaluate the performance of various methods using this catalogue.

Table 3.5 shows the performance of various methods. We can observe that these results are similar to those in Figure 3.4. Our COACH achieves the highest F-measure and coverage rate. Although MCODE has the highest precision and DPCLus has the highest recall, our COACH has both balanced high precision and recall and thus obtains the best overall performance with respect to F-measure.

3.3.7 Comparison with the CoreMethod

Recently, Leung *et al.* [Leun 09] also developed an approach called CoreMethod to detect protein complexes in PPI networks by identifying their cores and attachments separately. Next, we briefly introduce the CoreMethod and then compare it with our COACH method.

In the CoreMethod, the probability for two proteins to be in the same protein-complex core (denoted as the p-value) is mainly determined by two factors: whether these two proteins interact or not and the number of the common neighbors between them. The CoreMethod then calculates the p-values for all pairs of proteins (*i.e.*, pair-wise fashion) to identify protein-complex cores. However in our COACH method, preliminary cores are first detected by locally searching highly interactive proteins in each vertex's neighborhood graph. Protein-complex cores are then obtained by filtering out redundant preliminary cores (please see the Algorithm section). As such, both approaches for detecting protein-complex cores are quite different in computational essence. To detect cores, our COACH performs local search within vertex's neighborhood graphs while the CoreMethod computes the p-values between all the proteins in the whole PPI networks. Therefore, COACH is able to achieve significant computational speedup against the CoreMethod.

In addition, we notice that there is another difference between the two methods. Protein-complex cores detected by the CoreMethod are non-overlapping whereas our COACH method is able to detect the overlapping cores. Biologically, certain proteins

Table 3.6: Comparative analysis between CoreMethod and COACH.

Performance	DIP data		Krogan <i>et al.</i> 's data	
	CoreMethod	COACH	CoreMethod	COACH
Precision	0.128	0.382	0.163	0.432
Recall	0.598	0.582	0.535	0.453
F-measure	0.211	0.461	0.250	0.442
Coverage Rate	0.356	0.349	0.334	0.328

can serve as coordinators and thus can be involved in multiple cores. In fact in Gavin *et al.*'s Nature paper [Gavi 06], it has been shown that the different complex cores have some overlaps, demonstrating that our COACH method has matched well with the biological evidences than the CoreMethod.

Interestingly, both the CoreMethod and COACH select attachments to form protein complexes, based on the same strategy. That is, a protein will be included as an attachment to a core if more than half of proteins within the core are its interacting partners. Furthermore, we employed various evaluation measures, such as F-measure and Coverage rate to compare the overall quality of complexes predicted by these two methods. Table 3.6 shows the comparison results between the CoreMethod and our COACH method, in terms of precision, recall, F-measure and Coverage rate. On DIP data, the precision of COACH is 38.2%, which is almost as 3 times as that of the CoreMethod. The recall of the CoreMethod (59.8%) is slightly better than that of COACH (58.2%). As such, the resulting F-measure of COACH is significantly (25%) higher than that of the CoreMethod. Similarly on Krogan *et al.*'s data, the precision and F-measure are significantly better than those of the CoreMethod (the overall F-measure of COACH is 19.2% higher than CoreMethod), while the recall of the CoreMethod is higher than that of COACH. The Coverage rates of the two methods are comparable on these two PPI networks - CoreMethod is slightly (around 0.7%) higher than that of COACH method. However, the overall performance of COACH is much better than CoreMethod.

3.3.8 Running Time Analysis

In this section, we ran our COACH method on two kinds of random graphs (*i.e.*, Power-law random graphs and Geometric random graphs). In power-law graphs, the degree distribution follows a power-law, *i.e.*, $P(k) \sim k^{-\alpha}$, where $P(k)$ is the probability of a nodes with a degree of k and $\alpha > 0$. Geometric random graphs are constructed by dropping n points randomly into the unit square and adding edges between two nodes with distance less than a pre-defined threshold.

In our experiments, after a power-law graph is randomized, a geometric random graph will be generated with the same number of nodes and edges as the power-law graph. Given a fixed number of nodes, we generated 50 pairs of power-law graphs and

CHAPTER 3. COACH: DETECTING PROTEIN COMPLEXES FROM PPI NETWORKS

geometric random graphs and then calculated the average running time of COACH over them. The basic information of these random graphs, *i.e.*, the average number of edges, is shown in Figure 3.13. We also note that the average degrees ($2 \times \#Edges/\#Nodes$) of the random graphs slightly increase as the number of nodes increases.

Figure 3.13: The average number of edges in the random graphs in our experiments.

We ran our COACH method over two kinds of random graphs (*i.e.*, Power-law random graphs and Geometric random graphs) on the workstation with 3.4GHz Dual Core processors and 3GB RAM. Figure 3.14 thus demonstrates that our COACH method is efficient in large-scale graphs. For example, it takes about 40 seconds for COACH to run on a random network with 10,000 nodes and about 40,000 edges. Considering that there are only about 6,000 genes/proteins in yeast genome, it is really fast for COACH to run on yeast interactome.

3.4 Conclusions

Protein complexes are key molecular entities to perform cellular functions. The increasing amount of protein-protein interaction (PPI) data has enabled us to detect protein complexes from PPI networks. However, current computational methods only focus on detecting dense subgraphs in PPI networks as protein complexes but ignore their inherent organization. Hence, new approaches that can provide insights into the organization of protein complexes are greatly desired.

In this chapter, we proposed a core-attachment based method to detect protein complexes from PPI networks. We first mined the protein-complex cores from the neighborhood graphs and then formed protein complexes by including attachments into cores.

Figure 3.14: The running time of our COACH on the random graphs.

The evaluation and analysis of our predictions demonstrated the following advantages of our COACH method over the state-of-the-art techniques. First, Our proposed method is fundamentally different from existing methods. It provides insights into the inherent organization of protein complexes while existing methods mainly focus on detecting dense graphs. Proteins within the same protein-complex core detected by our method have high functional similarity and tend to be co-localized, indicating that protein cores are possible biological hearts of protein complexes. Second, COACH has achieved significantly higher F-measure and Coverage rate than existing methods. Thus, our predicted complexes match very well with benchmark complexes. In addition, COACH also performs very well in terms of other measures such as co-annotation, co-localization and p-values, indicating that COACH can predict protein complexes very accurately. Our identified complexes, therefore, could be probably the true complexes to help the biologists to get novel biological insights.

Chapter 4

Integrating Diverse Sources For Reliable Protein-Protein Interactions

4.1 Introduction

4.1.1 Background

Protein-protein interactions (PPIs) are at the core of the biological machinery of any living cells. PPIs are involved in almost every level of cellular functions, playing key roles in the transport machinery across the various biological membranes, the packaging of chromatin, the network of sub-membrane filaments, the regulatory mechanism of gene expression, and so on [Dona 03], while abnormal PPIs are key to various diseases. Research in PPIs will therefore allow the biologists to understand biological functions, uncover the underlying disease pathways, and provide the basis for new therapeutic approaches to benefit human beings [Li 09].

In recent years, many high-throughput experimental techniques, most notably yeast-two-hybrid and affinity purification, have been developed to enable comprehensive detection of PPIs. Numerous publicly available databases, such as DIP [Salw 04], BIND [Bade 03b] and BioGrid [Star 06], have been set up for researchers to access PPI data for biological and computational analysis.

However, the quality of current detected PPI data is far from satisfactory. The experimental conditions in which the detection methods are carried out may cause a bias towards detecting interactions that do not occur under physiological conditions. In other words, the experimental data may contain *false positive* PPIs that do not occur in the cell. At the same time, the high-throughput methods can also be biased against soluble or membrane proteins and fail to detect certain types of interactions such as weak

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

transient interactions and interactions that require post-translational modification. This results in *false negative* detection and low experimental coverage of the interactomes [Meri 02, Tara 08].

How serious is the PPI data quality situation - specifically, what are the false positive and false negative rates in current PPI data? Many researchers have attempted to answer this question [Bade 02, Bade 04, Hart 06, Gent 07, Meri 02]. By comparing the overlap between the PPIs collected from different large-scale biological experiments, or checking the consistency of the cellular function or location information between the interaction partners, it has been found that the quality of the current PPI data is alarmingly low, with accuracy ranging from only 10% to 50% [Gent 07, Meri 02], and coverage lower than 50%. This is the case even for the most studied and curated interactome of *Saccharomyces cerevisiae* (yeast) [Hart 06]. One can expect the accuracy and coverage of the PPIs to be even lower for many other species.

Given the experimental limitations in high throughput PPI screenings, researchers have recently begun to exploit the growing availability of various additional biological and computational resources to address the issues on *false positives* and *false negatives* computationally. The computational methods can be categorized into two classes. The first line of bioinformatics research focuses on exploiting the topological properties of the PPI networks to assess the reliability of PPI. For example, Interaction Generality (IG1) [Sait 02] was proposed to detect false positives created by “sticky” proteins. IG2 [Sait 03] was subsequently proposed to measure the reliability of PPI using five pre-defined network motifs to dissect the local network topology. IRAP, or Interaction Reliability by Alternative Path [Chen 06], made use of the assumption that an interaction associated with an alternative path of reliable interactions is likely to be reliable, while PathRatio [Pei 05] measured the reliability of an interaction by considering all the alternative paths connecting these two target proteins. Several methods [Gold 03, Chua 06, Liu 08] were also proposed to quantify the propensity of two given proteins to be interacting partners based on the number of their common neighbors.

Another direction of the research exploits the increasingly enriched genomic features to identify true positives and predict false negatives. Deane *et al.* [Dean 02] selected proteins pairs with similar gene expression profiles as true positives. Jansen *et al.* [Jans 03] employed a Bayesian network approach to combine genomic features such as gene expression profiles, co-essentiality, and co-localization, to predict novel interactions that were false negatives in the PPI networks. Kernel methods have also been used to detect false negatives [Ben 05, Ben 06]. Patil *et al.* [Pati 05] also used multiple genomic features, including sequence similarity, functional annotations and 3D structures, to filter out false positives.

4.1.2 Motivation

In this chapter, we aim to exploit the use of both topological properties of the PPI networks and the genomic features of proteins to identify reliable detected PPIs and predict putative interactions. As such, we can manually control the quality of PPI data and address the issue that our COACH (please refer to Chapter 3) does not take the reliability of interactions into consideration and works only on unweighed graphs.

We propose a method to integrate the diverse biological and computational resources. Our method consists of two steps: interaction binning (InterBIN) and interaction classification (InterCLASS). Since the diverse biological and computational evidences were individually scored in a different nature, we first group the protein pairs for each class of evidence into bins so that we can more accurately estimate the likelihood (Bin-Confidence score) that the protein pairs interact. Then, we employ machine learning methods such as support vector machines and Bayesian classifier to classify whether a detected or putative PPI is reliable or not using the Bin-Confidence scores as features. We conduct experiments to show that our proposed method is significantly better than the existing methods, and that integrating the various biological and computational sources of data and evidence can provide better results. We also show that more reliable PPI data collected will of course generate protein complexes with higher quality.

4.2 Method

In this section, we present our two-step method for integrating diverse sources of biological and computational evidences for reliable PPI detection. First, we introduce each of individual biological and computational evidences which are used to assign original raw scores for protein pairs to indicate their propensity to interact. Then, we describe our InterBIN method to group protein pairs based on their scores from each evidence, and then define the more accurate Bin-Confidence scores for each protein pair. Finally, we describe our InterCLASS method that exploits the integration of the evidences re-weighted using Bin-Confidence scores and infers reliable PPIs using machine learning methods such as Support Vector Machines (SVM) and Bayesian classifier (BC).

4.2.1 Individual Biological and Computational Evidences

In this study, given a protein pair (x, y) , we use each of the following biological and computational evidences to evaluate their propensity to interact. Each evidence f assigns a score $S_f(x, y)$ to (x, y) independently based on the nature by which the evidence was derived.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

- (i) *Topological properties.* As we have discussed in the previous section, there are a number of topology-based methods to assess the reliability of protein interactions. PathRatio [Pei 05] and IRAP [Chen 06] assume that the protein interactions with more alternative paths are more reliable. The main difference between them is that PathRatio considers all the alternative paths while IRAP [Chen 06] only considers the strongest non-reducible alternative paths. The FS-weight [Chua 06], on the other hand, considers the number of common neighbors of the protein pair (x, y) — if x and y interact with some common proteins (e.g, z), then (x, y) is assumed to share some physical or biochemical characteristics that allow them to bind to these common proteins. The more common proteins they interact with, the higher chance they interact.

In this chapter, we utilized all the above three topological measures and compared their effectiveness to infer reliable protein interactions. In particular, for FS-weight, we used the following simplified variant [Chua 06]:

$$S_{FS}(x, y) = \frac{2 \times |N_x \cap N_y|}{|N_x - N_y| + 2 \times |N_x \cap N_y| + 1} \quad (4.1)$$

$$\times \frac{2 \times |N_x \cap N_y|}{|N_y - N_x| + 2 \times |N_y \cap N_x| + 1}, \quad (4.2)$$

where N_x and N_y are the set of direct neighbors of the proteins x and y respectively.

- (ii) *Gene expression profiles.* The correlation between the similarity of expression patterns for a pair of genes and interaction of the proteins they encode has been demonstrated for various species [Grig 01]. Statistical results have shown that protein pairs encoded by co-expressed genes interact with each other more frequently than with random proteins. This has allowed biologists to exploit large-scale gene expression to assess the reliability of protein interaction screens.

Given a protein pair (x, y) , the proteins' propensity to interact can be measured by using the *Pearson Correlation Coefficient* between their encoded genes' expression profiles G_x and G_y as follows,

$$S_{GE}(G_x, G_y) = \frac{|\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})|}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4.3)$$

where n is the number of time points for the expression profiles and x_i is the i^{th} expression value of protein x 's expression profiles G_x , whose average expression value is $\bar{x}$.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

- (iii) *Protein domains and their interactions.* Protein domains are evolutionarily conserved modules of amino acid sub-sequences that are postulated as nature’s functional “building blocks” for constructing the vast array of different proteins. Protein functional domains are thus regarded as essential units for such biological functions as the participation in transcriptional activities and other intermolecular interactions. The existence of certain domains in the proteins orchestrates the propensity for the proteins to interact due to the underlying domain-domain interactions. Databases, such as the Protein families (Pfam) database and others, have compiled comprehensive information about domains [Li 08].

In this study, we use domain-domain interactions to evaluate the propensity of a protein pair to interact [Li 06b]. The propensity score between a protein pair (x, y) can be computed by the number of interacting domains that they contain:

$S_{PF}(x, y) = |\{(d_i, d_j) | d_i \in D_x, d_j \in D_y, d_i \text{ interacts with } d_j\}|$, D_x and D_y are the set of Pfam domains of proteins x and y respectively.

- (iv) *Protein sequences and their similarity.* One of the most fundamental and successful tools in the field of bioinformatics is sequence alignment. By aligning protein sequences with each another, we can evaluate how similar the protein sequences are. In most cases, close homologs ($> 30\%$ sequence identity) physically interact with each other [Zaki 08] [Chua 07b]. In particular, many of paralogs (close homologs within the same species) are known to interact with each other in heterodimer complexes. There are many well-known examples such as the spliceosomes and many transcription factors [Ispo 05]. Given two proteins x and y , the bit-scores from the BLAST results are used to show the sequence similarity between them, denoted as $S_{SS}(x, y)$.

In summary, the various biological and computational evidences mentioned above are widely used for assessing the reliability of protein interactions. Another commonly-used biological evidence is functional similarity — proteins that have the same molecular functions are more likely to interact since they are involved in the same biological processes. In this study, we intentionally exclude the use of functional information to predict reliable protein interactions — we keep it to evaluate our experimental results eventually.

4.2.2 InterBIN: Grouping Interactions by Their Original Scores

The original raw scores obtained from each evidence, (*e.g.*, FS-weight scores) have been directly exploited to identify reliable interactions previously [Pei 05, Chen 06]. However, since each of the evidences does not necessarily cover all interacting protein pairs as well

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

as useful biological knowledge, it is necessary to integrate all these information to make our PPI prediction more robust.

Integrating original raw scores from various evidences is not as straightforward as one may think. Each piece of individual evidence actually provides information of a rather different nature; the original raw scores assigned by each method can differ significantly in scale, range, and distribution. For example, out of the 17262 interactions in DIP data, 9265 interactions (*i.e.*, 53.7%) are with FS-weight scores less than 0.1 while 14099 interactions (*i.e.*, 81.7%) are with IRAP scores at least 0.9. It is clearly not a good idea to integrate various evidences by directly using their original raw scores.

To address this issue, we re-weight the evidence more robustly by estimating the likelihood that a protein pair (x, y) interacts given the observation that a evidence f assigns a score $S_f(x, y)$ to (x, y) . To do this, we can examine all protein pairs that are assigned with the same score $S_f(x, y)$ in the training data and compute the fraction of those protein pairs that interact or occur in positive training set P (how to construct positive and negative training data P and N will be described in next subsection). However, the scores assigned by each evidence f may not be discrete, and there may not be enough protein pairs that are assigned the same score. Thus, it is necessary that we first group the protein pairs with similar scores together.

Let us now introduce how to group interactions and calculate the Bin-Confidences for them using our method InterBIN. For a evidence f , let $S_f(x, y)$ be the original raw score of the protein pair (x, y) . Assuming that there are n protein pairs, they can be divided into different Bins based on their original raw scores in the following simple steps.

- (i) Sort all protein pairs based on their original raw scores.
- (ii) For the sorted protein pairs, $(x_1, y_1), \dots, (x_n, y_n)$, the first μ protein pairs (*i.e.*, $(x_1, y_1), \dots, (x_\mu, y_\mu)$) are inserted into the first group G_1^f . All the protein pairs that have the same score as the protein pair (x_μ, y_μ) will be also inserted into G_1^f . Let the size of G_1^f be $(\mu + k)$.
- (iii) Start from the protein pair $(x_{\mu+k+1}, y_{\mu+k+1})$ and repeat Step 2 until m groups are obtained finally. If the last group G_m^f has the size smaller than μ , it is merged into the group G_{m-1}^f . (Here, μ is an integer to represent the group size. To better show our experimental results, we have used the relative group size, μ/n , to replace μ in our experiment section, where n is the total number of protein pairs in the training set)

Using the above method, we divide all the protein interactions into different groups such that the interactions within the same group have similar original raw scores and

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

each group also has similar size. We can then estimate the confidence score for each group by the proportion of positive examples within it.

Basically, the more positive protein interactions a group G_k^f has, the more reliable a protein interaction within the group G_k^f is. For a protein pair (x, y) in G_k^f , we re-weight it with a score, denoted as $\text{Bin-Confidence}_f(x, y)$ in equation 4.4, which is the confidence score of G_k^f .

$$\text{Bin-Confidence}_f(x, y) = \frac{|\{(x_i, y_i) | (x_i, y_i) \in P, (x_i, y_i) \in G_k^f\}|}{|G_k^f|}, \quad (4.4)$$

where P is the positive training set.

4.2.3 InterCLASS: Integrating Evidences by Machine Learning Techniques

For all the protein pairs in a detected PPI network, InterBIN basically transforms their original scores from various evidences into standard Bin-Confidence scores. Next, we show how we incorporate the Bin-Confidence scores to build classifiers to identify reliable interactions.

In order to use machine learning to build an accurate PPI classifier, first, we need to compile a positive and a negative training set P and N . While experimentally derived protein interaction data are publicly available in a number of databases such as DIP [Salw 04] and BioGrid [Star 06] etc, we need to be careful in the selection of positive and negative examples due to their noisy nature (as discussed in Introduction previously). In this study, we select a protein pair as a positive example only if it was detected by small-scale experiments (not high throughput screenings) or by multiple (at least 3) wet-lab techniques. As for constructing a negative training set, we adopt the method used in [Pati 05] by considering those (potentially wrongly detected) protein pairs in the PPI databases whose two interacting partners are from different cellular locations. This is because proteins located in the different cellular components are unlikely to interact physically. We then consider all the other protein pairs in the PPI dataset (that do not belong to P and N) to be with an unknown status — we will decide if they interact using our machine learning-based classifiers discussed below.

