

IDENTIFYING PROTEIN COMPLEXES AND DISEASE GENES  
FROM BIOMOLECULAR NETWORKS

A Thesis Submitted to the  
College of Graduate Studies and Research  
in Partial Fulfillment of the Requirements  
for the degree of Doctor of Philosophy  
in the Division of Biomedical Engineering  
University of Saskatchewan  
Saskatoon

By  
Bolin Chen

©Bolin Chen, November 2014. All rights reserved.

## PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering  
Engineering Building  
57 Campus Dr.  
University of Saskatchewan  
Saskatoon, Saskatchewan  
Canada  
S7N 5A9

# ABSTRACT

With advances in high-throughput measurement techniques, large-scale biological data, such as protein-protein interaction (PPI) data, gene expression data, gene-disease association data, cellular pathway data, and so on, have been and will continue to be produced. Those data contain insightful information for understanding the mechanisms of biological systems and have been proved useful for developing new methods in disease diagnosis, disease treatment and drug design. This study focuses on two main research topics: (1) identifying protein complexes and (2) identifying disease genes from biomolecular networks.

Firstly, protein complexes are groups of proteins that interact with each other at the same time and place within living cells. They are molecular entities that carry out cellular processes. The identification of protein complexes plays a primary role for understanding the organization of proteins and the mechanisms of biological systems. Many previous algorithms are designed based on the assumption that protein complexes are densely connected sub-graphs in PPI networks. In this research, a dense sub-graph detection algorithm is first developed following this assumption by using clique seeds and graph entropy. Although the proposed algorithm generates a large number of reasonable predictions and its *f-score* is better than many previous algorithms, it still cannot identify many known protein complexes. After that, we analyze characteristics of known yeast protein complexes and find that not all of the complexes exhibit dense structures in PPI networks. Many of them have a star-like structure, which is a very special case of the core-attachment structure and it cannot be identified by many previous core-attachment-structure-based algorithms. To increase the prediction accuracy of protein complex identification, a multiple-topological-structure-based algorithm is proposed to identify protein complexes from PPI networks. Four single-topological-structure-based algorithms are first employed to detect raw predictions with clique, dense, core-attachment and star-like structures, respectively. A merging and trimming step is then adopted to generate final predictions based on topological information or GO annotations of predictions. A comprehensive review about the identification of protein complexes from static PPI networks to dynamic PPI networks is also given in this study.

Secondly, genetic diseases often involve the dysfunction of multiple genes. Various types of evidence have shown that similar disease genes tend to lie close to one another in various biomolecular networks. The identification of disease genes via multiple data integration is indispensable towards the understanding of the genetic mechanisms of many genetic diseases. However, the number of known disease genes related to similar genetic diseases is often small. It is not easy to capture the intricate gene-disease associations from such a small number of known samples. Moreover, different kinds of biological data are heterogeneous and no widely acceptable criterion is available to standardize them to the same scale. In this study, a flexible and reliable multiple data integration algorithm is first proposed to identify disease genes based on the

theory of Markov random fields (MRF) and the method of Bayesian analysis. A novel global-characteristic-based parameter estimation method and an improved Gibbs sampling strategy are introduced, such that the proposed algorithm has the capability to tune parameters of different data sources automatically. However, the Markovianity characteristic of the proposed algorithm means it only considers information of direct neighbors to formulate the relationship among genes, ignoring the contribution of indirect neighbors in biomolecular networks. To overcome this drawback, a kernel-based MRF algorithm is further proposed to take advantage of the global characteristics of biological data via graph kernels. The kernel-based MRF algorithm generates predictions better than many previous disease gene identification algorithms in terms of the area under the receiver operating characteristic curve (AUC score). However, it is very time-consuming, since the Gibbs sampling process of the algorithm has to maintain a long Markov chain for every single gene. Finally, to reduce the computational time of the MRF-based algorithm, a fast and high performance logistic-regression-based algorithm is developed for identifying disease genes from biomolecular networks. Numerical experiments show that the proposed algorithm outperforms many existing methods in terms of the AUC score and running time.

To summarize, this study has developed several computational algorithms for identifying protein complexes and disease genes from biomolecular networks, respectively. These proposed algorithms are better than many other existing algorithms in the literature.

## ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor Prof. Fang-Xiang Wu. Throughout my PhD studies, he always provides me with constant encouragement, research ideas and daily life suggestions. This thesis would not have been completed without his help.

I would also like to express my gratitude to other members of my advisory committee Prof. Wen-Jun Zhang, Prof. Tony Kusalik and Prof. Mark Keil for their assistance and great advice during the PhD program.

I also want to thank my group members Jinhong Shi, Lin Wu, Lizhi Liu, Yan Yan, Weiwei Fan, Jian Sun, Yichao Shen and Amin Mohammadbagheri for their help in both my life and research work.

In addition, I would also like to thank my dear friends Jia Sun, Lei Ren, Jun Liu, Bo Gui for making my student life joyful.

I would like to thank all my family for their continuous support and love.

Finally, I gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC), University of Saskatchewan and the China Scholarship Council (CSC) for the financial supports.

This thesis is dedicated to my family and Jia.

# CONTENTS

<b>Permission to Use</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Contents</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation and objectives . . . . .	2
1.3 Organization of the thesis . . . . .	3
1.4 Contributions of the primary investigator . . . . .	4
Bibliography . . . . .	5
<b>2 Identifying protein complexes and functional modules - from static PPI networks to dynamic PPI networks</b>	<b>7</b>
2.1 Introduction . . . . .	8
2.2 Protein interactions and PPI networks . . . . .	9
2.3 Identifying protein complexes based on topological structures of unweighted PPI networks . . . . .	13
2.4 Identifying protein complexes based on characteristics of weighted PPI networks . . . . .	16
2.5 Identifying protein complexes and/or functional modules by multiple data integrations . . . . .	21
2.6 Distinguishing between protein complexes and functional modules via dynamic PPI networks . . . . .	23
2.7 Evaluation methods . . . . .	25
2.8 Conclusions . . . . .	28
Bibliography . . . . .	29
<b>3 Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy</b>	<b>37</b>
3.1 Introduction . . . . .	38
3.2 Materials and methods . . . . .	40
3.2.1 The entropy-based algorithm . . . . .	40
3.2.2 The seed-selection strategy . . . . .	41
3.2.3 Investigation of the entropy-based algorithm . . . . .	42
3.2.4 Data source . . . . .	43
3.2.5 Accuracy evaluation approaches . . . . .	43
3.3 Results and discussions . . . . .	44
3.3.1 Properties of cliques and maximal cliques . . . . .	44
3.3.2 Modification strategies . . . . .	46
3.3.3 Strategies for the entropy-based algorithm . . . . .	48
3.3.4 The overall results of predictions . . . . .	50
3.4 Conclusions . . . . .	51
Bibliography . . . . .	53

<b>4</b>	<b>Not all protein complexes exhibit dense structures in <i>S. cerevisiae</i> PPI network</b>	<b>56</b>
4.1	Introduction . . . . .	57
4.2	Materials and methods . . . . .	57
4.2.1	Protein complexes and their relative neighbours . . . . .	58
4.2.2	The number of edges, density, relative density and radius . . . . .	58
4.3	Experiments and results . . . . .	59
4.3.1	Data source . . . . .	59
4.3.2	Statistical results . . . . .	59
4.3.3	The structures of protein complexes . . . . .	61
4.4	Algorithm and results . . . . .	62
4.4.1	The random-star algorithm . . . . .	63
4.4.2	Accuracy evaluation . . . . .	63
4.4.3	Predicted results . . . . .	64
4.5	Conclusions . . . . .	64
	Bibliography . . . . .	65
<b>5</b>	<b>Identifying protein complexes based on multiple topological structures in PPI networks</b>	<b>67</b>
5.1	Introduction . . . . .	68
5.2	Materials and methods . . . . .	69
5.2.1	Terminologies . . . . .	69
5.2.2	Deriving the weights for PPI networks . . . . .	70
5.2.3	A framework for identifying protein complexes based on multiple topological structures . . . . .	70
5.2.4	Data sources . . . . .	72
5.2.5	Evaluating predictions by GO annotations . . . . .	73
5.2.6	Evaluating predictions by gold standard protein complexes . . . . .	74
5.3	Results and discussions . . . . .	75
5.3.1	Edge weights in PPI networks . . . . .	75
5.3.2	Properties of known protein complexes . . . . .	76
5.3.3	Results of predicted protein complexes . . . . .	80
5.4	Conclusions . . . . .	84
	Bibliography . . . . .	85
<b>6</b>	<b>Identifying disease genes by integrating multiple data sources</b>	<b>89</b>
6.1	Introduction . . . . .	90
6.2	Methods . . . . .	91
6.2.1	The Bayesian labelling problem . . . . .	92
6.2.2	Gibbs distribution in MRF . . . . .	93
6.2.3	The MRF model for identifying human disease genes . . . . .	93
6.2.4	The Gibbs sampling . . . . .	95
6.2.5	Parameter estimation . . . . .	96
6.2.6	Estimating a prior probability . . . . .	98
6.2.7	Data sources . . . . .	98
6.2.8	Validation method and evaluation criteria . . . . .	99
6.2.9	Decision score and declaration of positives . . . . .	99
6.3	Results and discussions . . . . .	100
6.3.1	Stability and reliability of MRF methods . . . . .	100
6.3.2	Comparisons with the MRF-Deng method . . . . .	102
6.3.3	Integration of heterogeneous data sources . . . . .	102
6.3.4	Comparisons by using multiple data sources . . . . .	104
6.4	Conclusions . . . . .	105
	Bibliography . . . . .	107
<b>7</b>	<b>Disease gene identification by using graph kernels and Markov random fields</b>	<b>111</b>
7.1	Introduction . . . . .	112
7.2	Methods and materials . . . . .	113



7.2.1	Problem statement . . . . .	113
7.2.2	Markov random field . . . . .	114
7.2.3	Graph kernels . . . . .	116
7.2.4	Kernel-based MRF method . . . . .	117
7.2.5	Experimental design . . . . .	119
7.3	Results . . . . .	122
7.3.1	Stability and reliability of the kernel-based MRF method . . . . .	122
7.3.2	Comparisons between different kernels . . . . .	124
7.3.3	Comparisons with previous methods . . . . .	125
7.4	Conclusions . . . . .	126
	Bibliography . . . . .	127
<b>8</b>	<b>A fast and high performance algorithm for identifying human disease genes</b>	<b>131</b>
8.1	Introduction . . . . .	132
8.2	Methods and materials . . . . .	134
8.2.1	Problem formulation . . . . .	134
8.2.2	Logistic regression . . . . .	135
8.2.3	Feature vector constructions . . . . .	137
8.2.4	Prior probability estimation . . . . .	139
8.2.5	Decision score . . . . .	139
8.2.6	Validation method and evaluation criteria . . . . .	140
8.2.7	Algorithm . . . . .	141
8.3	Results and discussions . . . . .	141
8.3.1	Data sources . . . . .	141
8.3.2	Comparisons between different priors . . . . .	142
8.3.3	Comparisons between different feature vectors . . . . .	143
8.3.4	Comparing with previous algorithms . . . . .	144
8.4	Conclusions . . . . .	146
	Bibliography . . . . .	148
<b>9</b>	<b>Conclusions, contributions and future work</b>	<b>151</b>
9.1	Conclusions . . . . .	151
9.2	Contributions . . . . .	152
9.3	Future work . . . . .	153
<b>A</b>	<b>List of Publications</b>	<b>154</b>
<b>B</b>	<b>Copyright Permissions</b>	<b>156</b>

## LIST OF TABLES

3.1	The number of cliques and maximal cliques in the PPI network . . . . .	45
3.2	The average <i>f-score</i> , <i>precision</i> and <i>recall</i> for different experiments . . . . .	47
3.3	The number of seeds and their output clusters . . . . .	49
4.1	The average degree for IB and OB vertices . . . . .	61
5.1	Details of the PPI datasets . . . . .	72
5.2	The number of predictions of identification methods . . . . .	81
5.3	The evaluations of different identification methods . . . . .	83
5.4	The evaluations of unmatched predictions . . . . .	84

# LIST OF FIGURES

2.1	The organization of computational algorithms . . . . .	10
2.2	Experimental protein interactions and a combined PPI network . . . . .	12
2.3	The general procedure of hierarchical clustering algorithms . . . . .	18
2.4	Schematic of the agglomerative and divisive clustering methods . . . . .	19
2.5	Static PPI network and dynamic PPI network . . . . .	24
2.6	Comparison of the <i>precision</i> , <i>recall</i> , <i>f-score</i> and <i>overlapping score</i> . . . . .	26
3.1	The average <i>f-score</i> and the number of output clusters for each size seeds . . . . .	45
3.2	The average <i>f-score</i> after the first way of modification . . . . .	46
3.3	The average <i>f-score</i> after the second way of modification . . . . .	48
3.4	The difference of accuracy between uWVE and WVE . . . . .	49
3.5	The differences of the average <i>f-score</i> between CE and GE . . . . .	50
3.6	The predicted clusters in the PPI network . . . . .	51
4.1	Statistic results for protein complexes with no less than five proteins . . . . .	60
4.2	The structure of protein complexes . . . . .	62
5.1	A framework for identifying protein complexes based on multiple topological structures. . . . .	71
5.2	The cumulative distribution of edge weights in yeast and human PPI networks . . . . .	75
5.3	The size distribution of the yeast and human protein complexes. . . . .	76
5.4	The structure histogram for the yeast and human protein complexes . . . . .	77
5.5	The average number of within edges, the number of outgoing edges and the number of all edges . . . . .	78
5.6	The average $n_{GO}$ value distribution for MIPS protein complexes . . . . .	79
5.7	The functional homogeneity <i>p-value</i> distribution for MIPS protein complexes . . . . .	80
5.8	Comparison of the number of known protein complexes . . . . .	82
5.9	Comparison of algorithms on the yeast PPI network . . . . .	82
5.10	Comparison of algorithms on the human PPI network . . . . .	83
6.1	Analyses of stability and reliability of MRF methods . . . . .	101
6.2	The variation of estimated parameters for adjacent steps by using the IMRF <sub>1</sub> method . . . . .	102
6.3	Comparisons of IMRF <sub>1</sub> , IMRF <sub>2</sub> and MRF-Deng . . . . .	103
6.4	Comparisons of different data integration methods with IMRF <sub>2</sub> analysis . . . . .	104
6.5	ROC curves of cross-validation results of different methods by integrating five biological networks . . . . .	105
7.1	Analyses of stability and reliability of MRF methods . . . . .	123
7.2	Comparisons of different kernels by using the kernel-based MRF method . . . . .	124
7.3	ROC curves of cross-validation results of different methods . . . . .	125
8.1	The general idea of the proposed logistic-regression-based algorithm . . . . .	135
8.2	Comparisons between different priors of the logistic-regression-based algorithm . . . . .	143
8.3	Comparisons between different feature vectors of the logistic-regression-based algorithm . . . . .	144
8.4	Comparison of the computational time among different algorithms . . . . .	145
8.5	ROC curves of cross-validation results of different algorithms . . . . .	146

## LIST OF ABBREVIATIONS

AP/MS	affinity purification followed by mass spectrometry
Acc	accuracy
AN	adding neighbors
AUC	area under the ROC curve
CE	cluster entropy
CGI	combining gene expression and protein interaction
CMC	clustering-based on maximal cliques
ClusterONE	cluster with overlapping neighborhood expansion
CS	clique seeds
DIP	database of interacting proteins
DIR	data integration rank
DME	dense module enumeration
FPR	false positive rate
GE	graph entropy
GO	gene ontology
HPRD	human protein reference database
IB	inner boundary
KLR	kernel logistic regression
KSV	keep seed vertices
LCMA	local clique merging algorithm
LED	Laplacian exponential diffusion kernel
MAP	maximizing a posteriori probability
MCMC	Markov chain Monte Carlo
MCL	Markov clustering
MCS	maximal clique seeds
MED	Markov exponential diffusion
MLE	maximum likelihood estimation
MMR	maximum matching ratio
MRF	Markov random fields
nAN	not adding neighbors
NMF	nonnegative matrix factorization
OB	outer boundary
OMIM	online Mendelian inheritance in man
PCC	Pearson correlation coefficient
PPI	protein-protein interaction
PPV	positive predictive value
RNSC	restricted neighborhood search clustering
ROC	receiver operating characteristic
RSV	remove seed vertices
RWR	random walk with restart
SGD	<i>Saccharomyces</i> genome database
Sn	sensitivity
TPR	true positive rate
uWVE	unweighted vertex entropy
WVE	weighted vertex entropy
Y2H	yeast two-hybrid

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

“Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life”, wrote by Zuckerkandl and Pauling (1962) in their chapter on “Molecular disease, evolution, and genetic heterogeneity” [1]. Over 50 years later, we are still far from unraveling mechanisms of many cellular processes and genetic diseases.

Within living cells, proteins rarely function as isolated entities, but rather interact with other proteins (i.e. forming protein complexes) to perform biological functions [2, 3]. The studies of proteins and their interactions are essential for understanding their roles within cells. On the one hand, since it is proteins that are responsible for the execution of cellular functions, it is important to understand how proteins are organized at the molecular level. On the other hand, since protein interactions are intimately related to gene-phenotype associations, uncovering mechanisms by which genes related to a specific phenotype (typically a human genetic disease) reveals information of interactions between their corresponding gene products.

With advances in high-throughput measurement techniques, various kinds of large-scale biological data have been produced, such as protein-protein interaction (PPI) data [4, 5], gene expression data [6], cellular pathways [7], gene-disease associations [8], etc. Generally, those large-scale biological data can be formulated as different kinds of biomolecular networks, such as PPI networks, gene co-expression networks, pathway co-existence networks, and gene-disease association networks, where individual genes and/or proteins are vertices, and specific biological relationships between them are edges. The analysis of those biomolecular networks is essential for us to understand the mechanisms of various molecular systems, and it has proven useful for developing new methods in disease diagnosis, disease treatment and drug design.

In this study, two topics related to the analysis of large-scale biomolecular networks are focused on: (1) the identification of protein complexes and (2) the identification of disease genes.

## 1.2 Motivation and objectives

A protein complex is a group of proteins that interact with each other in a living cell, forming a single multi-molecular machine [9, 10]. It is molecular entities that are responsible for most cellular processes [2]. In PPI networks, protein complexes are often assumed to be cliques or densely connected sub-graphs [2, 11], where vertices tend to have frequent connections within individual complexes, but rarely have connections with components outside of complexes. Another assumption used to identify protein complexes is based on the core-attachment topological structure [12]. The core of a protein complex consists of a constant set of proteins, which are highly co-expressed and share high functional similarity, while attachments just assist the core to perform subordinate functions [13].

Disease genes are related to a specific disease, that is, the behaviors of these genes at the disease state are significantly different from them at the normal state. Typically, a specific disease is related to multiple genes or proteins [14, 15]. Hence, the topic of disease gene identification is to find a set of candidate genes that are strongly related to a specific disease. The mutation of disease genes may cause the disease, or an occurrence of this disease can lead to their mutations or abnormal expressions. Various kinds of evidence have shown that genes related to the same or similar diseases are often “close” in different molecular networks, such as encoding proteins that are members of the same protein complex, participating in the same biological pathway or involving the same single transduction [14]. A “guilt-by-association” assumption [16] and a multiple data integration strategy are often used by various algorithms to identify disease genes, where a gene is more likely to be regarded as a disease gene if it is ranked as “closer” to known disease genes in various molecular networks.

Although various algorithms [11, 17–20] have been proposed to identify protein complexes and/or disease genes from many kinds of molecular networks, the prediction accuracy of those algorithms is limited and still has room to be further improved. For the protein complex identification topic, one of the key objectives is to identify the different components in individual protein complexes. Hence, the relationship of a protein with itself is ignored in this study. Many algorithms employ a seed-growth-style heuristic to identify protein complexes, which randomly selects individual vertices as seeds to search for local optimum clusters. However, in many cases a single protein is not enough to grow into a meaningful complex, and in many other cases more than one protein is known in a complex of interest. Since protein complexes generally exhibit intricate connections in PPI networks, a single-topological-structure-based algorithm is often not enough to identify all kinds of protein complexes. For the disease gene identification topic, many “guilt-by-association” based algorithms only take edges of a candidate gene with known disease genes into consideration, ignoring edges of the gene with non-disease genes. They ignore the fact that many biomolecular networks are built independently from the description of gene-disease associations. Although such a gene may be “close” to disease

genes, it is “closer” to those non-disease genes. It is not meaningful to simply label it as a disease gene. In addition, different kinds of biomolecular networks are heterogeneous. There is no widely acceptable criterion available to standardize them into the same level. When a multiple data integration method combines useful information from multiple datasets, it integrates their noise as well, which may not yield better performance in terms of disease gene identification. Based on this motivation, the objectives of this study are described as follows:

1. Reviewing current algorithms for identifying protein complexes and discussing their advantages and disadvantages.
2. Developing novel algorithms to identify protein complexes based on their topological characteristics from PPI networks.
3. Developing improved multiple data integration algorithms that consider edges of candidate genes with not only disease genes, but also with non-disease genes.
4. Developing a fast and high performance algorithm for disease gene identification.

### 1.3 Organization of the thesis

The thesis is organized in a manuscript-based style. It is presented in the form of published or prepared manuscripts. At the beginning of each chapter, a brief introduction is included to describe the connection of the manuscript to the context of the thesis. A general discussion chapter is also provided at the end of the thesis. All manuscripts have been re-formatted to be consistent across the thesis.

The remainder of the thesis is organized as follows: Chapter 2 presents a comprehensive review of protein complex identification algorithms from static PPI networks to dynamic PPI networks. Chapter 3 introduces a dense sub-graph detection algorithm based on clique seeds and graph entropy. Chapter 4 studies topological characteristics of known protein complexes and claims that not all protein complexes exhibit dense structures in PPI networks. Chapter 5 gives a multiple-topological-structure-based algorithm to identify protein complexes from PPI networks. Chapter 6 proposes a flexible and stable MRF-based algorithm to identify disease genes by multiple data integration. Chapter 7 further improves the MRF-based algorithm by including three kinds of graph kernels. Chapter 8 generalizes the feature construction idea of the MRF-based algorithms and presents a fast and high performance logistic-regression-based algorithm for disease gene identification. Chapter 9 draws conclusions, contributions and future work of this thesis. The list of related publications is included in Appendix A, and the copyright permissions of included manuscripts are in Appendix B.

## 1.4 Contributions of the primary investigator

It is noted that all papers presented in this dissertation are co-authored. However, it is the mutual understanding of all authors that Bolin Chen, as the first author, is the primary investigator.



## BIBLIOGRAPHY

- [1] Zuckerkandl E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. *In Kasha M and Pullman B (editors). Horizons in Biochemistry. Academic Press, New York. 1962: 189-225.*
- [2] Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stümpflen V, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 2008, **36**(Database issue): D646-D650.
- [3] De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 2010, **6**(6): e1000807.
- [4] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005, **122**(6): 957-968.
- [5] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004, **32**(Database issue): D449-D451.
- [6] Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. A global map of human gene expression. *Nat Biotechnol* 2010, **28**(4): 322-324.
- [7] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, **28**(1): 27-30.
- [8] McKusick VA. Mendelian Inheritance in man and its online version, OMIM. *Am J Hum Genet* 2007, **80**(4): 588-604.
- [9] Terentiev AA, Moldogazieva NT, Shaitan KV. Dynamic proteomics in modeling of the living cell. Protein-protein interactions. *Biochemistry (Mosc)* 2009, **74**(13): 1586-1607.
- [10] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, **100**(21): 12123-12128.
- [11] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* 2012, **9**: 471-472.

- [12] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868): 141-147.
- [13] Yu L, Gao L, Kong C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* 2011, **11**(19): 3826-3834.
- [14] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA* 2007, **104**(21): 8685-8690.
- [15] Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007, **71**(1): 1-11.
- [16] Altshuler D, Daly M, Kruglyak L. Guilt by association. *Nat Genet* 2000, **26**(2): 135-137.
- [17] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol* 2008, **4**: 189.
- [18] Köhler, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008, **82**(4): 949-958.
- [19] van Dongen S. Graph clustering by flow simulation. *PhD thesis, University of Utrecht*, 2000.
- [20] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006, **43**(8): 691-698.

## CHAPTER 2

# IDENTIFYING PROTEIN COMPLEXES AND FUNCTIONAL MODULES - FROM STATIC PPI NETWORKS TO DYNAMIC PPI NETWORKS

*Published as:* Chen B, Fan W, Liu J and Wu FX. Identifying protein complexes and functional modules - from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics* 2014, **15**(2): 177-194.

This chapter gives a comprehensive review of algorithms for identifying protein complexes and/or functional modules from PPI networks. It first summarizes some issues and pitfalls when analyzing PPI networks. Then, based on the type of data source and/or the main assumption of an algorithm, the chapter groups those identification algorithms into four categories, and reviews them respectively. Evaluation methods are also reviewed in this chapter.

### Abstract

Cellular processes are typically carried out by protein complexes and functional modules. Identifying them plays an important role for our attempt to reveal principles of cellular organizations and functions. In this article, we review computational algorithms for identifying protein complexes and/or functional modules from protein-protein interaction (PPI) networks. We first describe issues and pitfalls when interpreting PPI networks. Then based on types of data used and main ideas involved, we briefly describe protein complex and/or functional module identification algorithms in four categories: (i) those based on topological structure of unweighted PPI networks; (ii) those based on characteristics of weighed PPI networks; (iii) those based on multiple data integrations; and (iv) those based on dynamic PPI networks. The PPI networks are modelled increasingly precise when integrating more types of data, and the study of protein complexes would benefit by shifting from static to dynamic PPI networks.

## 2.1 Introduction

Cellular processes are typically not carried out by individual proteins, but rather by groups of proteins that interact with each other [1–3]. Understanding those groups of proteins is a critical step towards unraveling the intricate molecular relationship within cells [4–6]. Given a set of protein interaction data, a protein-protein interaction (PPI) network can be constructed by taking individual proteins as vertices and pair-wise interactions between them as edges. Large-scale PPI data have provided maps of molecular networks for several organisms [7–10]. Although most of them are incomplete and inaccurate, they reveal important principles of protein organization within cells.

Generally, two types of protein organization are commonly studied: protein complexes and functional modules. A protein complex is a group of proteins that interact with each other at the same time and place, forming a single multi-molecular machine [11, 12], while a functional module consists of a group of proteins participating in a specific cellular process, but proteins may interact with each other at a different time and place [4, 11–13]. They have both close relationship and different biological meanings. On the one hand, functional modules often contain one or multiple protein complexes in specific time and space. Therefore, they often exhibit similar characteristics in PPI networks [13, 14]. On the other hand, they are grouped according to different criteria. Protein complexes are specific molecular entities whose proteins tend to be co-localized and co-expressed [13, 14], whereas functional modules are grouped according to individual cellular processes whose proteins carry out different biological functions within those processes [13].

It is important to distinguish between protein complexes and functional modules, as they are different protein organizations [12–14]. However, owing to the lack of temporal and spatial information for pair-wise protein interaction data, it is not easy to make this distinction. Most computational algorithms can detect sets of proteins grouped as either protein complexes or functional modules. However, they hardly distinguish between them unless other kinds of (typically dynamic) data are further incorporated [12, 14].

In this article, we review computational algorithms for identifying protein complexes and/or functional modules from static PPI networks to dynamic PPI networks. According to types of data used and main ideas involved, all algorithms are organized in four categories. To start with, those based on topological structures of unweighted PPI networks are regarded as the first category. They are often designed to detect sub-graphs with specific topological structures in a PPI network, such as cliques [12, 15–17], dense sub-graphs [18–24], core-attachment structures [25–29] and star-like structures [30]. Although the predictive accuracy is limited, algorithms in this category play fundamental roles in the identification of protein complexes and/or functional modules. The second category consists of algorithms that are based on characteristics of weighted PPI networks. Numerous topological indices, such as the local neighbourhood density [31], the number of

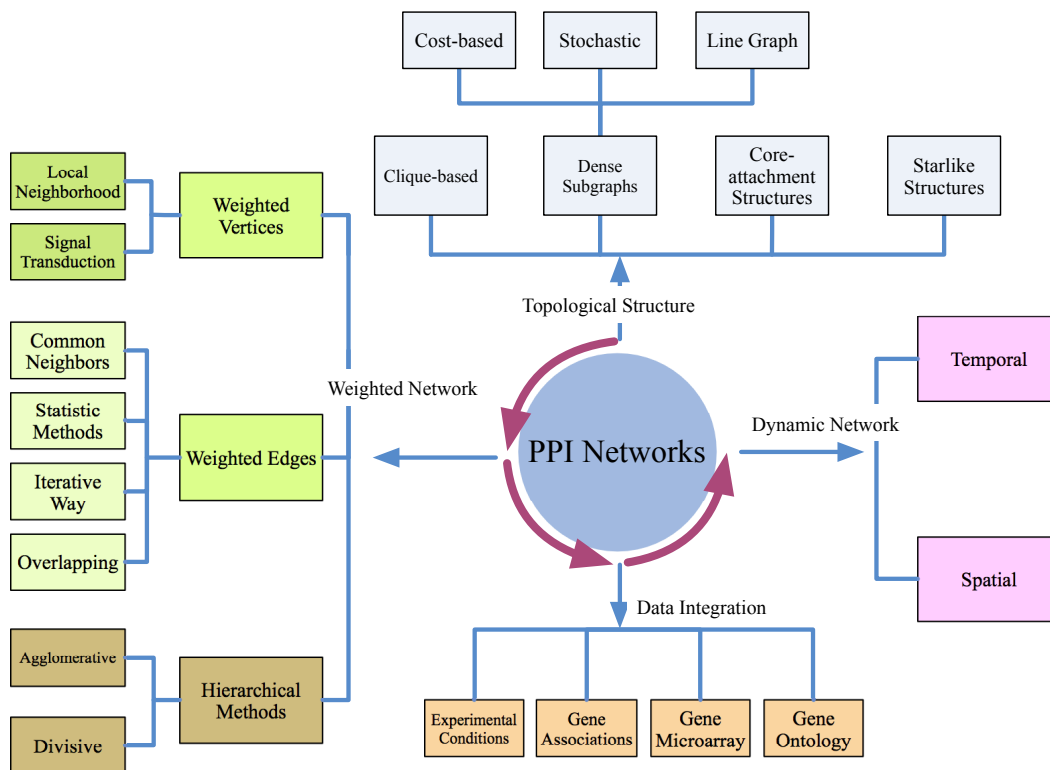
common neighbours [32–34], the edge-betweenness [35–37], the edge-clustering coefficient [38] and the shortest path [39], are used to assign weights to vertices or edges. Various hierarchical clustering approaches are also designed to partition a network into sub-graphs in this category [35–41]. Next, those involving ideas of multiple data integrations contribute to the third category. They use other sources of information, such as the experimental conditions [42, 43], gene expression profiles [44–48] and gene ontology (GO) [49, 50] to assign weights to edges of a PPI network, rather than using those topological indices. Generally, more biological meaningful results can be obtained by integrating more types of data. Finally, algorithms based on dynamic PPI networks are regarded as the fourth category. They also need to integrate multiple types of data, but they model PPI networks as dynamic systems [51–53], which are more reasonable for cellular systems. It is also possible to distinguish between protein complexes and functional modules in this situation.

A recent review article is also proposed by Srihari and Leong [54]. They provide an up-to-date survey, classification and evaluation of most key protein complex identification methods till 2012. The algorithms are organized as a chronology-based ‘bin-and-stack’ and a methodology-based ‘tree’ classification. Open challenges are also discussed in [54] for reconstructing accurate protein complexes. Another recent survey article is proposed by Li et al. [55]. They list majority protein complex detecting algorithms that have been developed till 2009. The surveyed algorithms are based on static network models, from pair-wise unweighted PPI networks to multiple data integrating. Both of them provide valuable insights for researches that have been done in this area. Differently, we focus on reviewing the related work according to different types of PPI networks in this article. Computational algorithms are organized according to the types of PPI networks being modelled and the main ideas involved in. By this way, one only needs to focus on what kinds of data are being used and accordingly select or design an appropriate algorithm. The organization of the article and the relationship of those computational algorithms are illustrated in Figure 2.1.

Before giving detail reviews of algorithms for identifying protein complexes and/or functional modules from PPI networks, it is important to make clear about characteristics of protein interactions and PPI networks, such as what protein interactions are, how PPI networks are established and what kinds of pitfalls should be aware when interpreting PPI networks.

## 2.2 Protein interactions and PPI networks

Protein interactions occur when two or more proteins bind together in a cell *in vivo* [2]. With advances in high-throughput proteomics technologies, such as yeast two-hybrid assay (Y2H) and affinity purification followed by mass spectrometry (AP/MS), etc., numerous PPI datasets have been produced for many organisms. The availability of those large-scale PPI data has led to the recent popularity of the study in PPI networks [56],



**Figure 2.1:** The organization of computational algorithms for identifying protein complexes and/or functional modules.

especially of those investigating principles of cellular organizations and functions. However, it should be careful when interpreting PPI data, especially to draw biologically relevant conclusions from reported PPI datasets.

The first issue is that most PPI datasets are not complete [56–60]. Various PPI databases are established by collecting reported PPIs from literature and experiments. However, experimental data reported in literature are only a small fraction of all biologically relevant PPIs. In addition, many databases have to manually collect PPIs from literature, which can obtain an even smaller fraction of the entire PPI space [3]. For example, for yeast PPIs which are extensively studied, only about 50% of them are reported [56, 57]. For human PPIs which cover much less, only about 10% of them are reported [56, 57].

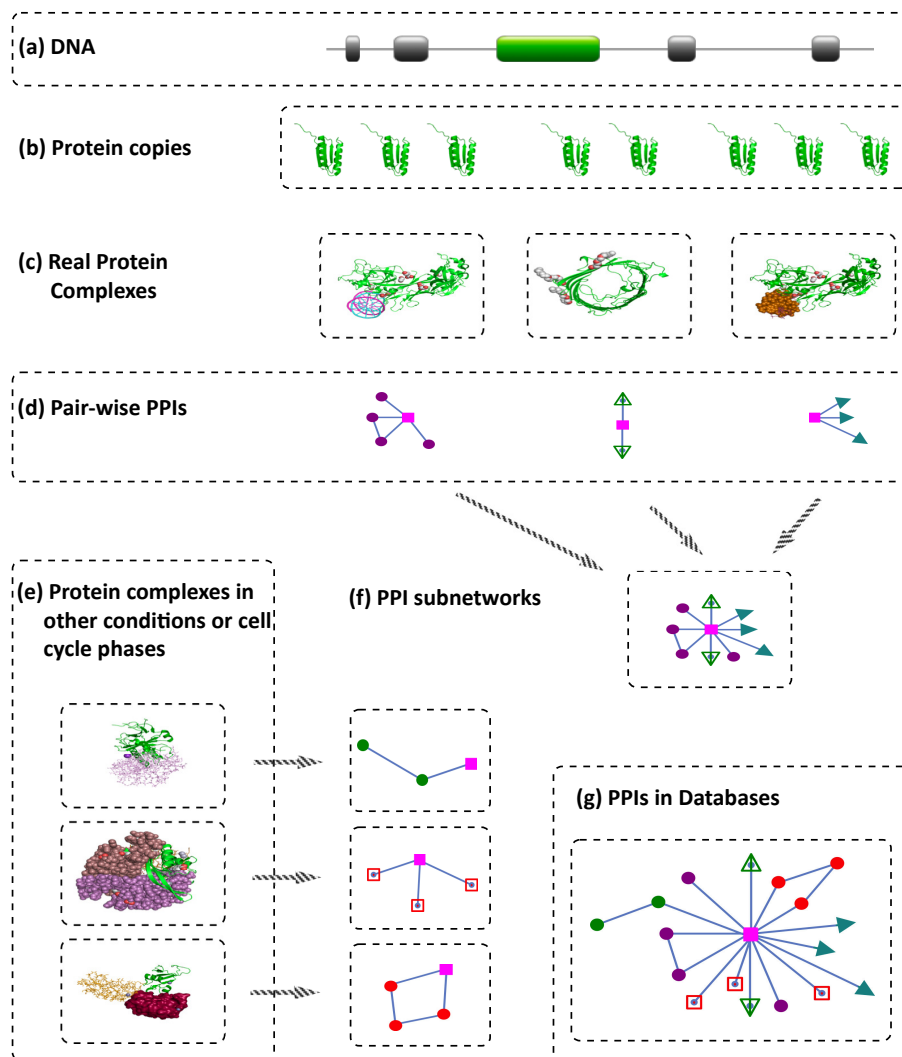
The second issue is that reported PPI data are not reliable. This problem arises from both the original experimental methods for identifying those interactions and the subsequent models for interpreting data generated using those experimental methods [61]. Data from both Y2H assay and AP/MS are subjected to high error rates if experiments are not performed under appropriate controls [62]. It is estimated that the reliability of Y2H assay, which reports pair-wise PPI data, does not exceed 50% [11]. The measurements are made under non-physiologic conditions, such that the observed interactions may not be present in the wild-type cells if two proteins are over expressed [62]. Although AP/MS detects protein interactions within

a native environment, it cannot distinguish whether binding of a prey to the bait is direct or indirect, due to the fact that AP/MS reports co-complex information of PPIs [2, 62]. Hence, an additional algorithm or model is needed to interpret co-complex observations into pair-wise interactions, which may introduce more noise, including both false-positive and false-negative PPIs [63].

The third issue is that reported PPI datasets are biased as a consequence of differences in the original detected interactions and the following handling methods [56]. On the one hand, some PPIs are more widely to be studied than others. Hence, the reported PPIs are biased towards proteins from particular cellular environments and towards proteins of more ancient, conserved and highly expressed ones [56]. On the other hand, the attempt of circumventing the problem of inaccurate data often makes the issue of biases more serious. Many algorithms [42–50] select only interactions that satisfy specific criteria by multiple validation and data integration. However, they introduce new biases into PPI networks because the validated datasets are further subsets of known PPIs [56]. The issue of biases is at least the same problematic as issues of data quality. It can alter the underlying structure of networks in unpredictable ways. As a result, the PPI networks we obtained may drastically be different from the real and complete networks.

Finally, there are several pitfalls associated with the form of pair-wise PPI networks. Firstly, such a PPI network is an integrated network. It is not the same as the real cellular interaction system. Within cells, one kind of protein should have numerous copies, each as a specific molecular entity. Some copies may interact with one group of proteins, while some others interact with other groups. When it comes to a pair-wise PPI network, a vertex of the network represents a collection of all that kind of protein, rather than those individual protein copies [64]. This is the reason why a hub vertex can bind hundreds of ‘proteins’ in a PPI network, while it is impossible in biological cells. Secondly, a PPI network is actually an integration of many subnetworks, including both local protein organizations and global PPI networks, measured under various experimental conditions and cell cycle phases. Those sub-networks are also collected from different experiments that done in various laboratories. As a consequence, the integrated PPI network contains interactions happening in various times and spaces, no matter whether they happen simultaneously or not, or whether they are exclusive or not. Thirdly, it is easier to include data into a database than to clean them out. Because there is currently no simple way to report ‘negative’ interactions [65], the previous inaccurate interactions will affect the quality of entire datasets for a long time. Figure 2.2 illustrates the general way about how various experimental PPIs are transformed into PPI networks in a database.

Although PPI data in various databases are problematic, biologically relevant conclusions can be drawn with an extra carefulness when interpreting PPI networks. Mackay et al. [61, 65] and Chatr-aryamontri et al. [42] have discussed thoroughly and clearly about PPI data in their *TiBS* letter and response articles. Mackay et al. [61] first analyse the reliability of reported PPI data, and they conclude that many reported PPIs might not occur as presented. Chatr-aryamontri et al. [42] later argue that sensible and biologically relevant



**Figure 2.2:** Experimental protein interactions and a combined PPI network. (a) A protein-coding gene in DNA. (b) Some protein copies that coded from the same gene. (c) Real protein complexes that involve that protein copies. (d) Pair-wise PPIs that were obtained from experimental analysis of protein complexes. (e) Protein complexes in other experimental conditions or cell cycle phases. (f) PPI sub-networks that were obtained by various experiments or done by different laboratories. Each set of data represents a local structure of the PPI network. (g) PPI dataset in databases, which are established by collecting reported PPIs from literature and experiments.



conclusions can be obtained by integrating various experimental data. Mackay et al. do agree with it, they still appeal researchers to pay more attention to pitfalls associated with PPI networks in [65]. However, with the improvement of various PPI databases [66–69], it is believed that results from interpreting PPI networks would become increasingly reliable and thus important in current research on protein science.

## 2.3 Identifying protein complexes based on topological structures of unweighted PPI networks

Protein complexes exhibit specific topological structures in PPI networks. Systematical cataloguing all those protein complexes and their interactions within living cells is one of the key topics in post-genomic biomedical research [70]. Although it is difficult to tell their identical structures, various attempts have been made based on topological structures, such as cliques [12, 15–17], dense sub-graphs [18–24], core-attachment structures [25–29] and star-like structures [30], to identify protein complexes from PPI networks. Actually, cliques are special cases of dense sub-graphs, while star-like structures are special cases of core-attachment structures. Taking cliques as a particular kind of category is due to the fact that many computational algorithms use cliques as candidates or components of protein complexes. Taking star-like structures as a particular kind of category is due to the fact that many core-attachment based algorithms only identify dense sub-graphs as cores, which may miss plenty of predictions that exhibit star-like structures.

Firstly, cliques are often used as candidates or components of protein complexes. To start with, Spirin and Mirny [12] propose an iterative algorithm to enumerate all cliques in a network. Starting from cliques of size  $n$ , one can enumerate all cliques of size  $n + 1$  by checking each adjacent vertex of previous cliques. If there is no vertex can be added to form a larger clique, one can also obtain a maximal clique simultaneously. It does not take too long time to enumerate all cliques in a PPI network because most of them are very sparse. However, simply using those cliques as candidates of protein complexes does not obtain a high accuracy. Alternatively, Li et al. [15] design the LCMA (Local Clique Merging Algorithm) to detect protein complexes by using cliques. They first locate local cliques in the network for each vertex, and then merge overlapped cliques as predictions of protein complexes according to their affinity to form maximal dense sub-graphs. Moreover, cliques can be used to construct a new graph for purposes of protein complex identification. One example based on this idea is CFinder [16, 17]. It first detects all  $k$ -cliques in a PPI network. Then, based on the definition that two  $k$ -cliques are accessible if they share  $k-1$  vertices, a  $k$ -clique accessibility graph can be constructed by taking individual  $k$ -cliques as vertices and the accessible relationships as edges. The connected components of the accessibility graph are then used to generate overlapping protein complexes, which are unions of all  $k$ -cliques that can be reached from each other through a series of adjacent  $k$ -cliques. Results of the CFinder are highly correlated to the value of the parameter  $k$ . Larger values of  $k$  tend to

reduce the number of adjacencies, and therefore may result in smaller protein complexes in the network.

Secondly, many algorithms [18–24] are designed to detect dense sub-graphs as candidates of protein complexes. This is due to the fact that proteins tend to exhibit strong interactions within a complex and weak interactions to proteins outside the complex [71]. However, there has not been a generally accepted quantitative definition for dense sub-graphs. They usually are described as sets of vertices within a network such that the connections between those vertices are denser than connections to the rest of the network [38]. Various cost-based methods [18–21], stochastic approaches [22, 23] and line-graph-based algorithm [24] are developed to identify dense sub-graphs in PPI networks.

In the first instance, cost-based methods usually define specific cost functions to calculate the cost of a partition in PPI networks. Local dense sub-graphs are obtained by optimizing those costs. The RNSC (Restricted Neighborhood Search Clustering) algorithm [18] is one of such methods. The cost function is calculated according to the number of invalid connections for each vertex. The algorithm starts from a random user-specified partition, and iteratively moves a vertex from one cluster to an adjacent cluster to decrease the total cost. It ends up with a partitioning of the network if some moves have been reached without decreasing the cost function. The output clusters are filtered according to criteria, such as the cluster size, the cluster density and the functional homogeneity. Cho et al. [19, 20] propose an entropy-based graph clustering algorithm that assigns a cost for each cluster. The vertex entropy is defined according to the connectivity of that vertex. The graph entropy is calculated by summing all vertex entropy in a graph. It is also a seed-growth style algorithm. Starting from a random seed vertex and its neighbours, the algorithm iteratively removes and adds vertices on the boundary of the cluster to minimize the graph entropy. The process of seed selection and optimal cluster generation is repeatedly performed for all candidate seeds. Chen et al. [21] suggest using cliques as initial seeds, rather than individual vertices. The entropy-based algorithm is used as an example to show how clique seeds can be used to increase the predictive accuracy of protein complex identification [21].

In the second instance, stochastic approaches handle the problem of dense sub-graph identification from a statistic point of view. One of such algorithm is called MCL (Markov CLustering) [22, 23]. It works by simulating random flows in a graph. The process takes a stochastic matrix as input, which represents the transition probabilities between all pairs of nodes. The self-loop of each vertex is added initially, and the loop weight is assigned as the maximum weight of all edges connected to the vertex. It changes the values of the transition matrix at each step according to the previous one until a stochastic condition is satisfied. Two processes, expansion and inflation, are interactively involved during the simulation. The expansion takes the  $e^{th}$  power of the stochastic matrix, while the inflation promotes the dense clusters and weakens the sparse clusters. Because greater path lengths are more common within clusters than between different clusters, the expected behaviour of random flows results in community structures of the original network. In practice, the

MCL algorithm converges very fast, and it is highly scalable in terms of predicting protein complexes from PPI networks.

In the third instance, the line-graph-based approach [24] gives another way to identify dense sub-graphs from PPI networks. It first transforms a PPI network into its line graph, and then applies the MCL algorithm on this new graph. The procedure of this transformation brings a number of advantages for graph clustering. First, it does not sacrifice any information of the original graph. Second, it amplifies the higher-order local neighbourhood of connections. Third, it is more highly structured than the original graph. The algorithm can produce overlapping sub-graphs in PPI networks.

Thirdly, core-attachment structures are commonly used to identify protein complexes from PPI networks. Gavin et al. [72] have demonstrated that a protein complex should generally contain a core and attachments. A core in a protein complex is formed by a constant set of proteins, which are highly co-expressed and share high functional similarity [11, 25]. The attachments surrounding the protein complex core assist in performing subordinate functions [25]. This property is also supported by other high-throughput protein data [73].

In terms of identification algorithms, Leung et al. [26] propose a method to identify cores and attachments of protein complexes separately. They use a *p-score* to evaluate how likely a potential core would be the core component of a complex, according to the number of interactions between the potential core and the rest of the networks. Then neighbours that have interactions with the majority of the core are added to form a protein complex. Wu et al. propose the COACH (core-attachment-based method) in [27]. They identify cores based on the neighbourhood graphs of vertices, and then adding attachments into these cores to form candidate clusters. Biologically meaningful clusters are then selected as final predictions. Other algorithms based on similar heuristic can be found in [25, 28, 29].

Last but not least, star-like structures are recently proposed to identify protein complexes from PPI networks. Chen et al. [30] investigate topological structures of known protein complexes in a *Saccharomyces cerevisiae* PPI network. They find that many protein complexes exhibit star-like structures. That is, proteins within individual complexes tend to have interactions with only one or a few hub proteins, while most proteins do not interact with each other. A random-star algorithm is also proposed to identify star-like structures in PPI networks [30].

## 2.4 Identifying protein complexes based on characteristics of weighted PPI networks

Interpreting large-scale PPI data is a challenging task because of the widespread of false positive (FP) [74]. To minimize the effect of those inaccurate data, various weighted strategies are used for identifying protein complexes in PPI networks. Although it is hard to assess the reliability of a single edge weight, Nepusz et al. [75] argue that taking into account network weights globally can greatly improve the detection of protein complexes. Therefore, weights should be used when available. Both weights of vertices and edges can be assigned to increase the reliability.

Bader and Hogue [31] propose the MCODE approach based on a strategy of weighting vertices. The algorithm is made up of three steps. At first, the weight of each vertex  $v$  is assigned based on the local neighbourhood density, which is defined from the density of the highest local  $k$ -core of  $v$  and the value of  $k$ . Then, starting from the vertex with the highest weight, a cluster is obtained by recursively including neighbour vertices whose weights are above a given threshold. Finally, the algorithm iteratively removes one-degree vertices to form a cluster in the ‘haircut’ process and adds connected vertices to the cluster if the neighbourhood density exceeds a given threshold in the ‘fluff’ process.

Hwang et al. [76] develop a weight strategy for PPI networks in a different way. They propose a STM (signal transduction model) for PPI networks, and demonstrate the signal transduction behavior of the perturbation by each vertex on a PPI network statistically. For each vertex  $v$ , the signal between  $v$  and  $w$  is modeled using the Erlang distribution, where  $w$  is any vertex in the network except  $v$ . Preliminary clusters in the network are formed by using this weighted relationship among all vertices, and then predicted clusters are generated by a merging process. It allows overlapping of output clusters, and can identify clusters with a large size, arbitrary shape, and low density. However, unexpected huge clusters may also be generated in the post-process of merging.

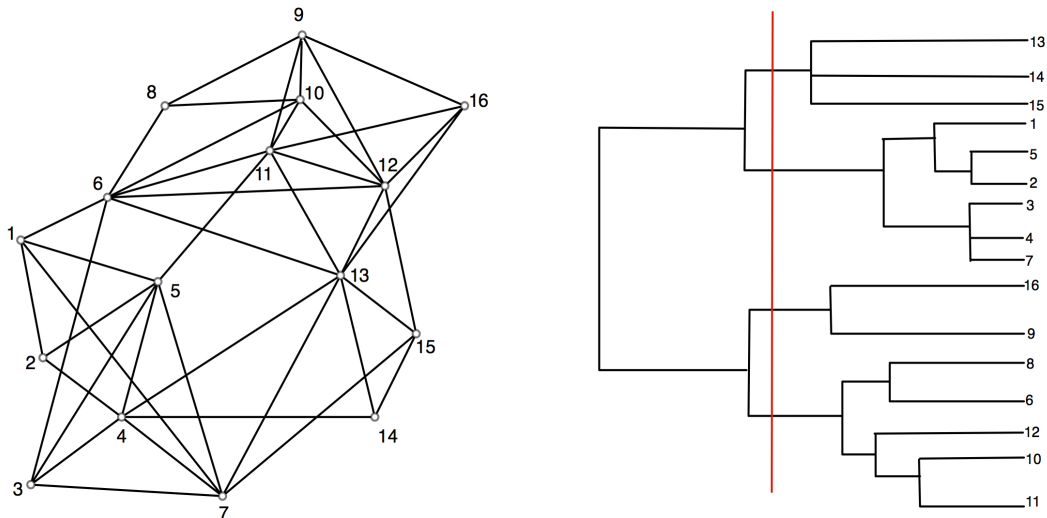
The number of common neighbours between two vertices is a kind of widely used information to assign weights to edges. Altaf-Ul-Amin et al. [32] design the DPCLus algorithm to assign weights to edges in this way. Then the weight of a vertex is assigned by summing weights of edges that are incident to it. A seed-growth style strategy is developed to generate clusters according to the edge weight and the vertex weight. Li et al. further modify the DPCLus algorithm and propose the IPCA in [33]. The rationale behind this algorithm is that most complexes have a very small diameter and a very small average vertex distance. They use the same process to assign weights to edges and vertices, but generate clusters based on a new criterion. The DPCLus identifies clusters that satisfy a density condition and certain cluster connectivity property, while the IPCA generates clusters that have a small diameter and satisfy a different cluster connectivity-density

property. Once a cluster is identified the DPCLust removes the cluster and recalculates the vertex weights based on the new remaining graph; while the IPCA computes the vertex weights based on the original graph only once. Kim and Tan [34] propose the miPALM (module inference by Parametric Local Modularity) that combines the parametric local modularity measure and the greedy search strategy to identify communities in PPI networks. It first assigns weight to edges by using the number of common neighbors and vertex degrees. Each triangle of the weighted network is ranked according to the parametric local modularity and expanded to candidate complexes by a recursive greedy search. Additional parameters are used to control the background neighborhood size around candidate complexes and to filter unreasonable results.

Statistic approaches are also involved by comparing the known PPI network with a random network of the same size. Samanta and Liang [74] rank the statistical significance of forming shared partnerships for all protein pairs in a PPI network and find that two proteins have close functional associations if they share a significantly larger number of common neighbors than random. They use *p-value* to rank all pairs of proteins in the PPI network and select only interactions with a *p-value* smaller than a threshold. Clusters are then generated in the weighted network. The algorithm is stable. Even adding 50% randomly generated interactions to the PPI dataset, it can still recover 89% of the original associations. Li and Liang [77] further use this heuristic by comparing a PPI network with truncated power-law preserving random networks, and find that the likelihood of two proteins sharing a common or related biological function can be enhanced if they share significantly more neighbors than random. They adopt this idea to investigate the functional relationship among proteins of a human PPI network.

Because a weighted PPI network is often more accurate than the initial pair-wise one, the weighted PPI network itself can be used to improve the weight in an iterative manner. Liu et al. [78] propose an iterative scoring method to reassign weights for edges and develop the CMC (Clustering-based on Maximal Cliques) method to identify protein complexes from PPI networks. The initial weight of an edge is calculated from the AdjustCD-distance, by using the neighborhood information and two penalty parameters. Although this iterative scoring method can effectively reduce the impact of random noise, more iterative steps do not necessarily perform better results. They suggest that two iterative steps is usually a safe choice. The CMC algorithm then generates all the maximal cliques from the weighted PPI network. Highly overlapped cliques are removed or merged to achieve the final predictions of complexes.

For any given weighted PPI network, Nepusz et al. [75] recently design the ClusterONE (Cluster with Overlapping Neighborhood Expansion) to detect overlapped protein complexes from the network. They argue that a meaningful candidate cluster representing a protein complex should have two structural properties. First, it should contain many reliable interactions within the cluster. Second, it should be well separated from the rest of the network. Based on this heuristic, they define a cohesiveness score for a group of vertices, which considers the total weight of edges within those vertices and total weight of edges between these vertices

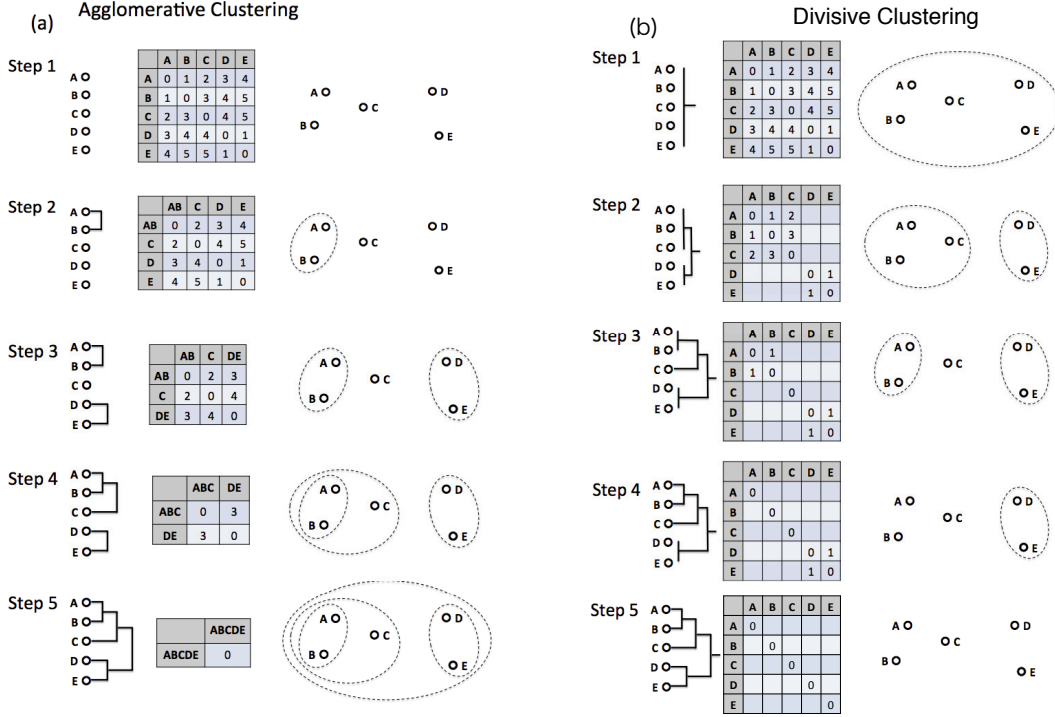


**Figure 2.3:** The general procedure of hierarchical clustering algorithms. The network is first transformed into a dendrogram, and then identify communities according to branches from the joint nodes.

and the rest of the network. A penalty term is also included to model the uncertainty of the undiscovered interaction data. Starting from each seed vertex, the algorithm iteratively generates high cohesiveness clusters by using a greedy procedure. After that, clusters are merged if the overlap score is above a specified threshold, and candidates that contain less than three proteins or whose density is below a given threshold are removed. The cohesiveness measures how likely a cluster is to form a protein complex, which provides an easy and efficient way to assess predictions for almost all algorithms. Wang et al. [79] propose the EPOF (Essential Protein and lOcal Fitness) by using essential proteins and the local vertex fitness. The fitness of a sub-graph is defined similar to the cohesiveness score used in [75], without the penalty term. Then the vertex fitness of  $v$  is defined for a sub-graph as the difference of the sub-graph fitness with and without the vertex  $v$ . A seed growth style algorithm is proposed in [79], where cliques that consist of only essential proteins and those do not contain any essential proteins are used as seeds, respectively.

Various hierarchical clustering algorithms also make contributions for identifying communities from different networks. Most of them can be used in PPI networks in terms of detecting protein complexes. Biological processes usually exhibit hierarchical structures in which proteins physically bind together as stable complexes [6]. The general procedure of hierarchical clustering algorithms is illustrated in Figure 2.3, where a weighted network is commonly transformed into a dendrogram [38]. The leaves of the dendrogram represent the vertices of the network, while the branches from joint nodes indicate groups of clusters in the network. Therefore, identifying hierarchical structure of clusters equals to designing a way to generate such dendrogram and assigning joint nodes for branches.

Typically, two methods of generating the dendrogram of a network are used: agglomerative and divisive. The



**Figure 2.4:** Schematic of the (a) agglomerative and (b) divisive clustering methods. In agglomerative clustering, the distance between two clusters is calculated by using the single-linkage method. In the divisive clustering, edges of high distance within a cluster are removed until the cluster breaks into two separated clusters.

agglomerative method starts from the state of all vertices in distinct clusters. The similarity of each pair of vertices in the network is calculated, which represents how closely the vertices are connected. Vertices and/or branches are iteratively organized into the hierarchical structure by merging the highest similar clusters step-after-step. In this method, the dendrogram is built from leaves to the root, where all vertices of the network in one cluster. In contrast, the divisive method builds the dendrogram in a reverse order. It first starts from all vertices in one cluster, and then subsequently splits the big cluster iteratively into smaller ones identified as clusters. In this manner, the dendrogram will down to the level of single vertices. In practical, additional information is needed to decide which branches of the dendrogram have real significance [38]. Figure 2.4 illustrates the basic idea behind the agglomerative and divisive clustering method.

Two issues are usually considered in a hierarchical clustering algorithm: assigning weights to edges for iteratively merging/splitting clusters and designing quantitative measures to evaluate output clusters.

For the first issue, Girvan and Newman [35] introduce a divisive algorithm, the G-N algorithm, based on the value of ‘edge betweenness’. The betweenness of an edge is defined as the number of all shortest paths running through it [35]. The rationale behind this idea is that a highly organized network is filled with densely inter-community edges and sparsely intra-community edges. Therefore, all shortest paths between

vertices of different clusters have to go through a few intra-community edges, thereby obtaining higher betweenness values. The algorithm iteratively removes the edges of the highest betweenness until a given network breaks into desired number of clusters. The G-N algorithm represents a major step in terms of identifying communities in networks, and is widely adopted to investigate functional associated communities in PPI networks in the past years. Dunn et al. [36] apply the G-N algorithm on a small set of human protein interactions to investigate biological functions involved in them. Newman [37] propose a new agglomerative algorithm to improve the computational efficient of G-N algorithm. At the same time, Radicchi et al. [38] also develop a fast algorithm to address the similar issue. They alternatively introduce an *edge-clustering coefficient* by considering the number of triangles that builds on edges. Edges connecting vertices in different communities are included in a few or no triangles and tend to have small values of edge-clustering coefficients. Rives and Galitski [39] use the all-pairs-shortest path matrix to defined an association for each pair of vertices in a network. The association is calculated by  $1/d^2$ , where  $d$  is the length of the shortest path. Then they develop an agglomerative algorithm based on the average linkage to reveal the modular organizations in yeast signaling networks. Wang et al. [40] propose a fast agglomerative algorithm, called HC-PIN, by using the number of common neighbours to calculate the clustering value of individual edges in a weighted PPI network. Cho and Zhang [41] introduce another way to use the hierarchical idea to identify functional hubs and modules in a network. They propose an algorithm by exploring two intrinsic topological features of PPI networks: the high modularity and the hub-oriented structures. A weighted PPI network is taken as input, and a path strength model is designed to measure the functional similarity between protein pairs. Then the network is converted into a hub-oriented hierarchical structure, and communities are generated by using the score of hub confidence.

For the other issue of designing the measure to evaluate clusters of hierarchical algorithms, various quantitative measures are proposed. Newman and Girvan [37, 80] introduce a measure called the *modularity*  $Q$  by comparing the observed fraction of edges inside a cluster with expected fraction of edges in the cluster. It is defined on the global sense. However, in many networks, sub-graphs are only locally connected. Based on this idea, Muff et al. [81] give a local version of the modularity measure,  $LQ$ , by considering only the immediate neighbors of a cluster, rather than the entire network. Kim and Tan [34] extend this idea by introducing a *coarseness* parameter. Li et al. [82] argue that the modularity  $Q$  has been exposed to resolution limits. The size of a detected community by  $Q$  depends on the size of the whole network, which may fail to identify modules smaller than a scale. Alternatively, they propose a modularity density, which they call  $D$  *value* based on the concept of average modularity degree. Zhang et al. [83] further extend the  $D$  *value* into a more general case, where a tuning parameter  $\lambda$  is introduced. They also adopt the simulated annealing algorithm to maximize the modularity density. Radicchi et al. [38] define the concepts of *strong community* and *weak community*, by considering the connections within a cluster and that toward vertices in the rest of the network. It gives a general criterion for deciding which detected sub-graphs are meaningful. Chen



and Yuan [44] extend the idea to a directed graph, and propose a quantitative measure in both strong and practical sense.

## 2.5 Identifying protein complexes and/or functional modules by multiple data integrations

It is believed that no single experimental approach can reach the sensitivity of 100% (i.e. no false negative) and the specificity of 100% (i.e. no false positive) [42]. The data emerging from individual ‘omic’ approaches should be viewed with caution [84]. Moreover, large-scale PPI data usually do not readily allow one to discriminate their various features [85], such as the interaction strength (affinity), the type of interactions (protein-protein interaction or protein-peptide interaction), and spatiotemporal existence (where and when the proteins are present and interact). However, it does not mean that we can do nothing to deal with these issues. The approaches multiple data integrations can achieve this goal to some extent.

Various kinds of data contain the information of protein interactions. Besides the high-throughput technologies, such as the Y2H assay and AP/MS, many other kinds of information, such as the reliability of experiments, the gene expression profiles (gene microarray, co-expression), the GO terms [86], the subcellular localization annotations [87], can be used to assess the reliability of PPIs and their biological features. Of course, the additional cautions should still be emphasized here. As Hakes et al. [56] remind that keeping only those highly reliable data may introduce new biases about the PPI data. Reasonable ways of data integration need to pay attention to the interpretation of PPI networks.

Firstly, the reliability of experimental technologies is used employed to evaluate PPIs. It is clear that, on the one hand, interactions observed at multiple times should be more likely to be true than those that have only been observed once, and on the other hand, the reliability of different experiment methods are not always the same. Therefore, one way to achieve the high reliability of PPIs is to assign different weights to interactions according to different times they are reported and different types of experiments they are derived from. Chatr-aryamontri et al. [42] have concluded that sensible, biologically relevant results can be obtained by integrating multiple interaction evidences. For instance, Tan et al. [43] first build interaction-specific networks independently from six groups of data. The integrated network is obtained by the weighted combination of individual networks. The MINT [66] is one of the databases that annotates various information such as detection methods, expression levels, protein tags, *in vivo* and *in vitro* conditions, the experimental role, post-translational modifications and so on [42], which can be used to evaluate the reliability of protein interactions from the experimental pointviews.

Secondly, genomic associations are believed to reflect functional associations between their proteins [45]. It is acknowledged that the strength of genomic association correlates with the strength of protein interactions. Various genomic contexts, such as gene fusions, gene co-occurrences, gene expression profiles, phenotype data and transcription factor binding data have been used to predict functional associations [88]. Tanay et al. [46] propose a biclustering algorithm that integrates genomic data to partition the molecular network of yeast. They use a weighting scheme on a bipartite graph to identify groups of genes with statistically significantly correlated behavior. Snel et al. [45] introduce a method that integrate genomic associations to identify functional modules. Among those kinds of genomic information, gene expression profiles are most commonly used for data integration. Genes with similar expression profiles tend to encode proteins that interact with each other [47]. Integrating PPIs and gene-expression data can generate a meaningful biological content in terms of identifying functional associations [48]. Chen and Yuan [44] use the abundant information of microarray expression profiles to assign weights to edges of the PPI network. The weight of an edge represents the dissimilarity between two associated expression profiles. They extend the idea of edge-betweenness to a ‘non-redundancy’ way. The shortest paths are not enumerated among all-against-all verities, but rather the non-redundancy ones. An extended G-N algorithm is also proposed to find functional modules in weighted PPI networks in [44].

Thirdly, the GO terms [86] contribute to another resource that can be used to assign weights to PPI networks. The weights of edges in PPI networks can be assigned by the semantic similarity of the relative GO terms. It is an effective way to identify protein complexes than the unweighted ones. Lubovac et al. [49] use two measures, called *weighted clustering coefficient* and *weighted average nearest-neighbors degree*, to assign weights to protein interactions. They are calculated from Lin’s similarity [89] of GO terms. The SWEMODE (Semantic WEights for MODule Elucidation) algorithm is developed to identify communities containing functionally similar proteins. It first ranks vertices in the network according to their weighted clustering coefficient. Those with the high rank are iteratively selected as seeds to generate densely connected clusters with high functional similarity according to the chosen parameters. Xu et al. [50] propose the OIIP approach to identify protein complexes from a weighted PPI network. The weight of an edge is assigned according to the annotation size of GO terms while the weight of a vertex is assigned by summing weights of incident edges. A seed-growth-style is applied on this network similarly to the way that the IPCA [33] algorithm used.

Finally, many algorithms try to integrate more kinds of information for PPI networks. Shi and Zhang [90] first use GO to build a weighted PPI network. Then a semi-supervised learning method is developed to learn features of protein complexes. There are 21 features of protein interactions employing in their multi-layer neural network model, all of which are used to identify protein complexes in the weighted networks. Georgii et al. [91] develop the DME (Dense Module Enumeration) to detect all clusters that satisfy a user-defined minimum density threshold in a given weighted network. The weight can be determined by any additional information, such as gene expression, phenotype data, evolutionary conservation and subcellular localization.

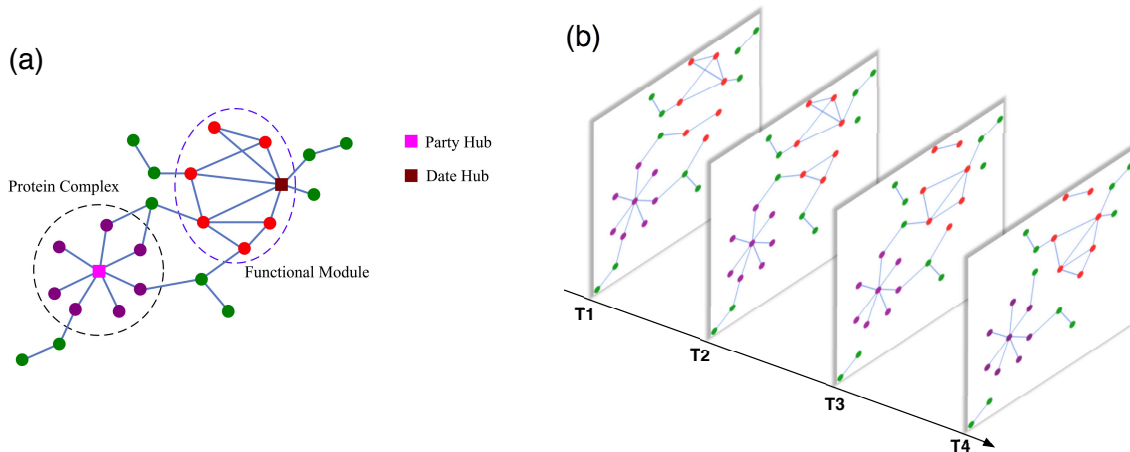
Luo et al. [92] propose a framework for discovering conditional co-regulated protein complexes by integrating transcription regulation data, gene expression data and PPI data. Jansen et al. propose a Bayesian approach to combine multiple types of data to reconstruct PPI network in [93].