Given the positive and negative training sets P and N , for each protein pair (x, y) in P and N , we gather the original score $S_f(x, y)$ for each evidence f . Next, we compute the $\text{Bin-Confidence}_f(x, y)$ for all the protein pairs (x, y) with respect to f . With them, we are now ready to build the final classifiers that integrate the diverse biological and computational evidence sources. In this study, we apply two different types of classifiers, namely, Support Vector Machines and Bayesian classifier, that are popular classifiers in many bioinformatics applications.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

4.2.3.1 Support vector machines

The SVM [Vapn 95] is a state-of-the-art classification technique in machine learning and it has been proven to be one of the best classifiers in many application domains such as text categorization, image recognition, protein remote homology detection, and so on. Assume that we have a training data $\{(x_i, y_i)\}_{i=1}^m$ and a hyperplane $y = \mathbf{w} \cdot \mathbf{x} - b$. $\frac{2}{\|\mathbf{w}\|}$ is thus the margin of this hyperplane. The basic principle of SVM is to solve the convex optimization problem that finds a hyperplane with the maximum margin as shown in equation (4.5). This hyperplane with the maximum margin can thus be used to classify new unseen samples as positive or negative class.

$$\min \quad \frac{\|\mathbf{w}\|^2}{2} \quad (4.5)$$

$$s.t. \quad y_i \mathbf{w} \cdot \mathbf{x}_i \geq 1 \quad i = 1, \dots, m \quad (4.6)$$

For each training example in P and N , we can build a feature vector $(f_1, f_2, \dots, f_\lambda)$ where $f_i (i = 1, 2, \dots, \lambda)$ is a score (either the original raw score, or the re-weighted Bin-Confidence computed by InterBIN) from the individual evidences and λ is the total number of biological evidences. In this way, SVM is able to train a classifier [Joac 98] exploiting all the evidences together to classify those detected protein pairs with unknown status.

4.2.3.2 Bayesian classifier

Given a protein pair (x, y) and its λ features $f_1, \dots, f_\lambda$, the probability that it is a negative interaction is denoted as $P\{false|f_1, \dots, f_\lambda\}$. We can assume that the λ features of a protein pair are conditionally independent and then apply Bayesian rule to calculate this probability. In this way, the probability of (x, y) to be positive, $P\{true|f_1, \dots, f_\lambda\}$, can be calculated as follows:

$$\begin{aligned} P\{true|f_1, \dots, f_\lambda\} &= 1 - P\{false|f_1, \dots, f_\lambda\} \\ &= 1 - \prod_{i=1}^{\lambda} (1 - P\{true|f_i\}), \end{aligned} \quad (4.7)$$

where the probability of (x, y) to be positive with respect to the feature f_i , $P\{true|f_i\}$, can be estimated by using either its original raw score or its Bin-Confidence score for f_i .

Basically, for a given protein pair, the Bayesian method will give it a high reliability score if there are certain evidences that show that it is a positive interaction.

4.3 Experimental Data and Evaluation Metrics

We will first introduce the experimental data used in our work. We will also introduce evaluation measures for the accuracy of various prediction methods. The validation for the identified false negatives and false positives using Gene Ontology terms will also be presented.

4.3.1 Experimental Data

Two publicly available yeast PPI data, DIP [Salw 04] and BioGrid [Star 06], were downloaded and used in our experiments. After removing the self-interactions, DIP consists of 17262 interactions and BioGrid consists of 71020 interactions. To train classifiers, we selected the positive training examples as follows. First, since all the 4314 interactions in DIP-core [Salw 04] are detected by small-scale experiments or by multiple wet-lab techniques, they are regarded as positive examples for the DIP data. Second, 13424 interactions in BioGrid which (1) are detected by more than 3 experiments, or (2) are in DIP-core, or (3) are high-quality interactions in Krogan *et al.*'s purification data [Krog 06] or Gavin *et al.*'s data [Gavi 06], are selected as positives. We selected the negative training examples as described previously — known interactions whose two interacting partners are observed with different cellular locations are considered as negatives [Pati 05]. Specifically, given two interacting proteins, the interaction between them is regarded to be negative if the semantic similarity of their GO terms (Cellular Component) is below a certain threshold. In our experiments, semantic similarity between GO terms are calculated by the method in [Wang 07] and the threshold is set as 0.4 [Ben 06]. We have selected 3892 negative examples for DIP data and 11557 negatives for BioGrid data respectively.

As to the genomic sources, interacting domains were downloaded from the following 3 databases, 3DID [Stein 05], iPFam [Finn 05] and DOMINE [Ragh 08]. Gene-expression data were downloaded from [Genea], while Gene-Ontology data were downloaded from [Geneb], and sequence data were downloaded from UniProt (the Universal Protein Resource, <http://www.uniprot.org/>).

4.3.2 Receiver Operating Characteristics

The Receiver Operating Characteristics (ROC) curve is a graphical plot of the sensitivity vs. 1-specificity for a classifier as the decision threshold varies. The best possible classifier would yield a point in the upper left corner or coordinate (0,1) in ROC curve, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). A completely random guess would give a point along a diagonal line from the left bottom to the top right corners. The Area Under the ROC Curve (AUC) is a widely-used

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

measure of the accuracy of a specific classifier. Given a classifier, it results in a perfect classification if its AUC is 1 and it works just like random guess if its AUC is 0.5.

In this chapter, in addition to our integrative approach (using SVM and BC), each original method for identifying reliable interactions can also be regarded as a classifier. For example, IRAP [Chen 06] can be used as a classifier — both the original IRAP score and the Bin-Confidence score for each protein pair can be considered as a decision value by the IRAP classifier. In other words, each original method for the various evidences can also have an AUC to show its accuracy.

4.3.3 Validation of False Positives/Negatives

Since a protein performs a certain cellular function by interacting with other proteins, the cellular functions of interacting proteins are likely to be similar. Functional similarity between proteins based on the GO terms are often exploited as computational validation of protein-protein interactions [Pei 05, Chen 06]. Assuming that each protein pair can be computed with a score which is the functional similarity between its two interacting partners, protein pairs with top decision values under a better classifier are expected to have higher average similarity scores.

Given two proteins x and y annotated by $\{g_{11}, \dots, g_{1m}\}$ and $\{g_{21}, \dots, g_{2n}\}$ respectively, their functional similarity defined in [Wang 07], $S_{GO}(x, y)$, is calculated as follows.

$$S_{GO}(x, y) = \frac{\sum_{1 \leq i \leq m} \max_{1 \leq j \leq n} \text{sim}(g_{1i}, g_{2j}) + \sum_{1 \leq j \leq n} \max_{1 \leq i \leq m} \text{sim}(g_{1i}, g_{2j})}{m + n} \quad (4.8)$$

where $\text{sim}(g_{1i}, g_{2j})$ is the semantic similarity between GO terms g_{1i} and g_{2j} defined in [Wang 07].

4.4 Experimental Results

We have performed comprehensive experiments to evaluate the proposed techniques. Since our method used Bin-Confidence scores instead of the original raw scores, in this section, we first show the benefits of using Bin-Confidence scores. Then, we show that our integrating method is very effective in identifying false negatives and false positives than existing methods.

4.4.1 Comparison between Bin-Confidence Scores and Original Scores

Linear SVM as well as SVM with polynomial kernels and Gaussian kernels (using SVM^{light} software [Joac 98]) have been used in our experiments. Since they have comparable

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Table 4.1: AUC of each method using original raw scores and Bin-Confidences, respectively.

Methods/Data	DIP data		BioGrid data	
	Original Scores	Bin-Confidence	Original Scores	Bin-Confidence
FS-weight	0.741	0.768	0.741	0.745
PathRatio	0.702	0.710	-	-
IRAP	0.686	0.723	0.580	0.594
Gene-expression	0.549	0.560	0.580	0.566
Interacting Domains	0.547	0.544	0.561	0.562
Sequence Similarity	0.576	0.569	0.523	0.529
InterSVM	0.776	0.804	0.749	0.768
InterBC	0.710	0.787	0.600	0.750

performance, the simplest (*i.e.*, linear SVM) is used in this chapter. For easy reference, let us call our integrative method using SVM and Bayesian Classifiers InterSVM and InterBC respectively.

After collecting the positive and negative training examples, we calculated the Area Under ROC Curve (AUC) for each method using both the original raw scores and Bin-Confidence score. The AUC for each method was obtained by performing a 5-fold cross validation. To group protein pairs and calculate their Bin-Confidence score, μ was set as 1% on DIP data and 0.5% on BioGrid data. For example, $\mu = 1\%$ means that each group has at least $n \times 1\%$ protein pairs and n is the total number of protein pairs in the training set. We will explain how such a parameter setting came about in the coming subsection.

Table 4.1 shows the AUC of each method using original raw scores and Bin-Confidences, respectively (the running time for PathRatio to perform on comprehensive BioGrid data was prohibitively high. Thus, we were unable to obtain the results of PathRatio on BioGrid data). From the results on the two different PPI datasets, we can draw the following two conclusions. First, integrating diverse individual evidences can improve the prediction accuracy. For example, FS-weight is the best-performed individual feature on DIP data, whose AUC is 0.768, while InterSVM and InterBC with Bin-Confidences have an AUC of 0.804 and 0.787, which are 3.6% and 1.9% higher than that of FS-weight respectively. Second, while grouping protein interactions and calculating Bin-Confidence scores may not improve each original individual method very much, it improves the accuracy of InterSVM and InterBC. As shown in Table 4.1, the AUC of InterBC with Bin-Confidence scores is 0.787 on DIP data and 0.750 on BioGrid data, which is 7.7% and 15.0% higher than that of InterBC using the original raw scores respectively, indicating that our InterBIN grouping technique makes our integrative approach more robust.

It is interesting to note that the genomic evidences appeared to have lower performance than the topological evidences in Table 4.1. One reason is that the original raw

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Figure 4.1: AUC for each method as the parameter varies on DIP data.

scores from the genomic evidences are too limited in coverage. For example, in DIP data, only 5663 out of 17262 interactions, *i.e.*, 32.8%, have non-zero original scores from the protein domain information, while majority of the interactions (16036 or 92.9%) have non-zero IRAP scores.

4.4.2 Effect of the Parameter μ

The parameter μ , which was used to determine the group size, has a direct effect on the Bin-Confidences and on the accuracy for methods using Bin-Confidences. Figure 4.1 and 4.2 show the AUC for each method on DIP data and BioGrid data respectively as μ varies.

A larger μ generally means a smaller number of groups. Therefore, there will be more interactions in each group having the same Bin-Confidence. As such, the resulting Bin-Confidence assignments can become too general as the large number of interactions within the same group cannot be distinguished. On the other hand, using a smaller μ risks getting less accurate Bin-Confidences due to the small number of samples in each group. Figure 4.1 and 4.2 confirm that group sizes that are either too large or too small will result in poor accuracy for each method.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Figure 4.2: AUC for each method as the parameter varies on BioGrid data.

With the DIP dataset, the individual evidences as well as InterBC achieved their decent AUC when μ was around 1%, while InterSVM achieved its highest AUC when μ was set as 0.05%. However, since there are only 8206 positive and negative examples in the DIP dataset, each group would contain only around 4 interactions if μ were to be set as 0.05%, resulting in inaccurate Bin-Confidences for interactions and the over-fitting of the SVM. As such, we recommend that μ is set as 1% for DIP data, and μ is set as 0.5% for BioGrid data.

4.4.3 False Positive/Negative Detection

Note that while protein-protein interactions have been detected and stored in the protein interaction databases such as DIP [Salw 04] and BioGrid [Star 06] etc, many of them may not be reliable (*i.e.*, with an unknown status that is neither in P nor in N). We can apply the classifiers learnt from the training data to classify those with unknown status to address the false positive detection issue.

Meanwhile, although the current PPI data are becoming more and more comprehensive, it is still far from complete due to inherent experimental limitations and biases. A recent paper [Huan 09] concluded that false negative is a serious issue in experimental detection of protein-protein interactions — the false negative rate of yeast two-hybrid

Figure 4.3: The average functional similarity of top-ranked interactions in DIP data.

experiments (which is currently the most widely used technique that detects *in vivo* pairwise interactions) to be around 50%. Given the generalization capability of the classifiers learnt from the training sets, they can also be used to detect false negative interactions by predicting putative interaction between undetected pairs of proteins. In this chapter, to avoid classifying all the possible undetected protein pairs for false negative elimination, we generate potential false negative candidates using those protein pairs with at least 2 common neighbors in the DIP database [Kell 05].

We applied our integrative methods which is learned from existing reliable interactions and non-interactions, to predict false negatives and false positives. The quality of the predicted false negatives and false positives is evaluated and validated by the corresponding functional similarity scores between the protein partners. Figures 4.3, 4.4 and 4.5 show the average similarity scores for top-ranked interactions (or predicted interactions). A point (x, y) in these figures means that top- x interactions (or predicted interactions) have an average similarity y .

4.4.3.1 False positive detection.

For those interactions in the PPI datasets that we have labeled with an unknown status, we rank them by their decision values from each classifier. For DIP data, the top-ranked interactions by InterBC and InterSVM were shown to have much higher functional similarity than those ranked by individual evidences, including FS-weight, PathRatio and

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Figure 4.4: The average functional similarity of top-ranked interactions in BioGrid data.

IRAP. (Genomic features, such as sequence similarity, gene expression correlation and interacting domain information, have poor performance as shown in Table 4.1. Therefore, we did not present the functional similarity for interactions ranked by these evidences in Figures 4.3, 4.4 and 4.5.) Therefore, those detected interactions that have higher ranks by InterSVM or InterBC are more likely to be true positives. Conversely, the detected interactions with lower ranks by InterSVM or InterBC tend to have lower similarity and are thus likely to be false positives.

In the BioGrid dataset, the overall performance of InterSVM and InterBC is still slightly better than that of FS-weight, as shown in Figure 4.4. However, the top 3000 interactions ranked by FS-weight have even higher functional similarity than those by InterSVM. This is due to the fact that IRAP suffers from a poor performance in BioGrid data as shown in Table 4.1 and Figure 4.4. As such, it is understandable that our integrative method cannot improve the performance much as it is also penalized by including those evidences with very poor performance.

4.4.3.2 False negative detection.

It is very common to take those protein pairs with at least a common neighbor as false negative candidates [Chen 08]. To avoid generating too many false negative candidates in those large PPI networks, we selected those protein pairs with at least 2 common

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

neighbors as candidates in this work. In DIP data, we finally collected 33,482 such candidates.

Figure 4.5 shows the average functional similarity of the top-ranked false negative candidates. The top-ranked candidates by InterSVM and InterBC are shown to have much higher functional similarity than those by FS-weight, PathRatio and IRAP. Thus, they are more likely to be novel interactions that are false negatives in the original datasets. In Figure 4.5, the average functional similarity of the top 2000 candidates by InterSVM and InterBC are 0.561 and 0.571 respectively — both are even higher than that of all interactions in our gold standard benchmark DIP core data (detected by small-scale experiments; all the protein interactions in the core have an average functional similarity 0.545), illustrating that our integrative approach has truly good results in predicting novel yet reliable protein interactions for eliminating false negatives in current PPI datasets.

We further analyzed the two sets of top 2000 ranked candidate interactions (from DIP) by InterSVM and InterBC respectively and found that they actually have a big overlap—there are 1574 interactions in common. The top 2000 candidates ranked by InterSVM also have 476 interactions that are confirmed to be reliable interactions in BioGrid data, while the top 2000 ranked by InterBC have 462 interactions that are similarly reliable. For example, the interaction between proteins YCR081W and YDR443C in both above two sets is believed to be reliable as it has been detected multiple times by TAP-MS [Gavi 06, Krog 06, Coll 07]. The high overlap between the predicted sets, as well as their large and almost equal numbers of confirmed interactions further affirm the usefulness of our integrative method for false negative identification.

4.4.4 Comparison with Patil’s Integration Method

Given an interaction, Patil *et al.* [Pati 05] proposed a naive bayesian method to check whether it is reliable by integrating the following three features: (1) whether this interaction has homologous interactions, *i.e.*, these two interacting proteins have homologs that interact in other species; (2) whether interacting proteins have at least one GO annotation in common and (3) whether interacting proteins have interacting Pfam domains. In our proposed method, we have utilized six commonly-used features to identify reliable interactions. Note that GO annotations (the second feature) typically are not used for predicting protein interactions since they are reserved to evaluate the performance of computational/experimental methods. In order to compare with Patil’s method, we also included two new features in our method, namely, the number of homologous interactions and the semantic similarity of two interacting proteins (measured by the semantic similarity of their GO annotations as show in equation 4.8).

Table 4.2 shows the AUC of each method (using 5-fold cross validation) on DIP data and BioGrid data, respectively. The performance for InterBC using original scores on two

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Figure 4.5: The average functional similarity of top-ranked false negative candidates generated from DIP data.

datasets is worse than that of Patil’s method. However, it outperforms Patil’s method when it exploits Bin-Confidence scores. Meanwhile, InterSVM with either original scores or Bin-Confidence outperforms Patil’s method on both datasets. We believe that the reasons why our proposed methods perform better than Patil’s method are as follows. First, we integrated more biological and computational features that are indeed beneficial, *e.g.*, FS-weight [Chen 06]. Second, Bin-Confidences as shown in both Tables 4.1 and 4.2 can effectively enhance the performance of the classifiers. Third, as each feature value in Patil’s method is binary (either 0 or 1), our probabilistic-based methods with fine granularity can better exploit each individual feature for identifying reliable interactions. Take homologous interactions as an example, Patil’s method only cares whether an interaction has homologous interactions or not. However, our methods care how many homologous interactions a given interaction has in other species, because an interaction with more homologous interactions would be considered to be more reliable in our methods.

4.5 Expanding and Purifying Interactions for Protein Complexes

In this section, we will expand and purify protein interactions in the PPI networks using our InterBIN and InterCLASS methods for identifying protein complexes.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Table 4.2: The comparison of AUC between Patil’s method and ours.

Methods/Data	DIP data		BioGrid data	
Patil’s method	0.810		0.784	
	Original Scores	Bin-Confidence	Original Scores	Bin-Confidence
InterSVM	0.852	0.865	0.797	0.818
InterBC	0.743	0.841	0.776	0.798

4.5.1 From COACH to COACH*

After our paper [Wu 09a] was published, we realized that two protein-complex cores predicted by our COACH may have big overlap (with NA score larger than our predefined threshold t , please refer to our Algorithm 1 on Page 35). To guarantee that any two cores have NA scores less than or equal to the threshold t , we modified our COACH to COACH* as follows.

First, when we get a new preliminary core, we will process it with our redundancy filtering procedure in COACH. Now, we do not process it, but preserve it in COACH*.

Second, after we collect all the preliminary cores, we assign each preliminary core a score indicating its *significance* in equation 4.9. We then rank these candidate cores based on their scores.

$$W(pc) = density(pc) \times |V_{pc}|, \quad (4.9)$$

where $pc = (V_{pc}, E_{pc})$ is a preliminary core and $density(pc)$ is the density of the preliminary core defined in equation 2.1 in section 2.3.

Third, we iteratively select a preliminary core with the highest score as a predicted protein-complex core and then remove all its neighboring preliminary cores (*i.e.*, those have NA scores at least t with the selected core). Finally, we collect all the cores to form protein complexes and we can guarantee that any two cores predicted by COACH* have NA scores less than our predefined threshold t .

4.5.2 Expanding PPI Networks by Adding False Negatives

We ranked false negative candidates predicted by our InterBC method and selected different number of top ranked candidates to expand current DIP data. Figure 4.6 shows the performance of our COACH* on DIP data when different number of reliable false negatives or random edges are added (evaluation measures are introduced in section 2.4 and we used CYC2008 catalogue as benchmark protein complexes). In this figure, red lines show the performance after adding random edges while black lines show that of adding reliable false negatives predicted by InterBC. We have the following conclusions from Figure 4.6. First, adding a small number of random edges does not affect the performance of our COACH*. Second, adding reliable false negatives can increase the recall

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

Figure 4.6: The performance of COACH* on the PPI network expanded by adding false negatives.

and coverage rate, indicating that many newly added edges are within known protein complexes. Third, as the PPI network expanded, both the number of predicted complexes and those correctly predicted complexes are increased. Therefore, the precision does not change much when different numbers of false negatives are added. Nevertheless, the overall F-measure still increases as more predicted complexes match benchmark complexes, indicating that the performance of COACH* is enhanced by adding reliable false negatives.