## 2.6 Distinguishing between protein complexes and functional modules via dynamic PPI networks

Although PPIs imply physical contact between proteins, it does not mean that all possible interactions occur in any cell at any time. PPIs are not static but dynamic [71, 94, 95]. They vary with time and space that mediate protein complexes to assemble and disassemble as cellular processes [11, 85]. It is thus crucial to understand a PPI network in a sense of dynamic, such as how the cellular system responds to cues of environment and how it changes during development or differentiation [95]. It is essential to shift the analysis of PPI networks from static to dynamic for further understanding of molecular systems [51].

The large-scale PPI datasets are unable to capture the dynamic properties of protein interactions [96]. The challenge now becomes how to grasp dynamic behavior of PPI networks and how to figure out which interactions occur simultaneously. The commonly used way is by projecting additional information onto PPI networks [85, 97]. The temporal dimensionality of PPI networks can be enhanced by linking protein complexes to time series of gene expression data [11], while the spatial information can be partly handled from the subcellular localization annotations [98]. For instance, de Lichtenberg et al. [98] use both those data to investigate the dynamics of protein complexes during the yeast cell cycle. They find that almost all complexes contain both dynamic and static subunits. Most of them cannot be identified through the analysis of any single type of experimental data, but only through integrative analysis of several types of data. Moreover, condition-specific co-expression information also gives a way to achieve the dynamic features of the networks. Lin et al. [99] integrate PPIs with biological annotations and gene expression profiles to reveal dynamic functional modules under conditions of dilated cardiomyopathy. They show that hub proteins tend to be differentially expressed in different biological states. However, Lu et al. [100] argue that hubs and superhubs tend to have similar gene expression profiles under conditions of experimental asthma, by comparing with peripheral vertices based on the GO classification. Moreover, Han et al. [101] investigate how hubs might contribute to robustness and other cellular properties in the yeast PPI network. They find two types of hubs: party hubs and date hubs. Party hubs interact with most of their partners simultaneously to function inside modules, whereas date hubs bind their partners at different times or locations to organize the proteome, and connect biological processes.

A generalized framework to identify communities in a dynamic network is introduced by Mucha et al. [52]. It



**Figure 2.5:** Static PPI network and dynamic PPI network. (a) A party hub, a data hub, a protein complex and a functional module in static PPI network. (b) The protein complex and functional module can be distinguished by checking their existence in individual slices. The party hub and date hub can also be identified by checking protein interactions in different time points.

can be used in time-dependent, multiscale, and multiplex networks that containing arbitrary slices. Each slice represents a network at a specific time point. In terms of identifying protein complexes from PPI networks, the time-course microarray data can be used to reconstruct such dynamic behaviour. The composition of protein complexes and/or functional modules may change during a cell cycle. Even in the same time, a ‘protein’ may also be involved in several different processes (by different protein copies). A party hub can be identified from the network in each slice, while the date hub can be detected by considering multiple slices. It is also possible to distinguish between protein complexes and functional modules from such dynamic PPI networks by checking whether detected communities are in individual slices or not. Figure 2.5 gives a simple schematic for reconstructing dynamic PPI networks from a small static PPI network.

It is proposed that different protein modules can be found in the vertices of dynamic PPI networks [11]. Permanent interactions are strong and stable, which give rise to protein complexes, while the transient interactions vary with cellular processes and form functional modules. The transient interactions are equally important, since they play a major role in signal transduction. Yu et al. [5] find the bottlenecks in protein interaction networks are key connectors that correspond to the dynamic properties. Komurov and White [102] conclude that static and dynamic modules in the eukaryotic protein interaction network have distinct properties. Static modules provide robustness to the cell against genetic perturbation or protein expression noise while dynamic modules are mainly responsible for condition-dependent regulation of cell behaviors. Taylor et al. [103] examine the dynamic structure of the human protein interaction network. They argue that, similar to the date hub and party hub, inter-modular hubs co-express with their partners in a tissue-restricted manner while intra-modular hubs co-express with their partners in most tissues.

In the framework of the dynamic network, Mucha et al. [52] develop a way to detect communities from such

multiple slice networks. Similarly to the situation in static networks, the way they proposed to quantify communities is by comparing the number of intra-community edges to what one would expect at random. Three types of connection are considered: intra-slice connections, inter-slice connections between only neighboring slices and inter-slice connections among all-to-all slices. A multiple adjacency matrix is also defined to handle the problem of community identification. Jin et al. [53] define a dynamic network module to be a set of proteins satisfying two conditions. First, it is connected in the static PPI networks. Second, the expression profiles of its vertices form certain structures in the temporal domain. Then they detect dynamic modules in temporal networks by a mining algorithm. Although most dynamic network modules are highly condition-specific, they further demonstrate that identifying frequent dynamic modules can significantly increase the signal to noise separation. Tang et al. [51] propose a time course PPI model to identify functional modules. Time series gene expression data are used to construct the network. Although the temporal parameters are not sufficiently accurate, they find that the functional modules from the time course network have much more significant biological meanings compared with the static PPI network and a pseudorandom network.

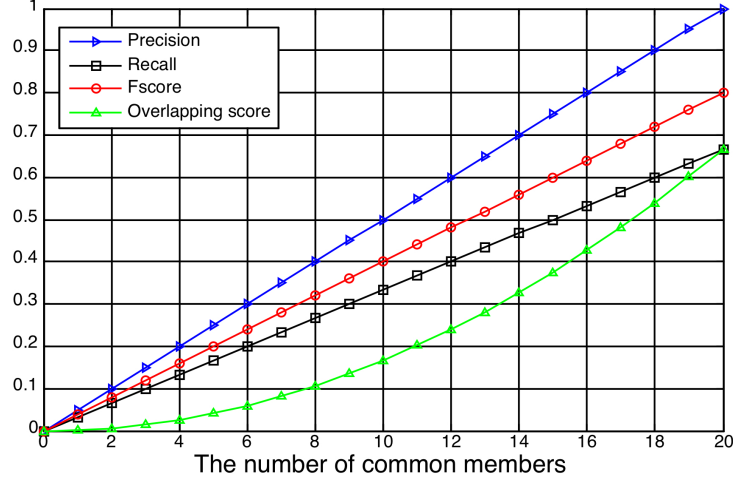
## 2.7 Evaluation methods

Evaluating a set of prediction algorithms is always a challenging problem, especially when there is no complete gold standard dataset available as a reference. The incompleteness of the datasets would introduce biases for any evaluation methods, which may mislead the comparison results. In addition, there is no widely accepted evaluation criterion. One algorithm may outperform the others in terms of one criterion, yet it may perform worse in terms of other criteria. Nevertheless, we believe that each algorithm has its own advantages and is helpful to spark novel ideas for the identification of protein complexes. Moreover, this article is to give a survey on how computational methods revolve with the available data. Hence, we summarize widely used criteria across this research area, without quantitatively comparing those algorithms. Typically, known gold standard protein complexes, GO annotations or localization annotations are often involved in those criteria.

Giving a set of gold standard protein complexes as references, two levels of comparison are conducted to perform such evaluation: (i) the comparison between a predicted protein complex and a reference protein complex and (ii) the comparison between a group of predicted protein complexes and a group of reference protein complexes.

Four measures are commonly used to compare the difference between a predicted complex  $X$  and a reference complex  $P$ , which are *precision*, *recall*, *f-score* and overlapping score. Suppose a predicted protein complex  $X$  is compared with a reference protein complexes  $P$ , the *precision* and *recall* are defined as follows:

$$precision = \frac{|X \cap P|}{|X|} \quad \text{and} \quad recall = \frac{|X \cap P|}{|P|}. \quad (2.1)$$



**Figure 2.6:** Comparison of the *precision*, *recall*, *f-score* and *overlapping score* with different common members. The predicted protein complex consists of 20 proteins, while the reference protein complex consists of 30 proteins.

The *f-score* is the harmonic mean of *precision* and *recall*, which is

$$f\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot |X \cap P|}{|X| + |P|}, \quad (2.2)$$

while the *overlapping score* is defined as:

$$\text{overlapping score} = \frac{|X \cap P|^2}{|X| \cdot |P|}, \quad (2.3)$$

which is the multiplying between *precision* and *recall*. The comparison of those metrics is shown in Figure 2.6.

After calculating the *f-score* and/or *overlapping score* for each predicted and reference protein complex pairs, the set of true-positive predictions can be obtained by selecting predictions with *f-score* or *overlapping score* larger than a threshold. Most researchers also use the term *precision* and *recall* to represent the true-positive rate and the positive predicted value, respectively, when comparing a group of predicted complexes with a group of references. To make distinction from previous ones, we use *Pr* and *Rc* to represent *precision* and *recall*, respectively, in this situation. They are defined as:

$$Pr = \frac{TP}{TP + FN} \quad \text{and} \quad Rc = \frac{TP}{TP + FP}. \quad (2.4)$$

The F-measure combines the *Pr* and *Rc*, which is defined as:

$$F\text{-measure} = \frac{2 \cdot Pr \cdot Rc}{Pr + Rc}. \quad (2.5)$$

Generally, the value of the average *f-score* or the *F-measure* can be used to measure the performance of an algorithm. However, because the number of predictions varies widely for different algorithms and the set of

reference protein complexes are commonly incomplete, it is unfair to use the average *f-score* or the *F-measure* to compare different algorithms. Moreover, a reference protein complex often partially matches with more than one predicted complex and *vice versa*. To handle this problem, Nepusz et al. [75] introduce a maximum matching ratio (MMR) to evaluate different predictions. A weighted bipartite graph is built where two sets of vertices represent the predicted and reference protein complexes, respectively, and weights of edges represent the *overlapping score* between predicted and reference proteins complexes. The MMR is calculated by the total weight of the maximum matching edges, divided by the number of reference complexes.

The other commonly used measure is called accuracy (*Acc*), which is the geometric mean of the clustering-wise sensitivity (*Sn*) and the clustering-wise positive predictive value (*PPV*). Given  $m$  predicted and  $n$  reference protein complexes, a confusion matrix  $T = [t_{ij}]$  is constructed, where  $t_{ij}$  denote the number of common proteins in the  $i^{th}$  reference and the  $j^{th}$  predicted complex. The *Sn* and *PPV* are defined as:

$$Sn = \frac{\sum_{i=1}^n \max_j t_{ij}}{\sum_{i=1}^n n_i} \quad \text{and} \quad PPV = \frac{\sum_{j=1}^m \max_i t_{ij}}{\sum_{j=1}^m \sum_{i=1}^n t_{ij}} \quad (2.6)$$

where  $n_i$  is the number of proteins in the  $i^{th}$  reference protein complex. The *Acc* is then defined as:

$$Acc = \sqrt{Sn \cdot PPV} \quad (2.7)$$

Because gold standard protein complex datasets are commonly incomplete, a predicted protein complex that does not match with any known complexes may belong to valid but still uncharacterized complexes. Hence, it is also important to analyze those unmatched predictions by using GO annotations and/or localization annotations.

For GO annotations [86], they are usually accepted as ground-truth and used for comparison and validation purposes. A prediction can be statically evaluated using the *p-value* defined by the following hypergeometric distribution:

$$p\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i} \cdot \binom{N-F}{n-i}}{\binom{N}{n}}, \quad (2.8)$$

where  $F$  is the number of proteins in a GO term,  $n$  is the number of proteins in the predicted complex,  $k$  is the number of proteins they have in common and  $N$  is the total number of proteins in a PPI network. The smaller the *p-value* is, the more statistically significant the protein complex is enriched by GO annotations.

The other kind of data used is localization annotations. This is motivated by the fact that a protein complex can be formed only when its constituents are to be found in the same cellular compartment [87]. The *co-localization score* of a single complex is defined as the maximum fraction of proteins in the complex that are found at the same localization. The *co-localization score* of a set of complexes is the mean *co-localization score* of all complexes in the set, weighted by the size of the complexes.

## 2.8 Conclusions

In this review, we focus on computational algorithms for identifying protein complexes and functional modules from PPI networks in terms of what kinds of data are used and what kinds of detection ideas are based on. Four categories of algorithms for interpreting PPI networks are surveyed from static ones to dynamic ones. The first category focuses on algorithms that based on only topological structures of a single static PPI network. They treat vertices and edges equally in a PPI network and sub-graphs such as cliques, dense sub-graphs, core-attachment structures and star-like structures are mined as predictions of protein complexes. The next category consists of algorithms that based on characteristics of weighted PPI networks. They are also based on a single static PPI network, but they use various topological indices of a network to assign weights to vertices and/or edges. Many hierarchical clustering algorithms also contribute to this category. The third category addresses algorithms that involving multiple data integrations. They use other experimental dependent and/or independent datasets to assign weights to PPI networks, such as the experimental conditions, gene expression profiles and GO annotations. More biological meaningful results can be achieved by using such data integration. The fourth category is made up of algorithms that involving dynamic PPI networks. They are reviewed from the general framework of dynamic systems to the time-course PPI networks. It is hard to say whether an algorithm is better than the other, as there is no generally accepted criterion to perform such comparison. In this review, we have summarized some evaluation measures to compare algorithms, which are widely used across this research area. It is believed that the PPI networks are modelled increasingly precise when integrating more types of data, and the study of protein complexes would benefit by shifting from static to dynamic PPI networks.

## Acknowledgements

Natural Sciences and Engineering Research Council of Canada (NSERC); National Science Foundation of China [61272274, 60970063]; Program for New Century Excellent Talents in Universities [NCET-10-0644]; Ph. D. Programs Foundation of Ministry of Education of China [20090141110026].



## BIBLIOGRAPHY

- [1] Butland G, Peregrín-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 2005, **433**(7025): 531-537.
- [2] De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 2010, **6**(6): e1000807.
- [3] Pellegrini M, Haynor D, Johnson JM. Protein interaction networks. *Expert Rev Proteomics* 2004, **1**(2): 239-249.
- [4] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 1999, **402**(6761 Suppl): C47-C52.
- [5] Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007, **3**(4): e59.
- [6] Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ, Koller D. A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. *Mole Cell Proteomics* 2009, **8**(6): 1361-1381.
- [7] Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrola S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of *Drosophila melanogaster*. *Science* 2003, **302**(5651): 1727-1736.
- [8] Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick

- ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, **303**(5657): 540-543.
- [9] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, **403**(6770): 623-627.
- [10] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, **437**(7062): 1173-1178.
- [11] Terentiev AA, Moldogazieva NT, Shaitan KV. Dynamic proteomics in modeling of the living cell. Protein-protein interactions. *Biochemistry (Mosc)* 2009, **74**(13): 1586-1607.
- [12] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, **100**(21): 12123-12128.
- [13] Lu H, Shi B, Wu G, Zhang Y, Zhu X, Zhang Z, Liu C, Zhao Y, Wu T, Wang J, Chen R. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochem Biophys Res Commun* 2006, **345**(1): 302-309.
- [14] Li M, Wu X, Wang J, Pan Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC Bioinformatics* 2012, **13**: 109.
- [15] Li XL, Tan SH, Foo CS, Ng SK. Interaction graph mining for protein complexes using local clique merging. *Genome Inform* 2005, **16**(2): 260-269.
- [16] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005, **435**(7043): 814-818.
- [17] Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 2006, **22**(8): 1021-1023.
- [18] King AD, Pržulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, **20**(17): 3013-3020.
- [19] Kenley EC, Cho YR. Detecting protein complexes and functional modules from protein interaction networks: a graph entropy approach. *Proteomics* 2011, **11**(19): 3835-3844.

- [20] Lian H, Song C, Cho YR. Decomposing protein interactome networks by graph entropy. *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, 2010: 585-589.
- [21] Chen B, Yan Y, Shi J, Zhang S, Wu FX. An improved graph entropy-based method for identifying protein complexes. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 2011: 123-126.
- [22] van Dongen S. Graph clustering by flow simulation. *PhD thesis, University of Utrecht*, 2000.
- [23] van Dongen S. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* 2008, **30**(1): 121-141.
- [24] Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. *Proteins* 2004, **54**(1): 49-57.
- [25] Yu L, Gao L, Kong C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* 2011, **11**(19): 3826-3834.
- [26] Leung HC, Xiang Q, Yiu SM, Chin FY. Predicting protein complexes from PPI data: a core-attachment approach. *J Comput Biol* 2009, **16**(2): 133-144.
- [27] Wu M, Li X, Kwok CK, Ng SK. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics* 2009, **10**: 169.
- [28] Ma X, Gao L Detecting protein complexes in PPI networks: The roles of interactions. *Systems Biology (ISB), 2011 IEEE International Conference on*, 2011: 52-59.
- [29] Wu M, Li XL, Kwok CK, Ng SK, Wong L. Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data. *J Comput Biol* 2012, **19**(9): 1027-1042.
- [30] Chen B, Shi J, Wu FX. Not all protein complexes exhibit dense structures in *S. cerevisiae* PPI network. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, 2012: 470-473.
- [31] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, **4**: 2.
- [32] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kruokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 2006, **7**: 207.
- [33] Li M, Chen JE, Wang JX, Hu B, Chen G. Modifying the DPCLUS algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics* 2008, **9**: 398.
- [34] Kim J, Tan K. Discover Protein Complexes in Protein-Protein Interaction Networks Using Parametric Local Modularity. *BMC Bioinformatics* 2010, **11**: 521.

- [35] Girvan M, Newman ME. Community structure in social and biological networks. *Proc Natl Acad Sci USA* 2002, **99**(12): 7821-7826.
- [36] Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 2005, **6**: 39.
- [37] Newman ME. Fast algorithm for detecting community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **69**(6 Pt 2): 066133.
- [38] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. *Proc Natl Acad Sci USA* 2004, **101**(9): 2658-2663.
- [39] Rives AW, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci USA* 2003, **100**(3): 1128-1133.
- [40] Wang J, Li M, Chen J, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**(3): 607-620.
- [41] Cho YR, Zhang A. Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. *BMC Bioinformatics* 2010, **11**(Suppl 3): S3.
- [42] Chatr-Aryamontri A, Ceol A, Licata L, Cesareni G. Protein interactions: integration leads to belief. *Trends Biochem Sci* 2008, **33**(6): 241-242.
- [43] Tan PP, Dargahi D, Pio F. Predicting protein complexes by data integration of different types of interactions. *Int J Comput Biol Drug Des* 2010, **3**(1): 19-30.
- [44] Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* 2006, **22**(18): 2283-2290.
- [45] Snel B, Bork P, Huynen MA. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci USA* 2002, **99**(9): 5890-5895.
- [46] Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* 2004, **101**(9): 2981-2986.
- [47] Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001, **29**(4): 482-486.
- [48] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 2008, **24**(13): i223-i231.
- [49] Lubovac Z, Gamalielsson J, Olsson B. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins* 2006, **64**(4): 948-959.

- [50] Xu B, Lin H, Yang Z. Ontology integration to identify protein complex in protein interaction networks. *Proteome Sci* 2011, **9**(Suppl 1): S7.
- [51] Tang X, Wang J, Liu B, Li M, Chen G, Pan Y. A comparison of the functional modules identified from time course and static PPI network data. *BMC Bioinformatics* 2011, **12**: 339.
- [52] Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP. Community structure in time-dependent, multiscale, and multiplex networks. *Science* 2010, **328**(5980): 876-878.
- [53] Jin R, McCallen S, Liu CC, Xiang Y, Almaas E, Zhou XJ. Identifying dynamic network modules with temporal and spatial constraints. *Pac Symp Biocomput* 2009: 203-214.
- [54] Srihari S, Leong HW. A survey of computational methods for protein complex prediction from protein interaction networks. *J Bioinform Comput Biol* 2013, **11**(2): 203-214.
- [55] Li X, Wu M, Kwok CK, Ng SK. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* 2010, **11**(Suppl 1): S3.
- [56] Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology - what's the connection? *Nat Biotechnol* 2008, **26**(1): 69-72.
- [57] Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006, **7**(11): 120.
- [58] de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C, Stumpf MP. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* 2006, **4**: 39.
- [59] Nesvizhskii AI. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* 2012, **12**(10): 1639-1655.
- [60] Thorne T, Stumpf MP. Graph spectral analysis of protein interaction network evolution. *J R Soc Interface* 2012, **9**(75): 2653-2666.
- [61] Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM. Protein interactions: is seeing believing? *Trends Biochem Sci* 2007, **32**(12): 530-531.
- [62] Adelmant G, Marto JA. Protein complexes: the forest and the trees. *Expert Rev Proteomics* 2009, **6**(1): 5-10.
- [63] Chu W, Ghahramani Z, Krause R, Wild DL. Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. *Pac Symp Biocomput* 2006: 231-242.
- [64] Tsai CJ, Ma B, Nussinov R. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem Sci* 2009, **34**(12): 594-600.

- [65] Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM. Response to Chatr-aryamontri et al.: Protein interactions: to believe or not to believe? *Trends Biochem Sci* 2008, **33**(6): 242-243.
- [66] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the molecular interaction database. *Nucleic Acids Res* 2007, **35**(Database issue): D572-D574.
- [67] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004, **32**(Database issue): D449-D451.
- [68] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND - The biomolecular interaction network database. *Nucleic Acids Res* 2001, **29**(1): 242-245.
- [69] Mewes H. W., Frishman D., Güldener U., Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M, Rudd S, Weil B. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002, **30**(1): 31-34.
- [70] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, **5**(2): 101-113.
- [71] Pereira-Leal JB, Levy ED, Teichmann SA. The origins and evolution of functional modules: lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1467): 507-517.
- [72] Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868): 141-147.
- [73] Pang CN, Krycer JR, Lek A, Wilkins MR. Are protein complexes made of cores, modules and attachments? *Proteomics* 2008, **8**(3): 425-434.
- [74] Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci USA* 2003, **100**(22): 12579-12583.
- [75] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* 2012, **9**: 471-472.
- [76] Hwang W, Cho YR, Zhang A, Ramanathan M. A novel functional modules detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol* 2006, **1**: 24.
- [77] Li H, Liang S. Local network topology in human protein interaction data predicts functional association. *PLoS ONE* 2009, **4**(7): e6410.

- [78] Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. *Bioinformatics* 2009, **25**(15): 1891-1897.
- [79] Wang J, Chen G, Liu B, Li M, Pan Y. Identifying protein complexes from interactome based on essential proteins and local fitness method. *IEEE Trans Nanobioscience* 2012, **11**(4): 324-335.
- [80] Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **69**(2): 026113.
- [81] Muff S, Rao F, Caffisch A. Local modularity measure for network clusterizations. *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **72**(5 Pt 2): 056107.
- [82] Li Z, Zhang S, Wang RS, Zhang XS, Chen L. Quantitative function for community detection. *Phys Rev E Stat Nonlin Soft Matter Phys* 2008, **77**(3 Pt 2): 036109.
- [83] Zhang S, Ning XM, Ding C, Zhang XS. Determining modular organization of protein interaction networks by maximizing modularity density. *BMC Syst Biol* 2010, **4**(Suppl 2): S10.
- [84] Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 2003, **19**(10): 551-560.
- [85] Levy ED, Pereira-Leal JB. Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol* 2008, **18**(3): 349-357.
- [86] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**(1): 25-29.
- [87] Friedel CC, Krumsiek J, Zimmer R. Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J Comput Biol* 2009, **16**(8): 971-987.
- [88] Huynen M, Snel B, Lathe W 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000, **10**(8): 1204-1210.
- [89] Lin D. An information-theoretic definition of similarity. *Machine Learning San Francisco (ICML), 1998 Proceedings of the Fifteenth International Conference on*, 1998: 296-304.
- [90] Shi L, Zhang A. Semi-supervised Learning Protein Complexes from Protein Interaction Networks. *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, 2010: 247-252.
- [91] Georgii E, Dietmann S, Uno T, Pagel P, Tsuda K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 2009, **25**(7): 933-940.

- [92] Luo F, Liu J, Li J. Discovering conditional co-regulated protein complexes by integrating diverse data sources. *BMC Syst Biol* 2010, **4**(Suppl 2): S4.
- [93] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003, **302**(5644): 449-453.
- [94] Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001, **292**(5518): 929-934.
- [95] Przytycka TM, Singh M, Slonim DK. Toward the dynamic interactome: it's about time. *Brief Bioinform* 2010, **11**(1): 15-29.
- [96] Hegde SR, Manimaran P, Mande SC. Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol* 2008, **4**(11): e1000237.
- [97] Chautard E, Thierry-Mieg N, Ricard-Blum S. Interaction networks: from protein functions to drug discovery. A review. *Pathol Biol (Paris)* 2009, **57**(4): 324-333.
- [98] de Lichtenberg U, Jensen LJ, Brunak S, Bork P. Dynamic complex formation during the yeast cell cycle. *Science* 2005, **307**(5710): 724-727.
- [99] Lin CC, Hsiang JT, Wu CY, Oyang YJ, Juan HF, Huang HC. Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst Biol* 2010, **4**: 138.
- [100] Lu X, Jain VV, Finn PW, Perkins DL. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol Syst Biol* 2007, **3**: 98.
- [101] Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 2004, **430**(6995): 88-93.
- [102] Komurov K, White M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol* 2007, **3**: 110.
- [103] Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 2009, **27**(2): 199-204.



## CHAPTER 3

# IDENTIFYING PROTEIN COMPLEXES IN PROTEIN-PROTEIN INTERACTION NETWORKS BY USING CLIQUE SEEDS AND GRAPH ENTROPY

*Published as:* Chen B, Shi J, Zhang S and Wu FX. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics* 2013, **13**(2): 269-277.

In the previous chapter, we have given a comprehensive review of algorithms to identify protein complexes and/or functional modules from various kinds of PPI networks. Many of those algorithms were developed based on the assumption that protein complexes exhibit densely connected sub-graphs in PPI networks.

Based on the assumption, this chapter proposes a dense sub-graph detection algorithm to identify protein complexes by using clique seeds and graph entropy. Numerical experiments show that the proposed entropy-based algorithm generates predictions with a high average *f-score*, which outperforms the original algorithm.

### **Abstract**

The identification of protein complexes plays a key role in understanding major cellular processes and biological functions. Various computational algorithms have been proposed to identify protein complexes from protein-protein interaction (PPI) networks. In this paper, we first introduce a new seed-selection strategy for seed-growth style algorithms. Cliques rather than individual vertices are employed as initial seeds. After that, a result-modification approach is proposed based on this seed-selection strategy. Predictions generated by higher-order clique seeds are employed to modify results that are generated by lower-order ones. The performance of this seed-selection strategy and the result-modification approach are tested by using the entropy-based algorithm, which is currently the best seed-growth style algorithm to detect protein complexes from PPI networks. In addition, we investigate four pairs of strategies for this algorithm in order to improve

its accuracy. The numerical experiments are conducted on a *Saccharomyces cerevisiae* PPI network. The group of best predictions consists of 1711 clusters, with the average *f-score* at 0.68 after removing all similar and redundant clusters. We conclude that higher-order clique seeds can generate predictions with higher accuracy and that our improved entropy-based algorithm outputs more reasonable predictions than the original one.

### 3.1 Introduction

One of the fundamental goals of systems biology is to understand a cell as an interacting system [1]. Various networks have been constructed for this purpose such as gene regulatory network [2], metabolic networks [3], protein-DNA interaction networks [4], and protein-protein interaction (PPI) networks [5, 6], and so on. Biomolecules are vertices of these networks, and molecular interactions between them are edges. Since the topology of molecular networks can reveal essential principles of most cellular processes and biological functions, it is vitally important to explore topological organizations and biological modules in those networks [7].

In this study, we focus on the identification of protein complexes from PPI networks. Protein complexes are essential molecular entities, which consist of groups of proteins that interact with each other at the same time and place [7, 8]. They are responsible for most biological processes in living cells [9]. However, it is still limited to detect them by experimental methods, especially those involved in high-throughput techniques [10]. Meanwhile, due to recently accumulated PPI data of diverse organisms, various computational algorithms have been developed for detecting protein complexes from PPI data. Those predicted results provide crucial complements to the limitations of experimental ones.

Most computational approaches are designed to identify densely connected sub-graphs (or clusters) from PPI networks. This is due to the fact that proteins in a complex tend to cooperatively interact with each other [9]. They usually display strong and frequent interactions within a complex while display weak and rare connections with proteins out of the complex [11–13]. Hence, identifying highly connected sub-graphs in a PPI network is the key issue for most computational methods.

Various algorithms have been proposed to identify highly connected sub-graphs in PPI networks. Spirin and Mirny [7] introduce an iterative method to enumerate all cliques as predictions in a PPI network. However, most PPI networks are incomplete, inaccurate, and sparsely connected. The constraint of fully connected sub-graphs is too strict in terms of detecting protein complexes from such networks. Alternatively, Bader and Hogue[14] design a seed-growth style algorithm, called MCODE, to generate clusters based on the local density. King et al. [15] propose the RNSC algorithm to partition a network into clusters by using a cost

function. Georgii et al. [16] develop the DME method to detect all densely connected protein sets according to a predetermined threshold. van Dongen [17] propose the Markov clustering (MCL) algorithm to generate clusters by random flows. Li et al. [18] introduce a seed-growth style algorithm, called local cliques merging algorithm (LCMA), to detect local cliques around seed proteins, and merge similar cliques as predictions. Cho’s group [11, 19] recently proposes the entropy-based algorithm by introducing novel binary entropy for vertices. It is also a seed-growth style algorithm, and can generate local optimum modules by minimizing the graph entropy.

Two issues about current computational algorithms should be emphasized here. Firstly, none of the algorithms has been widely applied to identify protein complexes from PPI networks due to their low accuracy. Among those algorithms, the entropy-based algorithm, the RNSC algorithm and the MCL algorithm are regarded as highly efficient [11, 20]. In [11], the entropy-based algorithm can generate more accurate predictions than the MCL, which is very promising in terms of detecting protein complexes. Secondly, many computational algorithms employ a seed-growth style heuristic, which typically start from a set of seed vertices to search the local optimum clusters. Individual proteins are usually employed as seeds. However, in many cases a single protein is not enough to grow into a meaningful complex while in many other circumstances more than one protein is known in a complex of interest. Hence, a new way to select seeds need to be investigated.

In this paper, we introduce a new seed-selection strategy for those seed-growth style algorithms to identify protein complexes from PPI networks. To be more precise, all cliques and maximal cliques are employed as initial seeds, rather than individual vertices. They are more reasonable than the individual ones in terms of detecting densely connected protein complexes in a PPI network. Based on this seed-selection strategy, we propose a result-modification approach to improve the accuracy of predictions. Results generated by higher-order cliques are employed to modify results that are generated by lower-order cliques. Here, the order of a clique is the number of its vertices.

We test this seed selection strategy by using the entropy-based algorithm proposed in [11], which is currently the best seed-growth style algorithm in terms of prediction accuracy. Although the algorithm is already quite efficient, there are still rooms for it to be improved. In this paper, we investigate the graph-entropy-based algorithm in details. The performance of the algorithm are compared with four pairs of strategies, which are (i) either adding neighbors (AN) or not adding neighbors (nAN) to initial seeds; (ii) either removing seed vertices (RSV) or keeping seed vertices (KSV) during the growth of a cluster; (iii) either using weighed vertex entropy (WVE) or unweighted vertex entropy (uWVE) to measure the vertex entropy; and (iv) either using the cluster entropy (CE) or the graph entropy (GE) as the cost function. We compare our results with Cho’s previous valuable work in [11]. The numerical results show that our improved entropy-based algorithm performs much better than the initial one and that the proposed seed-selection strategy can easily be used in various seed-growth style algorithms.

## 3.2 Materials and methods

A PPI network can be represented as an undirected simple graph  $G = (V(G), E(G))$ , where  $V(G)$  is the set of vertices (proteins), and  $E(G)$  is the set of edges (physical interactions between proteins).

### 3.2.1 The entropy-based algorithm

The concept of graph entropy introduced by Cho's group [11, 19] is based on a partition of a graph. Suppose that  $G' = (V', E')$  is a sub-graph of  $G$ , then  $V(G)$  can be divided into two sets:  $V(G')$  and  $V(G \setminus G')$ . For any given vertex  $v$  of  $G$ , the set of its neighbors consists of all vertices adjacent to  $v$ , and is denoted by  $N_G(v) = \{u | u \in V(G), (u, v) \in E(G)\}$ . Similarly, for the neighbors of  $v$  in  $G'$  and  $G \setminus G'$  we can define  $N_{G'}(v) = \{u | u \in V(G'), (u, v) \in E(G)\}$  and  $N_{G \setminus G'}(v) = \{u | u \in V(G \setminus G'), (u, v) \in E(G)\}$ , which consist of the set of vertices adjacent to  $v$ , but belong to  $V(G')$  and  $V(G \setminus G')$ , respectively. Obviously,  $N_G(v) = N_{G'}(v) \cup N_{G \setminus G'}(v)$ , and  $|N_G(v)| = |N_{G'}(v)| + |N_{G \setminus G'}(v)|$ , where  $|*|$  is the cardinality of the set.

Without loss of generality, let  $v \in V(G')$ , then for any vertex  $u \in N_G(v)$ , the probability of  $u \in N_{G'}(v)$  is

$$p_i(v) = \frac{|N_{G'}(v)|}{|N_G(v)|}, \quad (3.1)$$

and the probability of  $u \in N_{G \setminus G'}(v)$  is

$$p_o(v) = 1 - p_i(v) = \frac{|N_{G \setminus G'}(v)|}{|N_G(v)|}. \quad (3.2)$$

With above concepts, the vertex entropy of  $v$  defined in [11] is

$$e(v) = -p_i(v) \cdot \log(p_i(v)) - p_o(v) \cdot \log(p_o(v)). \quad (3.3)$$

The entropy of a graph can be calculated by summing the entropy of all vertices in the graph. Using this value as the cost function, the entropy-based algorithm proposed in [11, 19] can be described as follows:

1. Initialize a set  $S$  candidate seeds by including all vertices in the graph  $G$ ;
2. Select a vertex  $v_i \in S$  and form an initial cluster  $C_i = \{v_i\} \cup N_G(v_i)$ ;
3. If removing a vertex  $u$  on the inner boundary of  $C_i$  can decrease the graph entropy, then let  $C_i = C_i \setminus \{u\}$  until no vertex can be removed;
4. If adding a vertex  $w$  on the outer boundary of  $C_i$  can decrease the graph entropy, then let  $C_i = C_i \cup \{w\}$  until no vertex can be added;

5. Output the cluster  $C_i$ , and let  $S = S \setminus V(C_i)$ ;
6. Repeat steps (2) through (5), until all candidate seeds are tested.

In this algorithm, the inner boundary of  $C_i$  is the vertex set  $IB(C_i) = \{v | v \in V(C_i), N_G(v) \setminus V(C_i) \neq \emptyset\}$ , while the outer boundary of  $C_i$  is  $OB(C_i) = \{v | v \in V(G \setminus C_i), N_G(v) \cap V(C_i) \neq \emptyset\}$ . In paper [21], we have concluded that (i) *2-clique* and *3-clique* seeds can generate predictions with higher *f-score* than individual vertices; (ii) clusters generated by higher-order cliques can be used to modify results that are generated by lower-order cliques; and (iii) enumerating all possible seeds and growing them into clusters can increase the accuracy of predictions. In this paper, we propose a complete seed-selection strategy based on those ideas. It can easily be used in almost all seed-growth style algorithms for detecting protein complexes from PPI networks.

### 3.2.2 The seed-selection strategy

Since protein complexes are supposed to be densely connected sub-graphs in a PPI network, various seed-growth style algorithms employ individual vertices as seeds to generate predicted clusters. However, many protein complexes consist of a large number of proteins. Individual protein seeds are usually not enough to grow into meaningful predictions. In addition, under some circumstances more than one protein is known in a complex of interest and they can be treated as the initial seed together. Therefore, we suggest using cliques or maximal cliques as seeds, rather than only individual vertices.

It is not hard to generate all cliques in a PPI network. Spirin and Mirny [7] propose a complete enumeration method for this purpose. By adding a vertex to a *k-clique*, one can get a clique with  $k + 1$  vertices. All cliques can be enumerated quickly, as a PPI network is usually very sparse and cliques of order 1 and order 2 are naturally available as the vertex set and the edge set, respectively. Maximal cliques can be enumerated at the same time. A *k-clique* that cannot be enlarged by any vertex can be regarded as *maximal k-clique*.

The benefits of this seed-selection strategy are as follows. Firstly, cliques are fully connected sub-graphs in a PPI network. They are more reasonable to be used as seeds to generate candidate densely clusters. Secondly, various overlapped clusters are generated. Although some of them may not be reasonable, they can be modified by others. Thirdly, it is very easy for this seed-selection strategy to be used in other seed-growth style algorithms, especially to detect protein complexes in PPI networks.

Predictions generated by lower-order cliques can be modified by results that are generated by higher-order cliques. On the one hand, these higher-order output clusters can be added into the group of final predictions, by simply excluding same predictions and unreasonable clusters. On the other hand, different overlapped

seeds may generate various similar clusters. They can be merged together to form the more reasonable ones. By using this result modification approach, predictions are expected to have a higher accuracy .

### 3.2.3 Investigation of the entropy-based algorithm

Although the previous entropy-based algorithm in [11] can output predictions with a high accuracy, there are still rooms for being further improved. We use the proposed seed-selection strategy in the first step of the entropy-based algorithm. For the other part of the algorithm, we present four pairs of strategies as follows:

1. Either AN or nAN

When growing a seed in the second step of the algorithm, all neighbors of the seed are added to form the initial cluster. It is necessary and indispensable for a single vertex seed; otherwise the seed cannot be grown into a larger cluster. However, it should also be realized that those neighbors can bring a large number of noise vertices for the cluster, especially of those hub vertices. Alternatively, when it comes to clique or maximal clique seeds, it is not necessary for them to add all neighbors. We test these two different ways in this step and aim to obtain a better way to improve the algorithm.

2. Either RSV or KSV

When removing vertices on the inner boundary of a cluster in the third step, a vertex of the initial seed may be removed according to the decrease of the graph entropy. It is possible, because at least one vertex of the initial seed should appear on the inner boundary of the cluster after one vertex is removed. Whether such vertices are allowed to be removed or not will significantly affect the quality of output clusters. The RSV strategy allows removing any vertex of the inner boundary if it decreases of the graph entropy while the KSV strategy does not allow removing vertices of the initial seeds.

3. Either using WVE or uWVE

The vertex entropy defined in [11] is the unweighted vertex entropy. It does not take the degree of a vertex  $v$  into consideration and thus the value of vertex entropy is only depended on the distribution of its neighbors. A vertex can achieve a higher value of  $e(v)$  if most of its connections contribute to either inner degree or outer degree, no matter how many connections are there. It is obvious that the connectivity of a hub vertex may gain a small value of the vertex entropy if half of its neighbors contribute to a cluster. Here, we introduce a new measurement, called the weighed vertex entropy, which is  $e_w(v) = e(v) \cdot d(v)$  by multiplying the degree of the vertex. We investigate whether they have significant differences when predicting protein complexes.

4. Either using the CE or the GE

In terms of the cost function, two suggestions are proposed based on specific emphasis:

(a) The CE

$$e(G') = \sum_{v \in V(G')} e(v) \quad (3.4)$$

(b) The GE

$$e(G) = \sum_{v \in V(G)} e(v) \quad (3.5)$$

The value of  $e(v)$  equals to zero, if the vertex  $v$  belongs to neither the inner boundary nor the out boundary of the cluster. Hence, if we only focus on the local connectivity of a sub-graph, the cluster entropy is encouraged to be used, whose value equals to  $\sum_{v \in IB(C_i)} e(v)$ . Alternatively, if we care about the overall connectivity of the whole graph, the graph entropy is recommended to be used, which equals to  $\sum_{v \in IB(C_i)} e(v) + \sum_{v \in OB(C_i)} e(v)$ . It is the summation entropy of vertices on both inner and outer boundary of the clusters.

### 3.2.4 Data source

As yeast PPI data are widely studied, many manually curated protein complexes data are available as ‘gold standard’ [12, 22]. We test the seed-selection strategy and the improved entropy-based algorithm on a *Saccharomyces cerevisiae* PPI dataset. The file named as *Scere201010.txt* is downloaded from the Database of Interacting Proteins (DIP) [23]. After removing redundant data and interactions between *S. cerevisiae* and other species, we finally obtain a PPI network with 5154 proteins and 24848 interactions.

To evaluate the accuracy of predictions, we collect the information of known protein complexes from the database of MIPS [24], CYC2008 [22] and YHTP2008 [22]. Moreover, Spirin and Mirny [7] also catalogue protein complex data from the database of MIPS, BIND and Cellzome, which can be downloaded from <http://web.mit.edu/leonid/modules/>. After removing redundant complexes, we obtain 2220 protein complexes as ‘gold standard’. There are 2956 proteins overlapping with the PPI network.

### 3.2.5 Accuracy evaluation approaches

We use the *f-score* to evaluate the accuracy of predictions for protein complexes. Given a predicted cluster with  $m$  proteins and a known complex with  $n$  proteins, the number of proteins they have in common is denoted as  $k$ . Then the true positive predictive value, which is called *precision*, is defined as  $k/m$ , and the true positive rate, which is called *recall*, is defined as  $k/n$ . The *f-score* is defined as the harmonic mean of

*precision* and *recall*, as follows

$$f\text{-score} = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} = \frac{2 \cdot k}{m + n}. \quad (3.6)$$

The *f-score* measures the degree of match between an output cluster and a known protein complex. For each predicted cluster, we search for the best match from the ‘gold standard’ protein complex. The one reaching the maximum *f-score* is taken as a target prediction. We use the average *f-score* to evaluate the overall accuracy of a group of clusters, the performance of the seed-selection strategy and the improved algorithm.

### 3.3 Results and discussions

Numerical experiments are conducted on the yeast PPI network of the DIP. In this section, we summarize results of these experiments and give discussions at the same time.

#### 3.3.1 Properties of cliques and maximal cliques

Cliques and maximal cliques are first enumerated from the network by using the algorithm proposed in [7]. The largest clique found contains 12 vertices. Table 3.1 summarizes the number of cliques and maximal cliques according to the number of vertices. All cliques and maximal cliques are employed as initial seeds on the improved entropy-based algorithm. For the experimental conditions, each pair of contrary strategies illustrated in section 3.2.3 is successively tested. The specific experimental condition is given by the denotation of particular strategies’ abbreviations, especially in figures and tables when we illustrate the performance of each experiment.

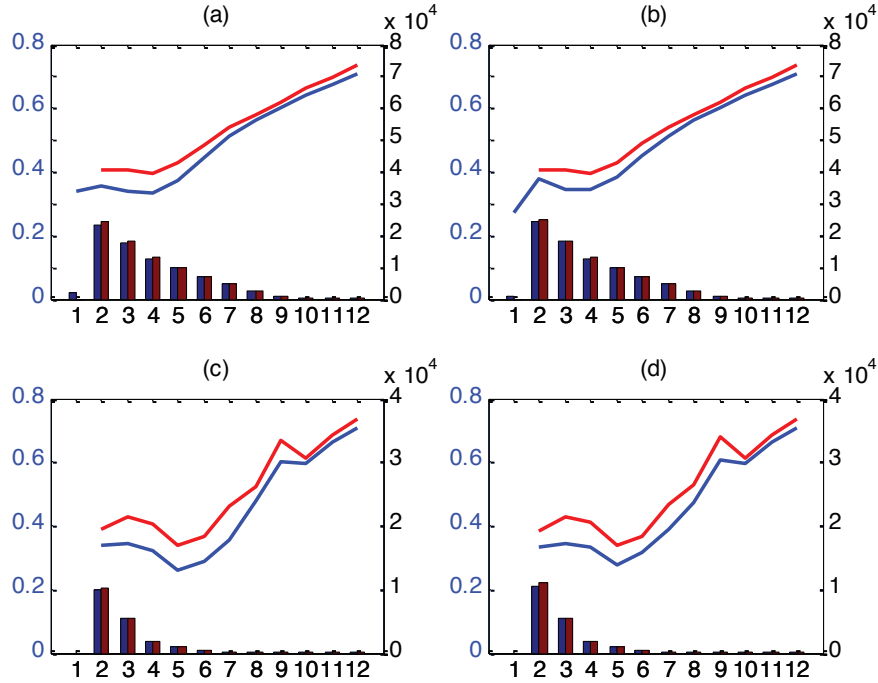
Our proposed seed-selection strategy performs better than the traditional ones. It is worth mentioning that individual vertices are the *1-cliques* of a network, and predictions of those *1-cliques* can be taken as comparisons of traditional seed-growth style algorithms. The overall performance of each group of these seeds is illustrated in Figure 3.1, in terms of the average *f-score* and the total number of predicted clusters. It can be seen from each subfigure that, the average *f-score* has a trend of ascent with the increase order of seed cliques. In contrast, the number of predicted clusters shows an opposite trend, which decreases dramatically when enlarging the order of seed cliques.

It can be seen that clique seeds perform better than maximal clique seeds by comparing results of subfigures between (a) and (c), or between (b) and (d) in Fig. 3.1, or by comparing the overall accuracy in Fig. 3.2. This can be understood by recalling the basic assumption that protein complexes are densely connected

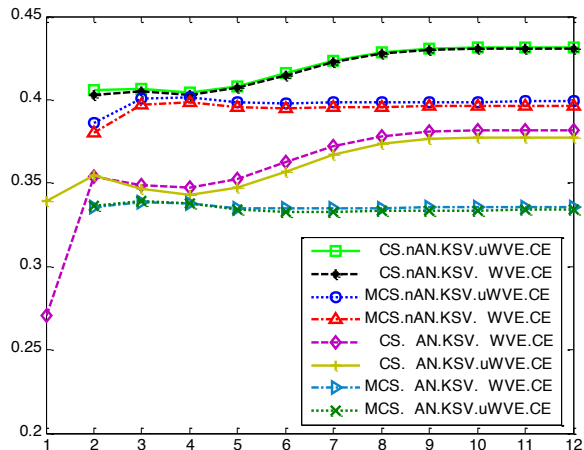


**Table 3.1:** The number of cliques and maximal cliques in the PPI network

Order	Cliques	Maximal cliques
1	5152	-
2	24847	10945
3	18291	5412
4	12952	1841
5	9782	850
6	7200	392
7	4872	106
8	2622	23
9	1000	26
10	248	12
11	35	12
12	2	2



**Figure 3.1:** The average  $f$ -score and the number of output clusters for each size seeds. The entropy-based algorithm has been implemented under the condition of (a) CS, KSV, uWVE, and CE; (b) CS, KSV, WVE, and CE; (c) MCS, KSV, uWVE, and CE; (d) MCS, KSV, WVE, and CE. The blue lines and left bars represent the cases under condition of AN, while the red lines and right bars represent the results under conditions of nAN. The left y-axis represents the accuracy for the line graph, while the right y-axis indicates the number of output clusters for the bar graph. The x-axis illustrates the order of seeds in each group.



**Figure 3.2:** The average  $f$ -score after the first way of modification. The process of modification follows as: at first, we take clusters that generated from  $2$ -cliques to modify the results that are generated from  $1$ -cliques. The average  $f$ -score after this modification are plotted at the position of 2. After that, the results are further modified by clusters generated from  $3$ -cliques, and plot the average  $f$ -score at the position of 3, and so on.

sub-graphs in PPI networks. Growing maximal cliques may result in randomly adding vertices to clusters, which tends to increase the rate of false positive for the predictions. This conjecture can also be seen from Table 3.2, where the values of recalls are almost the same for each pair of experiments, but the values of precisions differ greatly. The overall  $f$ -score of experiments by using maximal clique seeds is decreased due to the lower value of precision for each experiment.

The number of reasonable predictions is not as same as the number of clique seeds. This is mainly due to the fact that (1) some seeds may grow into disconnected clusters, with more than one connected components, and (2) some similar seeds could result in the same clusters according to the connectivity of the network. Those disconnected clusters should not be treated as meaningful predictions, and those redundant predictions should also be excluded from the final results. In the following of this paper, all results about the number of predictions are illustrated as the number of clusters that have been excluded disconnected and redundant ones.

### 3.3.2 Modification strategies

Although higher-order cliques generate clusters with a higher accuracy, it is not reasonable to take only those clusters as the final predictions. This is due to the fact that the number of output clusters decreases dramatically with the increase order of cliques. We cannot get enough predictions in this case. In addition, not all protein complexes contain higher-order cliques from biological point of view. However, the ascending

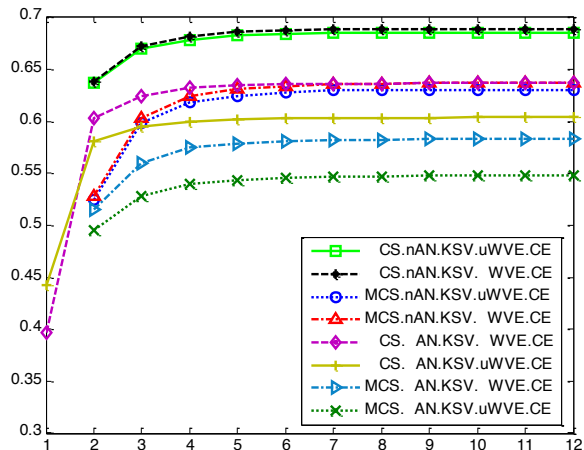
**Table 3.2:** The average *f-score*, *precision* and *recall* for different experiments

Strategies	<i>f-score</i>	<i>precision</i>	<i>recall</i>
CS. AN.KSV.uWVE.CE	0.377	0.485	0.403
MCS. AN.KSV.uWVE.CE	0.334	0.334	0.446
CS.nAN.KSV.uWVE.CE	0.432	0.594	0.393
MCS.nAN.KSV.uWVE.CE	0.399	0.457	0.406
CS. AN.KSV. WEV.CE	0.382	0.499	0.391
MCS. AN.KSV. WVE.CE	0.336	0.358	0.409
CS.nAN.KSV. WVE.CE	0.431	0.594	0.391
MCS.nAN.KSV. WVE.CE	0.396	0.460	0.399

For each prediction, the *precision* and *recall* are calculated first, and the *f-score* for this prediction is the harmonic mean of the *precision* and *recall* according to the formula. The values of the average *f-score*, the average *precision* and the average *recall* are calculated by taking the average values of *f-score*, *precision*, and *recall* over all predictions, respectively. Note that the value of the average *f-score* was not calculated by the harmonic means of the average *precision* and the average *recall*.

values of predictive accuracy inspire us to adopt modification strategies. Results generated by higher-order cliques are employed to modify those generated by lower-order seeds.

Two ways to modify results are adopted here. Firstly, since higher-order clique seeds bring a large number of novel predictions for protein complexes, we can just add all output clusters together as predictions by simply excluding same predictions and unreasonable clusters. Figure 3.2 illustrates such modifications in terms of the average *f-score*. We can see from the figure that the accuracy of predictions increases gradually for most experiments, from 0.40 to 0.44 for the group of best predictions, and from 0.27 to 0.38 for the group having the largest improvement. Secondly, when adding novel predictions from higher-order clique seeds, unreasonable or similar clusters should be excluded. On the one hand, adding all clusters from higher-order clique seeds dramatically increase the number of predictions. Only those reasonable ones are selected as final predictions according to their biological functions. On the other hand, different overlapped seeds may generate various similar clusters. Not all of them are necessarily predicted complexes. In order to test the performance of the seed-selection strategy, we select those meaningful predictions according to the known protein complex information. Figure 3.3 illustrates this way of modification in terms of the average *f-score*. We can see from the figure that the accuracy of predictions significantly increases at first three steps, and then remain stable at that level.



**Figure 3.3:** The average  $f$ -score after the second way of modification. After modified by clusters that are generated from higher-order clique seeds, the number of cliques increases dramatically and it is too large compared to the number of known protein complexes. By using known protein complex information, we can obtain approximately 1700 clusters for each experiment. The best overall accuracy among those results is 0.68 in terms of the average  $f$ -score, which is very good among current computational algorithms.

### 3.3.3 Strategies for the entropy-based algorithm

#### Neighbors of a cluster

We can also obtain different performance about whether adding neighbors for initial seeds from Figure 3.1. Two strategies are investigated here: AN and nAN. The blue lines in Figure 3.1 represent results of those experiments with AN while the red lines illustrates those with nAN. It clearly shows that only for  $1$ -clique seeds, those with AN performs better than those with nAN. On all other cases, those with AN not only decreases the number of predicted clusters, but also declines the average  $f$ -score of output clusters. This is reasonable because some high-degree seed vertices have too many neighbors. Adding all of them to the initial cluster may bring a large number of false positive proteins, which can also decrease the accuracy of final predictions. However, there is an important drawback if neighbors are not initially added to seed clusters. Most of them stop growing after only a few iterations and yet not get reasonable predictions for large protein complexes.

#### Seed vertices of a cluster

Seed vertices of a cluster should not be removed from the cluster during the implement of the algorithm. Table 3.3 summarizes the number of seed cliques and the number of predicted clusters under the conditions

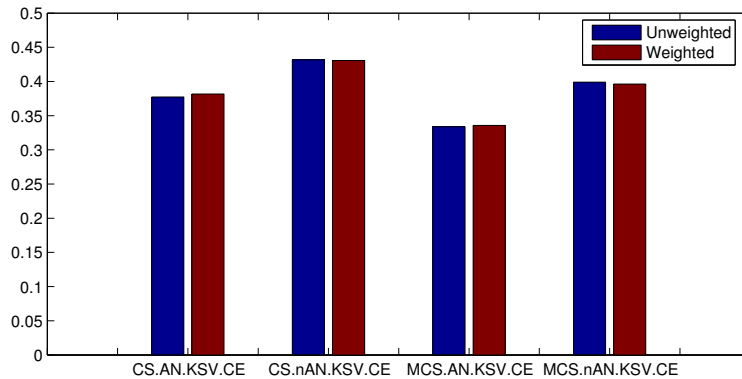
**Table 3.3:** The number of seeds and their output clusters

Strategies	Seeds number	KSV strategy	RSV strategy
CS. AN.uWVE.CE.	87003	78936	2109
CS.nAN.uWVE.CE.	87003	80973	93
MCS. AN. WVE.CE.	19621	19129	55
MCS.nAN. WVE.CE.	19621	19586	27

of RSV and KSV. All experiments we have done show similar characteristics. Here, we list four of them under specific experimental conditions. We can see from Table 3.3 that the number of meaningful clusters decreases dramatically under the condition of RSV. Only several clique seeds can grow into cliques as connected clusters. On the contrary, the condition of KSV generally works well. Most seeds can generate clusters as connected sub-graphs. Hence, it is convinced that the condition of KSV is necessary and indispensable when growing clusters.

### The vertex entropy

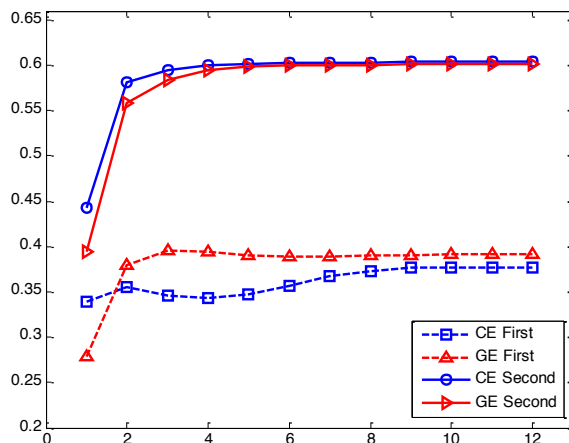
By comparing the overall accuracies in Fig. 3.4, it can be concluded that results do not show significant differences between strategies of WVE and uWVE. These trivial differences may be attributed to their limited discriminations according to the definitions. A new way to weight the vertex entropy or a new process to grow a cluster may result in better predictions, which can be further investigated.



**Figure 3.4:** The difference of accuracy between uWVE and WVE. Four groups of experiments have been performed to compare the different between the measurements of vertex entropy. Although the results are expected to be different at first, the computational results deny our conjecture.

## The cost function

For the cost function of the entropy-based algorithm, we test two different ways to perform the algorithm: CE and GE. The results are represented in Figure 3.5. Although the GE works better with the first modification way, it does not predict clusters with high accuracy by comparing their results with the second modification method. Therefore, it is hard to tell which one is better than the other. This is because these two cost functions focus on different measurements of clusters. The CE focuses only on the cluster itself while the GE focuses on both the cluster and the rest components of the network.

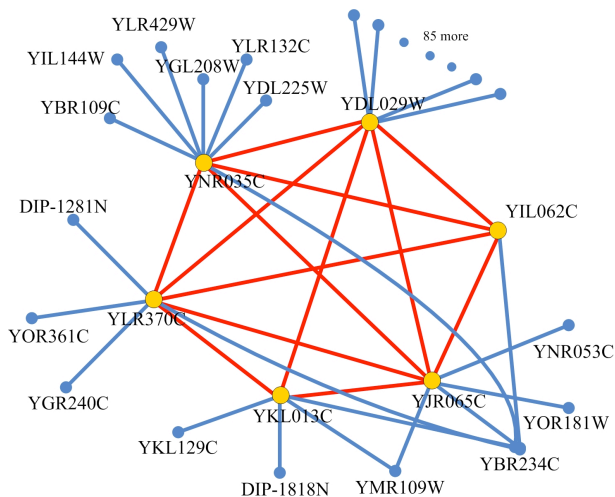


**Figure 3.5:** The differences of the average  $f$ -score between CE and GE. The experiments are performed under the strategies of CS, AN, KSV, and uWVE. The value of average  $f$ -score is shown after two ways of modifications, respectively.

However, it is worth mentioning that none of cost functions measures the quality of a cluster with consideration of the cluster size, or more specific, the number of vertices on the boundary. Some other kinds of cost function are encouraged to investigate in future studies such as the average (boundary) vertex entropy, the maximin vertex entropy, or the maximal vertex entropy.

### 3.3.4 The overall results of predictions

To summarize, the best strategies for the entropy-based algorithm is under conditions of CS, nAN, and KSV. We use known protein complex information to select predictions under these experimental conditions, and obtain 1711 clusters as the group of best predictions. The average  $f$ -score is about 0.44 for 1-clique seeds, and is about 0.68 after modifications by results of higher-order cliques. There are 549 of clusters exactly matching with ‘gold standard’ protein complexes in this group.



**Figure 3.6:** The predicted clusters in the PPI network. This is one of examples of predicted clusters. The Arp2p/Arp3p complex is exactly detected by our algorithm. It contains six proteins: YDL029w, YIL062c, YJR065c, YKL013c, YLR370c, and YNR035c. For the protein YDL029w, there are 85 more interactions according to the PPI network. However, we do not show all of them. The other proteins are neighbors of this protein complex.

Comparing with Cho’s previous work in [11], they obtained approximate 500 clusters as predictions, with the average *f-score* at about 0.44. It is almost the same with our results when using *1-clique* seeds (which are shown in Figure 3.3). The results indicate that the new seed selection strategy and our improvements for the entropy-based algorithm can increase the accuracy of predictions.

Figure 3.6 gives an example of the predicted clusters, which is exactly identified by our algorithm. The protein complex is called “Arp2p/Arp3p complex”, which consists of six proteins: YDL029w, YIL062c, YJR065c, YKL013c, YLR370c and YNR035c. The other vertices in Figure 3.6 are neighbors of this complex.

However, the accuracy of predicted results is still limited if we do not use known protein complex information. It can first be attributed to the large number of predictions. Since all cliques and maximal cliques in the PPI network are employed as seeds, too many output clusters are generated as candidate predictions. In practice, when there is no protein complex information available, it is better to use data, such as the Gene Ontology(GO), protein functions and localization, to exclude unreasonable clusters.

### 3.4 Conclusions

In this paper, we have first introduced a new seed-selection strategy for seed-growth style algorithms. Cliques in a PPI network are suggested as seeds rather than individual vertices. We have concluded that higher-order clique seeds not only generate more novel clusters with a higher accuracy, but also can be used to modify

results that are generated from lower-order clique seeds. In addition, we have investigated the performance of the entropy-based algorithm with four pairs of strategies and tested the proposed seed-selection strategy on the improved entropy-based algorithm.

Experiments are conducted on the *S. cerevisiae* PPI network of the DIP. Numerical results have indicated that cliques generate clusters with higher accuracy than individual vertices, and the best way to improve the entropy-based algorithm is under the condition nAN and KSV. The other two conditions about the vertex entropy and the cost function do not show significant differences based on our experiments.

However, there are still some issues that should be addressed along with this study in the future. First, novel methods should be investigated to improve results generated from lower-order cliques. Since the number of predictions decreases dramatically with the increase of clique orders, the improvement by modifying the results from those higher-order cliques is still limited. Secondly, a new way to eliminate redundant clusters should be investigated. In this paper, we pick the cluster with the largest *f-score* as a prediction for each known protein complex. However, it is impossible to make the novel predictions only according to those output clusters. The high accuracy of these predictions inspires us to use information, such as gene ontology, protein function and localization annotations to select those meaningful clusters.

## Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## Declarations

The authors have declared no conflict of interest.



## BIBLIOGRAPHY

- [1] Zhang S, Jin G, Zhang XS, Chen L. Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics* 2007, **7**(16): 2856-2869.
- [2] Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* 2008, **9**(10): 770-780.
- [3] Guimerá R, Nunes Amaral LA. Functional cartography of complex metabolic networks. *Nature* 2005, **433**(7028): 895-900.
- [4] Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ. A gene-centered *C. elegans* protein-DNA interaction network. *Cell* 2006, **125**(6): 1193-1205.
- [5] Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000, **18**(12): 1257-1261.
- [6] Pinkert S, Schultz J, Reichardt J. Protein Interaction Networks - More Than Mere Modules. *PLoS Computational Biology* 2010, **6**(1): e1000659.
- [7] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, **100**(21): 12123-12128.
- [8] Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waagele B, Schmidt T, Doudieu ON, Stümpflen V, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 2008, **36**(Database issue): D646-D650.
- [9] Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868): 141-147.

- [10] Pellegrini M, Haynor D, Johnson JM. Protein interaction networks. *Expert Rev Proteomics* 2004, **1**(2): 239-249.
- [11] Kenley EC, Cho YR. Detecting protein complexes and functional modules from protein interaction networks: a graph entropy approach. *Proteomics* 2011, **11**(19): 3835-3844.
- [12] Ruepp A, Waegle B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complex - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501.
- [13] Yu L, Gao L, Kong C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* 2011, **11**(19): 3826-3834.
- [14] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, **4**: 2.
- [15] King AD, Pržulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, **20**(17): 3013-3020.
- [16] Georgii E, Dietmann S, Uno T, Pagel P, Tsuda K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 2009, **25**(7): 933-940.
- [17] van Dongen S. Graph clustering by flow simulation. *PhD thesis, University of Utrecht*, 2000.
- [18] Li XL, Tan SH, Foo CS, Ng SK. Interaction graph mining for protein complexes using local clique merging. *Genome Inform* 2005, **16**(2): 260-269.
- [19] Lian H, Song C, Cho YR. Decomposing protein interactome networks by graph entropy. *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, 2010: 585-589.
- [20] Moschopoulos CN, Pavlopoulos GA, Iacucci E, Aerts J, Likothanassis S, Schneider R, Kossida S. Which clustering algorithm is better for predicting protein complexes? *BMC Res Notes* 2011, **4**:549.
- [21] Chen B, Yan Y, Shi J, Zhang S, Wu FX. An improved graph entropy-based method for identifying protein complexes. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 2011: 123-126.
- [22] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 2009, **37**(3): 825-831.
- [23] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004, **32**(Database issue): D449-D451.

- [24] Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KF, Münsterkötter M, Ruepp A, Spannagl M, Stümpflen V, Rattei T. MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 2008, **36**(Database issue): D196-D201.

## CHAPTER 4

# NOT ALL PROTEIN COMPLEXES EXHIBIT DENSE STRUCTURES IN *S. cerevisiae* PPI NETWORK

*Published as:* Chen B, Shi J and Wu FX. Not all protein complexes exhibit dense structures in *S. cerevisiae* PPI network. *Bioinformatics and Biomedicine (BIBM)*, 2012 *IEEE International Conference on:* 470-473.

In the previous chapter, we have proposed a dense sub-graph detection algorithm to identify protein complexes based on clique seeds and graph entropy. Although the algorithm can generate a large number of predictions with high *f-score*, there are still many protein complexes that cannot be correctly identified by that algorithm.

In this chapter, real topological characteristics of known yeast protein complexes are investigated on a DIP yeast PPI network, and it is concluded that not all protein complexes exhibit dense structures in PPI networks. Many of them have a star-like structure, which is a very specific core-attachment structure. The conclusion of this chapter gives a new direction for identifying protein complexes based on topological characteristics in PPI networks.

### Abstract

Various algorithms have been proposed to identify protein complexes from PPI networks, based on the assumption that protein complexes are densely connected sub-graphs. In this study, we conclude that most known protein complexes do not exhibit dense structures in *S. cerevisiae* PPI network, but maintain star-like structures in the network. Moreover, vertices of protein complexes are not sparsely connected with the rest components of the network. Many vertices tend to have more outgoing interactions than they have within protein complexes. Based on star-like properties of known protein complexes, we propose a random-star algorithm to identify protein complexes in PPI networks. Predictions are evaluated in terms of the average *f-score*. After excluding similar clusters, we finally obtained 744 predictions with the average *f-score* at 0.51.

## 4.1 Introduction

Protein complexes are essential molecular entities that carry out major cellular processes [1]. They consist of groups of proteins that physically bind together in living cells [2]. Understanding them is an essential step for our attempt towards unraveling the intricate biological systems [3].

Various computational algorithms have been proposed to identify protein complexes according to their topological structures in protein-protein interaction (PPI) networks. The most commonly used assumption is that protein complexes exhibit dense structures in PPI networks. Algorithms, such as the maximal clique algorithm [4], MCODE [5], RNSC [6], DEM [7], LCMA [8], MCL [9], and the graph-entropy-based algorithm [10], are proposed based on this assumption. The other assumption for protein complexes is the core-attachment structures [2]. Many core-attachment approaches [3][11] are developed to identify the cores and attachments of protein complexes, separately. Although most of those algorithms are efficient and helpful, their accuracy is still limited, which is due to not only the intricate connections of PPI networks, but also the unclear characteristics of protein complexes.

In this paper, we first investigate properties of protein complexes in a *S. cerevisiae* PPI network of DIP. We find that most protein complexes do not exhibit dense structures in the PPI network, but are sparsely connected in terms of both the density and the average degree. We introduce a cyclic-level model to describe the relationship between protein complexes and their surrounding neighbours. Statistic results show that protein complexes have distinct statistic characteristics, which indicates that they are identifiable in PPI networks. Moreover, we conclude that most protein complexes exhibit star-like structures in the PPI network. Proteins are more likely to have interactions with only one or more hub-proteins within complexes, and most of them tend to have frequently connections with proteins out of complexes. Based on this characteristic, we finally propose a random-star algorithm to identify protein complexes in PPI networks. Numerical experiments are conducted on the PPI network of DIP. Predicted results show that the algorithm can output protein complexes with high accuracy, which is very promising in predicting protein complexes.

## 4.2 Materials and methods

A PPI network can be represented as an undirected graph  $G = (V(G), E(G))$ , where  $V(G)$  is the set of vertices (individual proteins), and  $E(G)$  is the set of edges (protein interactions). Let  $H = (V(H), E(H))$  be a sub-graph of  $G$ , the neighbours of  $H$  can be defined as

$$N(H) = \{v | (u, v) \in E(G), u \in V(H), v \in V(G) \setminus V(H)\}. \quad (4.1)$$

Without loss of generality, in this paper we do not distinguish concepts of PPI networks, protein complexes and proteins from graphs, sub-graphs and vertices, respectively.