4.5.3 Purifying PPI Networks by Deleting False Positives

We ranked the whole DIP data using InterBC and those interactions with lower scores are considered to be false positives. Figure 4.7 shows the performance of COACH* when different numbers of false positives or random edges are removed. Random deleting of edges will decrease the recall and coverage rate obviously because many true interactions (*e.g.*, those within known protein complexes) may be removed. However, removed false positives are much less likely to be those within known complexes. Therefore, the recall and coverage rate of COACH* remain stable while false positives are deleted. In addition, as those false positives were deleted, we predicted less protein complexes and still achieved a stable recall, and thus we had a higher precision as we expected. Correspondingly, we can also improve the overall performance (F-measure) of COACH* by deleting false negatives.

Figure 4.7: The performance of COACH* on the PPI network purified by deleting false positives.

4.5.4 Concurrently Adding and Deleting Edges

We also tried a simple way to concurrently add and delete edges for DIP data. We added and deleted the same number of edges so that the number of edges in modified DIP PPI network was fixed. Similarly in Figure 4.8, red lines show the performance of COACH* by adding and deleting random edges while black lines show that by adding reliable false negatives and deleting predicted false positives. In Figure 4.6, we observed that the performance of COACH* remains unaltered when we added random edges. Therefore, we subsequently observed that the curves for random edges in Figure 4.8 are very similar to those in Figure 4.7. Meanwhile, the operation of concurrently adding false negatives and deleting false positives is observed to be effective as shown in Figure 4.8, which integrated the strengths obtained by doing them separately. In particular, concurrently adding and deleting edges can achieve a higher coverage rate and recall (achieved by only adding false negatives) and also increase the precision (achieved by solely deleting false positives).

When adding and deleting edges concurrently, we realize that we may add more edges to increase the quality of the modified PPI data if those added edges still have a higher quality than deleted edges. Therefore, it is more reasonable to add and delete different number of edges. An open issue to be discussed is how can we achieve the best performance by adding and deleting different numbers of edges?

Figure 4.8: The performance of COACH* on the PPI network modified by concurrently adding and deleting edges.

4.5.5 Summary

From the above results, we can better predict protein complexes on both conditions that reliable false negatives are added or reliable false positives are removed. Particularly, we can collect protein complexes with better recall and coverage rate when reliable false negatives are added, while our predicted protein complexes are more accurate when false positives are removed. Therefore, identifying reliable interactions is very important for detection of protein complexes and some other studies, for example, protein function prediction, network motif finding and so on.

4.6 Conclusions

Protein-Protein interactions play important roles in cellular processes and provide an invaluable resource for network-based biological knowledge discovery such as protein complex detection [Li 10a], protein function prediction [Chua 07b], and disease pathway discovery [Li 09]. While high-throughput experimental techniques seemed to provide us with many large-scale PPI datasets, their usefulness is diminished by high level of noise and incompleteness. To eliminate the abundant false positives and false negatives in the current PPI datasets, it is important to develop robust methods for identifying and predicting reliable interactions.

CHAPTER 4. INTEGRATING DIVERSE SOURCES FOR RELIABLE PROTEIN-PROTEIN INTERACTIONS

In this chapter, we integrate diverse sources to detect PPI for eliminating the false positives and negatives. We first estimate the Bin-Confidences for interactions by grouping them based on their original raw scores from each source. Using the robust Bin-Confidence scores of protein pairs from the various sources as features, we train machine learning models using SVM and BC to integrate the sources and construct classifiers to detect reliable interactions. Experimental results show that our integrative method outperforms the approaches merely using each of individual topological or genomic evidences. The accuracy of SVM and BC is also significantly improved by using the Bin-Confidences instead of the original raw scores. In addition, removing false positives or adding false negatives predicted by our method are demonstrated to be helpful to better detect protein complexes.

Chapter 5

CACHET: Detecting Protein Complexes Directly from TAP Data

5.1 Introduction

5.1.1 Background

It was observed that protein complexes generally correspond to dense subgraphs or cliques in the PPI networks [Tong 02, Spir 03], as proteins in the same complex tend to be highly interactive with each other. Numerous computational methods have been proposed to detect protein complexes from PPI networks based on this observation—for example, MCODE [Bade 03a], MCL [Pere 04], DPCLus [Alta 06], DECAFF [Li 07b], COACH [Wu 09a], and CMC [Liu 09], just to name a few. All these methods are based on experimental data that describe the pairwise interactions between two proteins. However, converting TAP data to accurate PPI data is still challenging. We briefly review the strategies for converting TAP data to PPI networks in recent studies. In TAP data, bait-prey relationships and even prey-prey relationships are regarded as candidate interactions and can be scored by various reliability measurements [Gavi 06, Coll 07, Zhan 08]. A threshold can thus be selected and the candidate interactions with scores above the threshold are then predicted as true interactions. Majority of the strategies determine this threshold based on the reference protein-complex data. In [Zhan 08] and [Pu 07], all the connections between proteins within the same complexes are regarded as known positive interactions and connections between proteins from distinct complexes are negative interactions (one more restriction—cellular localization, is used for defining negative interactions in [Pu 07]). The threshold will be selected when the best map is achieved between predicted interactions and known interactions in terms of F-measure [Zhan 08] and ROC curves [Pu 07]. In [Hart 07] and [Gavi 06], a set of complexes will be derived at each candidate threshold and the final threshold will be selected when the predicted complexes can best match the reference complexes.

5.1.2 Motivation

Obviously, it is not reasonable to assume that all protein pairs within a complex are positive interactions in the above methods ([Pu 07] and [Zhan 08]) when converting TAP co-complex information to pair-wise protein interaction data. Moreover, these strategies are biased with respect to known protein complexes. All these challenges will be avoided if we can detect protein complexes directly from TAP data. In addition, TAP describes co-complex relationships between proteins [Puig 01]. Directly exploiting this co-complex information for complex identification may thus be more promising and useful than the isolated binary interaction information in conventional PPI networks [Scho 05, Geva 11].

In this chapter, we introduce a novel method called CACHET to detect protein complexes directly from TAP data. Instead of transforming them into a pairwise PPI network, we model the TAP data as a bipartite graph, where one vertex set consists of the bait proteins, the other vertex set contains all the prey proteins, and the edges represent the bait-prey interactions. Meanwhile, protein complexes have been found to contain *core-attachment structures* in which the cores are the functional “hearts” of the protein complexes and the attachment proteins assist the core proteins to perform subordinate functions [Dezs 03, Gavi 06]. While there are also several recent studies detecting core-attachment structures from PPI networks [Wu 09a, Leun 09, Srih 10], here we detect protein complexes with the core-attachment structures directly from TAP data. Our proposed method will focus on the detection of high quality protein-complex cores by first computing the complete set of maximal bicliques from the bipartite graph and then extracting those non-redundant reliable bicliques. The protein complexes are then predicted by assembling attachments into the complex cores.

A newly proposed method called CODEC [?] also detects protein complexes directly from TAP data as bipartite graphs. Our CACHET method works differently from CODEC in the following aspects. Firstly, CODEC detects dense bipartite subgraphs as protein complexes by iteratively adding proteins into seed subgraphs. Our CACHET selects protein-complex cores from bicliques and then simultaneously assembles all the attachments into cores to form the complexes. If we regard our cores as seed graphs, our approach for adding proteins into seed graphs to form protein complex is much more efficient than CODEC in terms of computation. Secondly, CACHET and CODEC address the issue of false positives in TAP data differently. CODEC approximates the reliability of bait-prey links mainly by the degree of the vertices. In contrast, CACHET exploits three state-of-the-art reliability measurements, namely, Socio-Affinity (SA) [Gavi 06], Purification Enrichment (PE) [Coll 07] and Dice Coefficient (DC) [Zhan 08], which are demonstrated to assess the reliability of bait-prey relationships more accurately. As a result, our CACHET can detect protein complexes more efficiently as well as with more biological significance (*i.e.*, higher accuracy and homogeneity) than CODEC.

5.2 Methods

In this section, we first introduce some preliminaries in bipartite graph. We then introduce our proposed algorithm CACHET to detect protein complexes with core-attachment structures from TAP data. We focus on detecting high quality protein-complex cores. Firstly, we compute the *maximal bicliques* in the TAP bipartite graph, since proteins within protein-complex cores are highly interactive [Gavi 06]. Next, we detect *reliable* protein-complex cores from the computed bicliques. Given the presence of false positive bait-prey interactions, we perform a filtering process to extract reliable bicliques (Section 5.2.2). Since these reliable bicliques may have big overlaps with each other (share many common proteins), we eliminate redundancy among these potential cores by computing a maximal independent set of them (Section 5.2.3). The reliable and independent/non-redundant bicliques are thus predicted as the protein-complex cores. Finally, attachment proteins are included into the cores to form protein complexes with core-attachment substructures.

5.2.1 Preliminaries

As mentioned earlier, TAP data are represented as a bipartite graph $G = (U, V, E)$, where U is the set of vertices for the baits, V the set of vertices for the preys captured by the baits in U , and E the bait-prey relationships. In addition, every edge $(u, v) \in E$ is associated with a reliability score that estimates the likelihood of the bait $u \in U$ interacting with the prey $v \in V$. Next, we will introduce some definitions in a bipartite graph G .

From $G = (U, V, E)$, a biclique $G' = (U', V', E')$ is a bipartite graph such that for any two vertices $u \in U'$ and $v \in V'$, there is an edge $(u, v) \in E'$, where U' , V' , and E' are subsets of U , V , and E respectively. A biclique $G_1 = (U_1, V_1, E_1)$ is a maximal biclique if there is no other biclique $G_2 = (U_2, V_2, E_2)$ such that $U_1 \subseteq U_2$, $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$.

The **reliability** of the bipartite graph $G = (U, V, E)$, $R(G)$, is defined as the average reliability scores of all the edges in E ,

$$R(G) = \frac{\sum_{u \neq v, (u,v) \in E} w(u, v)}{\sum_{(u,v) \in E} \delta(u, v)}, \quad (5.1)$$

where $w(u, v)$ is the edge weight (reliability score of the edge (u, v)) and $\delta(u, v)$ is 0 if vertices u and v are identical and 1 otherwise. This is because a bait together with a specific antibody tends to tag the bait itself as a prey in pull-down experiments; as such, the edges in the form of (v, v) are not considered for calculating the reliability score in

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Figure 5.1: A example TAP bipartite graph. $\{v_1, v_2, v_3\}$ is the set of bait proteins and $\{v_1, v_4, v_5, v_6, v_7\}$ is the set of prey proteins. Please note that certain proteins (e.g., v_1) can serve as baits and they can also be tagged as preys by other baits.

Equation (5.1) [Geva 11]. Similarly, we can also define a reliability score for each vertex l in G called $ratio_G(l)$,

$$ratio_G(l) = \frac{\sum_{u \neq l, (u,l) \in E} w(u,l)}{\sum_{(u,l) \in E} \delta(u,l)}. \quad (5.2)$$

Note that only l 's direct neighbors are used to compute its $ratio$ score in Equation (5.2).

Neighborhood affinity score [Bade 03a] is adopted to measure similarity between two vertex sets. Given two bipartite graphs $G_1 = (U_1, V_1, E_1)$ and $G_2 = (U_2, V_2, E_2)$, we define their neighborhood affinity score, $NA(G_1, G_2)$ as follows:

$$NA(G_1, G_2) = \frac{|(U_1 \cup V_1) \cap (U_2 \cup V_2)|^2}{|U_1 \cup V_1| \times |U_2 \cup V_2|}. \quad (5.3)$$

Figure 5.1 provides an example TAP bipartite graph. The baits v_1 and v_2 together with the preys v_1 , v_4 and v_5 form a biclique G , with the reliability score $R(G) = \frac{0.7+0.6+0.65+0.9}{4} = 0.71$. The protein v_1 as a bait has a $ratio(v_1) = \frac{0.6+0.65}{2} = 0.625$, while it as a prey has a reliability 0.7. The reliability of this sample bipartite graph is 0.6 according to the Equation (5.1).

5.2.2 Mining Reliable Bicliques

We apply the FP-MBC algorithm [Li 07a] to compute the complete set of maximal bicliques. We require that a detected maximal biclique should have at least Min_Bait baits and Min_Prey preys. In this study, both Min_Bait and Min_Prey are set as 2. We consider a biclique G' to be reliable if its reliability score is greater than or equal to a threshold r , i.e., $R(G') \geq r$. Reliable bicliques are likely to be protein-complex cores since most of their bait-prey interactions are supposed to be true with respect to an appropriate threshold r .

Algorithm 4 shows our detailed algorithm to extract reliable bicliques. First, we compute the reliability scores for all the bait-prey edges in the whole TAP bipartite graph $G_{TAP} = (U_T, V_T, E_T)$. In this work, we exploit various reliability measurements,

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

such as Socio-Affinity (SA) [Gavi 06], Purification Enrichment (PE) [Coll 07] and Dice Coefficient (DC)[Zhan 08] for edge-weighting. Then, a biclique is preserved if it is already reliable (Lines 3-4). If a biclique is of low reliability, we iteratively delete a vertex with the minimum *ratio* score (Lines 7-8) to increase its reliability and finally obtain a reliable biclique if possible (Lines 10-11). Such a removal process is guaranteed to increase or at least keep the overall reliability of the biclique in each iteration (a theoretical proof is provided in Appendix). During this process, we also require that the vertices to be removed are selected from an available node set in order to guarantee that all the detected reliable bicliques satisfy the size constraint (*i.e.*, they should have at least *Min_Bait* baits and *Min_Prey* preys). For example, if a biclique $G' = (U', V', E')$ has a bait set with the minimum size (*i.e.*, $|U'| = \text{Min_Bait}$) and the prey set is still large enough (*i.e.*, $|V'| > \text{Min_Prey}$), we should select a vertex from the prey set and then remove it. The available node set in G' , denoted as $A(G')$, is defined in Equation (5.4):

$$A(G') = \begin{cases} U' \cup V', & \text{if } |U'| > \text{Min_Bait}, |V'| > \text{Min_Prey}; \\ V', & \text{if } |U'| = \text{Min_Bait}, |V'| > \text{Min_Prey}; \\ U', & \text{if } |U'| > \text{Min_Bait}, |V'| = \text{Min_Prey}; \\ \phi, & \text{Otherwise.} \end{cases} \quad (5.4)$$

Clearly, our iterative removal process (Lines 6-9) will stop in the following two conditions. If a biclique is reliable after the removal of some vertices, it will be preserved in Line 11. Otherwise, it will be discarded if it has already reached the minimum size (*i.e.*, $A(G') = \phi$) but it is still not reliable.

5.2.3 Mining Maximal Independent Set of Bicliques

Reliable bicliques may overlap with each other. Therefore, we need to select a maximal subset of independent/non-redundant bicliques as protein-complex cores. To filter the redundancy among bicliques, we construct a redundancy graph, denoted as $RG = (V_{RG}, E_{RG}, W)$, where the super vertex set $V_{RG} = \{G_i | G_i \in S\}$ consists of reliable bicliques in S generated by Algorithm 4. Although bicliques are themselves graphs, they are regarded here as super vertices in our redundancy graph. The super edge set $E_{RG} = \{(G_i, G_j) | G_i, G_j \in V_{RG}, NA(G_i, G_j) \geq 0.2\}$ links two super vertices if these two reliable bicliques are deemed to be similar or redundant. In this work, two reliable bicliques are considered to be similar when their *NA* score (see the Equation (5.3)) is at least 0.2, which is also used in many previous study [Bade 03a, Alta 06, Li 07b, Wu 09a] to evaluate whether a predicted complex matches a benchmark complex. The vertex weighting function W shows the *significance* of reliable bicliques. We compute W as

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Algorithm 4 Mining Reliable Bicliques

Input: The set of bicliques from the FP-MBC algorithm,
 r is the threshold for reliability score for reliable bicliques.

Output: The set of reliable bicliques, S .

- 1: Collect the reliability scores for all the bait-prey edges in G_{TAP}
- 2: **for each** biclique $G' = (U', V', E')$ **do**
- 3: **if** $R(G') \geq r$
- 4: insert G' into S
- 5: **else**
- 6: **while** $R(G') < r$ and $A(G') \neq \phi$ **do**
- 7: $v = \arg \min_{w \in A(G')} \text{ratio}_{G'}(w)$
- 8: $G' = G' - \{v\}$ //update G' by deleting v and its corresponding edges
- 9: **end while**
- 10: **if** $R(G') \geq r$
- 11: insert G' into S
- 12: **end for**

follows. First, we count the frequency of each vertex/protein in all the reliable bicliques to indicate their importance, since proteins that appear in multiple modules are more biologically important [Navl 10]. Equation 5.5 below computes $w(u)$ as the importance score of the vertex/protein u , where the frequency of u , $\text{freq}(u)$, is the number of reliable bicliques involving u :

$$w(u) = \frac{\text{freq}(u)}{\max_v \text{freq}(v)} \quad (5.5)$$

We refine the reliability score for bicliques in Equation (5.1) by considering the importance of vertices to evaluate the *significance* of reliable bicliques:

$$W(G') = \frac{\sum_{u \in U', v \in V', u \neq v} w(u, v) \times (w(u) + w(v))}{2 \times \sum_{u \in U', v \in V'} \delta(u, v)}, \quad (5.6)$$

where $G' = (U', V', E')$ is a reliable biclique, the edge weight $w(u, v)$ is the edge reliability which can be measured by various reliability schemes (*i.e.*, PE [Coll 07], SA [Gavi 06] and DC [Zhan 08]) and the vertex weight $w(u)$ is the vertex u 's importance as computed in Equation (5.5).

Given a redundancy graph $RG = (V_{RG}, E_{RG}, W)$, our aim is to mine an independent set (*IS*) in RG where any two reliable bicliques are not similar. This is a maximum independent set (*MIS*) problem. To conserve those more significant bicliques, we aim to maximize the weight sum of the super vertices in the resulting independent set. As

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

such, our *MIS* problem can thus be modeled as a maximum weighted independent set (*MWIS*) problem, as defined in Equation (5.7):

$$SC = \arg \max_{IS} \sum_{G_i \in IS} W(G_i), \forall G_j, G_k \in IS, (G_j, G_k) \notin E_{RG}. \quad (5.7)$$

Here, *IS* is an independent set of *RG* and the resulting set *SC*, the maximum weighted independent set, contains reliable and independent bicliques which is the set of our predicted protein-complex cores.

Given a super vertex $v \in V_{RG}$, $N_{RG}(v)$ is the set of v 's neighbors in *RG* and its degree in *RG*, denoted as $d_{RG}(v)$, is the cardinality of $N_{RG}(v)$, i.e., $|N_{RG}(v)|$. The weighted degree of v is define as $wd_{RG}(v) = \sum_{u \in N_{RG}(v)} W(u)$. As we know, maximum independent set (*MIS*) and maximum weighted independent set (*MWIS*) are well-known NP-Complete problems [Gare 79]. Nonetheless, many approximation algorithms have been proposed for efficiently computing them. We introduce three heuristics in Algorithm 5 for mining the *MWIS*.

Algorithm 5 Maximum Weighted Independent Set

Input: A weighted graph $RG = (V_{RG}, E_{RG}, W)$.

Output: *SC*, a maximum weighted independent set in *RG*.

```

1:  $SC = \phi, i = 0, T_i = RG$ 
2: while  $V(T_i) \neq \phi$  do
3:   select a vertex  $v_i$  based on one of the following criteria
4:    $H_1: v_i = \arg \max_{u \in V(T_i)} \frac{W(u)}{d_{T_i}(u) + 1}$ 
5:    $H_2: v_i = \arg \max_{u \in V(T_i)} \frac{W(u)}{wd_{T_i}(u)}$ 
6:    $H_3: v_i = \arg \max_{u \in V(T_i)} W(u)$ 
7:    $SC = SC \cup \{v_i\}, T_{i+1} = T_i - (N_{T_i}(v_i) \cup \{v_i\}), i = i + 1$ 
   //remove  $v_i$ , its neighbors and the corresponding edges from  $T_i$ .
8: end while
```

The first heuristic (H_1 in Line 4) [Saka 03] greedily selects the super vertex with the maximum $\frac{W(u)}{d_{T_i}(u)+1}$ value in the evolving graph T_i . The selected vertex generally has a relatively large weight and a low degree. In other words, it keeps those significant and less-connected bicliques. The second heuristic (H_2 in Line 5) [Kako 05] does the same considers the weighted degree in the denominator. The third heuristic (H_3 in Line 6) elects the super vertex with the maximum significance score. H_1 and H_2 are well lied on their performance (e.g., approximation ratio). For example, in a redundancy

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

graph $RG = (V_{RG}, E_{RG}, W)$, H_1 outputs an independent set with weight sum at least $\sum_{v \in V_{RG}} W(v)/(d_{RG}(v) + 1)$ [Saka 03]. H_3 is clearly most efficient to compute since H_1 and H_2 need to update the degree and weighted degree for super vertices in each iteration, while H_3 does not.