### 4.2.1 Protein complexes and their relative neighbours

We introduce a cyclic-level model to represent the relationship between a protein complex and the rest components of a PPI network. From inside to outside, they are (1) the core level, (2) the inner boundary (IB) level and (3) the outer boundary (OB) level, respectively. To be more precise, let  $P$  be a protein complex, then the core level consists of vertices that interact with proteins only in the complex,

$$Core(P) = \{v | v \in V(P), N(v) \subset V(P)\}, \quad (4.2)$$

while the IB level consists of vertices of the complex, but have interactions with proteins out of the complex,

$$IB(P) = \{u | (u, v) \in E(G), u \in V(H), v \in V(G) \setminus V(H)\}. \quad (4.3)$$

The OB level is made up of all proteins that have interactions with proteins in the complex, but are not components of the complex, which is

$$OB(P) = N(P). \quad (4.4)$$

The cyclic-level model provides a meticulous way to describe a protein complex in a PPI network. Specifically, edges of a vertex can be divided into three categories, which incident with vertices (1) in the inside level, (2) in the same level and (3) in the outside level, respectively. Then the degree  $d(v)$  of a vertex  $v$  is decomposed into three kinds of degrees:  $d_i(v)$ ,  $d_l(v)$  and  $d_o(v)$ , which represent the number of each kind of edges, respectively.

### 4.2.2 The number of edges, density, relative density and radius

Given two adjacent levels  $L_1$  and  $L_2$ , the set of edges that incident with vertices only in  $L_1$  is denoted by  $E(L_1)$ , and the set of edges that incident with vertices between  $L_1$  and  $L_2$  is denoted by  $E(L_1, L_2)$ . Therefore, the number of edges in  $L_1$  and between  $L_1$  and  $L_2$  are

$$m(L_1) = |E(L_1)| \text{ and } m(L_1, L_2) = |E(L_1, L_2)|, \quad (4.5)$$

respectively.

The density of  $L_1$  can be measured by the commonly used definition

$$Q(L_1) = \frac{2 \cdot m(L_1)}{n(L_1) \cdot (n(L_1) - 1)}, \quad (4.6)$$

where  $n(L_1)$  is the number of vertices in  $L_1$ . However, when it comes to the density of two adjacent levels, edges in both levels should not be counted. The density is given as

$$Q(L_1, L_2) = \frac{m(L_1, L_2)}{n(L_1) \cdot n(L_2)}. \quad (4.7)$$

The relative density of two levels or the relative density between a level and two adjacent levels are defined as

$$RQ(L_1|L_2) = \frac{Q(L_1)}{Q(L_2)} \text{ and } RQ(L_1|L_1, L_2) = \frac{Q(L_1)}{Q(L_1, L_2)}. \quad (4.8)$$

The concept of radius in the cyclic-level model is more important. Suppose each vertex of a sub-graph  $H$  is assigned an unit area  $S = \pi r^2$ , where  $r = 1$ , then the overall area of the sub-graph should be  $S(H) = \pi \sqrt{n(H)} \cdot r^2$ . It gives a quantitative definition about how large a sub-graph should cover if the density of a network is equally distributed. For a single level  $L_1$ , the radius is defined as  $r(L_1) = \sqrt{n(L_1)}$ , while for two adjacent levels  $L_1$  and  $L_2$ , the radius is defined as  $r(L_1, L_2) = \frac{\sqrt{n(L_1)} + \sqrt{n(L_2)}}{2}$ .

## 4.3 Experiments and results

### 4.3.1 Data source

Protein complex data are collected from the database of MIPS [12], CYC2008 [13], YHTP2008 [13], and from the paper of Spirin and Mirny [4]. After removing redundant protein complexes, we finally obtain 2,165 protein complexes as the *gold standard*. There are 870 protein complexes consisting of more than five proteins, and 462 of them consisting of more than ten proteins. However, 563 protein complexes are made up of only two proteins, and another 369 complexes contain only three proteins.

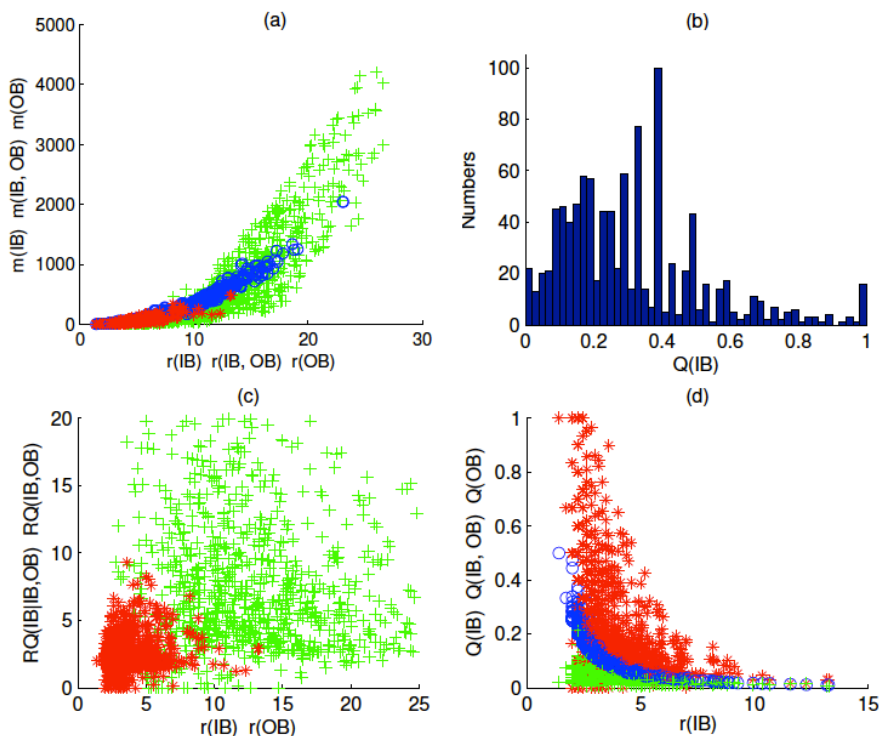
The PPI data are downloaded from the database of interacting proteins (DIP) [14]. The file, named as *Scere20120228.txt*, contains 5004 proteins and 22010 interactions after removing all redundant data, including interactions between *S. cerevisiae* and other species, interactions between the same proteins (loops) and the same interactions between two proteins (multiple edges).

### 4.3.2 Statistical results

We test known protein complexes data on the PPI network of DIP, and find that almost all vertices of protein complexes have interactions with proteins both in and out of the complexes. Only three protein complexes have core level. Therefore, in most cases the IB level consists of all vertices in a protein complex.

## The number of edges

The number of edges for each level of protein complexes shows distinct properties. Fig 4.1(a) illustrates  $m(IB)$ ,  $m(IB, OB)$  and  $m(OB)$  against to their relative radius. We can clearly see from Fig 4.1(a) that the value of  $m(IB)$  and  $m(IB, OB)$  rise gradually with the increase of the radius, while the value of  $m(OB)$  does not have such significant property. We also test the number of edges for the further outer boundary level of current OB ones. They do not show significant differences from characteristics of randomly selected sub-graphs in the PPI network, which indicates that structures of protein complexes are different from those of randomly selected ones.



**Figure 4.1:** Statistic results for protein complexes with no less than five proteins. The scatter diagrams are plotted according to the value of different statistics and their relative radius. (a) the number of edges; (b) the distribution of density; (c) the scatter diagram of relative density; (d) the scatter diagram of density.

## Density and relative density

It is hard to be convinced that protein complexes have dense structures. Fig 4.1(b) illustrates the density distribution for protein complexes that consist of more than five proteins. The densities for most protein complexes are less than 0.4, and the value of the average density is only 0.31.

However, protein complexes are still relatively denser than their neighbours. Fig 4.1(c) gives the scatter



**Table 4.1:** The average degree for IB and OB vertices

IB vertices	OB vertices
-	$\bar{d}_i(v) = 1.21$
$\bar{d}_l(v) = 2.30$	$\bar{d}_l(v) = 4.48$
$\bar{d}_o(v) = 18.02$	$\bar{d}_o(v) = 30.28$

diagram of relative density  $RQ(IB|IB, OB)$  and  $RQ(IB|OB)$ . We can see from Fig 4.1(c) that values of most relative densities are large than 1. In fact, the average value of  $RQ(IB|IB, OB)$  is 2.45, and the average value of  $RQ(IB|OB)$  is 10.54.

Moreover, if values of  $Q(IB)$ ,  $Q(IB, OB)$  and  $Q(OB)$  are compared vertically, which are plotted against to the value of  $r(IB)$ , they show clear boundaries for those densities. The scatter diagram is illustrated in Fig 4.1(d). From top to bottom, the red region is the values of  $Q(IB)$ , the blue region is the values of  $Q(IB, OB)$ , and the green region is the values of  $Q(OB)$ . This characteristic can be used to evaluate whether a predicted result is a protein complex.

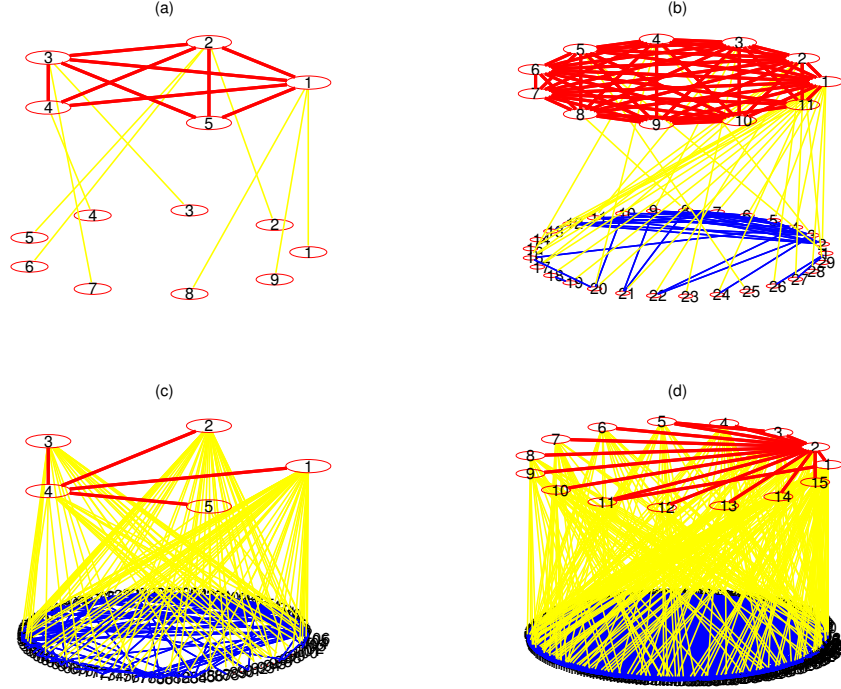
### The average degrees of protein complexes

The vertices of protein complexes tend to have more outgoing interactions than they have within the protein complexes. Table 4.1 summarizes the average degrees of both IB and OB vertices for all protein complexes with more than three vertices, in terms of  $\bar{d}_i(v)$ ,  $\bar{d}_l(v)$  and  $\bar{d}_o(v)$ .

It is hard to conclude that protein complexes are dense structures in PPI networks. The average degree of vertices within protein complexes is only 2.30. Considering that the average degree of a minimal connected sub-graphs almost equals to 2, the small value of the average degree of vertices within protein complexes indicates that the number of connections within them may only suffice for them to be connected sub-graphs.

### 4.3.3 The structures of protein complexes

We conclude that protein complexes exhibit star-like structures in *S. cerevisiae* PPI network. For all known protein complexes of yeast, we draw pictures of them and their relative neighbors. Although some of them are densely connected within themselves (such as the top red regions in Fig 4.2(a) and Fig 4.2(b)), there are approximately 70% of them tending to exhibit star-like structures (such as structures in Fig 4.2(c) and Fig 4.2(d)). Most protein complexes have one or more hub-proteins, where all other proteins only interact with



**Figure 4.2:** The structure of protein complexes. The top red region of each sub-graph represents a protein complex (IB vertices), while the bottom blue region illustrates relative OB vertices. The yellow lines indicate interactions between them. (a) a protein complex with dense inner connections and sparse outer connections; (b) a protein complex with dense inner and sparse outer connections; (c) a protein complex with star-like inner connections and dense outer connections; (d) a larger star-like protein complex.

them within complexes. It is noteworthy that proteins tend to have more outgoing interactions than they have within complexes, no matter whether they are hub-proteins or not.

## 4.4 Algorithm and results

Based on above statistic characteristics and star-like structures of known protein complexes, we propose a random-star algorithm to identify protein complexes from PPI networks. Since the overall degree of vertices in protein complexes are usually very large, we consider only those highly connected vertices in the PPI network. All vertices of the network are divided into three categories: (1) core vertices ( $d(v) \geq 50$ ), (2) important vertices ( $3 < d(v) < 50$ ) and (3) trivial vertices ( $d(v) \leq 3$ ). The upper threshold is assigned according to the average maximum degree, which is 49.6 for all known protein complexes. The lower threshold is selected based on the fact that they do not significantly affect the structure of protein complexes. However, they can be changed according to properties of a PPI network.

### 4.4.1 The random-star algorithm

The algorithm is described as follows:

**Input:** A PPI network  $G$ .

**Output:** A group of star-like clusters.

- 1: Initialize the random-times  $T$  and a threshold  $p$ .
- 2: **for**  $i = 1 : T$  **do**
- 3:   Initialize core vertices list  $L_{core}$ , important vertices list  $L_{impt}$ , and trivial vertices list  $L_{tvil}$ .
- 4:   **while**  $L_{core}$  is not empty **do**
- 5:     Randomly select a core vertex  $v \in L_{core}$  and let  $C = N(v) \setminus L_{tvil}$ ,  $L_{core} = L_{core} \setminus \{v\}$ .
- 6:     Randomly select a vertex  $u \in C$ .
- 7:     **while**  $u$  is not empty **do**
- 8:       Let  $C_1 = N_C(u)$  and  $C_2 = C \setminus (C_1 \cup \{u\})$ .
- 9:       Get a random number  $r \sim U(0, 1)$ .
- 10:       **if**  $r \geq p$  **then**
- 11:          Randomly select a vertex  $u \in C_1$ , and let  $C = C_1$ .
- 12:       **else**
- 13:          Randomly select a vertex  $u \in C_2$ , and let  $C = C_2$ .
- 14:       **end if**
- 15:     **end while**
- 16:     Output a cluster.
- 17:   **end while**
- 18: **end for**

### 4.4.2 Accuracy evaluation

We use  $f$ -score as the measure to evaluate the accuracy of predictions for protein complexes. It is defined as the harmonic mean of  $precision = k/n(H)$  and  $recall = k/n(P)$ , where  $n(H)$  and  $n(P)$  are the number of proteins in a predicted cluster  $H$  and a known protein complex  $P$  respectively, and  $k$  is the number of proteins they have in common. The  $f$ -score is defined as

$$f\text{-score} = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot k}{n(H) + n(P)}.$$

It is a measure that balances both the true positive predictive rate and the true positive rate.

### 4.4.3 Predicted results

We vary values of the probability threshold from 0.9 to 0.1, the overall accuracy of prediction increase gradually from only 0.28 to 0.43. The smaller the threshold, the more star-like clusters are generated.

Output clusters can be first excluded according to values of  $Q(IB)$ ,  $Q(IB,OB)$  and  $Q(OB)$ . After fitting boundaries of densities in Fig 4.1(d), we obtain the upper boundary line and the lower boundary line as

$$f_u = \frac{1.8}{r(IB)} \text{ and } f_l = \frac{0.9}{r(IB)},$$

respectively.

Since we randomly run 100 times for each core proteins (about 200 core proteins in the PPI network), the number of predictions far exceed the number of known protein complexes. After excluding similar predictions according to known protein complexes, we finally obtain 744 predictions of protein complexes, with the average *f-score* at 0.51. It indicates that our proposed random-star algorithm can be a promising method in terms of predicting protein complexes.

## 4.5 Conclusions

In this paper, we first analyze statistic properties of protein complexes in *S. cerevisiae* based on the PPI network of DIP. We have concluded that most protein complexes exhibit star-like structures in the PPI network, rather than the dense structures. Moreover, most proteins in those complexes have interactions with proteins out of complexes, and many of them even have more outgoing interactions than they have within complexes.

Based on statistic properties of protein complexes, we propose a random-star algorithm to generate star-like sub-graphs in PPI networks. Although it still need to be further improved, the best group of predictions still report protein complexes with high accuracy.

## Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## BIBLIOGRAPHY

- [1] Ruepp A, Waagele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complex - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501
- [2] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868): 141-147.
- [3] Yu L, Gao L, Kong C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* 2011, **11**(19): 3826-3834.
- [4] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, **100**(21): 12123-12128.
- [5] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, **4**: 2.
- [6] King AD, Pržulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, **20**(17): 3013-3020.
- [7] Georgii E, Dietmann S, Uno T, Pagel P, Tsuda K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 2009, **25**(7): 933-940.
- [8] Li XL, Tan SH, Foo CS, Ng SK. Interaction graph mining for protein complexes using local clique merging. *Genome Inform* 2005, **16**(2): 260-269.
- [9] van Dongen S. Graph clustering by flow simulation. *PhD thesis, University of Utrecht*, 2000.
- [10] Kenley EC, Cho YR. Detecting protein complexes and functional modules from protein interaction networks: a graph entropy approach. *Proteomics* 2011, **11**(19): 3835-3844.

- [11] Pang CN, Krycer JR, Lek A, Wilkins MR. Are protein complexes made of cores, modules and attachments? *Proteomics* 2008, **8**(3): 425-434.
- [12] Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KF, Münsterkötter M, Ruepp A, Spannagl M, Stümpflen V, Rattei T. MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 2008, **36**(Database issue): D196-D201.
- [13] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res*, 2009, **37**(3): 825-831.
- [14] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004, **32**(Database issue): D449-D451.

## CHAPTER 5

# IDENTIFYING PROTEIN COMPLEXES BASED ON MULTIPLE TOPOLOGICAL STRUCTURES IN PPI NETWORKS

*Published as:* Chen B and Wu FX. Identifying protein complexes based on multiple topological structures in PPI networks. *IEEE Transactions on Nanobioscience* 2013, **12**(3): 165-172.

In the previous chapter, we have investigated topological characteristics of known protein complexes in a DIP PPI network. We conclude that not all protein complexes exhibit dense structures in PPI networks. Many of them have a star-like structure in PPI networks. A random star algorithm has also been proposed to identify such star-like topological structures. Similar to the dense structure, a single star-like structure is also not enough to identify protein complexes with different topological structures.

In this chapter, we propose a multiple-topological-structure-based algorithm to identify protein complexes from PPI networks. Four kinds of raw topological sub-graphs are detected in PPI networks, which are cliques, dense sub-graphs, core-attachment structures and star-like structures. Then, those raw sub-graphs are merged and/or trimmed based on their topological information and/or GO annotations. Numerical experiments show that the proposed algorithm generates not only more reasonable predictions, but also with high prediction performance in terms of *f-score*. It outperforms many previous single-topological-structure-based algorithms for identifying protein complexes.

### Abstract

Various computational algorithms are developed to identify protein complexes based on only one of specific topological structures in protein-protein interaction (PPI) networks, such as cliques, dense sub-graphs, core-attachment structures and star-like structures. However, protein complexes exhibit intricate connections in a PPI network. They cannot be fully detected by only single topological structure. In this paper, we propose an algorithm based on multiple topological structures to identify protein complexes from PPI networks. In

the proposed algorithm, four single-topological-structure-based algorithms are first employed to identify raw predictions with specific topological structures, respectively. Those raw predictions are trimmed according to their topological information or GO annotations. Similar results are carefully merged before generating final predictions. Numerical experiments are conducted on a yeast PPI network of DIP and a human PPI network of HPRD. The predicted results show that the multiple-topological-structure-based algorithm can not only obtain a more predictions, but also generate results with high prediction performance in terms of *f-score*, matching with known protein complexes and functional enrichments with GO.

## 5.1 Introduction

Protein complexes are key molecular entities that carry out most cellular processes within cells [1]. Identifying them plays an important role for our attempts to reveal principles of cellular organizations and biological functions. However, it is still limited to detect protein complexes directly through experimental ways, especially of those involved high-throughput techniques. With advances of techniques, such as the yeast two-hybrid (Y2H) assay and affinity purification followed by mass spectrometry (AP/MS), enormous protein interaction data have been accumulated for various organisms [2, 3]. Although most of them are incomplete and inaccurate [4, 5], they reveal important principles of protein organizations within cells [6, 7].

Taking individual proteins as vertices and pair-wise interactions between them as edges, a group of protein interaction data can be modelled as a graph, called a protein-protein interaction (PPI) network. Various computational algorithms have been developed to identify protein complexes according to their topological structures in PPI networks. Among those algorithms, four kinds of topological structures are commonly assumed to be protein complexes, which are cliques, dense sub-graphs, core-attachment structures and star-like structures.

Firstly, cliques are employed due to the assumption that all proteins in a complex are interacted with each other. It is not too hard to enumerate all cliques and maximal cliques in a PPI network [8]. However, it is quiet arbitrary if only those fully connected sub-graphs are employed as predictions, and the condition of fully connected is also too strict for most protein complexes [9].

Secondly, for the dense sub-graphs, many bioinformatic analysis results of PPI networks have shown that proteins in a complex commonly display strong and frequent connections within the complex and weak and rare connections to proteins out of the complex [10, 11]. Algorithms, such as the MCODE [12], RNSC [13], DME [14], LCMA [15], MCL [16, 17] and the graph-entropy-based algorithm [9, 18] are designed to identify protein complexes based on this structure.



Thirdly, the core-attachment structure is suggested by Gavin et al. in [6]. They investigate properties of yeast protein complexes by using AP/MS, and find that a protein complex usually consists of two parts: core and attachment. The core proteins tend to have relatively more interactions among themselves, while the attachment proteins bind to proteins of the core to form a protein complex. This property is also supported by other high-throughput protein data [19]. Identification algorithms based on this assumption can be found in [20–22].

Finally, we claim that many known protein complexes exhibit star-like structures in a yeast PPI network [23]. Proteins in a complex tend to have interactions with only one hub-protein, rather than interact with each other to form a local dense sub-graph. Actually, the star-like structure can be viewed as a special case of core-attachment structure, where only one protein appears in the core part. Taking this particular topological structure as one of categories is due to the fact that many core-attachment-based algorithm only identify dense sub-graphs as cores, which may miss plenty of predictions that exhibit star-like structures. A random-star algorithm is also developed in [23].

Each kind of topological structure can generate a group of meaningful predictions. However, since protein complexes exhibit intricate connections in PPI networks, they cannot be fully detected by only one structure. Hence, in this paper, we propose an algorithm to identify protein complexes by multiple topological structures. To achieve this, we first learn characteristics of known protein complexes in a yeast and a human PPI network. The topological structure distribution is plotted and analyzed according to their sizes. The characteristics of known protein complexes are investigated from both topological and biological points of view, which can be used to evaluate how likely a prediction is a real protein complex. A sub-graph merging method is also developed to handle the problem of similar predictions exhibiting the same or different structures. The remainder of the paper is organized as follows. Section 5.2 describes materials and methods used in this paper. Section 5.3 addresses the computational results and discussions. Section 5.4 draws some conclusions.

## 5.2 Materials and methods

### 5.2.1 Terminologies

A PPI network can be represented as an undirected simple graph  $G = (V(G), E(G))$ , where  $V(G)$  is the set of vertices (individual proteins) and  $E(G)$  is the set of edges (protein interactions). The degree of a vertex  $v$  in  $G$ , denoted by  $d(v)$ , is the number of edges incident with  $v$ . The neighbors of  $v$  in  $G$ , denoted by  $N(v)$ , is the set of vertices adjacent to  $v$ .

Protein complexes are usually assumed to be sub-graphs of PPI networks. Let  $H = (V(H), E(H))$  be a sub-graph of  $G$ . The neighbors of  $H$  is defined as

$$N(H) = \{v | (u, v) \in E(G), u \in V(H), v \in V(G) \setminus V(H)\}. \quad (5.1)$$

The density of a sub-graph  $H$  is defined as

$$Q(H) = \frac{2 \cdot m_i(H)}{n(H) \cdot (n(H) - 1)}, \quad (5.2)$$

where  $n(H)$  is the number of vertices in  $H$ , and  $m_i(H)$  is the number of edges within it. Similarly, we define the density of area between  $H$  and its neighbors in  $G$  as

$$Q_o(H) = \frac{m_o(H)}{n(H) \cdot |N(H)|}, \quad (5.3)$$

where  $m_o(H)$  is the number outgoing edges between  $H$  and  $N(H)$  and  $|N(H)|$  is the cardinality of  $N(H)$ .

Without loss of generality, in this paper we do not distinguish concepts of PPI networks, protein complexes and proteins from graphs, sub-graphs and vertices, respectively.

## 5.2.2 Deriving the weights for PPI networks

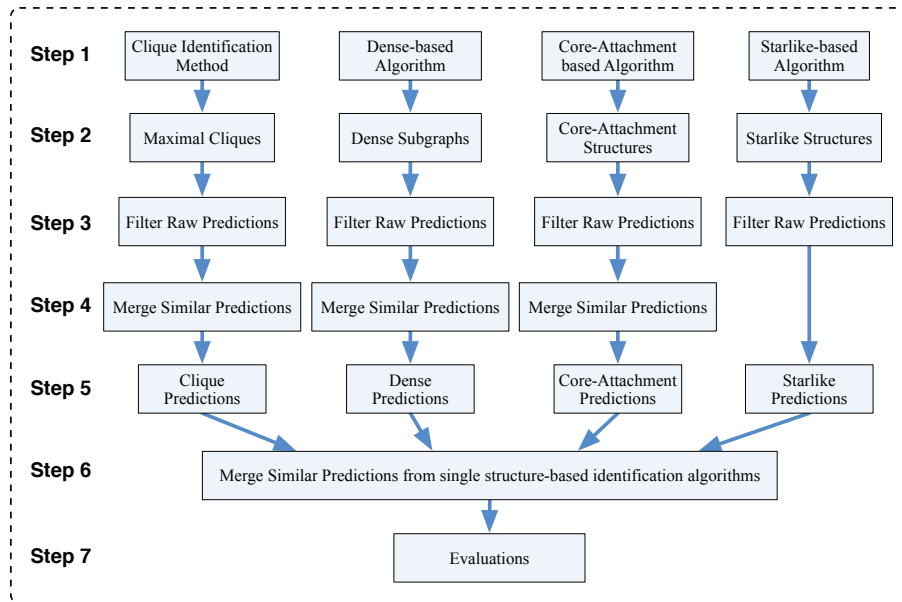
It is generally acknowledged that if two proteins have a significantly large number of common neighbors, they tend to have close functional relationship between them [24]. The probability that two proteins have  $k$  neighbors in common is defined by the *p-value* [24], *i.e.*,

$$P(N, n_1, n_2, k) = \frac{\binom{N}{k} \binom{N-k}{n_1-k} \binom{N-n_1}{n_2-k}}{\binom{N}{n_1} \binom{N}{n_2}}, \quad (5.4)$$

where  $n_1$  and  $n_2$  are the number of neighbors of those two proteins and  $N$  is the number of proteins in the PPI network. The numerator counts the number of distinct ways to select  $n_1$  and  $n_2$  vertices with  $k$  of them in common, while the denominator counts the number of distinct ways to select any  $n_1$  and  $n_2$  vertices from  $N$  vertices. The lower the *p-value* is, the more significant two adjacent vertices are closely functionally related. The value of  $P(N, n_1, n_2, k)$  can be used to assign weights for PPI networks, by its log form or its absolute value of the log form.

## 5.2.3 A framework for identifying protein complexes based on multiple topological structures

Protein complexes do not exhibit an universal structure in PPI networks, but have intricate connections within PPI networks. However, most computational algorithms focus on only one specific topological structure in



**Figure 5.1:** A framework for identifying protein complexes based on multiple topological structures.

a PPI network, such as cliques, dense sub-graphs, core-attachment structures and star-like structures. This is reasonable to some extent, but it is hard to identify all protein complexes based on only one single structure. To overcome this drawback, we propose an algorithm to identify protein complexes based on multiple topological structures. The framework is illustrated in Fig. 5.1.

From a particular single-structure-based algorithm, a number of clusters with that kind of structure are generated as raw predictions. They are trimmed according to their topological and/or biological informations. After that, similar predictions are merged before they are outputted as final predictions with that kind of topological structure. The multiple-topological-structure-based algorithm take all such final predictions as input. Similar results are further carefully merged before they are outputted to the evaluation step.

To test the performance of the multiple-topological-structure-based algorithm, Step 1 to 5 of the framework are performed as follows. For the clique identification method, the clique enumerating algorithm proposed in [8] are employed to generate all maximal cliques as raw predictions. For the dense-based algorithm, the graph-entropy-based algorithm introduced in [9, 18] is used, where all edges are employed as seeds to grow raw predictions. The core-attachment-based algorithm used in this study is proposed in [20], where a weighted PPI network is employed as input. Although it can also generate some star-like predictions, this happens only when those star hubs do not appear in any previous identified cores. Hence, a special star-like-based algorithm is necessarily to be performed for this particular kind of topological structure. Since the random-star algorithm introduced in [23] is quiet time-consuming, we propose a novel connected-star algorithm for this study. A step-by-step description of this algorithm is as follows. Step 1: assigning weight to each edge

**Table 5.1:** Details of the PPI datasets

	Databases	Proteins	Interactions	Average Degrees
Yeast	DIP	5 000	22 049	8.82
Human	HPRD	9 465	37 039	7.83

in a PPI network, by calculating the  $p$ -value by (5.4). Step 2: removing those edges whose weights are large than a threshold. Step 3: growing a star-like sub-graph by adding all its neighbors for every vertex in the remained network.

After obtaining raw predictions from each single-topological-structure-based algorithms, the next step addresses the issue of trimming raw predictions in the framework. This needs to evaluate how likely a predicted protein complex tends to be a true one. Details of this prediction trimming method and the subsequent prediction merging method are introduced in the following section.

We use the MCL algorithm [16, 17] as a comparison to evaluate the multiple-topological-structure-based algorithm. The MCL algorithm is generally regarded as an efficient method that can generate predictions with high accuracy. To optimize the performance of the MCL, we use the weighted PPI network as its input, and generate predictions under its default parameters (weights of edges are assigned by the absolute value of the log form in (5.4)).

## 5.2.4 Data sources

We test our method on a yeast PPI network and a human PPI network. The yeast PPI dataset is downloaded from the Database of Interacting Proteins (DIP) [25], named as *Scere20120228.txt*. The human PPI dataset is downloaded from the Human Protein Reference Database (HPRD) [26] under the package named as *HPRD\_Release9\_041310.tar.gz*. The details of those PPI datasets are shown in Table 5.1.

Data of yeast protein complexes are collected from the database of MIPS [27]. The data are organized hierarchically, where one protein complex may have sub-complexes as its descendants. The categories of MIPS are manually curated from the literature and thus have strong biological evidences, except the category 550 (predicted by computational methods). After excluding complexes consisting of a single and a pair of proteins, we finally obtain 143 manually curated protein complexes (denoted by *MIPS-Bio*) and 666 protein complexes from the category 550 (denoted by *MIPS-Com*). The number of all MIPS protein complexes (denoted by *MIPS-All*) is 807, as two of the computationally predicted ones are the same as the manually curated protein complexes. Data of human protein complexes are collected from the database of CORUM

[11]. There are 513 protein complexes remained after excluding those complexes consist of a single and a pair of proteins.

The yeast GO annotation dataset is downloaded from the *Saccharomyces* Genome Database (SGD) database [28]. The submission data is 12/15/2012. The human GO annotation dataset is downloaded from the Gene Ontology database [29]. The submission data is 12/10/2012. Generally, the number of proteins annotated under one GO term may vary from one to several thousands. When they are employed to analyze the enrichment of protein complexes, those large GO terms may bring noise information, since randomly selected protein groups may also co-appear in them. Hence, we only select those GO terms that annotate no more than 50 proteins, which is also the size comparable to majority protein complexes.

### 5.2.5 Evaluating predictions by GO annotations

The gene ontology project [29] has been developed to describe gene products in three vocabulary domains: cellular component, molecular function and biological process. It is accepted as ground-truth and used for comparison and validation purposes [30]. A group of proteins annotated under a GO term usually have similar biological functions.

Proteins in a complex usually have high functional homogeneity [31], which means they tend to be annotated by a same GO term. The maximal number of a group of proteins  $H$  that have a same GO term is

$$n_{GO}(H) = \max\{|H \cap F_i|\}, \quad i = 1, \dots, r, \quad (5.5)$$

where  $F_i$  is the set of proteins in the  $i^{th}$  GO term and  $r$  is the total number of GO terms. However, since the size of protein complexes varies differently, simply using  $n_{GO}$  for all protein complexes may lead to misleading conclusions. Hence, we use the GO annotation rate

$$r_{GO}(H) = \frac{n_{GO}(H)}{n(H)} \quad (5.6)$$

to represent how possible proteins in a complex are co-annotated by GO terms.

Moreover, since the number of proteins annotated by various GO terms is not uniform distributed, it is also useful to define the functional homogeneity *p-value* to measure the probability that a given set of  $n$  proteins is enriched by a given GO term by chance. The *p-value* is defined by the hypergeometric distribution as

$$p\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i} \binom{N-|F|}{n-i}}{\binom{N}{n}}, \quad (5.7)$$

where  $|F|$  is the number of proteins in a GO term,  $k$  is the number of proteins they have in common,  $N$  is the total number of proteins in a PPI network. The smaller the *p-value* is, the more statistically significant the protein complex is enriched by GO annotations.

## 5.2.6 Evaluating predictions by gold standard protein complexes

We use *f-score* as the measure to evaluate the accuracy of predicted protein complexes. Suppose a predicted protein complex  $H$  is compared with a known protein complex  $P$ , the *precision* and *recall* is defined as follows:

$$precision = \frac{k}{n(H)} \quad \text{and} \quad recall = \frac{k}{n(P)},$$

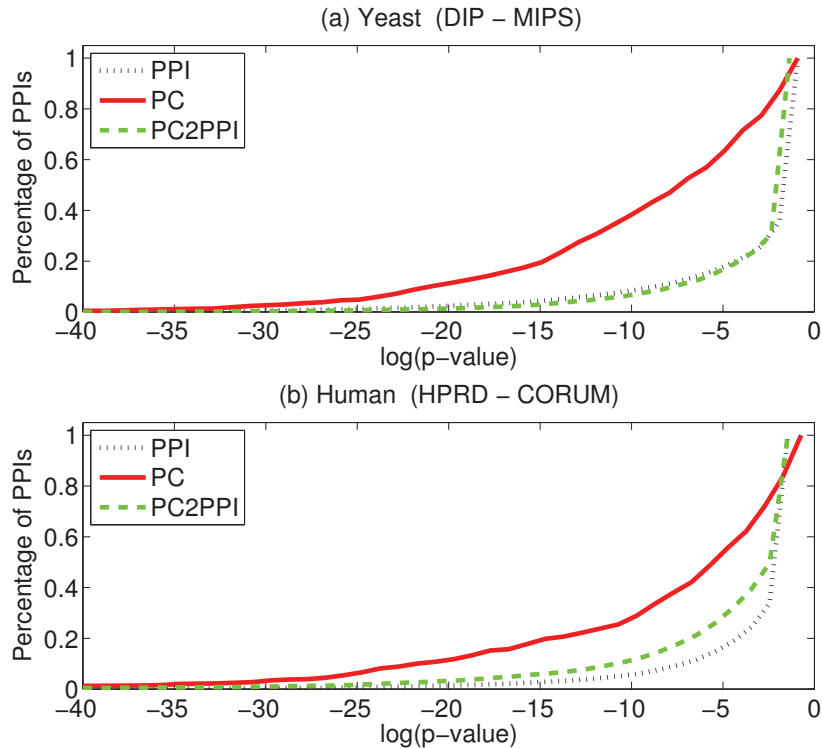
where  $n(H)$  and  $n(P)$  are the number of proteins in complex  $H$  and  $P$ , respectively, and  $k$  is the number of proteins they have in common. The value of *f-score* is defined as

$$f\text{-score} = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot k}{n(H) + n(P)}. \quad (5.8)$$

Generally, the value of the average *f-score* is used to measure the performance of an algorithm by comparing its predictions with a set of gold standard protein complexes. However, it may miss another important information, which is the number of true positive predictions. It is hard to say an algorithm that generate only several correct complexes is better than one that can obtain hundreds of meaningful predictions, even the average *f-score* of the previous one is much higher than the later one. Hence, we compare different algorithms by counting the number of true positive predictions by assigning different *f-score* cutoffs.

Moreover, a predicted protein complex may overlap with more than one gold standard protein complexes and *vice versa*. It is unfair to use the average *f-score* to evaluate different predictions, since several overlapped predictions may match with the same gold standard complex to achieve the best *f-score*. Similar to the evaluation method used in [32], we use the maximum matching ratio (MMR) to evaluate a group of predictions when compared with a set of gold standard protein complexes. To calculate the MMR, a weighted bipartite graph is built, where two sets of vertices represent the predicted and gold standard protein complexes, respectively, and weights of edges represent the *f-scores* between individual complexes. The MMR used in this paper is the average mean of the maximum bipartite matching in the graph.

Since gold standard protein complex datasets are commonly incomplete, a predicted protein complex that does not match with any known protein complexes may belong to a valid but still uncharacterized complex [32]. Hence, it is also important to analyze those unmatched predictions. In this paper, we select 100 predictions that have minimal *f-scores*, and evaluate them by using the functional homogeneity *p-value* and the average GO annotation rate.

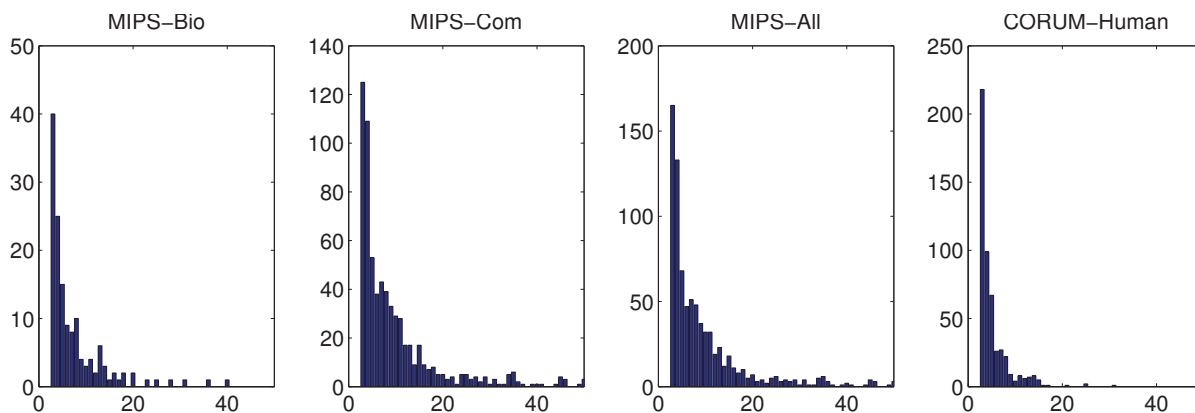


**Figure 5.2:** The cumulative distribution of edge weights in yeast and human PPI networks. In the legend, the PPI represents edges of the entire PPI network, the PC represents edges of within protein complexes and the PC2PPI represents edges from protein complexes to other proteins in the PPI network.

## 5.3 Results and discussions

### 5.3.1 Edge weights in PPI networks

Weights of edges are calculated as (5.4). To test the differences of weights, we draw the cumulative distribution of weights for edges within protein complexes, edges from protein complexes to other proteins and edges in PPI networks. The details are shown in Fig. 5.2. Statistical results indicate that edges within protein complexes are clearly different from other edges in PPI networks. They tend to obtain lower  $p$ -values than others. Another interested information in Fig. 5.2 is that most PPI edges (about 70%) obtain the  $p$ -value larger than  $10^{-3}$  for both yeast and human PPI networks. A good cutoff should be  $10^{-4}$  for keeping most protein complex edges and removing the most noise edges. We use this cutoff in the connected-star algorithm and the following prediction merging method.



**Figure 5.3:** The size distribution of the yeast and human protein complexes.

### 5.3.2 Properties of known protein complexes

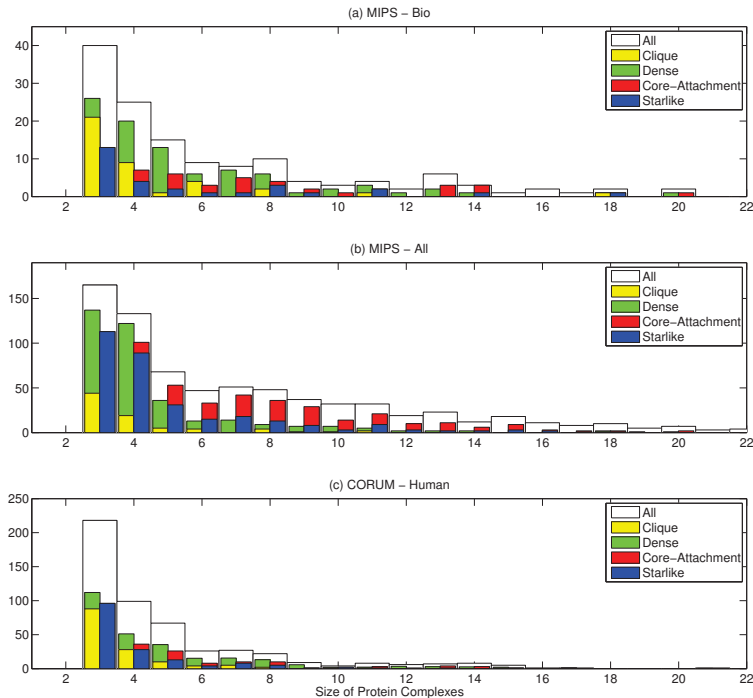
#### The size of protein complexes

The size of a protein complex is the number proteins in the complex. Since properties of protein complexes may vary with their sizes, it is reasonable to analyze different characteristics of known complexes according to their sizes. Fig. 5.3 illustrates the size distribution of three groups of yeast protein complexes in MIPS and one group of human protein complexes in CORUM. We can see from Fig. 5.3 that most protein complexes consist of less than 20 proteins. Hence, we only focus on those protein complexes for analyzing topological informations and GO annotations.

#### Structures of known protein complexes

As we mentioned before, protein complexes exhibit different topological structures in PPI networks. In Fig. 5.4, we draw histograms for protein complexes that display cliques, dense sub-graphs, core-attachment structures and star-like structures in PPI networks. If a protein complex is fully connected, it is regarded as a clique. If the density of a protein complex is large than 0.5, it is regarded as a dense sub-graph. If a sub-graph has some one degree vertices connect to a core part, it is regarded to have the core-attachment structure. If and only if there is only one hub protein interact with all other proteins in a complex, it is regarded as a star-like sub-graph. Obviously, cliques are special cases of dense sub-graphs, while star-like structures are special cases of the core-attachment structures. Therefore, we draw cliques as parts of those dense bars, and star-like structures as parts of those core-attachment bars. The statistic results indicate that dense sub-graphs tend to be dominated for small protein complexes, while core-attachment structures tend





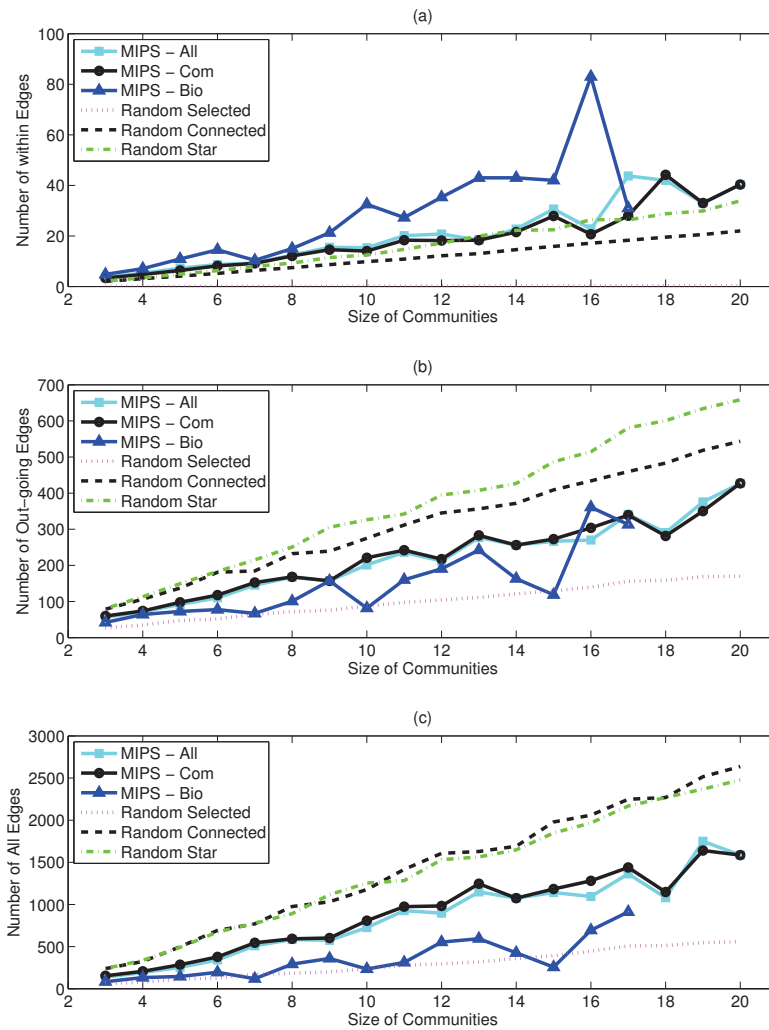
**Figure 5.4:** The structure histogram for the yeast and human protein complexes. (a) The manually curated protein complexes in MIPS. (b) All protein complexes in MIPS, including both biological and computational ones. (c) Human protein complexes in CORUM. The height of individual bars represent the number of protein complexes exhibit individual topological structures.

to take the majority part for those large ones.

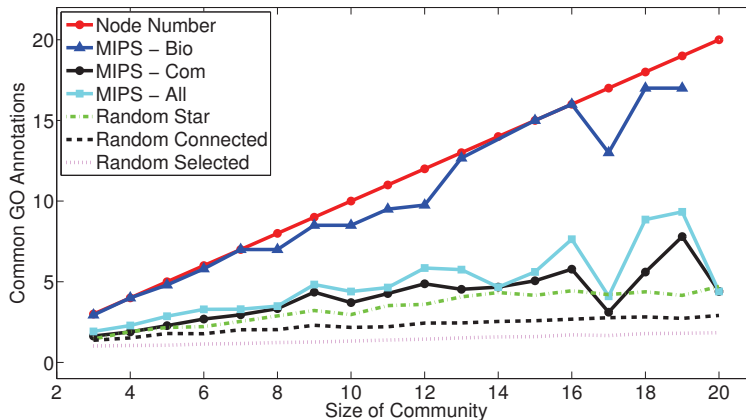
### Topological connectivities of known protein complexes

It is generally acknowledged that protein complexes exhibit specific topological structures in PPI networks. They should have some significant topological differences compared with randomly selected sub-graphs. Although plenty of works have been done in this area, there is still no widely accepted conclusions. In this paper, we simply count the number of within edges, the number of outgoing edges and the number of edges between neighbors for a sub-graph. Those connectivity information of known protein complexes are compared with randomly selected any sub-graphs, randomly selected connected sub-graphs and randomly selected star-like sub-graphs in the PPI network. The results are shown in Fig. 5.5.

As we have claimed in paper [23], protein complexes are relative dense sub-graphs in PPI networks. They tend to have more within edges (see Fig. 5.5(a)) and less outgoing edges (see Fig. 5.5(b)), compared with those random selected connected and star-like sub-graphs. It is reasonable, since it is very common for those randomly selected connected or star-like sub-graphs have edges to bridge different protein complexes, and



**Figure 5.5:** The average number of within edges, the number of outgoing edges and the number of all edges (sum of within edges, outgoing edges and edges within neighbors) for MIPS protein complexes and randomly selected sub-graphs in the PPI network.



**Figure 5.6:** The average  $n_{GO}$  value distribution for MIPS protein complexes. The MIPS-Bio complexes obtain the highest  $n_{GO}$  value, which is close to the number of vertices in the complexes. Although the MIPS-Com and the MIPS-All complexes do not have very high  $n_{GO}$  value, they still tend to be co-annotated in some GO terms, compare with those randomly selected sub-graphs.

thereby bringing more outgoing edges in PPI networks. This property can be used to trim raw predictions of computational algorithms, by checking each prediction’s outgoing edges and within edges between its neighbors as

$$1.5 \cdot \bar{d} \cdot n(H) \leq m_o(H) \leq 2.5 \cdot \bar{d} \cdot n(H), \quad (5.9)$$

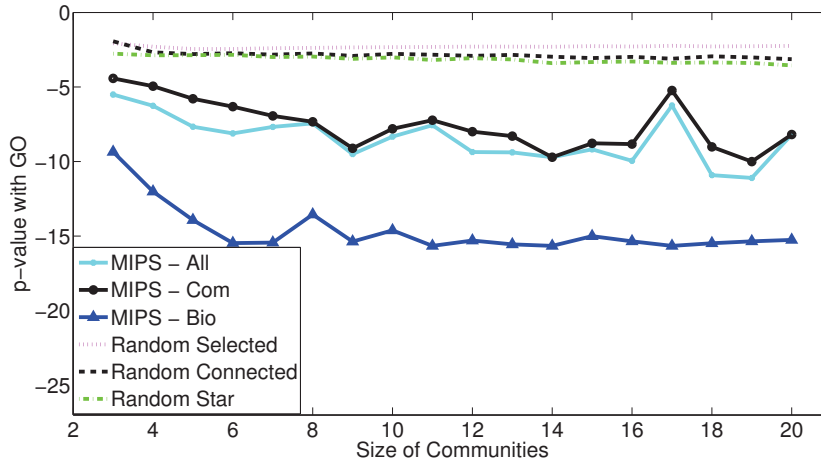
$$1.5 \cdot m_o(H) \leq m_i(H \cup N(H)) \leq 2.5 \cdot m_o(H), \quad (5.10)$$

where  $H$  is a predicted complex,  $\bar{d}$  is the average degree of the PPI network. The upper and lower bounds of (5.9) and (5.10) are empirical values that are obtained from Fig. 5.5.

### Biological properties of known protein complexes

GO annotations are employed to analyze enrichments of protein complexes. As it is shown in Fig. 5.6, known protein complexes tend to be co-annotated under the same GO terms. The average GO annotation rate for *MIPS-Bio* and *MIPS-All* protein complexes are 0.9656 and 0.5399, respectively.

The functional homogeneity  $p$ -value distribution of MIPS protein complexes is plotted in Fig. 5.7. It suggests that known protein complexes obtain much lower  $p$ -values than those randomly selected sub-graphs. The average  $p$ -value of randomly selected sub-graphs is about  $10^{-3}$ , while most known protein complexes obtain their  $p$ -value lower than  $10^{-7}$ . We use this  $p$ -value as another criterion to trim raw predictions. The threshold is selected as  $10^{-6}$ , which is twice the magnitude lower than those randomly selected sub-graphs.



**Figure 5.7:** The functional homogeneity  $p$ -value distribution for MIPS protein complexes. The lower the  $p$ -value is, the more significant a protein complex is enriched in GO annotations.

### 5.3.3 Results of predicted protein complexes

We use the clique enumerating algorithm in [8], the entropy-based algorithm in [9, 18], the core-attachment algorithm in [20] and the connected-star algorithm proposed in this paper to generate raw predictions that exhibit clique, dense, core-attachment and star-like structures, respectively. Those predictions are filtered by topological empirical information (5.9) and (5.10) or the functional homogeneity  $p$ -value, respectively.

Similar predictions are merged subsequently. To do this, the overlap score is calculated as follows:

$$w(A, B) = \frac{|A \cap B|^2}{|A| \cdot |B|},$$

where  $A$  and  $B$  are two predictions. The merging may be performed one after another or concurrently. In this study, a group of predictions are merged concurrently if their overlap score are larger than a threshold. Here, we do not simply merge constituent proteins together to form a large prediction, but merge them similar to generate a core-attachment structure in two steps. First, overlapped proteins of those predictions are detected as the core part, while all others that appear only in one prediction are regarded as candidate attachments. Then, an attachment protein is added only if the edge weight to the core part is less than a threshold ( $10^{-4}$  is used here). We do not merge star-like predictions, since the merged ones will not exhibit star-like structures any more.

The advantage of this method is obvious. The merged predictions are not so large, and noisy proteins are also possible to be eliminated during the merging. However, it needs a threshold to perform such merging. We use 0.33 here, which is able to merge a prediction of three vertices with another prediction of four vertices if they have two proteins in common. The number of predictions and how they are changed on the yeast PPI

**Table 5.2:** The number of predictions of identification methods

Methods	Raw Predictions		After Filter	After Merge
Cliques	7267	TP	426	189
	7267	GO	1348	220
Dense	22049	TP	949	444
	22049	GO	1433	339
CoreAttach	1702	TP	221	214
	1702	GO	494	456
Star-like	990	TP	151	N/A
	990	GO	543	N/A

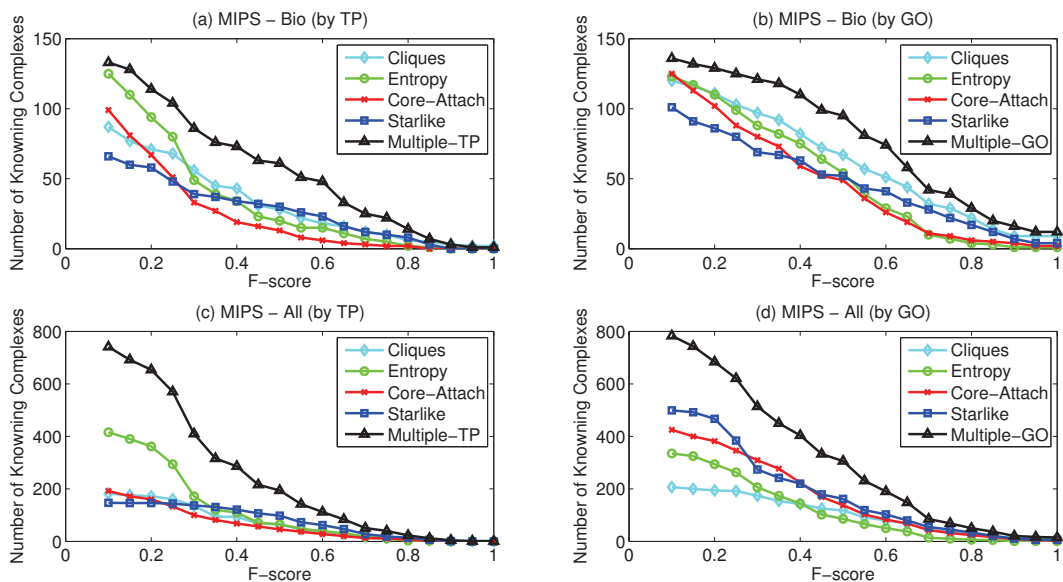
network is shown in Table 5.2.

The multiple-topological-structure-based algorithm takes all final predictions of individual algorithms as input and merges similar predictions just like to merge predictions with the same structure. The only difference is that we use a larger merging threshold as 0.5. Since those predictions exhibit different topological structures, they will only be merged if they overlap significantly.

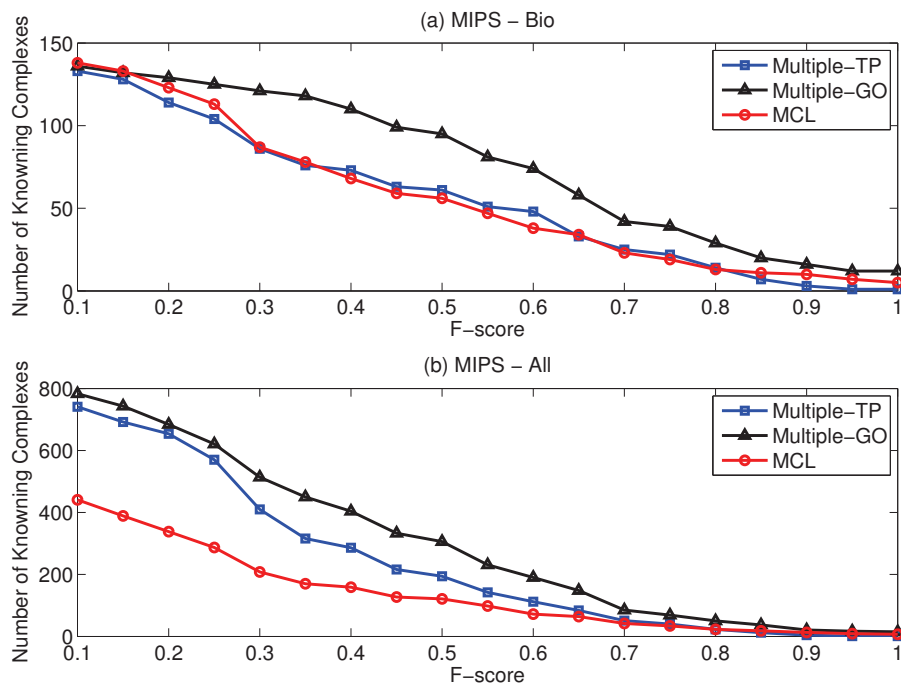
The multiple-topological-structure-based algorithm is denoted as *Multiple-TP* and *Multiple-GO* if topological information or the GO annotations is used to trim raw predictions, respectively. Fig. 5.8 compares differences between the multiple-topological-structure-based algorithm and those single-topological-structure-based algorithms in terms of the true positive prediction numbers on the yeast PPI network. The benchmark datasets are *MIPS-Bio* complexes and *MIPS-All* complexes, respectively. Fig. 5.9 compares differences between the multiple-topological-structure-based algorithm and the MCL algorithm by the same evaluation method. Fig. 5.10 compares such differences on the human PPI network, where CORUM protein complexes are employed as benchmarks. Table 5.3 summarizes the differences between the multiple-topological-structure-based algorithm and the MCL algorithm by using the average MMR *f-score*, the GO annotation rate and the functional homogeneity *p-value* on both datasets.

We can see that the multiple-topological-structure-based algorithm not only generate more true positive predictions, but also predict results with higher accuracy. Besides, the method that trims by GO annotations works better than the one using topological information.

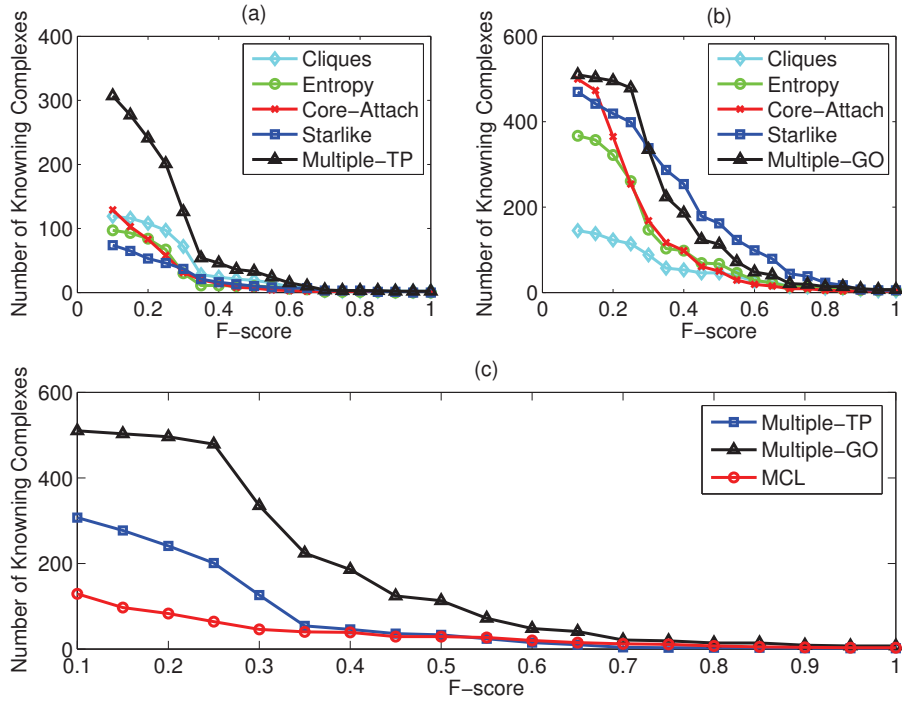
To evaluate predictions that do not match with any known complexes, we selecte the minimal 100 predictions according to their *f-score* compared with *MIPS-Bio* complexes and CORUM complexes for the yeast and human PPI network, respectively. The average GO annotation rate and the functional homogeneity *p-value*



**Figure 5.8:** Comparison of the number of known protein complexes matched by the predicted protein complexes on the yeast PPI network. Fig(a) and (b) use the *MIPS-Bio* complexes as benchmarks, while Fig(c) and (d) use the *MIPS-All* complexes as benchmarks. The predictions of Fig(a) and (c) are filtered by topological informations, whereas Fig(b) and (d) are filtered by GO annotations.



**Figure 5.9:** Comparison of the multiple-topological-structure-based algorithm and the MCL algorithm on the yeast PPI network. Fig(a) use the *MIPS-Bio* complexes as benchmarks, while Fig(b) use the *MIPS-All* complexes as benchmarks.



**Figure 5.10:** Comparison of the multiple-topological-structure-based algorithm, the single-topological-structure-based algorithms and the MCL algorithm. CORUM protein complexes are used as benchmarks. (a) algorithms that filter by topological informations. (b) algorithms that filter by GO annotations. (c) comparison between Multiple-TP, Multiple-GO and the MCL algorithm.

**Table 5.3:** The evaluations of different identification methods

Methods		Multiple-TP	Multiple-GO	MCL
Yeast	Raw Predictions	998	1558	624
	After Merge	790	986	N/A
	MMR (MIPS-Bio)	0.4479	0.5736	0.4478
	MMR (MIPS-All)	0.3433	0.3876	0.3420
	GO annotation rate	0.4797	0.5922	0.5750
	$\log(p\text{-value})$	-5.9395	-9.6022	-4.8791
Human	Raw Predictions	859	3222	443
	After Merge	656	1855	N/A
	MMR (CORUM)	0.2702	0.3147	0.2464
	GO annotation rate	0.4799	0.5267	0.4168
	$\log(p\text{-value})$	-6.1613	-8.8606	-8.0436

**Table 5.4:** The evaluations of unmatched predictions

	Methods	Multiple-TP	Multiple-GO	MCL
Yeast	GO annotation rate	0.4627	0.5596	0.3905
	$\log(p\text{-value})$	-4.1435	-5.9824	-3.9320
Human	GO annotation rate	0.5261	0.6999	0.4666
	$\log(p\text{-value})$	-5.5655	-8.1496	-7.4016

is summarized in Table 5.4. The results for those unmatched or slightly matched predictions can also obtain high GO annotation rates and low  $p$ -values, which indicates that the proposed multiple-topological-structure-based algorithm has very good performance in terms of identifying novel protein complexes.

## 5.4 Conclusions

In this paper, we have proposed a multiple-topological-structure-based algorithm to identify protein complexes with different structures. To test its performance, the algorithm has been applied to a yeast PPI network and a human PPI network. The experimental results have shown that the proposed algorithm can identify more protein complexes, compared with those single-topological-structure-based algorithms. Moreover, the predicted results of the proposed algorithm match with known protein complexes very well. In addition, the proposed algorithm performs better than the MCL algorithm in terms of the number of true positive predictions, the average MMR  $f$ -score and the functional enrichment with GO annotations. The unmatched or slightly matched predictions can also obtain the high GO annotation rate and the functional homogeneity  $p$ -value, which means the proposed algorithm is also promising in terms of identifying novel predictions.

## Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## Declarations

The authors have declared no conflict of interest.



## BIBLIOGRAPHY

- [1] Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868): 141-147.
- [2] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, **403**(6770): 623-627.
- [3] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, **437**(7062): 1173-1178.
- [4] Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006, **7**(11): 120.
- [5] Nesvizhskii AI. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* 2012, **12**(10): 1639-1655.
- [6] Gavin AC, Aloy P, Grandi P, Krause R, Bösch M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Höfert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, **440**(7084): 631-636.
- [7] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A,

- Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, **440**(7084): 637-643.
- [8] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, **100**(21): 12123-12128.
- [9] Kenley EC, Cho YR. Detecting protein complexes and functional modules from protein interaction networks: a graph entropy approach. *Proteomics* 2011, **11**(19): 3835-3844.
- [10] Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegle B, Schmidt T, Doudieu ON, Stümpflen V, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 2008, **36**(Database issue): D646-D650.
- [11] Ruepp A, Waegle B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complex - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501.
- [12] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, **4**: 2.
- [13] King AD, Pržulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, **20**(17): 3013-3020.
- [14] Georgii E, Dietmann S, Uno T, Pagel P, Tsuda K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 2009, **25**(7): 933-940.
- [15] Li XL, Tan SH, Foo CS, Ng SK. Interaction graph mining for protein complexes using local clique merging. *Genome Inform* 2005, **16**(2): 260-269.
- [16] van Dongen S. Graph clustering by flow simulation. *PhD thesis, University of Utrecht*, 2000.
- [17] van Dongen S. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* 2008, **30**(1): 121-141.
- [18] Chen B, Shi J, Zhang S, Wu FX. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics* 2013 **13**(2): 269-277.
- [19] Pang CN, Krycer JR, Lek A, Wilkins MR. Are protein complexes made of cores, modules and attachments? *Proteomics* 2008, **8**(3): 425-434.

- [20] Leung HC, Xiang Q, Yiu SM, Chin FY. Predicting protein complexes from PPI data: a core-attachment approach. *J Comput Biol* 2009, **16**(2): 133-144.
- [21] Wu M, Li X, Kwok CK, Ng SK. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics* 2009, **10**: 169. Jun. 2009.
- [22] Yu L, Gao L, Kong C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* 2011, **11**(19): 3826-3834.
- [23] Chen B, Shi J, Wu FX. Not all protein complexes exhibit dense structures in *S. cerevisiae* PPI network. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, 2012: 470-473.
- [24] Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci USA* 2003, **100**(22): 12579-12583.
- [25] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004, **32**(Database issue): D449-D451.
- [26] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database–2009 update. *Nucleic Acids Res* 2009, **37**(Database issue): D767-D772.
- [27] Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter M, Pagel P, Strack N, Stümpflen V, Warfsmann J, Ruepp A. Human protein reference database–2009 update. *Nucleic Acids Res* 2004, **37**(Database issue): D41-D44.
- [28] Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, Cherry JM Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res* 2008, **36**(Database issue): D577-D581.
- [29] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**(1): 25-29.
- [30] Mete M, Tang F, Xu X, Yuruk N. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics* 2008, **9**:S19.

- [31] Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, Li G, Chen R. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res* 2003, **31**(9): 2443-2450.
- [32] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* 2012, **9**: 471-472.

## CHAPTER 6

# IDENTIFYING DISEASE GENES BY INTEGRATING MULTIPLE DATA SOURCES

*Published as:* Chen B, Wang JX, Li M and Wu FX. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics* 2014, **7**(Suppl 2): S2.

In the previous three chapters, we have introduced three algorithms for identifying protein complexes from PPI networks, which based on densely connected structures, star-like structures, and multiple topological structures. The identification of protein complexes plays an essential role for understanding the organization of proteins within living cells. It is also important for studying the mechanisms of many biological systems.

On the other hand, various kinds of evidence from multiple biomolecular networks analyses have shown that human disease genes are often related to mutations of multiple genes. They are functionally related and tend to lie closely together on different kinds of biomolecular networks. From this chapter, we propose three disease gene identification algorithms by integrating multiple kinds of biomolecular networks.

In this chapter, the disease gene identification problem is first formulated as a two-class classification problem. Then a set of class labels of individual genes is modelled as a Markov random field (MRF), which follows a Gibbs distribution. A global-characteristic-based parameter estimation method and an improved Gibbs sampling process are proposed to generate the final predictions. The method is not only flexible in terms of integrating different kinds of biological data, but also reliable in terms of identifying meaningful disease genes.

### **Abstract**

Now multiple types of data are available for identifying disease genes. Those data include gene-disease associations, disease phenotype similarities, protein-protein interactions, pathways, gene expression profiles,

etc. It is believed that integrating different kinds of biological data is an effective method to identify disease genes. In this paper, we propose a multiple data integration method based on the theory of Markov random field (MRF) and the method of Bayesian analysis for identifying human disease genes. The proposed method is not only flexible in easily incorporating different kinds of data, but also reliable in predicting candidate disease genes. Numerical experiments are carried out by integrating known gene-disease associations, protein complexes, protein-protein interactions, pathways and gene expression profiles. Predictions are evaluated by the leave-one-out method. The proposed method achieves an AUC score of 0.743 when integrating all those biological data in our experiments.

## 6.1 Introduction

Many human genetic diseases or disorders are resulted from mutations of multiple genes [1]. The identification of those disease genes is not only important in understanding genetic disease mechanisms, but is also helpful in developing new methods in diagnostics and therapeutics [2].