5.2.4 Detecting Protein-Complex Attachments

In Section 5.2.3, we have already detected independent reliable bicliques as protein-complex cores. Let us now identify the attachment proteins. For a protein-complex core $G_c = (U_c, V_c, E_c)$ detected from the TAP bipartite graph $G_{TAP} = (U_T, V_T, E_T)$, we define its bait-attachments and prey-attachments respectively as follows. Given a prey protein $p \notin V_c$ (the prey set), we construct a temporary graph $TG = (U_c, \{p\}, E_{tg})$ where $E_{tg} = \{(u, p) | u \in U_c, (u, p) \in E_T\}$. The prey p is then a bait-attachment of G_c if it satisfies the following two constraints. First, it should interact with more than half of proteins in U_c (i.e., $|E_{tg}| > 0.5 * |U_c|$). As we know, G_c is a biclique and every prey in V_c interacts with all the baits in U_c . Here, we basically relax the connectivity constraint for selecting bait-attachments because the attachments are generally less interactive than the proteins in the cores [Gavi 06]. Second, the edges between p and p 's neighbors in U_c should have an average reliability score larger than the threshold r (i.e., the reliability of TG , $R(TG) \geq r$). The prey-attachments are defined in the same way.

The resulting protein complexes are formed by combining protein-complex cores and their bait-attachments and prey-attachments. Figure 5.2 shows the diagram of our CACHET working on a sample TAP graph. First (the step 1), a bipartite graph G is constructed from the TAP data. Here, $\{v_1, v_2, v_3, v_4\}$ is the bait set and $\{v_5, v_6, v_7, v_8, v_9, v_{10}\}$ is the prey set. Second (the step 2), 3 maximal bicliques are detected by the FP-MBC algorithm [Li 07a], namely, G_1 , G_2 and G_3 . Third (the step 3), we extract reliable bicliques using our algorithm 4. G_1 and G_2 with reliability score 0.773 and 0.75 respectively are considered to be reliable (assuming that the parameter r in algorithm 4 is set as 0.7) while G_3 with reliability 0.48 is thus discarded. Fourth (the step 4), we detect protein-complex cores by solving maximum weighted independent set (MWIS) problem and form protein complexes by combining cores and their attachments. Here, two reliable bicliques G_1 and G_2 have a high neighborhood affinity score (0.64) and are thus considered to be similar. Since G_1 has a larger *significance* score (please see the equation 6 in our main manuscript) than G_2 , G_1 is finally predicted as a protein-complex core and the bait protein v_3 (in grey) is selected as its prey-attachment to form a protein complex.

5.3 Experimental Results

We first introduce the evaluation metrics and the datasets used in our experiments and then show our experimental results.

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Figure 5.2: The diagram of our CACHET method working on a sample TAP graph.

5.3.1 Evaluation Metrics and Experimental Data

We still use Precision, Recall, F-measure and p-values defined in Section 2.4 to evaluate the performance of our CACHET. Recall that a predicted complex and a real complex are considered to be matching if their NA score is above 0.2. However, it might be concerned that this value 0.2 is too low and liberal. Therefore, we also tried different values of ω from 0.2 to 0.5 for fair evaluation. Please note that 0.5 is quite a high value. For example, given two complexes both with 6 proteins, they are considered to be matched with each other only if they share at least 5 common proteins when ω is 0.5.

Currently, the evaluation measures are considered to be at complex-level. These complex-level evaluations are intuitive for us to see which benchmark complexes can be correctly predicted. However, pairs of matching complexes may have different NA scores that indicate different matching degrees. This issue is not taken into consideration in the above definitions. Therefore, we also evaluate the predicted complexes using protein-level evaluation metrics. In particular, we assume that there are m predicted complexes in P and n real complexes in B (i.e., $m = |P|$, $n = |B|$). Protein-level Precision and Recall are defined as follows to evaluate various methods [Liu 09]. F-measure is still the harmonic mean of Protein-level Precision and Recall.

$$P - Precision = \frac{\sum_{j=1}^m \max_{NA(b_i, p_j) \geq 0.2} overlap(b_i, p_j)}{\sum_{j=1}^m |V_{p_j}|} \quad (5.8)$$

$$P - Recall = \frac{\sum_{i=1}^n \max_{NA(b_i, p_j) \geq 0.2} overlap(b_i, p_j)}{\sum_{i=1}^n |V_{b_i}|}. \quad (5.9)$$

Here, $b_i = \{V_{b_i}, E_{b_i}\} \in B$, $p_j = \{V_{p_j}, E_{p_j}\} \in P$ and $overlap(b_i, p_j)$ is the number of common proteins between b_i and p_j , i.e., $|V_{b_i} \cap V_{p_j}|$. V_{b_i} and V_{p_j} are the sets of proteins in b_i and p_j , respectively. From Equations (5.8) and (5.9), it is clear that the higher the P-Precision, the more proteins of predicted complexes can match with the proteins in benchmark complexes.

We applied our CACHET on two public large-scale TAP data, namely, Gavin *et al.*'s data [Gavi 06] and Krogan *et al.*'s data [Krog 06]. Gavin *et al.*'s data contain 1993 baits, 2670 preys and 19277 bait-prey links and Krogan *et al.*'s data contain 1999 baits, 5202 preys and 47204 bait-prey links. We also generated a dataset which combined both Gavin *et al.*'s data and Krogan *et al.*'s data. Let's denote Gavin *et al.*'s data and Krogan *et al.*'s data as $G_{gavin} = (U_g, V_g, E_g)$ and $G_{krogan} = (U_k, V_k, E_k)$. We thus obtained the

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

combined dataset, denoted as $G_{combined} = (U_c, V_c, E_c)$, where $U_c = U_g \cup U_k$, $V_c = V_g \cup V_k$ and $E_c = E_g \cup E_k$. The combined dataset contain 2850 baits, 5337 preys and 63631 bait-prey links. All the statistics about these three datasets are shown in Table 5.1. The benchmark protein complexes were downloaded from Wodaklab [Pu 09]. The catalogue is called CYC2008 and it contains 408 manually curated protein complexes. Gene Ontology was downloaded from <http://geneontology.org>.

Table 5.1: Details of TAP datasets.

Details	Gavin <i>et al.</i> 's data	Krogan <i>et al.</i> 's data	Combined data
# Baits	1993	1999	2850
# Preys	2670	5202	5337
# Links	19277	47204	63631

5.3.2 Converting TAP Data to PPI Networks

Existing state-of-the-art algorithms detect protein complexes from PPI networks while our CACHET identifies them directly from TAP data. To apply existing methods (*e.g.*, MCODE [Bade 03a], MCL [Pere 04] and COACH [Wu 09a]), we have to transform the TAP data to PPI networks and then run them in these PPI networks.

Currently, many strategies, including supervised and unsupervised methods, have been proposed to convert the TAP data to PPI networks [Zhan 08, Hart 07, Pu 07, Frie 08, Krog 06]. The method in [Zhan 08] is similar to that in [Pu 07] and both of them are demonstrated to be effective to construct PPI networks for the detection of protein complexes. Meanwhile, it is much easier to implement than those in [Hart 07, Frie 08, Krog 06]. Due to these facts, we applied it to collect PPI networks from TAP data in this thesis as follows. In TAP data, bait-prey relationships and even prey-prey relationships are regarded as candidate interactions and can be scored by various reliability measurements [Gavi 06, Coll 07, Zhan 08]. A threshold can thus be selected and the candidates with scores above the threshold are then predicted as true interactions. Meanwhile, all the connections between proteins within the same complexes are regarded as known positive interactions and connections between proteins from distinct complexes are negative interactions. The threshold will thus be selected when the best map is achieved between predicted interactions and known interactions in terms of F1-measure. Here, F1-measure is conventionally defined based on TP, FP, TN and FN which are all introduced in Table 5.2.

Based on the above method, we used CYC2008 catalogue [Pu 09] as benchmark complexes and collected the PPI networks converted from three TAP datasets using three scoring schemes, namely, Socio-Affinity (SA) [Gavi 06], Purification Enrichment (PE)

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Table 5.2: Contingency table from [Zhan 08] to evaluate the matching between predicted interactions and known interactions.

	Known positives	Known negatives
Above the threshold	True positive (TP)	False positive (FP)
Below the threshold	False negative (FN)	True negative (TN)
precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$ and F1-measure = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$		

Table 5.3: PPI networks converted from TAP data			
Data	Scoring schemes	# nodes	# edges
Gavin <i>et al.</i> 's data	SA	1817	7000
	PE	1758	8900
	DC	2019	9200
Krogan <i>et al.</i> 's data	SA	3275	4500
	PE	1401	7200
	DC	3395	16300
Combined data	SA	2895	10700
	PE	1886	12100
	DC	2595	9000

[Coll 07] and Dice Coefficient (DC) [Zhan 08]. Table 5.3 shows the details of each constructed PPI network. For example, we obtained a PPI network with 1817 proteins and 7000 interactions from Gavin *et al.*'s data using Socio-Affinity (SA) scores.

5.3.3 Comparative Evaluation of Predicted Protein Complexes

We focus on presenting our results on Gavin *et al.*'s data in this subsection. We will show the results of CACHET on Krogan *et al.*'s data and the combined data later in subsection 5.3.7 and 5.3.8, respectively.

5.3.3.1 Performance of various *MWIS* heuristics

We first computed the results using three different *MWIS* heuristics with the parameter r set as 0.7 (see Algorithm 4 for r —we will perform the sensitivity study of r in subsection 5.3.5). The second column of Table 5.4 shows the number of protein complexes predicted by different *MWIS* heuristics. H_1 and H_2 predicted nearly the same number of protein complexes and achieved similar Precision. Compared with H_1 and H_2 , H_3 obtained fewer protein complexes but achieved a higher Precision, a similar Recall and a higher F-measure. Given that CACHET with H_3 detects protein complexes more efficiently from the time-complexity perspective, we only report the results using H_3 in our CACHET hereafter.

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Table 5.4: Performance for various *MWIS* heuristics using SA scores on Gavin *et al.*'s data ($\omega = 0.2$)

Methods	# complexes	Precision	Recall	F-measure
H_1	437	0.643	0.439	0.521
H_2	438	0.651	0.446	0.529
H_3	369	0.678	0.441	0.534

5.3.3.2 F-measure for various methods

We compared our CACHET with the existing state-of-the-art methods, namely, MCODE [Bade 03a], MCL [Pere 04, Krog 06, Frie 08], DC-CM [Zhan 08], COACH in our chapter 3 [Wu 09a] and CODEC [Geva 11] (CODEC has two different weighting schemes, denoted as CODEC_{w_0} and CODEC_{w_1} . See the original paper [Geva 11] for details.) in terms of F-measure and p-values.

Figure 5.3 shows the F-measure of various methods using three different reliability schemes on Gavin *et al.*'s data. Our CACHET with SA scores achieves the highest overall F-measure. From the perspective of using individual scores, CACHET also achieves the highest F-measure when using SA and DC scores and performs equally well as MCL when using PE scores. Particularly, CACHET with SA scores has a F-measure 53.4%. MCL with PE scores, which is the best performer by combining existing methods and scoring schemes, has a F-measure 48.6%. Our CACHET with SA scores is still 4.8% higher than MCL with PE scores.

As SA scores are most widely used, Table 5.5 shows the details of the comparison among various methods using SA scores (some methods are independent of reliability measurements, such as DC-CM, CODEC and the original results “Gavin” and “Gavin-Core” from [Gavi 06]). The F-measure of our CACHET is 53.4%, which is 7.9%, 8.3% and 11.4% higher than that of MCL, COACH and CODEC_{w_1} respectively, the top three performers of the existing methods. Furthermore, CACHET is shown to perform better than CODEC in terms of all three measures although they also model TAP data as a bipartite graph. We also show the F-measure of various methods when ω is varied from 0.2 to 0.5 in Figure 5.4. CACHET consistently outperformed existing methods. The superiority of our CACHET is independent on the values of ω , indicating that its predicted complexes really match the benchmark complexes very well.

5.3.3.3 P-values for protein complexes predicted by various methods

A protein complex is statistically significant if its p-value is less than $10^{-\alpha}$ —the larger the α , the more likely that the complexes are genuine. Figure 5.5 shows the proportion of

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Figure 5.3: Comparison among various methods on Gavin *et al.*'s data using different reliability schemes.

Figure 5.4: F-measure of various methods when ω is varied from 0.2 to 0.5.

significant complexes under different values of α . In the figure ¹, CACHET and CODEC

¹Please note that the singleton complexes (those with only one protein) are not considered for p-value calculation [Wu 09a].

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Table 5.5: Performance for various methods on Gavin *et al.*'s data ($\omega = 0.2$ using SA scores)

Methods	# complexes	Precision	Recall	F-measure
MCODE	136	0.632	0.257	0.366
MCL	302	0.483	0.429	0.455
DC-CM	851	0.249	0.424	0.314
CODEC w_0	1082	0.276	0.343	0.306
CODEC w_1	1005	0.417	0.424	0.420
COACH	346	0.540	0.387	0.451
Gavin	491	0.267	0.309	0.287
Gavin-Core	478	0.329	0.414	0.367
CACHET	369	0.678	0.441	0.534

(with both weighting schemes) are both shown to achieve the highest proportion of significant complexes when $\alpha = 2$. However, as the value of α increases, the superiority of our CACHET becomes obvious. For example when α was set as 4, CACHET has 77.0% significant complexes, 10.7% higher than the second best (*i.e.*, CODEC w_1 with 66.3%). Furthermore, CACHET has 46.8% significant complexes, 13.0% higher than the second best (*i.e.*, COACH with 33.8%) when α was set as 10. This shows that complexes predicted by CACHET are enriched by proteins with same functions and thus more likely to be true complexes.

Figure 5.5: The proportion of significant complexes predicted by various methods.

5.3.4 Evaluation of Predicted Protein-Complex Cores

Proteins within cores generally have higher functional similarities and are more co-localized than the attachment proteins [Dezs 03, Gavi 06]. We can calculate a functional similarity score for a core (or a protein complex) by averaging the semantic similarities among all bait-prey pairs [Wang 07]. Table 5.6 shows the average functional similarity for all cores and protein complexes.

Table 5.6: Average functional similarity within protein-complex cores and protein complexes.

	Biological Process(BP)		Cellular Component(CC)	
	in cores	in complexes	in cores	in complexes
H_1	0.656±0.008	0.601±0.018	0.779±0.006	0.762±0.012
H_2	0.675±0.005	0.639±0.016	0.796±0.006	0.772±0.008
H_3	0.645±0.010	0.582±0.016	0.756±0.007	0.724±0.008

The functional similarities in Table 5.6 were calculated using two sub-ontologies of GO (biological process and cellular component). The protein-complex cores generated by all three *MWIS* heuristics are found to have higher average functional similarities than the corresponding protein complexes. This demonstrates that our detected cores are more functionally consistent and thus more biologically important.

We also compare the quality of our predicted cores with those detected by previous studies [Gavi 06, Zhan 08]. Each protein pair have a score that denotes the similarity between these two proteins (*e.g.*, functional similarity based on GO terms or gene expression correlation). Similarly, a core can be assigned a score that is the average similarity among all protein pairs within the core ². To evaluate the quality of a set of predicted cores, we define *inner-similarity* as the average score for all these cores. Table 5.7 shows the inner-similarity scores of cores detected by various methods.

In Table 5.7, CACHET is observed to achieve the highest inner-similarity no matter what measures are used for similarity between protein pairs. Cores predicted by COACH also have higher inner-similarity than those predicted by other two previous studies. That is, the cores detected by our inside-out methods have higher quality than previous studies which detect protein complexes first and select significant proteins within complexes as core proteins.

²Since previous studies do not distinguish baits and preys within cores, we here calculate the score for a core by averaging the similarity scores among all protein pairs instead of bait-prey pairs for fair comparison. In Table 5.7, “BP” in second column means that the inner-similarity of cores is calculated based on GO similarity using “Biological Process” sub-ontoloty while “GE” means using gene expression profiles.

Table 5.7: Inner-similarity of cores predicted by various methods on Gavin *et al.*'s data

Methods	BP	CC	MF	GE
Gavin	0.480	0.701	0.454	0.279
DC-CM	0.432	0.657	0.408	0.280
COACH	0.518	0.711	0.477	0.285
CACHET	0.616	0.756	0.530	0.311

Figure 5.6 shows two examples to illustrate the organization of our predicted complexes. In (A), our identified core was actually reported in [Gavi 06] and the predicted complex also matched well with the histone deacetylase complex (all 8 proteins in our identified complexes were involved in the benchmark histone deacetylase complex). The protein YPL139C was identified as both bait-attachment and prey-attachment and it is required to bind the protein YNL330C in the core for its activity. The protein YNL097C, which is predicted as another bait-attachment, has been shown to be a possible component of the histone deacetylase complex [Gavi 02, Gavi 06]. In (B), the identified complex contained 14 proteins, which cover 10 of the 12 proteins in the benchmark DNA-directed RNA polymerase II complex [Cram 00]. The other four proteins in grey (YDL115C, YGR005C, YGR186C and YML010W) are novel to our benchmark complex. However, YDL115C has already been demonstrated to belong to the DNA-directed RNA polymerase II complex [Krog 06]. The other 3 proteins identified as attachments were all annotated with the term GO:0006366 (transcription from RNA polymerase II promoter), which is also closely related to the function of this complex. This example indicates that our CACHET may be used to detect the potential novel knowledge that are absent in current reference data. Our identified core with the 7 proteins was also validated in [Gavi 06].

5.3.5 Effect of The Parameter r

The reliability parameter r in Algorithm 4 determines the number and quality of protein-complex cores in our CACHET. Here we investigate how the variation of r affects the performance of our CACHET. Figure 5.7 shows the Precision, Recall and F-measure of CACHET on Gavin *et al.*'s data under different values of r .

As we increase the reliability value of r , fewer false positive bait-prey links are involved in our identified complexes, as expected. The predicted complexes are thus more likely to be true complexes, as reflected by ascending trend of the precision of CACHET when r becomes larger, as shown in Figure 5.7. On the other hand, when r becomes too large, CACHET may filter some true positive bait-prey interactions. As such, it will predict a small number of complexes with a low recall although it achieves a very high precision.

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Figure 5.6: The organization of our identified protein complexes. Each complex consists of four rows as shown in this figure. The two rows in the middle (the second and third rows with rectangles as proteins) form the protein-complex core and other two rows (with circles as proteins) are attachments. The second and third rows are baits and preys in the cores, respectively. Correspondingly, the first and fourth rows represent bait-attachments and prey-attachments, respectively. The predicted complexes are histone deacetylase complex in (A) and DNA-directed RNA polymerase II complex in (B).

Figure 5.7: Performance of our CACHET on Gavin *et al.*'s data under different values of r .

Overall, the resulting curve for the F-measure in Figure 5.7 has three ranges. It increases significantly in the first range ($r \in [0.55, 0.65]$), remains stable in the second range ($r \in [0.65, 0.8]$) and decreases drastically in the last range. As such, in our experiments, we set $r = 0.7$ on Gavin *et al.*'s data since it can effectively filter false positive interactions

while retaining most of true bait-prey relationships.

5.3.6 Impact of Indexing Reliability and Filtering Redundancy

Recall that indexing the reliability of bait-prey links to identify reliable bicliques (denoted as Step-I, please see Section 5.2.2) and filtering redundancy among reliable bicliques (denoted as Step-II, please see Section 5.2.3) are two key steps in CACHET. We have performed experiments to run CACHET without each step respectively. In Table 5.8, CACHET_{woI} and CACHET_{woII} refer to CACHET without Step-I and Step-II, respectively. CACHET_{woI} is observed to achieve a much lower precision than CACHET and CACHET_{woII}, indicating that detecting reliable bicliques can filter false positives effectively and thus increase the reliability of resulting protein complexes. Although CACHET_{woII} has the highest recall, it predicts protein complexes with huge redundancy (e.g., $4576/188 = 24.3$ predicted complexes match a real one on average). In summary, these two steps in CACHET are essential to filter potential false positive interactions and reduce redundancy among protein-complexes.