Genes associated with similar disorders are often functionally related, supporting the existence of distinct disease-specific functional modules [3–5]. A “guilt-by-association” [6] assumption is often used by various algorithms to identify disease genes. If a gene is ranked as “close” to known disease genes, it would be likely regarded as related to the same disease. The principle is largely supported by many biological data sources, such as protein-protein interactions (PPIs) [7–11], pathways [12–15], gene expression profiles [16–18], etc. Lage et al. [19] rank disease genes from a constructed phenome-interactome network by using PPIs and phenotype similarities. Wu et al. [5] develop a tool called CIPHER to predict disease genes based on a global concordance between a PPI network and a phenotype network. Hwang et al. [20] use a similar coherence score between a gene network and a phenotype network. Vanunu et al. [21] design a method called PRINCE that predicts disease genes and protein complexes associated with diseases at the same time. Li et al. [22] analyze human disease and disease relationships from a pathway-based point of view. Ma et al. [23] employs the Markov random fields (MRF) theory to prioritize genes associated with a specific phenotype or trait by using gene expression profiles and PPI data.

Multiple data integration is another commonly used methodology that collects evidences of gene disease associations from different data sources. Köhler et al. [24] propose a random walk with restart (RWR) algorithm that predicts disease genes by using *a mixed PPI network*. Zhang et al. [25] develop a Bayesian regression approach to explain similarities between disease phenotypes by using diffusion kernels of one or several PPI networks. Chen et al. [26] define a data integration rank (DIR) score by taking a *max* instead of *average* to capture the most informative evidence among a set of integrated data sources. The DIR algorithm

potentially yields better performance than many other data integration methods [26].

However, challenges still exist because of the following reasons. Firstly, there are many levels of controls along paths from genotypes to phenotypes [26]. Genes have to be transcribed and then be translated into proteins, and proteins interact with many other molecules to perform cellular functions [26–28], resulting in the complex relationship between genotypes and phenotypes [29]. Secondly, different biological data are heterogeneous. They describe relationships of molecular entities in various levels. No widely acceptable criterion is available to standardize them into the same scale. An inappropriate integration method combines noise as well, which often decreases the prediction accuracy. Thirdly, many “guilt-by-association” methods only take edges of a candidate gene with known disease genes into account, ignoring edges of the gene with many other vertices in a biological network. They ignore the fact that the biological network, let’s say a PPI network or a gene co-expression network, is built independently for describing a specific biological relationship of proteins or genes. It may have no direct relationship with gene disease associations.

In this paper, we introduce a multiple data integration method for disease gene identifications, which considers comprehensive characteristics of a set of heterogeneous datasets to capture the complex relationship between genotypes and phenotypes. The method is based on the theory of MRF and the method of Bayesian analysis. Two previous algorithms of Deng et al. [30] and Ma et al. [23] have been proposed to integrating multiple datasets by using the MRF theory for yeast protein function predictions. Their method cannot be directly employed to identify human disease genes. Predictions of the method of Deng et al. [30] become unreliable due to the following scale problem. Human genome consists of around 21,000 genes [31], while most diseases are associated by mutations of only a few genes. Even merging similar diseases into classes, the associated genes of individual disease classes is still not enough to estimate parameters correctly by using Deng’s method. The method by Ma et al. [23] mainly uses gene expression profiles to group genes with similar characteristics. PPI data are only employed to calibrate predictions. It is not clear how to integrate more kinds of biological data by using their method. In paper [32], we have developed a basic modified MRF model for human disease gene prioritization. In this study, we will further improve it by introducing a new parameter estimation strategy and a new Gibbs sampling strategy. The improved MRF algorithm is not only stable in terms of parameter estimation, but also reliable in terms of its prediction accuracy.

## 6.2 Methods

In this paper, we first briefly describe how the problem is formulated as a Bayesian labelling problem. The labelling configuration assumes to follow a Gibbs distribution. After that, a MRF model is introduced to solve this problem by integrating multiple kinds of biological data, including known gene-disease associations,

protein complexes, PPIs, pathways and gene expression profiles.

### 6.2.1 The Bayesian labelling problem

Let  $L = \{L_1, L_2, \dots, L_k\}$  be a set of  $k$  labels and  $S = \{S_1, S_2, \dots, S_r\}$  be a set of  $r$  sites. A *labelling problem* [33] is defined as assigning each site  $S_i$  with a label in  $L$ .

Let  $F = \{F_1, F_2, \dots, F_r\}$  be a family of random variables defined on  $S$ , in which each random variable  $F_i$  takes value  $f_i$  of  $L$ . We use the notation  $F = f$  to represent the joint event that  $\{F_1 = f_1, \dots, F_r = f_r\}$ , where  $f = \{f_1, \dots, f_r\}$  is called a *configuration* of  $F$ . The set of all configurations is denoted as  $\mathcal{F}$ .

The relationship of sites is determined by a neighborhood system  $\mathcal{N} = \{N_i \mid \forall i \in S\}$ , where  $N_i$  is the set of sites neighboring  $i$ .

A family of random variables  $F$  is said to be a MRF on  $S$  w.r.t.  $\mathcal{N}$  if and only if the following two conditions are satisfied:

1. Positivity:  $P(f) > 0, \forall f \in \mathcal{F}$ ,
2. Markovianity:  $P(f_i \mid f_{S \setminus \{i\}}) = P(f_i \mid f_{N_i})$ .

The Markovianity indicates that the probability of a local event  $f_i$  conditioned on all other events is equivalent to that conditioned on only events of its neighbors. Hence, the joint probability  $P(f)$  of the random field can be uniquely determined by local conditional probabilities.

Let  $\mathbf{r}$  be an observation of  $F$ . Suppose we know both the prior probability distribution  $P(f)$  of configuration  $f$  and the conditional probability distribution  $P(\mathbf{r} \mid f)$  of the observation  $\mathbf{r}$  given the configuration  $f$ . The best estimation of  $f$  is the one maximizing a posteriori probability (MAP), which is

$$P(f \mid \mathbf{r}) = P(\mathbf{r} \mid f)P(f) / P(\mathbf{r}), \quad (6.1)$$

where  $P(\mathbf{r})$  is the probability that we get the observation  $\mathbf{r}$ .

The *Bayesian labelling problem* [33] is that given a set of observation  $\mathbf{r}$ , find the MAP configuration of labelling  $f^* = \arg \max_{f \in \mathcal{F}} P(f \mid \mathbf{r})$ . Here, as  $P(\mathbf{r})$  is not a function of  $f$ , it does not affect the MAP estimation of  $f$ .



## 6.2.2 Gibbs distribution in MRF

It is usually hard to specify a prior probability of a MRF for a real problem. Fortunately, the Hammersley-Clifford theorem [34] provides a solution for this. According to the theorem,  $F$  is a MRF on  $S$  w.r.t.  $\mathcal{N}$  if and only if the probability distribution of  $P(F = f)$  of the configuration is a Gibbs distribution w.r.t.  $\mathcal{N}$ . The Gibbs distribution has a form of

$$P(f) = Z^{-1} \cdot e^{-U(f)/T}, \quad (6.2)$$

where  $Z = \sum_{f \in \mathcal{F}} e^{-U(f)/T}$  is a normalizing constant,  $T$  is a global control constant that is often assumed to be 1, and  $U(f)$  is the energy function calculated as follows

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) = \sum_{\{i\} \in \mathcal{C}_1} V_1(f_i) + \sum_{\{(i,j)\} \in \mathcal{C}_2} V_2(f_i, f_j) + R_n(f), \quad (6.3)$$

where  $V_i(f)$  is the energy potential of  $C_i$  (the set of  $i^{\text{th}}$  order cliques) in the neighborhood system  $\mathcal{N}$ ,  $R_n(f)$  represents those higher-order terms. A special case of MRF is the Ising model that only considers up to the second-order of cliques [35].

Given a configuration  $f$ , let the conditional probability distribution of observation  $\mathbf{r}$  have the same exponential form

$$P(\mathbf{r}|f) = Z_{\mathbf{r}}^{-1} \cdot e^{-U(\mathbf{r}|f)}. \quad (6.4)$$

Then the posterior probability of the Gibbs distribution has form

$$P(f|\mathbf{r}) = Z_E^{-1} \cdot e^{-U(f|\mathbf{r})}, \quad (6.5)$$

where the posterior energy is [33]

$$U(f|\mathbf{r}) = U(f) + U(f|\mathbf{r}). \quad (6.6)$$

Based on this, suppose the collection of whole human genes  $G = \{g_1, g_2, \dots, g_N\}$  is the site set, and  $\{1, 0\}$  is the label set, where 1 represents a gene is a disease gene and 0 otherwise. The problem of human disease gene identification is actually to find the best configuration of  $G$  according to what is currently known about human diseases.

## 6.2.3 The MRF model for identifying human disease genes

Suppose human genome consists of a set of  $N$  genes  $G = \{g_1, g_2, \dots, g_N\}$ . Some of them are already known to be associated with genetic diseases, while associations of most other genes are still not known. Without loss of generality, let  $g_1, g_2, \dots, g_n$  be genes that have not yet been known to be associated with genetic diseases, and  $g_{n+1}, g_{n+2}, \dots, g_{n+m}$  be currently known disease genes. Obviously, we have  $N = n + m$ . Let

$\{D_1, D_2, \dots, D_M\}$  be a set of human diseases, where  $D_i$  consists of the set of genes that are already known associated with the  $i^{th}$  disease.

For a specific disease, let  $X = (X_1, X_2, \dots, X_{n+m})$  be the random variables defined on all genes, where  $X_i = 1$  represents gene  $g_i$  to be a associated gene of the disease and  $X_i = 0$  otherwise.

Consider those individual genes. Let  $(\pi_1, \pi_2, \dots, \pi_{n+m})$  be a set of probabilities, where  $\pi_i$  represents the probability that  $X_i = 1$ . Let  $x = (x_1, x_2, \dots, x_{n+m})$  be observations of  $X$ . The probability distribution of configuration  $x$  is proportional to

$$\begin{aligned} \prod_{i=1}^{n+m} \pi_i^{x_i} (1 - \pi_i)^{1-x_i} &= \prod_{i=1}^{n+m} \left( \frac{\pi_i}{1 - \pi_i} \right)^{x_i} (1 - \pi_i) = \exp \left[ \sum_{i=1}^{n+m} \alpha_i x_i + \sum_{i=1}^{n+m} \log(1 - \pi_i) \right] \\ &\propto \exp \sum_{i=1}^{n+m} \alpha_i x_i \end{aligned} \quad (6.7)$$

where  $\alpha_i = \log \frac{\pi_i}{1 - \pi_i}$ , and  $\sum_{i=1}^{n+m} \log(1 - \pi_i)$  is a constant.

Next, consider pair-wise relationships between genes. Suppose we have  $K$  biological networks  $H = (H^1, \dots, H^K)$ , where vertices represent genes. Given a  $H^k$ , edges of  $H^k$  represent a specific kind of biological relationship between those genes. Let  $x$  be the observation labels of  $X$ . According to  $x$ , edges of  $H^k$  can be classified into three categories: (1) edges that between two 1-labelled vertices, (2) edges that between a 1-labelled vertex and a 0-labelled vertex, and (3) edges that between two 0-labelled vertices. Let  $N_{11}^k, N_{10}^k$  and  $N_{00}^k$  denote the number of edges in each category of  $G^k$ , respectively. Then

$$N_{11}^k = \sum_{\{(i,j)\} \in E(H^k)} x_i x_j, \quad (6.8)$$

$$N_{10}^k = \sum_{\{(i,j)\} \in E(H^k)} (1 - x_i) x_j + x_i (1 - x_j), \quad (6.9)$$

$$N_{00}^k = \sum_{\{(i,j)\} \in E(H^k)} (1 - x_i)(1 - x_j). \quad (6.10)$$

The probability that we have such a kind of biological network  $H^k$  conditional on those observed labels  $x$  follows as

$$P(H^k | x, \theta^k) \propto e^{\beta^k N_{10}^k + \gamma^k N_{11}^k + \kappa^k N_{00}^k}, \quad (6.11)$$

where  $\theta^k = (\beta^k, \gamma^k, \kappa^k)$  are weights of these three kinds of edges for  $H^k$ . One of three parameters in  $\theta^k$  is redundant. Without loss of generality, let  $\kappa^k = 1$ . Similarly, for  $K$  biological networks, the probability that we observe them conditional on the observed labels follows as

$$P(H^1, \dots, H^K | x, \theta^1, \dots, \theta^K) \propto \sum_{k=1}^K e^{\beta^k N_{10}^k + \gamma^k N_{11}^k + N_{00}^k}. \quad (6.12)$$

Based on the Ising model, the energy function can be written in terms of  $x$  as

$$U(x|\theta) = - \sum_{i=1}^{n+m} \alpha_i x_i - \sum_{k=1}^K (\beta^k N_{10}^k + \gamma^k N_{11}^k + N_{00}^k) \quad (6.13)$$

where  $\theta = (\alpha_i, \beta^1, \gamma^1, \dots, \beta^K, \gamma^K)$  are parameters. In the terminology of MRF [30],  $U(x|\theta)$  defines a Gibbs distribution of the entire networks

$$P(x|\theta) = \frac{1}{Z(\theta)} \times e^{-U(x|\theta)}, \quad (6.14)$$

where  $Z(\theta)$  is the normalized constant that is calculated by summing over all configurations  $\mathcal{X}$ :

$$Z(\theta) = \sum_{x \in \mathcal{X}} e^{-U(x|\theta)}.$$

## 6.2.4 The Gibbs sampling

The Gibbs distribution (6.14) gives a prior probability distribution of the configuration for all genes. In the study of identifying human disease genes, the objective is to find the posterior probability of  $X_1, X_2, \dots, X_n$  conditional on known disease genes

$$P(X_1, X_2, \dots, X_n | X_{n+1}, X_{n+2}, \dots, X_{n+m}).$$

To achieve this, consider the following posterior probability distribution of an individual gene  $X_i$

$$P(X_i = 1 | X_{[-i]}, \theta)$$

where  $X_{[-i]} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+m})$  represents labels of all other genes except  $X_i$ ,  $\theta$  are the parameters. According to the Bayes' theorem [36] and the Gibbs distribution (6.14), we have

$$\begin{aligned} P(X_i = 1 | X_{[-i]}, \theta) &= \frac{P(X_i = 1, X_{[-i]} | \theta)}{P(X_i = 1, X_{[-i]} | \theta) + P(X_i = 0, X_{[-i]} | \theta)} \\ &= \frac{e^{-U(X_i=1, X_{[-i]} | \theta)}}{e^{-U(X_i=1, X_{[-i]} | \theta)} + e^{-U(X_i=0, X_{[-i]} | \theta)}} = \frac{e^{T(i)}}{e^{T(i)} + 1}. \end{aligned} \quad (6.15)$$

where

$$\begin{aligned} U(X_i = 1, X_{[-i]} | \theta) &= U(X_{[-i]} | \theta) - \alpha_i - \sum_{k=1}^K (\beta^k M_0^k - \gamma^k M_1^k), \\ U(X_i = 0, X_{[-i]} | \theta) &= U(X_{[-i]} | \theta) - \sum_{k=1}^K (\beta^k M_1^k - M_0^k), \end{aligned} \quad (6.16)$$

according to equation (6.13), and

$$T(i) = -U(X_i = 1, X_{[-i]} | \theta) + U(X_i = 0, X_{[-i]} | \theta) = \alpha_i + \sum_{k=1}^K [(\beta^k - 1)M_0^k + (\gamma^k - \beta^k)M_1^k]. \quad (6.17)$$

Here  $M_0^k$  and  $M_1^k$  are the number of neighbors of the gene  $g_i$  labelled with 0 and 1 on network  $H^k$ ,  $k = 1, \dots, K$ , respectively.

Equation (6.15) provides a method to update the label  $X_i$  according to all other labels. Suppose parameters  $\theta = (\alpha_i, \beta^1, \gamma^1, \dots, \beta^K, \gamma^K)$  of the model are given, together with prior observed labels of all genes. Using equation (6.15), we can update labels for all unknown genes. Repeating this procedure a number of times until all posterior probabilities of labels are stabilized. This is the essential procedure of the Gibbs sampling.

### 6.2.5 Parameter estimation

In practice, we do not know parameters of the model and they need to be estimated according to those known informations. Ideally, the maximum likelihood estimation (MLE) method is a good choice to estimate  $\theta$  in equation (6.14). However, the normalizing part  $Z(\theta)$  is also a function of  $\theta$ , which is the main difficulty for using the MLE method directly. Deng et al. [30] using a pseudo-likelihood method to estimate parameters in the MRF model. Specifically, the following pseudo-likelihood function is derived from equation (6.15), which is

$$\log \frac{P(X_i = 1 | X_{[-i]}, \theta)}{1 - P(X_i = 1 | X_{[-i]}, \theta)} = T(i) \quad (6.18)$$

The parameter estimation can be done by a *binary logistic regression*, where dependent variables in equation (6.18) are categorical labels and independent variables are  $M_0^1, M_1^1, \dots, M_0^K, M_1^K$  of the  $K$  biological networks. The standard MATLAB function *glmfit()* can be employed to perform such binary logistic regression.

The pseudo-likelihood method used by Deng et al. [30] is valuable. However, there is an important potential problem [32, 37], which may result in unreasonable predictions with their original method. The parameter estimation of Deng et al. [30] is conducted on only known labelled vertices of biological networks. However, a known vertex with labelling 1 may have plenty of unknown vertices with labelling 0 in a biological network and vice versa. A neglect of those unknown vertices may result in inaccurate estimated parameters, which makes predictions problematic. This problem becomes serious with the increasing number of unknown vertices [37]. Kourmpetis et al. [37] alternatively introduce a Bayesian MRF model to estimate parameters and update labels at the same time. An adaptive Markov Chain Monte Carlo (MCMC) algorithm is employed to perform the estimation by using another scaling parameter, a  $Z$  matrix and a multivariate normal distribution.

In this study, we introduce a new method to simultaneously estimate parameters and update labels. Suppose a prior probability of  $\pi_i$  for each unknown vertex is known. A set of prior labels of unknown vertices can be assigned according to this probability. Then the pseudo-likelihood parameter estimation method is performed on all labeled vertices, including those known labelled ones and those unknown prior labelled ones. Using these estimated parameters to update labels for all unknown vertices, and then using the updated labels to

re-estimate parameters until both of them are stable. The step-by-step description of this procedure is given as follows.

1. Initialization:

Let  $t = 0$ , and initialize labels of all vertices  $(X_1^{(0)}, X_2^{(0)}, \dots, X_{n+m}^{(0)})$

2. Estimating parameters:

$$\theta^{(t)} \Leftarrow (X_1^{(t)}, X_2^{(t)}, \dots, X_{n+m}^{(t)}) ;$$

3. Gibbs sampling:

$$X_1^{(t+1)} \Leftarrow (\theta^{(t)}, X_2^{(t)}, \dots, X_{n+m}^{(t)})$$

$$X_2^{(t+1)} \Leftarrow (\theta^{(t)}, X_1^{(t+1)}, X_3^{(t)}, \dots, X_{n+m}^{(t)})$$

$$X_3^{(t+1)} \Leftarrow (\theta^{(t)}, X_1^{(t+1)}, X_2^{(t+1)}, X_4^{(t)}, \dots, X_{n+m}^{(t)})$$

$\vdots$

$$X_n^{(t+1)} \Leftarrow (\theta^{(t)}, X_1^{(t+1)}, \dots, X_{n-1}^{(t+1)}, X_{n+1}^{(t)}, \dots, X_{n+m}^{(t)})$$

$$X_{n+1}^{(t+1)} \Leftarrow X_{n+1}^{(t)}$$

$\vdots$

$$X_{n+m}^{(t+1)} \Leftarrow X_{n+m}^{(t)}$$

4. Let  $t = t + 1$ , and go to 2, until stabilized.

During the Gibbs sampling procedure, a “burn-in period” and a “lag period” often need to be specified. The “burn-in period” is the period that a Markov process takes to become stabilized. Simulation results in this period are discarded to reduce the effect of initial prior probabilities. The “lag period” is the period that needs to reduce the dependence of the Markov process. The posterior probabilities in this period are estimated by averaging simulation results during individual lag steps.

In this study, the “burn-in period” takes 100 steps while the “lag period” takes 90 steps. Simulation results are averaged every 10 steps in the “lag period”. There is 1000 steps in total for simulations. For convenience, predictions made by the original MRF model of Deng et al. [30] is denoted as “MRF-Deng”, while predictions of our improved MRF method is denoted as “IMRF<sub>1</sub>” hereafter. A second improved MRF method is also given in the following by adding a new period at last in simulations, which is called “prediction period”. It takes the average estimated parameters in the “lag period” as parameters and fixes them hereafter in simulations. The input probabilities of unknown vertices are also obtained by the average posterior probabilities in the “lag period”. The Markov process runs another 100 steps in this period. The average posterior probabilities

in the “prediction period” are outputted as final predictions, and predictions of this method is denoted as “IMRF<sub>2</sub>”.

### 6.2.6 Estimating a prior probability

Now, the only problem left is to estimate the prior probability of  $\pi_i$ . Similarly as the method used in Deng et al. [30], we also estimate them according to known protein complexes. Since genes that encode proteins in a same complex tend to associated with similar diseases. For a gene  $g_i$  that encodes protein in a complex, let

$$\hat{\pi}_i = A/B \tag{6.19}$$

be the prior probability, where  $A$  is the number of disease genes for a specific disease in the complex, and  $B$  is the number of all disease genes in the complex. If a gene appears in multiple protein complexes, we use the maximum value as the prior probability for the gene.

For those genes that do not belong to any protein complex, let

$$\hat{\pi}_i = C/D \tag{6.20}$$

as the prior probability, where  $C$  is the number of all currently known disease genes for the specific disease, and  $D$  is the total number of genes in human genome.

### 6.2.7 Data sources

The gene-disease association data are obtained from Goh et al. [3], which contain 1284 disorders and 1777 disease genes. These data are originally collected from the Morbid Map list of the Online Mendelian Inheritance in Man (OMIM) [38]. Disorders are manually classified into 22 primary disease classes, including a ‘multiple’ class and a ‘unclassified’ class. In this study, we consider only those disease classes that consist of at least 30 genes. We also exclude the ‘multiple’ class, the ‘unclassified’ class, the ‘cancer’ class and the ‘neurological’ class due to the class evidence and the class heterogeneity [3]. The final dataset consists of 815 genes in 12 disease classes.

The protein complex data are collected from the database of CORUM [39] and PCDq [40]. There are 1677 and 1103 protein complexes in the dataset that consist of at least two proteins, respectively. There are in total 3881 proteins in those protein complexes.

The PPI datasets are derived from the database of HPRD (Release 9) [9], BioGrid (Release 3.2.108) [10] and IntAct (downloaded on Jan 26, 2014) [11], respectively. Duplicated edges between the same pair of vertices

and edges connecting to itself are deleted. Each dataset is processed independently, and three PPI networks are obtained finally. The HPRD PPI network consists of 9465 vertices and 37039 edges. The BioGrid PPI network consists of 15298 vertices and 127612 edges. The IntAct PPI network consists of 13449 vertices and 63825 edges.

The pathway datasets are obtained from the database of KEGG [12], Reactome [13], PharmGKB [14] and PIN [15], There are 280, 1469, 99 and 2679 pathways in datasets, respectively. There are in total 8614 proteins in those pathways. A pathway co-existence network is constructed by taking individual proteins/genes as vertices. Edges are constructed between two vertices, if they co-exist in any pathway.

The human gene expression profiles are obtained from BioGPS (GSE1133) [16, 17], which contain 79 human tissues in duplicates, measured using the Affymetrix U133A array. Pair-wise Pearson correlation coefficients (PCC) are calculated and a pair of genes are linked by an edge if the PCC value is larger than 0.5, similar to the method used in [3, 26].

Hence, five biological networks are constructed by collecting data from various databases. All protein IDs are mapped onto the form of the gene symbol. In order to test the performance of multiple data integration of our methods, we select those genes that appears at least four times in the five networks. The final datasets consist of 7311 human genes, 815 out of which are known associated with 12 disease classes.

## 6.2.8 Validation method and evaluation criteria

The accuracy of predictions is validated by the leave-one-out method. For each known disease gene with at least one annotated interaction partner in a biological network, we assume it is an unknown gene and predict its posterior probability by our proposed methods. We use the receiver operating characteristic (ROC) curve to show the relationship between the true positive rate and the false positive rate by varying the threshold for declaring positives. The area under the ROC curve (AUC) is also employed to show an overall measure of the performance. The negative control set consists of known disease genes that do not belong to current disease class, and they are also validated by using the leave-one-out method.

## 6.2.9 Decision score and declaration of positives

One can directly use the posterior probabilities obtained by the Gibbs sampling to select candidate disease genes. The greater the probability is for a gene, the more likely it is to associated with specific disease. However, different disease classes consist of different numbers of known disease genes, and thus the prediction results may not be good if a global threshold is used for all classes. Hence, we propose to use a percentage as

a decision score to generate the final predictions. All the ROC curves and the AUC scores of our “IMRF<sub>1</sub>” and “IMRF<sub>2</sub>” method are calculated according to the decision score hereafter.

## 6.3 Results and discussions

We first analyze the performance of the IMRF<sub>1</sub> and IMRF<sub>2</sub> algorithms in terms of stability and reliability, and then compare our method with the original MRF-Deng method [30], the RWR algorithm [24] and the DIR algorithm [26]. These three algorithms are selected elaborately.

Firstly, since ideas of our improved methods (IMRF<sub>1</sub> and IMRF<sub>2</sub>) are initially inspired by the MRF-Deng method, the direct comparison illustrates how much improvement can be made results from our methods.

Secondly, we compare our methods with the RWR algorithm to show which manner of multiple data integration is better. The RWR algorithm is a typical data integration method that uses a mixed network, where vertices and edges of several biological networks are simply merged together, while our methods integrate different networks separately.

Finally, the DIR algorithm has a very good performance among multiple data integration methods, which also integrates different networks separately. It is the same with our methods in terms of the data integration method.

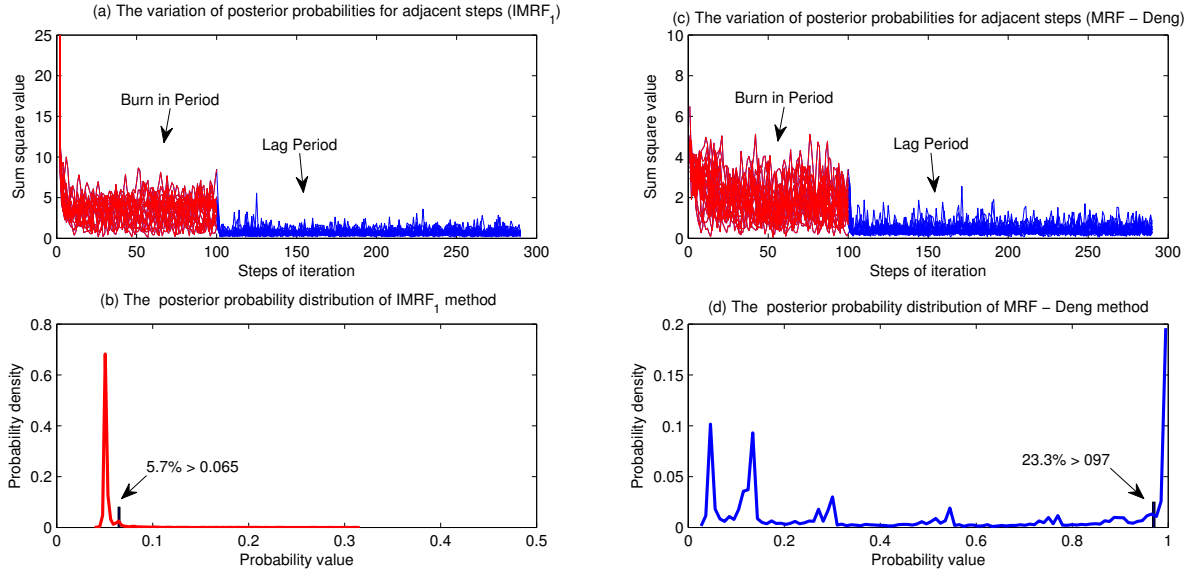
### 6.3.1 Stability and reliability of MRF methods

We first investigate the stability and reliability MRF methods, by analyzing Markov processes of the IMRF<sub>1</sub> method and the MRF-Deng method.

Parameters of the MRF-Deng method are estimated from subnetworks of known vertices. This is feasible to be used for predicting protein functions of yeast in [30], since each function class consists of at least hundreds known vertices, which is possible for estimating reasonable parameters.

However, for disease gene identifications, only dozens of disease genes are available for individual disease classes. The estimated parameters of the MRF-Deng method becomes unreliable. This can be seen by analyzing characteristics of Figure 6.1. In a Gibbs sampling process, it stops until all Markov processes and parameters are stabilized. However, stabilized Markov processes and parameters do not indicate they converge to expected results. It is also stabilized if most vertices are labelled with 1. Take the Figure 6.1 (a) and the Figure 6.1 (c) for example, the variation of posterior probability distributions by using the MRF-





**Figure 6.1:** Analyses of stability and reliability of MRF methods (by using single HPRD PPI network for endocrine disease class). (a) The variation of posterior probabilities for adjacent steps of the IMRF<sub>1</sub> method. (b) The posterior probability distribution of IMRF<sub>1</sub> method. There are only 5.7% of unknown vertices are predicted with probability larger than 0.065, which means only a small number significant vertices are predicted with higher probabilities. (c) The variation of posterior probabilities for adjacent steps of the MRF-Deng method; (d) The posterior probability distribution of MRF-Deng method. There are almost 23.3% of unknown vertices are predicted with probability larger than 0.97, which means too many vertices are predicted with very high probabilities.

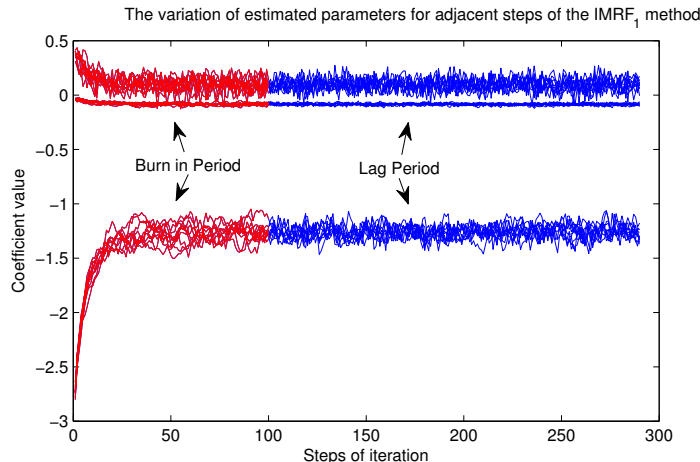
Deng method is smaller than the IMRF<sub>1</sub> method. It seems the performance of the MRF-Deng method is better. However, if we look at Figure 6.1 (b) and Figure 6.1 (d), we find that there are 23.3% vertices with probabilities larger than 0.97. This is commonly unreasonable in practices, since it contains too many false positive predictions. The predictions of the IMRF<sub>1</sub> is reasonable. Most unknown vertices are ranked with a very low probability by using the IMRF<sub>1</sub> method. Only 5.7% unknown vertices are ranked with probabilities larger than 0.065, and only a few significant vertices are predicted with higher probabilities.

Here, the variation of posterior probabilities for two adjacent steps is calculated from

$$Q(t) = \sum_{i=1}^n (P_i(t) - P_i(t-1))^2, \quad (6.21)$$

where  $P_i(t)$  is the posterior probability  $P(X_i = 1 | X_{[-i]}, \theta)$  of  $g_i$  obtained in the  $t^{\text{th}}$  iteration.

Figure 6.2 illustrates the variation of estimated parameters for adjacent steps by using the IMRF<sub>1</sub> method. We can see that all parameters converge very fast, but noise still exists and cannot be reduced by increasing iteration steps. This inspires us to add a “prediction period” for Gibbs sampling processes. The “prediction period” takes the average estimated parameters in the “lag period” as parameters and fixes them hereafter in simulations. The input probabilities of unknown vertices are also obtained by taking the average posterior



**Figure 6.2:** The variation of estimated parameters for adjacent steps by using the IMRF<sub>1</sub> method (by using single HPRD PPI network for endocrine disease class). There are three coefficients in the model. From top to bottom, they are coefficients of  $M_1$ ,  $M_0$  and the constant  $\alpha$ , respectively.

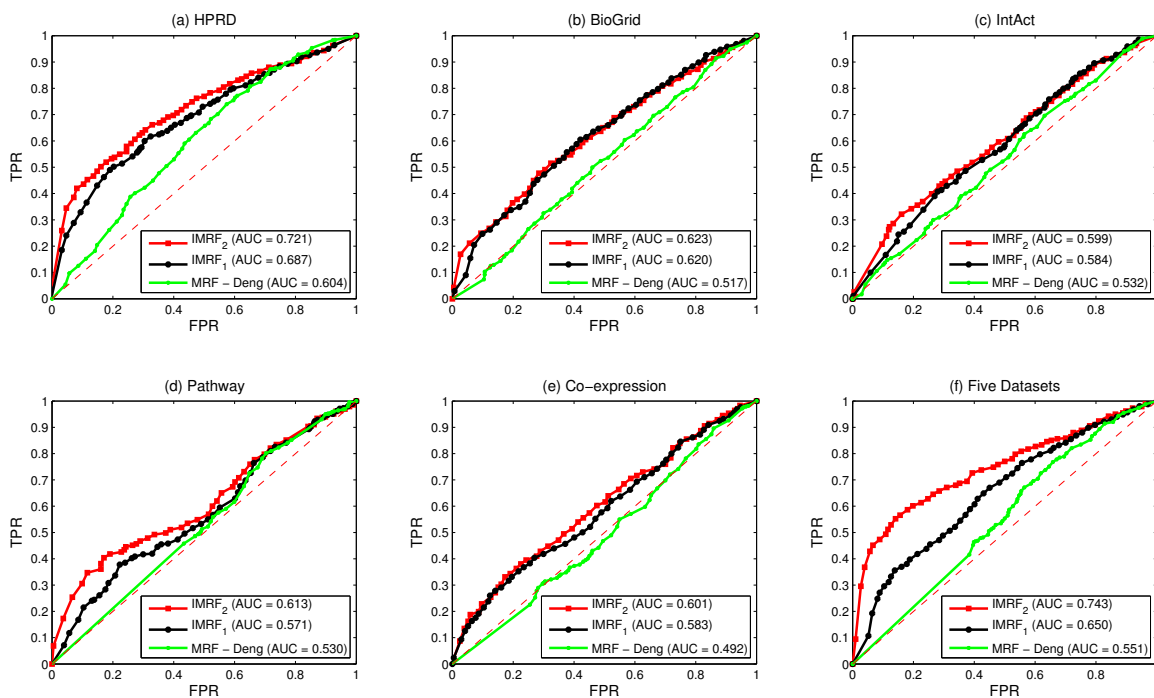
probabilities in the “lag period”.