Table 5.8: Performance for CACHET on Gavin *et al.*'s data ($\omega = 0.2$)

Methods	# complexes	N_{cp}	N_{cb}	Precision	Recall	F-measure
CACHET _{woI}	3126	163	174	0.052	0.427	0.093
CACHET _{woII}	9862	4576	188	0.464	0.461	0.462
CACHET	369	250	180	0.678	0.441	0.534

5.3.7 Running CACHET on Krogan *et al.*'s Data

We used a pairwise PPI network containing 3275 proteins and 4500 interactions from Krogan *et al.*'s TAP data using SA scores as shown in Table 5.3. We then obtained the results for MCL, MCODE and COACH from this PPI network (denoted as Krogan-I network). Table 5.9 shows that the F-measure of our CACHET is 50.9%, which is 13%, 14.8%, 16.4% and 20% higher than COACH, MCODE, CODEC_{w1} and MCL, respectively. Although COACH and MCODE were able to achieve very high precisions, they did so with extremely low recalls due to their small numbers of predicted complexes. As such, their F-measures were much less than that of CACHET.

Krogan *et al.* [Krog 06] also provided a larger reliable PPI network with 3581 proteins and 14077 interactions, which is constructed from their original TAP data based on machine learning algorithms (denoted as Krogan-II network). Table 5.10 shows the comparison between our CACHET and existing methods. Our CACHET performs the best as it achieves the highest F-measure. In particular, MCODE achieves the highest

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Table 5.9: Performance for various methods on Krogan-I network

Methods	# complexes	Precision	Recall	F-measure
MCODE	160	0.606	0.257	0.361
MCL	981	0.215	0.547	0.309
CODEC w_0	8348	0.151	0.475	0.230
CODEC w_1	2973	0.312	0.387	0.345
COACH	192	0.698	0.260	0.379
CACHET	447	0.638	0.422	0.509

Table 5.10: Performance for various methods on Krogan-II network

Methods	# complexes	Precision	Recall	F-measure
MCODE	86	0.640	0.150	0.242
MCL	825	0.233	0.525	0.322
CODEC w_0	8348	0.151	0.475	0.230
CODEC w_1	2973	0.312	0.387	0.345
COACH	570	0.470	0.375	0.417
CACHET	447	0.638	0.422	0.509

precision but it also has the lowest recall, resulting in a low F-measure. MCL and CODEC w_0 have a high recall and very low precision and thus they also have a low F-measure. Our CACHET has both balanced precision and recall and thus achieves the highest F-measure.

5.3.8 Running CACHET on the Combined Data

Currently, Gavin *et al.* [Gavi 06] and Krogan *et al.* [Krog 06] have already provided comprehensive TAP data. However, they just produced quite different descriptions of the yeast interactome due to the fact that they exploited different purification techniques and utilized different number of baits in their purifications. Therefore, combining these two data sources will definitely improve the coverage of the data for better identifying protein complexes [Hart 07, Pu 07, Frie 08].

Assuming that $G_{gavin} = (U_g, V_g, E_g)$ and $G_{krogan} = (U_k, V_k, E_k)$ denote Gavin *et al.*'s and Krogan *et al.*'s data respectively, $G_{combined} = (U_c, V_c, E_c)$ is the resulting combined data, where $U_c = U_g \cup U_k$, $V_c = V_g \cup V_k$ and $E_c = E_g \cup E_k$. To index the reliability scores of edges in E_c , it is a possible solution to re-compute these scores based on the purification records in the combined data. Here, we adopt the strategy used in [Coll 07] to assign reliability scores for bait-prey links in the combined data based on their original scores in individual datasets. For an edge $e \in E_c$, $w_k(e)$ and $w_g(e)$ are its scores in Krogan *et al.*'s data and Gavin *et al.*'s data respectively. $w_c(e)$, its score in the combined data,

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

is calculated the same as PE scores by a weighted sum of $w_k(e)$ and $w_g(e)$ in Equation (5.10) instead of a straight sum, suggested by [Coll 07].

$$w_c(e) = 0.5 \times w_k(e) + w_g(e). \tag{5.10}$$

We compared our CACHET with existing methods, *i.e.*, MCODE, MCL, COACH, the bootstrapping method (BT) [Frie 08], Pu *et al.*'s method [Pu 07] and Hart *et al.*'s method [Hart 07]. Since last three methods are independent of scoring schemes, we first show the F-measure of CACHET, MCODE, MCL and COACH using various scoring schemes in Figure 5.8. The result is similar to the case on Gavin *et al.*'s data. In terms of using individual scoring schemes, CACHET achieves the highest F-measure when using SA and DC scores and performs equally well as MCL when using PE scores. In particular, CACHET with SA scores achieves the highest F-measure than other methods with any scoring schemes. For example, CACHET with SA scores has a F-measure 62.2%, 6.2% higher than MCL with PE scores, which is the best performer by combining existing methods and scoring schemes.

Figure 5.8: Comparison among various methods on the combined data using different reliability schemes.

CACHET with SA scores predicted 489 protein complexes from the combined data (using the same reliability threshold 0.7). Compared with the results from individual data, CACHET achieved higher Precision, Recall and F-measure on the combined data as shown in Table 5.11, indicating that CACHET can effectively exploit the biological knowledge from multiple data. Table 5.11 shows the comparison details between our

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

CACHET using SA scores and other methods on the combined data. As shown in Table 5.12, while MCL, BT, Pu and Hart obtained relatively high recalls, our CACHET gained much higher Precision, and thus obtained the highest overall F-measure. The higher Precision indicates that the complexes predicted by CACHET are more accurate and also implies that our unmatched complexes are more likely to be novel true complexes. We also showed the F-measures of the above methods with different values of ω used in Figure 5.9 and observed that our CACHET consistently achieved the best overall performance even with more stringent matching criteria.

Table 5.11: CACHET on combined data using SA scores ($\omega = 0.2$).

Data	# complexes	N_{cp}	N_{cb}	Precision	Recall	F-measure
Gavin	369	250	180	0.678	0.441	0.534
Krogan	447	285	172	0.638	0.422	0.509
Combined	489	382	211	0.781	0.517	0.622

Table 5.12: Performance of various methods using SA scores on combined data ($\omega = 0.2$).

Method	# complexes	N_{cp}	N_{cb}	Precision	Recall	F-measure
CACHET	489	382	211	0.781	0.517	0.622
BT	409	204	244	0.499	0.598	0.544
Pu	400	201	241	0.503	0.591	0.543
Hart	390	203	242	0.521	0.593	0.554
MCODE	265	123	149	0.462	0.365	0.408
MCL	1076	236	270	0.219	0.662	0.329
COACH	439	221	174	0.502	0.426	0.462

From the perspective of functional homogeneity, we also show the proportion of statistically significant complexes from the combined data under different values of α in Figure 5.10. Our CACHET is observed to consistently achieve the highest proportion of statistically significant complexes. For example, when α is set as 2, CACHET has 89.8% significant complexes, 6.4%, 18% and 23.3% higher than COACH, Hart and Pu (top-3 performers), respectively.

5.3.9 Protein-Level Evaluations

Table 5.13 shows the comparison among various methods based on protein-level evaluation metrics in equations 5.8 and 5.9. To run MCODE, MCL and COACH, the PPI networks are obtained using SA scores (please see Table 5.3 for details). We can observe that CACHET and COACH achieved similar F-measure and performed better than other

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Figure 5.9: Performance of our CACHET on Gavin *et al.*'s data under different values of ω .

Figure 5.10: Performance of our CACHET on Gavin *et al.*'s data under different values of ω .

methods on both two datasets. In addition, Table 5.14 shows the comparison among various methods on the combined data and CACHET also obtained the highest Precision and F-measure.

CHAPTER 5. CACHET: DETECTING PROTEIN COMPLEXES DIRECTLY FROM TAP DATA

Table 5.13: Protein-level evaluation for various methods

Methods	Gavin <i>et al.</i> 's data			Krogan <i>et al.</i> 's data		
	P-Precision	P-Recall	F-measure	P-Precision	P-Recall	F-measure
MCODE	0.460	0.302	0.365	0.602	0.220	0.322
MCL	0.382	0.432	0.406	0.213	0.395	0.277
DC-CM	0.212	0.472	0.293	-	-	-
CODEC w_0	0.087	0.432	0.145	0.039	0.437	0.072
CODEC w_1	0.151	0.464	0.228	0.091	0.394	0.148
COACH	0.443	0.435	0.439	0.704	0.270	0.390
CACHET	0.468	0.415	0.440	0.423	0.374	0.397

Table 5.14: Protein level evaluation for various methods using SA scores on combined data.

Method	# complexes	P-Precision	P-Recall	F-measure
CACHET	489	0.739	0.472	0.576
BT-409	409	0.482	0.575	0.524
Pu-400	400	0.439	0.576	0.498
Hart-390	390	0.503	0.547	0.524
MCODE	265	0.402	0.391	0.396
MCL	1076	0.319	0.573	0.410
COACH	439	0.314	0.5	0.386

5.4 Conclusions

In this chapter, we have proposed a novel algorithm CACHET to detect protein complexes with core-attachment structures from TAP data. CACHET first computes all the maximal bicliques from the TAP bipartite graph. It then detects high quality protein-complex cores from the computed maximal bicliques, by extracting reliable bicliques and removing redundancy among them. Finally, CACHET forms protein complexes by incorporating attachments into cores.

CACHET significantly outperformed existing methods in terms of F-measure and functional homogeneity of predicted complexes. Experimental evaluation on our identified protein complexes and cores also demonstrated the following advantages of our CACHET. First, as CACHET detects protein complexes directly from TAP data, it is able to properly exploit the co-complex information in the TAP data and avoids the step of converting the TAP data into a binary PPI network. Second, the protein complexes predicted by CACHET are with core-attachment structures, which provide useful information for revealing the inherent organization of the detected protein complexes.

Chapter 6

Conclusions, Future Work and Discussions

Protein complexes play important roles in cellular organizations and functions. To detect protein complexes more accurately, we consider their core-attachment structures and present methods that detect cores first and then attachments to form complexes. Our inside-out methods provide insight into their functional organizations and also predict protein complexes more accurately compared with existing methods. To conclude this thesis, we will first briefly review our main contributions. We then have some recommendations for further studies on disease-related protein complex detection, which may have many practical usefulness in disease diagnostic, drug discovery and gene therapy etc.

6.1 Review of Main Contributions

In this thesis, we focus on detecting protein complexes from protein interaction data. We also present an approach to identify reliable interactions for improving the quality of our predicted protein complexes. We summarize our main contributions as follows.

First, we propose two novel algorithms to detect core-attachment structures for two different kinds of protein interaction data. COACH is devised to detect protein complexes from PPI networks while CACHET is designed to detect them from TAP bipartite graphs. The main difference between our methods and previous studies [Gavi 06, Zhan 08] is that they detect protein complexes first and then identify cores while we detect protein-complex cores first and then select attachments to form protein complexes. After collecting all protein complexes, they [Gavi 06, Zhan 08] evaluate the significance of each protein in the complexes based on the redundancy of their data sources and select significant proteins as core proteins. In [Zhan 08], complexes are first detected by clique merging algorithm and core proteins of a given complex are then defined as those proteins involved in more than $\frac{2}{3}$ of the cliques whose merging results in this complex. However,

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

the predicted cores formed by significant proteins may not be significant any more. For example, those predicted cores may not be reliable (*i.e.*, some inner-edges are with low reliability) or dense (*i.e.*, the core is with low inner-connectivity). Contrarily, we adopt an inside-out strategy to predict protein complexes. We first detect candidate protein-complex cores from local neighborhood graph. We subsequently evaluate the significance of these candidate cores as well as filtering those redundant candidates. Finally, we predict those non-redundant candidates with high significance scores as protein-complex cores. As such, our predicted protein-complex cores are more significant than those predicted by previous studies [Gavi 06, Zhan 08]. The quality evaluations of protein-complex cores by functional information and gene expression profiles also demonstrated that cores predicted by our inside-out strategies are more significant than those by previous studies as shown in Table 5.7. Besides, our detected protein complexes are more accurate than those predicted by existing state-of-the-art algorithms through comprehensive evaluations.

Second, we propose different measures to evaluate the significance of candidate cores for different methods. In COACH, the significance of a candidate core is mainly determined by its density and size. That is intuitive since larger candidates with higher density tend to be more important and significant. In CACHET, the significance score of a candidate core is defined in a more complicated manner. We take both its edge weights and node weights into considerations. The edge weights represent their reliability scores while the node weights indicate the importance of proteins based on the frequency of these proteins involved in all detected reliable bicliques. In comparison, the significance score in CACHET are statistically and biologically more meaningful than that in COACH. This point is also indicated in our Table 5.7 where cores predicted by CACHET have higher quality than those by COACH.

Third, the quality of predicted protein complexes is highly related to the reliability of source protein-protein interactions. Therefore, we propose a method to identify reliable interactions for improving the quality of our predicted protein complexes. We exploit well-known classification models, such as SVM and Bayesian classifiers, to identify reliable interactions by integrating both topological and genomic evidences. As such, we can expand current PPI network by adding reliable false negatives or/and purify it by deleting false positives. We then detect protein complexes from the modified PPI network. We observe that deleting false positives will result in complexes that are more accurate (achieve higher precision, please see the evaluation metrics in section 2.4) and adding reliable false negatives will increase the coverage of predicted complexes and also predict more real complexes (obtain higher recall and coverage rate).

Fourth, CACHET provides an easier way to detect protein complexes directly from TAP data. We model TAP data as a bipartite graph, where two node sets are baits and preys and edges represent bait-prey relationships. In traditional PPI networks, we

detect protein complexes based on the observation that densely connected subgraphs are likely to be protein complexes. Here in TAP bipartite graphs, we similarly consider that densely connected bipartite subgraphs as protein complexes. In particular, CACHET will collect reliable bicliques after computing all maximal bicliques. Subsequently, non-redundant, significant, reliable bicliques will be selected as protein-complex cores by maximum weighted independent set algorithms. Finally, protein complexes are formed by combining core proteins and attachments. CACHET predicts protein complexes by processing collected bicliques and thus is much more convenient than traditional methods which need to convert TAP data to PPI networks.

6.2 Future Work

Currently, many methods can generate a set of protein complexes on a PPI dataset. Therefore, we can generate a large amount of protein complexes using these methods from different PPI datasets. As such, analyzing these redundant complexes or ensemble clustering is a possible direction to exploit the superiorities of each method and generate complexes with higher accuracy and better coverage [Asur 07].

Proteins within the same protein complexes tend to have similar biological patterns, e.g., functions and knockout phenotypes. Therefore, protein complexes can be computationally used for predicting protein functions [Alta 06], predicting phenotypic effects of genes [Fras 07] and so on. For example, if there is a function-unknown protein within a protein complex, it will be predicted with a function which is shared by other proteins in this protein complex [Alta 06]. In [Fras 07], the best predictor of a given protein's knockout phenotype is believed to be the knockout phenotype of other proteins that are present in the same protein complex as the given protein.

The benchmark complexes are shown with high functional homogeneity, e.g., 94.4% of the 408 protein complexes in CYC2008 catalogue are with p-value (please refer to the Equation 2.12 in Section 2.4.3) less than 0.01. The predicted complexes of our COACH [Wu 09a] and CACHET [Wu 11] also have high functional homogeneity. For example, 83.4% of the complexes predicted by COACH on DIP data and 90.0% of the complexes predicted by CACHET on the combined TAP data are with p-value less than 0.01. In addition, the protein complexes predicted by our COACH and CACHET match very well with the benchmark complexes using various evaluation measures including precision, recall and F-measure. Therefore, our predicted complexes can also be used for the above purposes.

Besides above applications of our predicted protein complexes, we are now targeting to exploit how we can utilize them for other further studies. Recently, we exploited the underlying modular organization of the protein interactome for prediction and prioritization of disease genes [Yang 11]. We first built a protein-complex network where nodes

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

were protein-complexes and edges represented the overlap between complexes. We then predicted disease genes based on this protein-complex network instead of simple PPI networks and our method was found to be able to detect and prioritize disease genes better due to that our protein-complex network can better capture the modularity in the biological interaction networks. In this section, we will still focus on identifying disease-related genes and protein complexes, which is more closely related to our human and human health. We will briefly review existing works on this topic in this section and then have some recommendations for this topic in the coming section. Note that we use gene and protein (gene product) interchangeably in this section.

6.2.1 Neighborhood Counting

Physically interacting proteins tend to be involved in the same biological process and the mutations in their corresponding genes may lead to the same or similar phenotypes. Based on this observation, Oti *et al.* [Oti 06] proposed a simple method to predict disease genes based on protein-protein interaction data.

Given a phenotype p and a protein g , if g 's interacting partners are p 's causing genes, g will be predicted as a candidate causing gene of p . In the field of protein function prediction, there is a well-known method called *Neighborhood Counting* (NC) that believes a protein will have a function if many of its neighbors have this function [Schw 00]. Therefore, we also call Oti's method as NC here. Linghu *et al.* [Ling 09] also extended this NC method to prioritize disease genes. Given a gene g and a phenotype p , their association score is defined as follows.

$$s(p, g) = \sum_{g' \in N(p)} W(g, g'), \quad (6.1)$$

where $N(p)$ is the set of known disease genes for the phenotype p and $W(g, g')$ is the functional similarity score between g and g' (the edge weight in their built functional-linkage network). They integrated 16 genomic features (32 sub-features) and built a naive Bayes classifier to collect the functional similarity scores among 21,657 proteins. Gene-phenotype pairs with high association scores will then be predicted as reliable gene-phenotype associations. If we regard phenotypes as the functional annotations of genes, identifying disease genes is the same as predicting protein functions. Therefore, many approaches for function prediction (*e.g.*, using indirect neighbors for function prediction [Chua 06]) can be also applied for detecting disease genes.

6.2.2 Random Walk with Restart (RWR)

Random walk with restart (RWR) [Kole 08] simulates a random walker, either starts on a seed node or on several seed nodes and moves to their direct neighbors randomly at each step. Finally, RWR will reach a steady state and each node can then be ranked by the probability that the random walk can reach this node. Therefore, we can consider that RWR is propagating the influence from the seed nodes to the remaining nodes in the graph or measures the proximity of other nodes to the seed nodes.

Let p_0 be the initial probability vector and p_t be a vector in which the i -th element shows the probability of that the random walker is at node i at step t . The probability vector at step $t + 1$ is as follows.

$$p_{t+1} = (1 - \gamma) \times M \times p_t + \gamma \times p_0, \quad (6.2)$$

where M is the transition matrix of the graph and each element M_{ij} is the transition probability from node i to node j . We generally use the normalized adjacency matrix as the transition matrix.

Given a phenotype, $p_0(j)$ is set as 1 if node j is a disease gene for this phenotype and $p_\infty(i)$ is the probability of node i also being the disease gene. We can obtain the probability vector at the steady state (p_∞) with the following two methods. First, mathematically we can have that $p_\infty = \gamma \times (I - (1 - \gamma) \times M)^{-1} \times p_0$ if the matrix $(I - (1 - \gamma) \times M)$ has an inverse. Second, practically the power method can get p_∞ efficiently by performing iterations until the difference between p_{t+1} and p_t is below a threshold, for example, 10^{-10} .

6.2.3 CIPHER

Wu *et al.* [Wu 08b] proposed a method called CIPHER to prioritize disease genes integrating the following networks, namely, human phenotype network, protein network and gene-phenotype network.

CIPHER calculates the concordance score between a given protein and a target phenotype. A gene that has a high concordance score with a target phenotype will be predicted as a disease gene of this phenotype. In particular, we assume that there are n phenotypes and we have a gene g and a phenotype p . The phenotype similarity profile for p is a n -dimension vector $S(p) = (s(p_1, p), \dots, s(p_n, p))$, where $s(p_i, p)$ is the similarity between phenotypes p_i and p calculated through text mining [Drie 06]. The gene closeness profile for g is also a n -dimension vector $\Phi(g) = (\phi(p_1, g), \dots, \phi(p_n, g))$, where $\phi(p_i, g)$ is the similarity between gene g and the phenotype p_i defined as follows.

$$\phi(p_i, g) = \sum_{g' \in N(p_i)} \exp\{-L^2(g, g')\}, \quad (6.3)$$

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

where $N(p_i)$ is the set of genes connecting to the phenotype p_i in the gene-phenotype network. $L(g, g')$ is the topological distance between two genes in the protein network. CIPHER has two versions: (1) CIPHER-SP defines $L(g, g')$ as the length of the shortest path between g and g' , and (2) CIPHER-DN is a modified version of CIPHER-SP where $L(g, g') = \infty$ for indirect neighbors. Finally, the concordance score between the gene g and phenotype p is the Pearson correlation between $S(p)$ and $\Phi(g)$.

6.2.4 Random Walk with Restart on Heterogeneous Networks (RWRH)

RWR algorithm [Kole 08] only propagates the influence in the protein interaction network. Li *et al.* [Li 10b] extended RWR to RWRH that walks on a heterogeneous network, which combined protein interaction network, phenotype network and gene-phenotype network. The only difference between RWR and RWRH is that they used different networks. Therefore, we still iterate the equation 6.2 to reach the steady state although the transition matrix here is different. The transition matrix now is $M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_P \end{bmatrix}$, where M_G and M_P are protein interaction network and phenotype network, M_{GP} describes the gene-phenotype relationships (M_{PG} is its reverse). Given a phenotype, RWR can not rank genes if this phenotype has no known causing genes. RWRH can also predict disease genes for this phenotype if its phenotype neighbors still have known causing genes. In addition, Hwang *et al.* [Hwan 10] also proposed a label propagation algorithm working the same heterogeneous network.

6.2.5 PRINCE

Oron *et al.* recently proposed a method called PRINCE [Vanu 10] to associate genes and protein complexes with diseases. Given a phenotype, PRINCE first computes the prioritization scores for all the genes and then predicts those genes with high scores as candidate causing genes of this phenotype. The prioritization scores are calculated via network propagation which is same as the RWR algorithm in equation 6.2. However, PRINCE defines the initial vectors in a different manner. In RWR, given a phenotype x , the initial probability of gene g , $p_0(g)$ is 1 if g is a causing gene of x or 0 otherwise. In PRINCE, $p_0(g)$ will be a non-zero value if g is a causing gene of another phenotype which is similar to x . In particular, $p_0(g)$ in PRINCE is defined as follows.

$$p_0(g) = L(S(x, y)), \quad (6.4)$$

$$L(m) = \frac{1}{1 + \exp\{cm + d\}}, \quad (6.5)$$

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

where $S(x, y)$ is the similarity score between phenotypes x and y and c and d are parameters in the logistic function L .

Given a phenotype and the prioritization scores for proteins in the PPI network, PRINCE will subsequently detect the protein complexes associated with this phenotype. PRINCE starts with top 100 scoring proteins as complex seeds and then filters all the seeds with score less than a predefined threshold or removes those already predicted as members of previously detected complexes. For each seed, its high scored neighbors will be added. To this end, a initial list of putative complexes are collected. Next, PRINCE will refine the complexes to guarantee that all the member proteins are densely connected. PRINCE is the first study to detect disease-related protein complexes by grouping genes causing the same or similar diseases.

6.2.6 AlignPI

Wu *et al.* [Wu 09b] proposed an alignment-based framework called AlignPI to predict causative genes and disease-related protein modules. After collecting human protein interaction networks (*e.g.*, Human Protein Reference Database) and phenotype network, AlignPI applied the network comparison toolkit (<http://chianti.ucsd.edu/nct/index.php>) to identify bi-modules, which consist of disease modules and protein modules connected by the gene-disease association linkages. AlignPI is a good attempt to perform heterogeneous alignment for human interactions and phenome although it still has following limitations. First, the linkages connecting phenome network and protein network are still rather limited. Therefore, the number of detected bi-modules are limited, too. For example, AlignPI only obtained 39 bi-modules in their experiments. The coverage of bi-modules is thus very questionable. Second, integrating diverse sources to build a larger functional linkage network for proteins would be very helpful for identifying disease genes and complexes [Ling 09]. However, it would not be helpful in AlignPI framework because the bottleneck of AlignPI is still the limited number of gene-disease association linkages.

6.2.7 RegMOD

Qiu *et al.* [Qiu 10] devised a method called RegMOD to prioritize disease genes and detect disease-associated modules by integrating protein interaction data and gene expression data. RegMOD has the following three steps.

First, RegMOD calculates the active scores for proteins which measure the extent of genes responding to a specific disease. In particular, given a gene x , its expression profile can be divided into two sample sets and its active score, $w(x)$, can be calculated by the Signal to Noise Ratio (SNR) as follows.

$$w(x) = \frac{\mu_{x1} - \mu_{x2}}{\sigma_{x1} + \sigma_{x2}}, \quad (6.6)$$

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

where μ_{x1} and μ_{x2} are the means of the expression levels of gene x in two sample sets and σ_{x1} and σ_{x2} are the standard deviations.

Second, RegMOD also defines an active function to predict the underlying active scores for proteins based on the topological properties of the PPI network. Given a gene x , its predicted active score, $f(x)$, is defined as follows.

$$f(x) = \sum \beta_y k(x, y) + b, \quad (6.7)$$

where $k(x, y)$ is the kernel function to describe the similarity between genes x and y . The diffusion graph kernel is finally adopted in this paper for node similarities.

Third, support vector regression (SVR) is formulated for fitting the predicted active scores (*i.e.*, $f(x)$) to the observed scores (*i.e.*, $w(x)$) and estimating the parameters (*i.e.*, β) in equation 6.7. After f is fitted, each gene is assigned a new active score. RegMOD then extracts subgraphs consisting of nodes with high active scores as active modules.

In this paper, the kernel function is applied to measure the node similarities. Actually, the weight between two nodes, which is the Pearson correlation coefficient of their gene expression profiles, is already showing their functional similarity. The following two questions should be asked. How much does it benefit from using the complicated kernel function rather than the original edge weights? How much would it benefit from integrating more sources (*e.g.*, domain, sequence, even functional GO data and so on) rather than gene expression data?

6.3 Discussions

We next summarize existing works for identifying disease-related genes and protein complexes and list some key issues involved in detecting disease-related protein complexes to complete this thesis.

6.3.1 Summary

All above studies for identifying disease genes are based on the reasonable assumption that functionally similar proteins may result in the same phenotype or disease. Then all these methods just transfer the information from the neighborhoods [Oti 06, Ling 09, Wu 08b], indirect neighborhoods via shortest paths [Wu 08b] and even the proteins in the whole protein network via flow propagations [Kole 08, Li 10b, Hwan 10, Vanu 10] to the target protein.

We can thus divide these studies for identifying disease genes into two categories. The first category, denoted as “1-to-1”, refers to checking whether a gene is a causing gene of a given phenotype. Furthermore, the information will only be propagated from

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

those genes causing the given phenotype to our target gene in these methods. NC, RWR and CIPHER are in this category. The other category, denoted as “m-to-1”, means that multiple phenotypes will be taken into consideration for checking whether a gene is a causing gene of a given phenotype. As we know, the mutation of functionally similar genes would results in the same or similar phenotypes. Therefore, it makes more sense to propagate the information from those genes causing similar phenotypes of the given phenotype to our target gene. As such, “m-to-1” methods are reasonably expected to obtain better results than “1-to-1” methods. RWRH and PRINCE are in this category.

“1-to-1” and “m-to-1” methods focus on prioritizing independent genes. However, mutations in different loci may lead to the same disease in many cases. This genetic heterogeneity may reflect the underlying molecular mechanism that disease-causing genes form some kind of functional modules [Lage 07, Lage 08]. As such, the latest trend becomes to capture the linkage between a group of phenotypes (a phenotype module) and a group of proteins (a protein module). This type of studies currently include PRINCE, AlignPI and RegMOD, and are here denoted as “m-to-m”.

6.3.2 Possible Solutions

The existing “m-to-m” methods have their own limitations. Given a phenotype, PRINCE will select its several nearest neighbors in the phenotype network and then map these phenotypes to PPI network for protein complexes. Strictly speaking, these phenotypes just form a local neighborhood graph (please see the definition in section 2.3) and may not be a phenotype module. In addition, the number of detected bi-modules by AlignPI is very limited and we need such bi-modules with a better coverage. Therefore, we need more accurate extraction of phenotype modules in our new “m-to-m” methods.

Biological networks are modular, *i.e.*, they have many modules which refer to groups of physically or functionally linked molecules (*e.g.*, proteins, genes and so on) working together for distinct functions. Furthermore, the modularity in these networks is believed to be the main contributor to the robustness and stability of biological systems [Bara 04, Hint 08]. As we know, protein interaction network involves protein complexes (protein modules) and gene co-expression network has co-expression modules. Phenotype network is also revealed to contain phenotype modules with well-characterized modularity that is highly correlated with current diseases [Jian 08]. Therefore, it will provide deep insights into human diseases and the organization of phenotypes by clustering the phenotype network and decomposing it into phenotype modules.

6.3.2.1 How to Cluster the Phenotype Network

Van Driel *et al.* obtained pairwise similarity scores for 5080 phenotypes in OMIM database by text mining [Drie 06]. In particular, each phenotype is characterized by

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

a vector of standardized and weighted phenotypic feature terms using MeSH terms from OMIM database. The similarity score between two phenotypes is then defined as the cosine similarity between their feature vectors. Therefore, clustering methods, for example, hierarchical clustering, can be applied directly to generate overlapping or non-overlapping clusters (phenotype modules) from this 5080×5080 score matrix.

We can also build a smaller phenotype network by removing pairs with similarity scores less than a pre-defined threshold. We can similarly exploit those methods that are devised for protein-complexes in PPI networks (*e.g.*, graph clustering approaches based on network topology in section 2.3.3.1) to detect phenotype modules in the phenotype network. In addition, we should investigate the topological properties and then adjust the parameters involved in these methods before applying them. For example, if the phenotype network is denser than a PPI network, the optimal inflation parameter should be set larger than that in the PPI network [Enri 02]. In addition, we should also investigate the effect of varying the similarity threshold.

There is another issue to be discussed. Current similarity scores obtained by van Driel *et al.* are not perfect although pervious studies have verified the reliability of these scores [Lage 07, Drie 06]. We may further refine the current similarity scores to obtain more precise results. For example, novel phenotype-phenotype associations can be validated if they share common disease genes [Ling 09, Goh 07]

6.3.2.2 How to Evaluate Predicted Phenotype Modules

It is an important step to develop various methods for phenotype modules. However, it is more important that detected phenotype modules are of high quality. As such, measures for evaluating predicted modules are needed.

Recall that p-values of protein complexes in equation 2.12 in section 2.4 demonstrate their functional homogeneity based on the functional information of proteins within them. A similar method is presented to evaluate the quality of phenotype modules in [Jian 08]. There are 21 disease classes in [Goh 07] and each phenotype is assigned in at least one disease class. Given a phenotype module, we can thus calculate its enrichment with respect to a disease class using equation 2.12. A phenotype module with a low p-value also indicates that it is enriched by phenotypes from the same disease class and thus it is believed to have high quality. An obvious limitation of this measure is that the number of disease classes is too small and these disease classes are too general. As the disease classes become more specific, this measure would become more effective.

6.3.2.3 How to Map Phenotype Modules to Protein Complexes

After we collected phenotype modules, we have the following two strategies to detect disease-related protein complexes in human protein interaction network.

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

The first strategy is called “clustering-then-mapping”. We first decompose human PPI network into protein clusters (*i.e.*, protein complexes). We then define an association score between a phenotype module and a protein complex based on the phenotype-genes associations connecting them. Given a phenotype module, we can thus select its most associated (or top- k) protein complexes as its candidate causing modules. Generally, a candidate module of a phenotype module can be further refined by including those proteins that are not assigned into it but closely associated with it as well as the phenotype module.

Another strategy is called “mapping-then-clustering”. Given a phenotype module, we assign a score to each protein indicating their closeness to this phenotype module. In node-weighted PPI network, many clustering algorithms, for example, DPCLUS [Alta 06], can detect modules (*i.e.*, protein complexes) with high edge density and node weights as candidate causing modules.

A key step in the above strategies is to define association scores for proteins or protein complexes with respect to a given phenotype module. Intuitively, we would define the association scores based on the number of known phenotype-gene relationships linking a protein or a protein complex to the phenotype module. However, we currently have a limited number of known phenotype-gene associations. Therefore, predicting novel disease genes will increase the number of phenotype-gene relationships for better identifying disease-related protein complexes. Meanwhile, disease-related protein complexes may also be exploited to predict new disease genes. This iterative process would provide us a clearer elucidation of gene-phenotype relationships for clinic disease to a greater extent.

How to cluster human PPI networks is another key issue. The topological properties of human PPI networks should be investigated before clustering them. In addition, a single PPI network may be not enough for detecting protein complexes with good coverage. For example, HPRD database used in [Wu 09b] contains only 8,919 genes while there are more than 20,000 genes in human genome. Therefore, it would provide us protein complexes with better coverage and higher quality by combining diverse human PPI networks or constructing larger human protein network via integrating diverse genomic features [Ling 09].

6.3.2.4 How to Evaluate Disease-Related Protein Complexes

It is important to design computational methods to cluster human protein networks for protein modules. However, it is more important to evaluate the quality of these detected protein modules. As such, people would know which methods can generate protein complexes with better quality.

Currently, it is easy for us to evaluate predicted human protein complexes. On one hand, we already have a reference database for human protein complexes, called

CHAPTER 6. CONCLUSIONS, FUTURE WORK AND DISCUSSIONS

CORUM [Ruep 07]. We can similarly evaluate predicted human protein complexes by mapping them to CORUM complexes using existing measures, such as Precision, Recall, F-measure, Sensitivity, PPV and accuracy introduced in our section 2.4. On the other hand, we also have comprehensive Gene Ontology annotations for human proteins [Cons 06]. Therefore, functional homogeneity (p-values) in Section 2.4 can also be utilized for evaluation.

However, it is a bit hard to evaluate disease-related protein complexes. To meet this challenge, we plan to employ the following strategies. First, proteins within disease complexes should also share common functions and have correlated gene expression profiles. Therefore, we should also evaluate the functional homogeneity and expression coherency for disease complexes. Second, although we don't have a well-established database for disease complexes, we can also collect those disease complexes with high quality predicted by previous studies [Lage 08, Vanu 10] for evaluation. For example, there are 351 non-redundant disease complexes in [Lage 08]¹. In addition, 139 disease complexes with high quality are reported in [Vanu 10]. It is now possible to evaluate our predicted disease complexes by mapping them to these previously predicted ones.

¹http://www.cbs.dtu.dk/suppl/dgf/disease_complexes/index.php

References

- [Abdu 98] M. A. Abdulrahim and M. Misra. “A graph isomorphism algorithm for object recognition”. *Pattern Analysis and Applications*, Vol. 1, No. 3, pp. 189–201, 1998.
- [Adam 06] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek. “CFinder: Locating Cliques and Overlapping Modules in Biological Networks”. *Bioinformatics*, Vol. 22, No. 8, pp. 1021–1023, 2006.
- [Alex 09] R. P. Alexander, P. M. Kim, T. Emonet, and M. B. Gerstein. “Understanding Modularity in Molecular Networks Requires Dynamics”. *Science Signaling*, Vol. 2, No. 81, p. 44, 2009.
- [Alon 95] N. Alon, R. Yuster, and U. Zwick. “Color-Coding”. *Journal of ACM*, Vol. 42, No. 4, pp. 844–856, 1995.
- [Aloy 04] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. R.B. “Structure-Based Assembly of Protein Complexes in Yeast”. *Science*, Vol. 303, No. 5666, pp. 2026–2029, 2004.
- [Alta 06] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya. “Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks”. *BMC Bioinformatics*, Vol. 7, p. 207, 2006.
- [Asur 07] S. Asur, D. Ucar, and S. Parthasarathy. “An Ensemble Framework for Clustering Protein-Protein Interaction Networks”. In: *17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB)*, pp. 29–40, 2007.

REFERENCES

- [Ausi 99] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation*. Springer Verlag, 1999.
- [Bade 02] G. D. Bader and C. W. Hogue. “Analyzing Yeast Protein-Protein Interaction Data Obtained from Different Sources”. *Nature Biotechnology*, Vol. 20, No. 10, pp. 991–997, 2002.
- [Bade 03a] G. Bader and C. Hogue. “An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks”. *BMC Bioinformatics*, Vol. 4, p. 2, 2003.
- [Bade 03b] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue. “BIND: the Biomolecular Interaction Network Database”. *Nucleic Acids Research*, Vol. 31, No. 1, pp. 248–250, 2003.
- [Bade 04] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. “Gaining Confidence in High-Throughput Protein Interaction Networks”. *Nature Biotechnology*, Vol. 22, No. 1, pp. 78–85, 2004.
- [Balc 95] D. Balciunas and H. Ronne. “Three Subunits of the RNA Polymerase II Mediator Complex are Involved in Glucose Repression”. *Nucleic Acids Res*, Vol. 23, No. 21, pp. 4421–4425, 1995.
- [Bara 04] A. L. Barabasi and Z. N. Oltvai. “Network Biology: Understanding the Cell’s Functional Organization”. *Nature Reviews Genetics*, Vol. 5, No. 2, pp. 101–113, 2004.
- [Barr 09] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muerter, and R. Edgar. “NCBI GEO: Archive for High-Throughput Functional Genomic Data”. *Nucleic Acids Research*, Vol. 37, No. Database-Issue, pp. 885–890, 2009.
- [Bata 03] V. Batagelj and M. Zaversnik. “An $O(m)$ Algorithm for Cores Decomposition of Networks”. *CoRR*, Vol. cs.DS/0310049, 2003.
- [Bell 98] W. Bell, W. Sun, S. Hohmann, S. Wera, A. Reinders, C. De Virgilio, A. Wiemken, and T. J. M. “Composition and Functional Analysis of the *Saccharomyces Cerevisiae* Trehalose Synthase Complex”. *Journal of Biological Chemistry*, Vol. 273, pp. 33311–33319, 1998.

REFERENCES

- [Ben 05] A. Ben-Hur and W. S. Noble. “Kernel Methods for Predicting Protein-Protein Interactions”. In: *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 38–46, 2005.
- [Ben 06] A. Ben-Hur and W. S. Noble. “Choosing Negative Examples for the Prediction of Protein-Protein Interactions”. *BMC Bioinformatics*, Vol. 7(Suppl 1):S2, 2006.
- [Berg 04] J. Berg, M. Lassig, and A. Wagner. “Structure and Evolution of Protein Interaction Networks: A Statistical Model for Link Dynamics and Gene Duplications”. *BMC Evolutionary Biology*, Vol. 4, p. 51, 2004.
- [Bhar 05] N. Bhardwaj and H. Lu. “Correlation between Gene Expression Profiles and Protein-Protein Interactions within and across Genomes”. *Bioinformatics*, Vol. 21, No. 11, pp. 2730–2738, 2005.
- [Boyl 04] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. “GO::TermFinder—Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes”. *Bioinformatics*, Vol. 20, No. 18, pp. 3710–3715, 2004.
- [Broh 06] S. Brohee and J. van Helden. “Evaluation of Clustering Algorithms for Protein-Protein Interaction Networks”. *BMC Bioinformatics*, Vol. 7, p. 488, 2006.
- [Bu 03] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. “Topological Structure Analysis of the Protein-Protein Interaction Network in Budding Yeast”. *Nucleic Acids Res*, Vol. 31, No. 9, pp. 2443–2450, 2003.
- [Chen 06] J. Chen, W. Hsu, M. Lee, and S. Ng. “Increasing Confidence of Protein Interactomes using Network Topological Metrics”. *Bioinformatics*, Vol. 22, No. 16, pp. 1998–2004, 2006.
- [Chen 08] P.-Y. Chen, C. M. Deane, and G. Reinert. “Predicting and Validating Protein Interactions Using Network Structure”. *PLoS Comput Biol*, Vol. 4, No. 7, 07 2008.

REFERENCES

- [Cho 07] Y. Cho, W. Hwang, M. Ramanathan, and A. Zhang. “Semantic Integration to Identify Overlapping Functional Modules in Protein Interaction Networks”. *BMC Bioinformatics*, Vol. 8, p. 265, 2007.
- [Chua 06] H. N. Chua, W. K. Sung, and L. Wong. “Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions”. *Bioinformatics*, Vol. 22, No. 13, pp. 1623–1630, 2006.
- [Chua 07a] H. N. Chua, K. Ning, W. K. Sung, H. W. Leong, and L. Wong. “Using Indirect Protein-Protein Interactions for Protein Complex Prediction”. In: *International Conference on Computational Systems Bioinformatics (CSB)*, pp. 97–109, 2007.
- [Chua 07b] H. N. Chua, W.-K. Sung, and L. Wong. “An Efficient Strategy for Extensive Integration of Diverse Biological Data for Protein Function Prediction”. *Bioinformatics*, Vol. 23, No. 24, pp. 3364–3373, 2007.
- [Cole 83] T. F. Coleman and J. J. More. “Estimation of Sparse Jacobian Matrices and Graph Coloring Problems”. *SIAM Journal on Numerical Analysis*, Vol. 20, No. 1, pp. 187–209, 1983.
- [Coll 07] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan. “Toward a Comprehensive Atlas of the Physical Interactome of *Saccharomyces Cerevisiae*”. *Mol Cell Proteomics*, Vol. 6, No. 3, pp. 439–450, 2007.
- [Cons 06] G. O. Consortium. “The Gene Ontology (GO) Project in 2006”. *Nucleic Acids Res*, Vol. 34, No. Database issue, pp. 322–326, 2006.
- [Cout 07] F. Couto, M. Silva, and P. Coutinho. “Measuring Semantic Similarity between Gene Ontology Terms”. *Data Knowl. Eng.*, Vol. 61, No. 1, pp. 137–152, 2007.
- [Cram 00] P. Cramer, D. A. Bushnell, J. Fu, A. L. Gnatt, B. Maier-Davis, N. E. Thompson, R. R. Burgess, A. M. Edwards, P. R. David, and R. D. Kornberg. “Architecture of RNA Polymerase II and Implications for the Transcription Mechanism”. *Science*, Vol. 288, No. 5466, pp. 640–649, 2000.

REFERENCES

- [Davi 05] C. F. David and T. V. Lyubomir. “Targeting Protein-Protein Interactions for Cancer Therapy”. *Journal of Molecular Medicine*, Vol. 83, No. , pp. 955–963, 2005.
- [Dean 02] C. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. “Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations”. *Molecular Cell Proteomics*, Vol. 1, No. 5, pp. 349–356, 2002.
- [Dezs 03] Z. Dezso, Z. D. Oltvai, and A. L. Barabasi. “Bioinformatics Analysis of Experimentally Determined Protein Complexes in the Yeast *Saccharomyces cerevisiae*”. *Genome Res.*, Vol. 13, pp. 2450–2454, 2003.
- [Dona 03] I. Donaldson, J. Martin, and B. Bruijn. “PreBIND and Textomy—Mining the Biomedical Literature for Protein-Protein Interactions using a Support Vector Machine”. *BMC Bioinformatics*, Vol. 4, No. 11, 2003.
- [Dost 07] B. Dost, T. Shlomi, N. Gupta, E. Ruppin, V. Bafna, and R. Sharan. “QNet: A Tool for Querying Protein Interaction Networks”. In: *Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 1–15, 2007.
- [Drie 06] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen. “A Text-Mining Analysis of the Human Phenome”. *European Journal of Human Genetics*, Vol. 14, No. 5, pp. 535–542, 2006.
- [Dutk 07] J. Dutkowski and J. Tiuryn. “Identification of Functional Modules from Conserved Ancestral Protein-Protein Interactions”. *Bioinformatics*, Vol. 23, No. 13, pp. 149–158, 2007.
- [Dwig 02] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry. “*Saccharomyces* Genome Database Provides Secondary Gene Annotation using the Gene Ontology”. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 69–72, 2002.
- [Eise 98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. “Cluster Analysis and Display of Genome-Wide Expression Patterns”. *Proceedings of the National Academy of Sciences*, Vol. 95, No. 25, pp. 14863–14868, 1998.

REFERENCES

- [Enri 02] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. “An Efficient Algorithm for Large-Scale Detection of Protein Families”. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 1575–1584, 2002.
- [Este 96] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [Fahr 03] B. Fahrenkrog and U. Aeby. “The Nuclear Pore Complex: Nucleocytoplasmic Transport and Beyond”. *Nat Rev Mol Cell Biol*, Vol. 4, No. 10, pp. 757–766, 2003.
- [Feng 08] J. Feng, R. Jiang, and T. Jiang. “A Max-Flow Based Approach to the Identification of Protein Complexes Using Protein Interaction and Microarray Data”. In: *International Conference on Computational Systems Bioinformatics (CSB)*, pp. 51–62, 2008.
- [Finn 05] R. D. Finn, M. Marshall, and A. Bateman. “iPfam: Visualization of Protein-protein Interactions in PDB at Domain and Amino Acid Resolutions”. *Bioinformatics*, Vol. 21, No. 3, pp. 410–412, 2005.
- [Form 01] T. Formosa, P. Eriksson, J. Wittmeyer, J. Ginn, Y. Yu, and D. J. Stillman. “Spt16-Pob3 and the HMG Protein Nhp6 Combine to Form the Nucleosome-Binding Factor SPN”. *EMBO Journal*, Vol. 20, No. 13, pp. 3506–3517, 2001.
- [Frae 01] P. Fraering, I. Imhof, U. Meyer, J. Strub, A. van Dorsselaer, C. Vionnet, and A. Conzelmann. “The GPI Transamidase Complex of *Saccharomyces Cerevisiae* Contains Gaa1p, Gpi8p, and Gpi16p”. *Molecular Biology of the Cell*, Vol. 12, No. 10, pp. 3295–3306, 2001.
- [Fras 07] H. B. Fraser and J. B. Plotkin. “Using protein complexes to predict phenotypic effects of gene mutation”. *Genome Biology*, Vol. 8, No. 11, p. R252, 2007.
- [Frie 08] C. C. Friedel, J. Krumsiek, and R. Zimmer. “Boostrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast”. In: *12th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 3–16, 2008.

REFERENCES

- [Gall 89] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. "A Fast Parametric Maximum Flow Algorithm and Applications". *SIAM J. Comput.*, Vol. 18, No. 1, pp. 30–55, 1989.
- [Ganc 98] J. M. Gancido. "Yeast Carbon Catabolite Repression". *Microbiol Mol Biol Rev*, Vol. 62, No. 2, pp. 334–361, 1998.
- [Gare 79] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.
- [Gavi 02] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, and *et al.* "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes". *Nature*, Vol. 415, No. 6868, pp. 141–147, 2002.
- [Gavi 06] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, and *et al.* "Proteome Survey Reveals Modularity of the Yeast Cell Machinery". *Nature*, Vol. 440, No. 7084, pp. 631–636, 2006.
- [Genea] "Gene Expression Data from Eisen's Lab". <http://rana.lbl.gov/EisenData.htm>.
- [Geneb] "Gene Ontology Database". <http://www.geneontology.org/GO.database.shtml>.
- [Gent 07] R. Gentleman and W. Huber. "Making the Most of High-Throughput Protein-Interaction Data". *Genome Biology*, Vol. 8, No. 10, p. 112, 2007.
- [Geva 11] G. Geva and R. Sharan. "Identification of Protein Complexes from Co-Immunoprecipitation Data". *Bioinformatics*, Vol. 27, No. 1, pp. 111–117, 2011.
- [Goh 07] K.-I. Goh, , M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. "The Human Disease Network". *Proceedings of the National Academy of Sciences*, Vol. 104, No. 21, pp. 8685–8690, 2007.
- [Gold 03] D. Goldberg and F. P. Roth. "Assessing Experimentally Derived Interactions in a Small World". *Proceedings of the National Academy of Sciences*, Vol. 100, No. 8, pp. 4372–4376, 2003.

REFERENCES

- [Gran 98] P. A. Grant, D. Schieltz, M. G. Pray-Grant, D. J. Steger, J. C. Reese, J. R. Yates, and J. L. Workman. "A Subset of TAF(II)s are Integral Components of the SAGA Complex Required for Nucleosome Acetylation and Transcriptional Stimulation". *Cell*, Vol. 94, No. 1, pp. 45–53, 1998.
- [Grig 01] A. Grigoriev. "A Relationship between Gene Expression and Protein Interactions on the Proteome Scale". *Nucleic Acids Research*, Vol. 29, No. 17, pp. 3513–3519, 2001.
- [Groc 07] J. Grochow and M. Kellis. "Network Motif Discovery using Subgraph Enumeration and Symmetry-Breaking". In: *11th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 92–106, 2007.
- [Guzd 96] S. N. Guzder, P. Sung, L. Prakash, and S. Prakash. "Nucleotide Excision Repair in Yeast is Mediated by Sequential Sssembly of Repair Factors and not by a Pre-Assembled Repairosome". *The Journal of Biological Chemistry*, Vol. 271, No. 15, pp. 8903–8910, 1996.
- [Hamm 05] T. Hamma, S. L. Reichow, G. Varani, and A. R. Ferre-D'Amare. "The Cbf5-Nop10 Complex is a Molecular Bracket That Organizes Box H/ACA RNPs". *Nature Structural and Molecular Biology*, Vol. 12, No. 11, pp. 1101–1107, 2005.
- [Hart 06] G. T. Hart, A. K. Ramani, and E. M. Marcotte. "How Complete Are Current Yeast and Human Protein-Interaction Networks?". *Genome Biology*, Vol. 7, No. 11, p. 120, 2006.
- [Hart 07] G. T. Hart, I. Lee, and E. Marcotte. "A High-Accuracy Consensus Map of Yeast Protein Complexes Reveals Modular Nature of Gene Essentiality". *BMC Bioinformatics*, Vol. 8, No. 1, p. 236, 2007.
- [Hint 08] A. Hintze and C. Adami. "Evolution of Complex Modular Biological Networks". *PLoS Comput Biol*, Vol. 4, No. 2, p. e23, 02 2008.
- [Hirs 06] E. Hirsh and R. Sharan. "Identification of Conserved Protein Complexes based on a Model of Protein Network Evolution". *Bioinformatics*, Vol. 23, No. 2, pp. 170–176, 2006.

REFERENCES

- [Ho 02] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, and *et al.* “Systematic Identification of Protein Complexes in *Saccharomyces Cerevisiae* by Mass Spectrometry”. *Nature*, Vol. 415, No. 6868, pp. 180–183, 2002.
- [Hort 06] R. Horton, L. A. Moran, G. Scrimgeour, M. Perry, and D. Rawn. *Principles of Biochemistry*. Pearson Prentice-Hall, Inc., 2006.
- [Hu 05] H. Hu, X. Yan, Y. Huang, J. Han, and X. Zhou. “Mining Coherent Dense Subgraphs Across Massive Biological Networks for Functional Discovery”. *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Vol. 21, No. 1, pp. 213–221, 2005.
- [Huan 09] H. Huang and J. Bader. “Precision and Recall Estimates for Two-Hybrid Screens”. *Bioinformatics*, Vol. 25, 2009.
- [Hube 03] L. A. Huber. “Is proteomics heading in the wrong direction?”. *Nat Rev Mol Cell Biol*, Vol. 4, No. 1, pp. 74–80, 2003.
- [Huh 03] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O’Shea. “Global Analysis of Protein Localization in Budding Yeast”. *Nature*, Vol. 425, No. 6959, pp. 686–691, 2003.
- [Hwan 06] W. Hwang, Y. R. Cho, A. Zhang, and M. Ramanathan. “A Novel Functional Module Detection Algorithm for Protein-Protein Interaction Networks”. *Algorithms for Molecular Biology*, Vol. 1, p. 24, 2006.
- [Hwan 10] T. Hwang and R. Kuang. “A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery”. In: *SIAM International Conference on Data Mining*, pp. 583–594, 2010.
- [Idek 02] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. “Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks”. In: *12th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 233–240, 2002.
- [Ispo 05] I. Ispolatov, A. Yuryev, I. Mazo, and S. Maslov. “Binding Properties and Evolution of Homodimers in Protein-Protein Interaction Networks.”. *Nucleic Acids Res*, Vol. 33, No. 11, 2005.

REFERENCES

- [Ito 01] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. “A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome”. *Proceedings of the National Academy of Sciences*, Vol. 98, No. 8, pp. 4569–4574, 2001.
- [Jans 03] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. “A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data”. *Science*, Vol. 302, No. 5644, pp. 449–453, 2003.
- [Jian 08] X. Jiang, B. Liu, J. Jiang, H. Zhao, M. Fan, J. Zhang, Z. Fan, and T. Jiang. “Modularity in the Genetic Disease-Phenotype Network”. *FEBS Letters*, Vol. 582, No. 17, pp. 2549–2554, 2008.
- [Jin 09] R. Jin, S. Mccallen, C. Liu, Y. Xiang, E. Almaas, and X. H. Zhou. “Identify Dynamic Network Modules with Temporal and Spatial Constraints”. In: *Pacific Symposium on Biocomputing*, 2009.
- [Joac 98] T. Joachims. “Making Large-Scale SVM Learning Practical”. *Advances in Kernel Methods: Support Vector Machines*, 1998.
- [Jung 08] S.-K. Jung, W.-H. Jang, and H.-Y. Hur. “Protein Complex Prediction based on Mutally Exclusive Interactions in Protein Interaction Network”. *Genome Informatics*, Vol. 21, pp. 77–88, 2008.
- [Kako 05] A. Kako, T. Ono, T. Hirata, and M. M. Halldórsson. “Approximation Algorithms for the Weighted Independent Set Problem”. In: *Graph-Theoretic Concepts in Computer Science, 31st International Workshop, WG 2005*, pp. 341–350, 2005.
- [Kell 03] R. M. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. “Conserved Pathways within Bacteria and Yeast as Revealed by Global Protein Network Alignment”. *Proceedings of the National Academy of Sciences*, Vol. 100, No. 20, pp. 11394–11399, 2003.
- [Kell 05] R. Kelley and T. Ideker. “Systematic Interpretation of Genetic Interactions using Protein Networks”. *Nature Biotechnology*, Vol. 23, pp. 561–566, 2005.

REFERENCES

- [Kim 06] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein. “Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights”. *Science*, Vol. 314, No. 5807, pp. 1938–1941, 2006.
- [King 04] A. King, N. Przulj, and I. Jurisica. “Protein complex prediction via cost-based clustering”. *Bioinformatics*, Vol. 20, No. 17, pp. 3013–3020, 2004.
- [Kole 08] S. Köler, S. Bauer, D. Horn, and P. N. Robinson. “Walking the Interactome for Prioritization of Candidate Disease Genes”. *American Journal of Human Genetics*, Vol. 82, No. 4, pp. 949–958, 2008.
- [Krog 06] N. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, and *et al.* “Global Landscape of Protein Complexes in the Yeast *Saccharomyces Cerevisiae*”. *Nature*, Vol. 440, No. 7084, pp. 637–643, 2006.
- [Lage 07] K. Lage, E. O. Karlberg, S. Z. M., P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tüer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak. “A Human Phenome-Interactome Network of Protein Complexes Implicated in Genetic Disorders”. *Nature Biotechnology*, Vol. 25, No. 3, pp. 309–316, 2007.
- [Lage 08] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak. “A Large-Scale Analysis of Tissue-Specific Pathology and Gene Expression of Human Disease Genes and Complexes”. *Proceedings of the National Academy of Sciences*, Vol. 105, No. 52, pp. 20870–20875, 2008.
- [Leun 09] H. Leung, Q. Xiang, S. Yiu, and F. Chin. “Predicting Protein Complexes from PPI Data: A Core-Attachment Approach”. *Journal of Computational Biology*, Vol. 16, No. 2, pp. 133–144, 2009.
- [Li 05] X. L. Li, S. H. Tan, C. S. Foo, and S. K. Ng. “Interaction Graph Mining for Protein Complexes Using Local Clique Merging”. *Genome Informatics*, Vol. 16, No. 2, pp. 260–269, 2005.
- [Li 06a] X. L. Li, Y. C. Tan, and S. K. Ng. “Systematic Gene Function Prediction from Gene Expression Data by Using a Fuzzy Nearest-Cluster Method”. *BMC Bioinformatics*, Vol. 7, No. Suppl 4, p. S23, 2006.

REFERENCES

- [Li 06b] X.-L. Li, S.-H. Tan, and S.-K. Ng. “Improving Domain-based Protein Interaction Prediction using Biologically-Significant Negative Dataset”. *International Journal of Data Mining and Bioinformatics (IJDMB)*, Vol. 1, No. 2, 2006.
- [Li 07a] J. Li, G. Liu, H. Li, and L. Wong. “Maximal Biclique Subgraphs and Closed Pattern Pairs of the Adjacency Matrix: A One-to-One Correspondence and Mining Algorithms”. *IEEE Trans. Knowl. Data Eng.*, Vol. 19, No. 12, pp. 1625–1637, 2007.
- [Li 07b] X. Li, C. Foo, and S. Ng. “Discovering Protein Complexes in Dense Reliable Neighborhoods of Protein Interaction Networks”. In: *International Conference on Computational Systems Bioinformatics (CSB)*, pp. 157–168, 2007.
- [Li 08] X.-L. Li, J.-X. Lee, B. Veeravalli, and S.-K. Ng. “Inferring Transcription Factor Interactions Using a Novel HV-SVM Classifier”. *International Journal of Computational Biology and Drug Design*, Vol. 1, No. 1, pp. 59–73, 2008.
- [Li 09] X. Li and S. Ng. *Biological Data Mining in Protein Interaction Networks*. IGI Global, Pennsylvania, USA, 2009.
- [Li 10a] X. L. Li, M. Wu, C. K. Kwok, and S. K. Ng. “Computational Approaches for Detecting Protein Complexes from Protein Interaction Networks: A Survey”. *BMC Genomics*, Vol. S1, p. S3, 2010.
- [Li 10b] Y. Li and J. C. Patra. “Genome-Wide Inferring Gene-Phenotype Relationship by Walking on the Heterogeneous Network”. *Bioinformatics*, Vol. 26, No. 9, pp. 1219–1224, 2010.
- [Ling 09] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, , and C. DeLisi. “Genome-Wide Prioritization of Disease Genes and Identification of Disease-Disease Associations from an Integrated Human Functional Linkage Network”. *Genome Biology*, Vol. 10, No. 9, p. R91, 2009.
- [Liu 01] H. Liu, Y. Chiang, J. Pan, J. Chen, C. Salvatore, D. Audino, V. Badarinarayana, V. Palaniswamy, B. Anderson, and C. DenisDagger. “Characterization of CAF4 and CAF16 Reveals a Functional Connection between

REFERENCES

- the CCR4-NOT Complex and a Subset of SRB Proteins of the RNA Polymerase II Holoenzyme". *Journal of Biological Chemistry*, Vol. 276, No. 10, pp. 7541–7548, 2001.
- [Liu 08] G. Liu, J. Li, and L. Wong. "Assessing and Predicting Protein Interactions Using Both Local and Global Network Topological Metrics". *Genome Informatics*, Vol. 22, No. 138-149, 2008.
- [Liu 09] G. M. Liu, H. N. Chua, and L. Wong. "Complex Discovery from Weighted PPI Networks". *Bioinformatics*, Vol. 25, No. 15, pp. 1891–1897, 2009.
- [Lord 03] P. Lord, R. Stevens, A. Brass, and C.A.Goble. "Semantic Similarity Measures as Tools for Exploring the Gene Ontology". In: *Pacific Symposium on Biocomputing*, pp. 601–612, 2003.
- [Lubo 06] Z. Lubovac, J. Gamalielsson, and B. Olsson. "Combining Functional and Topological Properties to Identify Core Modules in Protein Interaction Networks". *PROTEINS:Structure, Function, and Bioinformatics*, Vol. 64, pp. 948–959, 2006.
- [Mara 07] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos. "Growing Functional Modules from a Seed Protein via Integration of Protein Interaction and Gene Expression Data". *BMC Bioinformatics*, Vol. 8, p. 408, 2007.
- [Marc 99] E. Marcotte, M. Pellegrini, H. Ng, D. Rice, T. Yeates, and D. Eisenberg. "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences". *Science*, Vol. 285, No. 5428, pp. 751–753, 1999.
- [McCr 00] S. McCraith, T. Holtzman, B. Moss, and S. Fields. "Genome-wide Analysis of Vaccinia Virus Protein-Protein Interactions". *Proceedings of the National Academy of Sciences*, Vol. 97, No. 9, p. 4879C4884, 2000.
- [Meri 02] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork. "Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions". *Nature*, Vol. 417, No. 6887, pp. 399–403, 2002.
- [Metete 08] M. Mete, F. Tang, X. Xu, and N. Yuruk. "A Structural Approach for Finding Functional Modules from Large Biological Networks". *BMC Bioinformatics*, Vol. 9(Suppl 9), No. 1, p. SI9, 2008.

REFERENCES

- [Mewe 04] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. "MIPS: Analysis and Annotation of Proteins from Whole Genomes". *Nucleic Acids Research (Database issue)*, Vol. 32, No. 1, pp. 41–44, 2004.
- [Mira 05] J. J. Miranda, P. De Wulf, P. Sorger, and S. C. Harrison. "The Yeast DASH Complex Forms Closed Rings on Microtubules". *Nat Struct Mol Biol*, Vol. 12, No. 2, pp. 138–143, 2005.
- [Navl 08] S. Navlakha, R. Rastogi, and N. Shrivastava. "Graph Summarization with Bounded Error". In: *ACM SIGMOD Conference on Management Of Data*, pp. 419–432, 2008.
- [Navl 09] S. Navlakha, M. C. Schatz, and C. Kingsford. "Revealing Biological Modules via Graph Summarization". *Journal of Computational Biology*, Vol. 16, No. 2, pp. 253–264, 2009.
- [Navl 10] S. Navlakha and C. Kingsford. "Exploring Biological Network Dynamics with Ensembles of Graph Partitions". In: *Proc. 15th Intl. Pacific Symposium on Biocomputing (PSB)*, pp. 166–177, 2010.
- [Ng 03] S. K. Ng and S. H. Tan. "Discovering Protein-Protein Interactions". *Journal of Bioinformatics and Computational Biology*, Vol. 18, No. 39, pp. 1–30, 2003.
- [Oti 06] M. Oti, B. Snel, M. A. Huynen, and B. H. G. "Predicting Disease Genes using Protein-Protein Interactions". *Journal of Medical Genetics*, Vol. 43, No. 8, pp. 691–698, 2006.
- [Pall 05] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society". *Nature*, Vol. 435, No. 7043, pp. 814–818, 2005.
- [Pati 05] A. Patil and H. Nakamura. "Filtering High-Throughput Protein-Protein Interaction Data using a Combination of Genomic Features". *BMC Bioinformatics*, Vol. 6, p. 100, 2005.

REFERENCES

- [Pei 05] P. Pei and A. Zhang. "A Topological Measurement for Weighted Protein Interaction Network". In: *International Conference on Computational Systems Bioinformatics (CSB)*, pp. 268–278, 2005.
- [Pell 99] M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates. "Assigning Protein Functions by Comparative Analysis: Protein Phylogenetic Profiles". *Proceedings of the National Academy of Sciences*, Vol. 96, No. 8, pp. 4285–4288, 1999.
- [Pere 04] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. "Detection of Functional Modules from Protein Interaction Networks". *PROTEINS:Structure, Function, and Bioinformatics*, Vol. 54, pp. 49–57, 2004.
- [Przu 04] N. Przulj, D. A. Wigle, and I. Jurisica. "Functional Topology in a Network of Protein Interactions". *Bioinformatics*, Vol. 20, No. 3, pp. 340–348, 2004.
- [Pu 07] S. Pu, J. Vlasblom, A. Emili, J. Greenblatt, and S. J. Wodak. "Identifying Functional Modules in the Physical Interactome of *Saccharomyces Cerevisiae*". *Proteomics*, Vol. 7, No. 6, pp. 944–960, 2007.
- [Pu 09] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodar. "Up-to-date Catalogues of Yeast Protein Complexes". *Nucleic Acids Research*, Vol. 37, No. 3, pp. 825–831, 2009.
- [Puig 01] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin. "The Tandem Affinity Purification Method: a General Procedure of Protein Complex Purification". *Methods*, Vol. 24, No. 3, pp. 218–229, 2001.
- [Qi 08] Y. Qi, F. Balem, C. Faloutsos, J. Klein-Seetharaman, and Z. Bar-Joseph. "Protein Complex Identification by Supervised Graph Local Clustering". In: *18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 250–258, 2008.
- [Qiu 10] Y.-Q. Qiu, S.-H. Zhang, X.-S. Zhang, and L. Chen. "Detecting Disease Associated Modules and Prioritizing Active Genes based on High Throughput Data". *BMC Bioinformatics*, Vol. 11, p. 26, 2010.

REFERENCES

- [Ragh 08] B. Raghavachari, A. Tasneem, T. M. Przytycka, and R. Jothi. "DOMINE: a Database of Protein Domain Interactions". *Nucleic Acids Research*, Vol. 36, No. Database-Issue, pp. 656–661, 2008.
- [Rain 01] J. Rain, L. Selig, H. Reuse, V. Battaglia, C. Reverdy, and *et al.* "The Protein-Protein Interaction Map of Helicobacter Pylori". *Nature*, Vol. 409, No. 6817, p. 211C215, 2001.
- [Resn 95] P. Resnik. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In: *International Joint Conferences on Artificial Intelligence*, pp. 448–453, 1995.
- [Ried 97] S. E. Rieder and S. D. Emr. "A Novel RING Finger Protein Complex Essential for a Late Step in Protein Transport to the Yeast Vacuole". *Mol. Biol. Cell*, Vol. 8, pp. 2307–2327, 1997.
- [Ruep 07] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegle, T. Schmidt, O. N. Doudieu, V. Stupflen, and H. W. Mewes. "CORUM: the Comprehensive Resource of Mammalian Protein Complexes". *Nucleic Acids Res*, Vol. 36(Database issue), No. 1, pp. D646–50, 2007.
- [Sait 02] R. Saito, H. Suzuki, and Y. Hayashizaki. "Interaction Generality, a Measurement to Assess the Reliability of a Protein-Protein Interaction". *Nucleic Acids Res*, Vol. 30, No. 5, pp. 1163–1168, 2002.
- [Sait 03] R. Saito, H. Suzuki, and Y. Hayashizaki. "Construction of Reliable Protein-Protein Interaction Networks with a New Interaction Generality Measure". *Bioinformatics*, Vol. 19, No. 6, pp. 756–763, 2003.
- [Saka 03] S. Sakai, M. Togasaki, and K. Yamazaki. "A Note on Greedy Algorithms for the Maximum Weighted Independent Set Problem". *Discrete Applied Mathematics*, Vol. 126, No. 2-3, pp. 313–322, 2003.
- [Salw 04] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. "The Database of Interacting Proteins: 2004 Update". *Nucleic Acids Research*, Vol. 30, No. 1, pp. 449–451, 2004.

REFERENCES

- [Scho 05] D. Scholtens, M. Vidal, and R. Gentleman. “Local Modeling of Global Interactome Networks”. *Bioinformatics*, Vol. 21, No. 17, pp. 3548–3557, 2005.
- [Schw 00] B. Schwikowski, P. Uetz, and S. Fields. “A Network of Protein-Protein Interactions in Yeast”. *Nature Biotechnology*, Vol. 18, No. 12, pp. 1257–1261, 2000.
- [Schw 05] M. Schwarz, R. Knauer, and L. Lehle. “Yeast Oligosaccharyltransferase Consists of Two Functionally Distinct Sub-Complexes, Specified by Either the Ost3p or Ost6p Subunit”. *FEBS Letters*, Vol. 579, No. 29, pp. 6564–6568, 2005.
- [Seal 00] D. F. Seals, G. Eitzen, N. Margolis, W. T. Wickner, and A. Price. “A Ypt/Rab Effector Complex Containing the Sec1 Homolog Vps33p is Required for Homotypic Vacuole Fusion”. *Proceedings of the National Academy of Sciences*, Vol. 97, No. 17, pp. 9402–9407, 2000.
- [Seid 83] S. B. Seidman. “Network Structure and Minimum Degree”. *Social Networks*, Vol. 5, No. , pp. 269–287, 1983.
- [Shar 04] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. “Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data”. In: *Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 282–289, 2004.
- [Shar 05] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. “Conserved Patterns of Protein Interaction in Multiple Species”. *Proceedings of the National Academy of Sciences*, Vol. 102, No. 6, pp. 1974–1979, 2005.
- [Shas 02] D. Shasha, J. T. L. Wang, and R. Giugno. “Algorithmics and Applications of Tree and Graph Searching”. In: *Symposium on Principles of Database Systems (PODS)*, pp. 39–52, 2002.
- [Spir 03] V. Spirin and L. Mirny. “Protein Complexes and Functional Modules in Molecular Networks”. *Proceedings of the National Academy of Sciences*, Vol. 100, No. 21, pp. 12123–12128, 2003.

REFERENCES

- [Spri 03] E. Sprinzak, S. Sattath, and H. Margalit. “How Reliable are Experimental Protein-Protein Interaction Data?”. *Journal of Molecular Biology*, Vol. 327, No. 5, pp. 919–923, 2003.
- [Srih 10] S. Srihari, K. Ning, and H. W. Leong. “MCL-CAw: a Refinement of MCL for Detecting Yeast Complexes from Weighted PPI Networks by Incorporating Core-Attachment Structure”. *BMC Bioinformatics*, Vol. 11, No. 1, p. 504, 2010.
- [Star 06] C. Stark, B. Breitkreutz1, T. Reguly1, L. Boucher, A. Breitkreutz1, and M. Tyers. “BioGRID: a General Repository for Interaction Datasets”. *Nucleic Acids Research*, Vol. 34, No. Database-Issue, pp. 535–539, 2006.
- [Stei 05] A. Stein, A. Panjkovich, and P. Aloy. “3Did Update: Domain-Domain and Peptide-Mediated Interactions of Known 3D Structure”. *Nucleic Acids Res.*, Vol. 37(Database issue), pp. 300–304, 2005.
- [Tara 08] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. Serna, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick. “An in vivo Map of the Yeast Protein Interactome”. *Science*, Vol. 320, No. 5882, pp. 1465–1470, 2008.
- [Tomi 06] E. Tomita, A. Tanaka, and H. Takahashi. “The Worst-Case Time Complexity for Generating All Maximal Cliques and Computational Experiments”. *Theor. Comput. Sci.*, Vol. 363, No. 1, pp. 28–42, 2006.
- [Tong 02] A. Tong, B. Drees, G. Nardelli, G. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. “A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules”. *Science*, Vol. 295, No. 5553, pp. 321–324, 2002.
- [Ucar 06] D. Ucar, S. Asur, Ü. V. Çatalyürek, and S. Parthasarathy. “Improving Functional Modularity in Protein-Protein Interactions Graphs Using Hub-Induced Subgraphs”. In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 371–382, 2006.

REFERENCES

- [Uetz 00] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, and *et al.* “A Comprehensive Analysis of Protein-Protein Interactions in accharomyces Cerevisiae”. *Nature*, Vol. 403, No. 6770, p. 623C627, 2000.
- [Ulitt 07] I. Ulitsky and R. Shamir. “Identification of Functional Modules using Network Topology and High-Throughput Data”. *BMC Systems Biology*, Vol. 1, No. 1, p. 8, 2007.
- [Ullm 76] J. R. Ullmann. “An Algorithm for Subgraph Isomorphism”. *Journal of ACM*, Vol. 23, No. 1, pp. 31–42, 1976.
- [Van 00] S. Van Dongen. *Graph Clustering by Flow Simulation*. University of Utrecht, Netherlands, 2000.
- [Vanu 10] O. Vanunu, O. Magger, E. Ruppim, and T. S. R. Sharan. “Associating Genes and Protein Complexes with Disease via Netwrok Propagation”. *PLoS Computational Biology*, Vol. 6, No. 1, pp. 1–9, 2010.
- [Vapn 95] V. N. Vapnik. “The Nature of Statistical Learning Theory”. *Springer*, 1995.
- [Wang 07] J. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen. “A New Method to Measure the Semantic Similarity of GO Terms”. *Bioinformatics*, Vol. 23, No. 10, pp. 1274–1281, 2007.
- [Watt 98] D. J. Watts and S. H. Strogatz. “Collective Dynamics of ‘Small-World’ Networks”. *Nature*, Vol. 393, No. 6684, pp. 440–442, 1998.
- [Wu 08a] M. Wu, X. Li, and K. Kwoh. “Algorithms for Detecting Protein Complexes in PPI Networks: An Evaluation Study”. In: *(Supplementary paper presented at) International Conference on Pattern Recognition in Bioinformatics (PRIB); 2008 Oct 15-17; Melbourne, Australia.*, pp. p.135–146, 2008.
- [Wu 08b] X. B. Wu, R. Jiang, M. Q. Zhang, and S. Li. “Network-based Global Inference of Human Disease Genes”. *Molecular Systems Biology*, Vol. 4, No. 189, pp. 1–11, 2008.
- [Wu 09a] M. Wu, X. L. Li, C. K. Kwoh, and S. K. Ng. “A Core-Attachment based Method to Detect Protein Complexes in PPI Networks”. *BMC Bioinformatics*, Vol. 10, p. 169, 2009.

REFERENCES

- [Wu 09b] X. Wu, Q. Liu, and R. Jiang. “Align Human Interactome with Phenome to Identify Causative Genes and Networks Underlying Disease Families”. *Bioinformatics*, Vol. 25, No. 1, pp. 98–104, 2009.
- [Wu 11] M. Wu, X. L. Li, C. K. Kwoh, S. K. Ng, and L. Wong. “Discovery of Protein Complexes with Core-Attachment Structures from Tandem Affinity Purification (TAP) Data”. *J Comput Biol*, Vol. 18, pp. 1–16, 2011.
- [Xena 02] I. Xenarios, L. Salwinski, X. Duan, P. Higney, S. Kim, and D. Eisenberg. “DIP, the Database of Interacting Proteins: a Research Tool for Studying Cellular Networks of Protein Interactions”. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 303–305, 2002.
- [Yang 11] P. Yang, X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng. “Inferring Gene-Phenotype Associations via Global Protein Complex Network Propagation”. *PLoS One*, Vol. to appear, No. , p. , 2011.
- [Yook 04] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi. “Functional and Topological Characterization of Protein Interaction Networks”. *Proteomics*, Vol. 4, No. 4, pp. 928–942, 2004.
- [Yoon 06] J. Yoon, A. Blumer, and K. Lee. “An algorithm for Modularity Analysis of Directed and Weighted Biological Networks based on Edge-Betweenness Centrality”. *Bioinformatics*, Vol. 22, No. 24, pp. 3106–3108, 2006.
- [Zaki 08] N. Zaki. “Prediction of Protein-Protein Interactions Using Pairwise Alignment and Inter-Domain Linker Region”. *Engineering Letter*, Vol. 16, No. 4, 2008.
- [Zhan 05] Y. Zhang, F. N. Abu-Khzam, N. E. Baldwin, E. J. Chesler, M. A. Langston, and N. F. Samatova. “Genome-Scale Computational Approaches to Memory-Intensive Applications in Systems Biology”. In: *Supercomputing*, p. 12, IEEE Computer Society, 2005.
- [Zhan 08] B. Zhang, B. Park, T. V. Karpinets, and N. F. Samatova. “From Pull-Down Data to Protein Interaction Networks and Complexes with Biological Relevance”. *Bioinformatics*, Vol. 24, No. 7, pp. 979–986, 2008.
- [Zhan 10] Y. Zhang and Y. Li. “The Expanding Mi-2/NuRD Complexes: A Schematic Glance”. *Proteomics Insights*, Vol. 3, No. , pp. 79–109, 2010.

REFERENCES

Appendix

Proposition In a biclique $G = (U, V, E)$, the removal of a vertex $v \in A(G)$ with the minimum ratio score will not decrease the reliability of G .

Proof: In G , we define the neighbors of a vertex v_i (without loss of generality, we assume $v_i \in V$), $N_G(v_i) = \{u | u \in U, (v_i, u) \in E, u \neq v_i\}$. For v_i , we assume that x_i is weight sum of edges between v_i and its neighbors and y_i is the number of v_i 's neighbors, i.e., $|N_G(v_i)|$. Therefore, we can have following equations.

$$ratio_G(v_i) = \frac{x_i}{y_i}, \quad (6.8)$$

$$R(G) = \frac{\sum_{v_i \in U \cup V} x_i}{\sum_{v_i \in U \cup V} y_i} \quad (6.9)$$

$$R(G - \{v_j\}) = \frac{(\sum_{v_i \in U \cup V} x_i) - x_j}{(\sum_{v_i \in U \cup V} y_i) - y_j}. \quad (6.10)$$

Here, $G - \{v_j\}$ is the graph obtained by removing the vertex v_j and the corresponding edges from G . For the sum $\sum_{v_i \in U \cup V} x_i$ in equations (6.9) and (6.10), we will use $\sum_i x_i$ for short. In addition, we assume $v_j \in V$ and then have following two equations.

$$R(G) = \frac{\sum_{v_i \in V} x_i}{\sum_{v_i \in V} y_i} \quad (6.11)$$

$$R(G - \{v_j\}) = \frac{(\sum_{v_i \in V} x_i) - x_j}{(\sum_{v_i \in V} y_i) - y_j}. \quad (6.12)$$

Now we assume v_1 is the vertex with the minimum ratio score in $A(G)$, i.e., $\forall v_i \in A(G), ratio_G(v_1) \leq ratio_G(v_i)$. Then we can easily prove $R(G) \leq R(G - \{v_1\}) = \frac{(\sum_i x_i) - 2 \cdot x_1}{(\sum_i y_i) - 2 \cdot y_1}$ when $A(G) = U \cup V$.

$$\begin{aligned} & \forall i, \frac{x_1}{y_1} \leq \frac{x_i}{y_i} \\ \Rightarrow & \forall i, x_1 y_i \leq x_i y_1 \\ \Rightarrow & x_1 \sum_i y_i \leq y_1 \sum_i x_i \\ \Rightarrow & \frac{x_1}{y_1} \leq \frac{\sum_i x_i}{\sum_i y_i} \\ \Rightarrow & \frac{\sum_i x_i}{\sum_i y_i} \leq \frac{(\sum_i x_i) - 2 \cdot x_1}{(\sum_i y_i) - 2 \cdot y_1} \end{aligned} \quad (6.13)$$

We can similarly prove $R(G) \leq R(G - \{v_1\})$ when $A(G) = U$ or V . Now, we claim that v_1 's removal will increase or at least keep the reliability of G . $\square$

REFERENCES

List of Publications

- (i) Min Wu, Xiaoli Li and Chee-Keong Kwoh, "Algorithms for Detecting Protein Complexes in PPI Networks: An Evaluation Study", The Third IAPR International Conference on *Pattern Recognition in Bioinformatics* (PRIB 2008).
- (ii) Min Wu, Xiaoli Li and Chee-Keong Kwoh, "Semantic labeling to identify reliable interactions in the protein interaction network", The 2nd International Conference on *Bioinformatics and Biomedical Engineering* (ICBBE 2008).
- (iii) Min Wu, Xiaoli Li, Chee-Keong Kwoh and See-Kiong Ng, "A Core-Attachment based Method to Detect Protein Complexes in PPI Networks", *BMC Bioinformatics* 2009, vol 10:169 [Impact Factor: 3.43].
- (iv) Xiaoli Li, Min Wu, Chee-Keong Kwoh and See-Kiong Ng, "Computational Approaches for Detecting Protein Complexes from Protein Interaction Networks: A Survey", *BMC Genomics* 2010, vol 11(S1):S3 [Impact Factor: 3.76].
- (v) Min Wu, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh and See-Kiong Ng, "Integrating Diverse Biological and Computational Sources for Reliable Protein-Protein Interactions", Proceedings of Ninth International Conference of *Bioinformatics (In-CoB)*, 2010. Also included in *BMC Bioinformatics* 2010, vol 11(Suppl 7):S8 [Impact Factor: 3.43].
- (vi) Min Wu, Xiaoli Li, Chee-Keong Kwoh, See-Kiong Ng and Limsoon Wong, "Discovery of Protein Complexes with Core-Attachment Structures from TAP Data", accepted by *Journal of Computational Biology*, 2011.
- (vii) Zhipeng Xie, Xiaoli Li, Chee-Keong Kwoh and Min Wu, "Construction of Co-Complex Score Matrix for Protein Complex Prediction from AP-MS Data", accepted by *ISMB/ECCB* 2011.
- (viii) Peng Yang, Xiaoli Li, Min Wu, Chee-Keong Kwoh, See-Kiong Ng, "Inferring Gene-Phenotype Associations via Global Protein Complex Network Propagation", accepted by *PLoS ONE*, 2011.