### 6.3.2 Comparisons with the MRF-Deng method

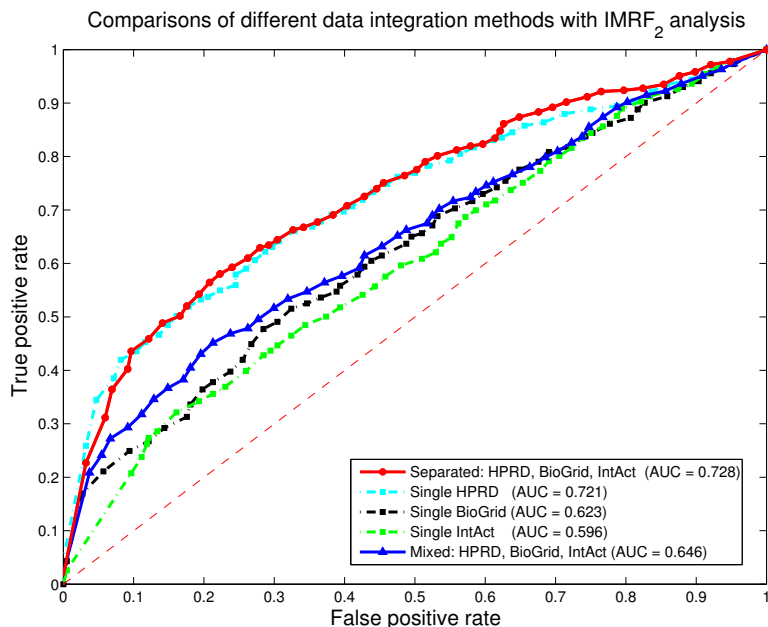
Our improved methods are significantly better than the MRF-Deng method in terms of identifying disease genes. Figure 6.3 illustrates comparisons of the MRF-Deng method, the IMRF<sub>1</sub> method and the IMRF<sub>2</sub> method in terms of ROC curves. Predictions of the IMRF<sub>1</sub> method is significantly better than that by using the MRF-Deng method, but is a little worse than the IMRF<sub>2</sub> method, no matter using single biological network or using integrated biological networks. In terms of informativeness of each biological network, the HPRD PPI network (shows in Figure 6.3 (a)) is the most informative data source, which obtains the highest AUC value in all three methods.

### 6.3.3 Integration of heterogeneous data sources

Different biological datasets are commonly heterogeneous. When information in those data is integrated, noise is also integrated. Hence, an inappropriate method may result in a set of worse predictions than using only single dataset. Generally, various data integration methods can be divided into two categories: (1) by using a mixed network and (2) by using several separated networks. Generally, separated networks contain more information than the mixed network, since it is very easy to generate the mixed network from several separated networks but not vice versa. One advantage of the MRF model is that it takes the whole network



**Figure 6.3:** Comparisons of IMRF<sub>1</sub>, IMRF<sub>2</sub> and MRF-Deng by using five single biological datasets separately and by integrating them together. (a) Comparisons by using single HPRD PPI network. (b) Comparisons by using single BioGrid PPI network. (c) Comparisons by using single IntAct PPI network. (d) Comparisons by using single pathway co-existence network. (e) Comparisons by using single gene co-expression network. (f) Comparisons by integrating the above five networks. The red lines are ROC curves by using the IMRF<sub>2</sub> method. The black lines are ROC curves by using the IMRF<sub>1</sub> method. The green lines are ROC curves by using the IMRF-Deng method. AUC values are listed in parentheses.



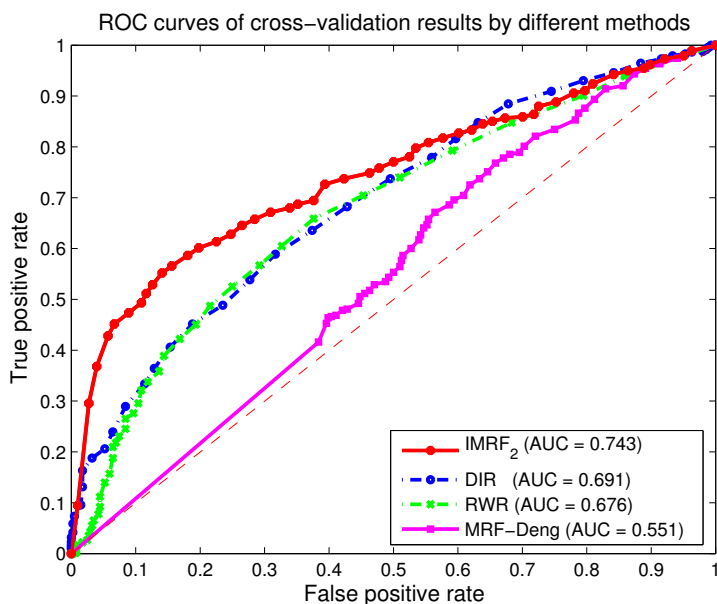
**Figure 6.4:** Comparisons of different data integration methods with IMRF<sub>2</sub> analysis by using three PPI networks. The red solid line represents the ROC curve by integrating three PPI networks. The cyan dash-dot line represents the ROC curve by using single HPRD PPI networks. The black dash-dot line represents the ROC curve by using single BioGrid PPI networks. The green dash-dot line represents the ROC curve by using single IntAct PPI networks. The blue solid line represents the ROC curve by using the mixed PPI network. AUC values are listed in parentheses.

into consideration, which potentially yields better performance than those using mixed network ones.

In Figures 6.4, we use the most stable IMRF<sub>2</sub> method to compare the differences between different kinds of data integration methods. The separated network method achieves the best performance among all predictions, while the mixed network method achieves only modest performance. It seems that the mixed network method combines informations of individual datasets together with their noise, which does not improve its performance by integrating multiple datasets.

### 6.3.4 Comparisons by using multiple data sources

The IMRF<sub>2</sub> method is compared with the RWR algorithm, the DIR algorithm and the MRF-Deng algorithm, respectively. Figure 6.5 illustrates ROC cross-validation results by integrating all five biological networks. The IMRF<sub>2</sub> method achieves the highest AUC score at 0.743, followed by the DIR algorithm (AUC = 0.691) and the RWR algorithm (AUC = 0.676). The MRF-Deng method achieves the AUC score only at 0.551. It also shows that the separated network interaction method performs better than the mixed network RWR



**Figure 6.5:** ROC curves of cross-validation results of different methods by integrating five biological networks. The red solid line represents the ROC curve by using the IMRF<sub>2</sub> method. The blue dash-dot line represents the ROC curve by using the DIR method. The green dash-dot line represents the ROC curve by using the RWR method. The Magenta solid line represents the ROC curve by using the MRF-Deng method. AUC values are listed in parentheses.

method.

## 6.4 Conclusions

In this paper, we have presented an improved multiple data integration method for prioritizing human disease genes, which is based on the theory of MRF and the method of Bayesian analysis. The presented method is both flexible in terms of integrating different kinds of biological data and reliable in terms of prioritizing human disease genes. Compared to the MRF-Deng method [30], two strategies have been developed to significantly improve the performance of the MRF method for disease gene identifications.

Firstly, parameters of our improved MRF methods are estimated according to all labelled vertices in integrated biological networks, instead of estimating them according to only known vertices. Moreover, parameters are updated together with sampling labels during iterations, instead of using fixed parameters. The improved parameter estimation method makes our MRF methods more stable and more reliable.

Secondly, a new “prediction period” is added to Gibbs sampling process. Parameters of this period is obtained

by taking average parameters in the previous “lag period” and is fixed during iterations of this period. The input probability is also obtained by taking average of posterior probabilities in the “lag period”. This strategy significantly improves the prediction accuracy of our method.

Predictions when integrating known gene-disease associations, protein complexes, PPIs, pathways and gene expression profiles achieve the AUC score of 0.743, which is better than the RWR method and the DIR method by using the same datasets.

## **Acknowledgement**

The publication costs for this article were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the National Natural Science Foundation of China under Grant No.61232001 and No. 61370024, and the Program for New Century Excellent Talents in University (NCET-12-0547).

## **Authors' contributions**

FXW and BC initiated this study and designed algorithms and experiments. BC performed the experiments, analyzed the results, and drafted the manuscript. FXW, JXW and ML revised the manuscript. All authors have read and approved the final manuscript.

## **Declarations**

The authors declare that they have no competing interests.

## BIBLIOGRAPHY

- [1] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006, **43**(8): 691-698.
- [2] Sun PG, Gao L, Han S. Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci* 2011, **7**(1): 61-73.
- [3] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA* 2007, **104**(21): 8685-8690.
- [4] Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007, **71**(1): 1-11.
- [5] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol* 2008, **4**: 189.
- [6] Altshuler D, Daly M, Kruglyak L. Guilt by association. *Nat Genet* 2000, **26**(2): 135-137.
- [7] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, **437**(7062): 1173-1178.
- [8] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005, **122**(6): 957-968.
- [9] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database - 2009 update *Nucleic Acids Res* 2009, **37**(Database issue): D767-D772.

- [10] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006, **34**(Database issue): D535-539.
- [11] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct - open source resource for molecular interaction data. *Nucleic Acids Res* 2007, **35**(Database issue): D561-565.
- [12] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, **28**(1): 27-30.
- [13] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007, **8**(3): R39.
- [14] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012, **92**(4): 414-417.
- [15] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. *Nucleic Acids Res* 2009, **37**(Database issue): D674-D679.
- [16] Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009, **10**(11): R130.
- [17] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004, **101**(16): 6062-6067.
- [18] Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. A global map of human gene expression. *Nat Biotechnol* 2010, **28**(4): 322-324.
- [19] Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007, **25**(3): 309-316.
- [20] Hwang T, Zhang W, Xie M, Liu J, Kuang R. Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics* 2011, **27**(19): 2692-2699.
- [21] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010, **6**(1): e1000641.



- [22] Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships *PLoS One* 2009, **4**(2): e4346.
- [23] Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007, **23**(2): 215-221.
- [24] Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008, **82**(4): 949-958.
- [25] Zhang W, Sun F, Jiang R. Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach. *BMC Bioinformatics* 2011, **12**(Suppl 1): S11.
- [26] Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. In silico gene prioritization by integrating multiple data sources. *PLoS One* 2011, **6**(6): e21137.
- [27] Chen B, Shi J, Zhang S, Wu FX. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics* 2013, **13**(2): 269-277.
- [28] Chen B, Wu FX. Identifying protein complexes based on multiple topological structures in PPI networks. *IEEE Trans Nanobioscience* 2013, **12**(3): 165-172.
- [29] Strohmaier R. Maneuvering in the complex path from genotype to phenotype. *Science* 2002, **296**(5568): 701-703.
- [30] Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol* 2004, **11**(2-3): 463-475.
- [31] Bentley DR. The human genome project - an overview. *Med Res Rev* 2000, **20**(3): 189-196.
- [32] Chen B, Wang J, Wu FX. Prioritizing human disease genes by multiple data integration. *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on* 2013: 621.
- [33] Li SZ. Markov random field models in computer vision. In *Proceedings of the European Conference on Computer Vision* 1994: 361-370.
- [34] Besag J. Spatial interaction and the statistical analysis of lattice systems. *J Royal Statist Soc B* 1974, **36**(2): 192-236.
- [35] Kamberova G. Markov random field models: a Bayesian approach to computer vision problems. *Department of Computer & Information Science Technical Reports, University of Pennsylvania* 1992.
- [36] Suess EA, Trumbo BE. Introduction to probability simulation and Gibbs sampling with R. *Springer New York* 2010.

- [37] Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov random field analysis for protein function prediction based on network data. *PLoS One* 2010, **5**(2): e9293.
- [38] McKusick VA. Mendelian Inheritance in man and its online version, OMIM. *Am J Hum Genet* 2007, **80**(4): 588-604.
- [39] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501.
- [40] Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, Kanaya S, Imanishi T. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. *BMC Syst Biol* 2012, **6**(Suppl 2): S7.

## CHAPTER 7

# DISEASE GENE IDENTIFICATION BY USING GRAPH KERNELS AND MARKOV RANDOM FIELDS

*Published as:* Chen B, Li M, Wang JX, Wu FX. Disease gene identification by using graph kernels and Markov random fields. *SCIENCE CHINA Life Sciences* 2014, **57**(11): 1054-1063.

In the previous chapter, a MRF-based algorithm has been proposed to identify disease genes by integrating five types of biomolecular networks. The prediction accuracy is better than many previous algorithms in terms of the AUC score. However, the Markovianity characteristic of the MRF-based algorithm means it only considers direct neighbors in biomolecular networks, ignoring indirect neighbors.

In this chapter, a graph-kernel-based MRF algorithm is proposed by combining advantages of graph kernels and the previously proposed MRF-based algorithm. Three kinds of kernels are designed to formulate the distant relationships among biomolecules by considering global topological characteristics. Pair-wise kernel values are then used to assign weights to multiple biomolecular networks, respectively. After that, a weighted MRF algorithm is developed to identify disease genes by integrating those weighted biomolecular networks.

### **Abstract**

Genes associated with similar diseases are often functionally related. This principle is largely supported by many biological data sources, such as disease phenotype similarities, protein complexes, protein-protein interactions, pathways, gene expression profiles, etc. Integrating multiple types of biological data is an effective method to identify disease genes for many genetic diseases. To capture the gene-disease-associations based on biological networks, a kernel-based MRF method is proposed by combining graph kernels and the Markov random field (MRF) method. In the proposed method, three kinds of kernels are employed to describe the overall relationships of vertices in five biological networks, respectively, and a novel weighted MRF method is developed to integrate those data. In addition, an improved Gibbs sampling procedure

and a novel parameter estimation method are proposed to generate predictions from the kernel-based MRF method. Numerical experiments are carried out by integrating known gene-disease associations, protein complexes, protein-protein interactions, pathways and gene expression profiles. The proposed kernel-based MRF method is evaluated by the leave-one-out cross validation paradigm, achieving an AUC score of 0.771 when integrating all those biological data in our experiments, which indicates that our proposed method is very promising compared with many previous methods.

## 7.1 Introduction

The availability of large-scale biological networks provides an opportunity to comprehensively identify disease genes of many genetic diseases, by synergizing evidences from multiple types of data sources. Various algorithms [1–6] have been developed to identify human disease genes based on the strategy of multiple data integration.

However, challenges still exist due to the following reasons. Firstly, there are many levels of controls along paths from genotypes to phenotypes [7], resulting in the indirect relationship between genotypes and phenotypes [8]. Secondly, different biological data are heterogeneous and describe relationships of molecular entities in various levels. It is not a trivial task to design a good algorithm that combines those data appropriately. Thirdly, many data integration methods simply assume that disease genes of similar diseases exhibit dense clusters in the integrated networks, but ignore the fact that those networks are built independently from the description of gene-disease association relationships.

The Markov random field (MRF) model proposed by Deng et al. [9, 10] for predicting yeast protein functions provides a good framework to integrate multiple biological networks. The issue of protein function prediction is formulated as a Bayesian labelling problem, where the function labels follow a Gibbs distribution. A binary logistic regression is employed to estimate parameters from known observations, and a Gibbs sampling approach is developed to generate final predictions. Advantages of the MRF method include its simplicity, its ability to explore contributions of direct neighbors, and its flexibility to integrate multiple types of datasets.

Although the issue of yeast protein function prediction is similar to the issue of human disease gene identification, the method of Deng et al. [9, 10] cannot be directly employed to identify human disease genes. Parameters of the MRF model cannot be estimated precisely due to the limited observations of human disease genes, which makes predictions of their method unreliable. Kourmpetis et al. [11] then propose a Bayesian MRF method to estimate parameters iteratively together with the update of posterior probabilities of function labels during the Gibbs sampling process. However, their method uses another predefined scaling parameter  $\gamma$ , a  $Z$  matrix and a multivariate normal distribution to perform the estimation, which makes the

method a little complex. Ma et al. [5] propose a Combining Gene expression and protein Interaction data (CGI) method to identify genes responsible for similar phenotypes or traits, motivated from the MRF model of Deng et al. [9, 10]. Similarity metric defined by the diffusion kernel is also compared with those by direct neighbors and shortest paths, where predictions from the diffusion kernel are greatly improved. However, the CGI method mainly uses gene expression profiles to group genes with similar characteristics. Protein interaction data are only employed to calibrate predictions. It is not clear how to integrate other types of biological data by using their method. Lee et al. [12] develop a kernel logistic regression (KLR) method for predicting yeast protein functions by combining advantages of both the MRF model and diffusion kernels. Although its predictive accuracy is higher than the original MRF method of Deng et al. [9, 10], the parameter estimation problem still exists if the KLR method is employed to identify human disease genes. Other forms of MRF methods can be found in [13–15].

Graph kernels, on the other hand, have shown their powers for interpreting complex relationships of vertices in biological networks [16–18]. A kernel-based algorithm often yields better performance than those using direct neighbors or shortest paths under the same condition. In paper [19, 20], we have developed a modified MRF model for human disease gene prioritization. In this study, we further propose a kernel-based MRF algorithm for identifying disease genes from multiple types of data by combining the MRF model and graph kernels. The kernel-based MRF algorithm is different from the methods proposed in [19, 20] in the following four aspects. Firstly, a novel weighted MRF method is developed for incorporating different graph kernels. Secondly, a new parameter estimation method is designed for the kernel-based MRF method based on global characteristics of biological networks. Thirdly, an improved Gibbs sampling strategy is proposed which takes the weight value of neighbors into consideration, rather than simply counting the number of neighbors attributed specific values. Finally, the kernel-based MRF method is extended to integrate multiple types of data sources, such as protein-protein interaction (PPI) networks, pathway co-existence networks and gene co-expression networks. We show that the kernel-based MRF algorithm can significantly improve the accuracy of disease gene identification compared with many existing methods.

## 7.2 Methods and materials

### 7.2.1 Problem statement

Suppose a human genome consists of a set of  $N$  genes  $\{g_1, g_2, \dots, g_N\}$ . Some of them are already known to be associated with  $r$  genetic diseases, while associations of most others are still not known and need to be determined. Let  $\{D_1, D_2, \dots, D_r\}$  be those  $r$  associated diseases. Each  $D_i$  consists of a set of known disease genes of the  $i^{th}$  disease. Hence, the number of all those known disease genes equals to  $m = |D_1 \cup D_2 \cup \dots \cup D_r|$ ,

where  $|*|$  is the cardinality of the set. Without loss of generality, let  $g_{n+1}, g_{n+2}, \dots, g_{n+m}$  be those known disease genes, and  $g_1, g_2, \dots, g_n$  be all others, where  $N = n + m$ .

For a specific disease, let  $x = (x_1, x_2, \dots, x_{n+m})$  be a vector of binary variables (i.e., taking values zero or one) defined on all genes, where  $x_i = 1$  represents gene  $g_i$  to be a disease gene of the disease and  $x_i = 0$  otherwise. The purpose of disease gene identification is to predict values of  $x^{miss} = (x_1, x_2, \dots, x_n)$  from current known values  $x^{obs} = (x_{n+1}, x_{n+2}, \dots, x_{n+m})$ . To achieve this, a vector of random variables  $X = (X_1, X_2, \dots, X_N)$  is defined corresponding to  $x$ , where  $P(X_i = x_i)$  represents the probability that  $X_i = x_i$ . The objective is to find the posterior probability of  $X_1, X_2, \dots, X_n$  conditional on known disease genes

$$P(X_1, X_2, \dots, X_n | X_{n+1}, X_{n+2}, \dots, X_{n+m}). \quad (7.1)$$

## 7.2.2 Markov random field

Let  $G = (V, E)$  be a graph with  $N$  vertices and  $X = (X_1, X_2, \dots, X_N)$  be a vector of random variables defined on  $V$ . The vector  $X$  is said to be a MRF on  $G$  if and only if the following two conditions are satisfied:

1. Positivity:  $P(X_i) > 0, \forall X_i \in \mathcal{X}$ ,
2. Markovianity:  $P(X_i | X_{[-i]}) = P(X_i | X_{N(i)})$ .

where  $\mathcal{X}$  are the set of all possible outcomes of  $X_i$ ,  $X_{[-i]}$  is the collection of random variable  $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$ , and  $X_{N(i)}$  is the collection of all  $X_j$  for  $j \in N(i)$ , where  $N(i)$  is the set of all neighbors of vertex  $i$  in  $G$ . The neighborhood structure of graph  $G$  is denoted as  $\mathcal{N}$ . The Markovianity indicates that the probability of  $X_i$  is conditionally independent of all other  $X_k$  except the value of its neighbors. A joint event  $\{X_1 = x_1, \dots, X_N = x_N\}$ , abbreviated as  $X = x$ , is a realization of  $X$ , where  $x = (x_1, x_2, \dots, x_N)$  is called a *configuration* of  $X$ .

One of the key features that facilitate the practical usage of MRF is its equivalence with the Gibbs random field, which is established by the Hammersley-Clifford theorem [21, 22]. According to the theorem,  $X$  is a MRF on  $V$  w.r.t.  $\mathcal{N}$  if and only if the probability distribution of  $P(X)$  follows a *Gibbs distribution*. The Gibbs probability has a form of

$$P(X = x) = Z^{-1} \cdot e^{-U(x)/T}, \quad (7.2)$$

where  $Z = \sum_{x \in \mathcal{X}} e^{-U(x)/T}$  is a normalizing constant called *partition function*,  $T$  is a global control constant called *temperature*, which is often assumed to be 1 unless otherwise stated, and  $U(x)$  is called the *energy function*, which can be decomposed as a sum over all cliques in  $G$  [23], in the form

$$U(x) = \sum_{c \in \mathcal{C}} V_c(x) = \sum_{\{i\} \in \mathcal{C}_1} V_1(x_i) + \sum_{\{i,j\} \in \mathcal{C}_2} V_2(x_i, x_j) + R_n(x), \quad (7.3)$$

where  $V_i(x)$  is the *clique potential* of  $C_i$  (the set of  $i^{\text{th}}$  order cliques in  $G$ ),  $R_n(x)$  represents those higher-order terms. A special case of MRF is the Ising model that only considers up to the second-order of cliques [24], which is also the same as many previous MRF methods did in [10, 11, 20].

The practical valuation of the Hammersley-Clifford theorem is that it gives a simple way to specify the probability  $P(X)$  by using those clique potentials. Suppose  $X_i$  is a binary random variable (i.e., taking values zero or one). Let  $V_1(x_i) = -\alpha \cdot x_i$ ,  $V_2(1, 1) = -\beta_{11}$ ,  $V_2(1, 0) = V_2(0, 1) = -\beta_{10}$  and  $V_2(0, 0) = -\beta_{00}$ . Let  $N_{11}$ ,  $N_{10}$  and  $N_{00}$  be the number of edges whose two endpoints have both the attribute values of 1, one attribute value of 1 and the other value of 0, and both the attribute values of 0, respectively. Then the energy function (7.3) of the MRF can be written as

$$\begin{aligned} U(x) &= -\alpha \sum_{i \in V} x_i - \beta_{11} N_{11} - \beta_{10} N_{10} - \beta_{00} N_{00} \\ &= -\alpha \sum_{i \in V} x_i - \beta_{11} \sum_{\{i,j\} \in E} x_i x_j - \beta_{10} \sum_{\{i,j\} \in E} [x_i(1-x_j) + (1-x_i)x_j] - \beta_{00} \sum_{\{i,j\} \in E} (1-x_i)(1-x_j). \end{aligned} \quad (7.4)$$

where  $\theta = (\alpha, \beta_{11}, \beta_{10}, \beta_{00})$  are parameters.

To generate predictions from a MRF model, the value of parameter  $\theta$  is necessary, which is generally unknown. The most natural approach to estimate  $\theta$  is through the maximum likelihood estimation (MLE) method. However, the MLE method is often intractable in this situation, since the normalizing partition function  $Z$  is also a function of parameters. Fortunately, the pseudo-likelihood approach and the Gibbs sampling process provide a solution for estimating parameters and generating predictions from a MRF model.

Firstly, for estimating parameters, suppose parameters  $\theta = (\alpha, \beta_{11}, \beta_{10}, \beta_{00})$  of equation (7.4) are given. Then fixing the value of  $X_i$ , the energy function of equation (7.4) can be rewritten as

$$U(X_i = 1, X_{[-i]} | \theta) = U(X_{[-i]} | \theta) - \alpha - \beta_{11} \sum_{j \in N(i)} x_j - \beta_{10} \sum_{j \in N(i)} (1 - x_j), \quad (7.5)$$

and

$$U(X_i = 0, X_{[-i]} | \theta) = U(X_{[-i]} | \theta) - \beta_{10} \sum_{j \in N(i)} x_j - \beta_{00} \sum_{j \in N(i)} (1 - x_j), \quad (7.6)$$

respectively, where  $U(X_{[-i]} | \theta)$  represents the energy contributed by all cliques do not contain vertex  $i$ .

Hence, according to the Bayes' theorem [25] and equation (7.2), (7.5) and (7.6), we have

$$\begin{aligned} P(X_i = 1 | X_{[-i]}, \theta) &= \frac{P(X_i = 1, X_{[-i]} | \theta)}{P(X_i = 1, X_{[-i]} | \theta) + P(X_i = 0, X_{[-i]} | \theta)} \\ &= \frac{e^{-U(X_i=1, X_{[-i]} | \theta)}}{e^{-U(X_i=1, X_{[-i]} | \theta)} + e^{-U(X_i=0, X_{[-i]} | \theta)}}. \end{aligned} \quad (7.7)$$

The log-odds of the probability  $P(X_i = 1 | X_{[-i]}, \theta)$  is

$$\log \frac{P(X_i = 1 | X_{[-i]}, \theta)}{1 - P(X_i = 1 | X_{[-i]}, \theta)} = \alpha + \beta_{11} M_{i1} + \beta_{00} M_{i0}, \quad (7.8)$$

where  $\beta_1 = (\beta_{11} - \beta_{10})$ ,  $\beta_0 = (\beta_{10} - \beta_{00})$ , and  $M_{i1}$ ,  $M_{i0}$  are the number of neighbors of gene  $i$  whose  $x_j$  are attributed with value of 1 and 0 in  $G$ , respectively. Those parameters of the MRF method can be estimated by using the standard MATLAB function `glmfit()`.

Secondly, for the Gibbs sampling process, it is a type of Markov Chain Monte Carlo (MCMC) algorithm. Given a set of probabilities  $X^{(t)}$  at time  $t$ , it iteratively updates the value of  $X$  according to the univariate conditional distribution  $P(X_i = 1|X_{[-i]}, \theta)$  as follows:

$$\begin{aligned}
X_1^{(t+1)} &\Leftarrow P(X_1|X_2^{(t)}, \dots, X_n^{(t)}, X^{obs}, \theta) \\
X_2^{(t+1)} &\Leftarrow P(X_2|X_1^{(t+1)}, X_3^{(t)}, \dots, X_n^{(t)}, X^{obs}, \theta) \\
X_3^{(t+1)} &\Leftarrow P(X_3|X_1^{(t+1)}, X_2^{(t+1)}, X_4^{(t)}, \dots, X_n^{(t)}, X^{obs}, \theta) \\
&\vdots \\
X_n^{(t+1)} &\Leftarrow P(X_n|X_1^{(t+1)}, \dots, X_{n-1}^{(t+1)}, X^{obs}, \theta)
\end{aligned} \tag{7.9}$$

where  $X^{obs} = (X_{n+1}, \dots, X_{n+m})$ . The Gibbs sampling process always uses the most recent values of  $X_i$  to update successive variables. The sequence  $X^{(1)}$ ,  $X^{(2)}$ ,  $X^{(3)}$ ,  $\dots$  clearly forms a Markov chain.

### 7.2.3 Graph kernels

Kernels provide a general framework to represent data in the form of pair-wise similarities. Generally, two mathematical conditions need to be satisfied that make a function  $k$  serving as a kernel: (1) it must be symmetric ( $k(x_i, x_j) = k(x_j, x_i)$ ) and (2) positive semi-definite. Mathematically, for any kernel function  $k$  on a space  $\mathcal{X}$ , there exists a Hilbert space  $\mathcal{H}$  and a mapping  $\phi : \mathcal{X} \rightarrow \mathcal{H}$ , such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, \text{ for any } x_i, x_j \in \mathcal{X}, \tag{7.10}$$

where  $\langle u, v \rangle$  represents the dot product in the Hilbert space between any two points  $u, v \in \mathcal{H}$ .

The definition of a kernel is a critical component of any kernel method, since it defines how an algorithm “see” the data. The graph representation of a biological network is often used to describe local topological relationships, which is often not enough to capture the distant relationships among biomolecules. A graph-kernel-based representation provides a solution for this by considering global topological structures [16, 18].

One of the most commonly used graph kernel of  $G$  is the Laplacian exponential diffusion kernel (LED) [16], where the kernel matrix is defined as

$$K_{LED} = e^{-\lambda L} = \lim_{m \rightarrow \infty} \left( I - \frac{\lambda L}{m} \right)^m = I - \lambda L + \frac{(\lambda L)^2}{2!} - \frac{(\lambda L)^3}{3!} + \dots \tag{7.11}$$

where  $L = D - A$  is the Laplacian matrix of the graph  $G$ ,  $A$  is the adjacency matrix, and  $D$  is a diagonal matrix with the  $i^{th}$  diagonal element  $d(i)$  being the degree of the vertex  $i$  and all off-diagonal elements being



0. The parameter  $\lambda$  controls the magnitude of the diffusion, which is often chosen as a very small number. In this study, we take  $\lambda = 0.04$  as [7] suggested.

A diffusion kernel defines the similarity of biomolecule pairs by considering all pair-wise relationships within a network. However, diffusion kernels between different biological networks may not be comparable when a method needs to integrate multiple data sources. To overcome this problem, Chen et al. [7] propose a measure, called *DKPC*, to normalize pair-wise similarities based on their relative strengths among all similarities within a network. The *DKPC* value between a vertex pair  $i$  and  $j$  is defined as

$$DKPC(i, j) = \frac{|\{(s, t) | K_{st} \geq K_{ij}\}|}{|\{(s, t) | K_{st} \geq 0\}|}, \quad (7.12)$$

where  $K_{ij}$  is a similarity value of vertex pair  $i$  and  $j$  in a kernel matrix. A smaller value of  $DKPC(i, j)$  indicates that two vertices  $i$  and  $j$  are more similar. However, in the  $K_{LED}$  matrix, it uses larger values to represent relationships of vertices are more similar. To be consistent, we use the complementary value  $\overline{DKPC}_{ij} = 1 - DKPC(i, j)$  to represent the normalized similarity between vertex pair  $i$  and  $j$  that is obtained from the *DKPC* method hereafter.

Generally, the above two kernels are strongly related to the degree of individual vertices, where the kernel value between two high degree vertices is significantly different from that between two low degree vertices. To make the strength of individual vertices comparable, we propose a Markov exponential diffusion (MED) kernel in this study by replacing the Laplacian matrix  $L$  in (7.11) with a Markov matrix  $M$ , which consists of nonnegative real numbers with each row and column summing to 1. The MED kernel matrix is defined as follows:

$$K_{MED} = e^{-\lambda M} = \lim_{m \rightarrow \infty} \left( I - \frac{\lambda M}{m} \right)^m = I - \lambda M + \frac{(\lambda M)^2}{2!} - \frac{(\lambda M)^3}{3!} + \dots \quad (7.13)$$

where  $M = (N \cdot I - D + A)/N$  and  $N$  is the number of vertices in the network.

#### 7.2.4 Kernel-based MRF method

Let  $K_{N \times N}$  be a kernel matrix derived from a biological network, where all diagonal elements are set to be zero (since the purpose of disease gene identification is to obtain a set of novel candidate genes according to the knowledge of known disease genes, the similarity metrics between a gene and itself are neglected). Let  $p = (p_1, p_2, \dots, p_N)$  be a vector of probabilities, where  $p_i$  represents the probability of  $X_i = 1$  conditional on all other variables  $X_{[-i]}$  given the parameter  $\theta$ . We propose the kernel-based MRF method in three steps as follows.

Firstly, let  $x = (x_1, x_2, \dots, x_N)$  be a set of configuration obtained from  $p$ . In the KLR method of Lee et al.

[12], the weighted number of neighbors whose  $x_j$  values are attributed with 1 and 0 for gene  $i$  are defined as

$$M_{i1}^w = \sum_{j=1}^N K_{ij} \cdot x_j \text{ and } M_{i0}^w = \sum_{j=1}^N K_{ij} \cdot (1 - x_j), \quad (7.14)$$

respectively, where  $K_{ij}$  is the entry in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of the matrix  $K_{N \times N}$ . It should be noticed that the value of  $M_{i1}^w$  and  $M_{i0}^w$  are highly depended on values of all  $x_j$ , which are randomly generated from those  $p_j$ . To reduce the dependence, the  $x_j$  in equation (7.14) can be replaced directly by using those  $p_j$ . Thus, the improved weighted number of neighbors can be written as

$$M_{i1}^{w'} = \sum_{j=1}^N K_{ij} \cdot p_j \text{ and } M_{i0}^{w'} = \sum_{j=1}^N K_{ij} \cdot (1 - p_j). \quad (7.15)$$

The log-odds of the probability  $P(X_i = 1|X_{[-i]}, \theta)$  in weighted form is then defined as

$$\log \frac{P(X_i = 1|X_{[-i]}, \theta)}{1 - P(X_i = 1|X_{[-i]}, \theta)} = \alpha + \beta_1 M_{i1}^{w'} + \beta_0 M_{i0}^{w'}. \quad (7.16)$$

Secondly, an improved Gibbs sampling method is proposed that can iteratively estimate and update parameters  $\theta$  simultaneously with the change of posterior probabilities. Suppose a prior probability of  $p^{(0)}$  is given for all vertices. A set of prior configuration  $x^{(0)} = (x_1^{(0)}, \dots, x_N^{(0)})$  can be randomly generated according to the prior probability. Then the pseudo-likelihood parameter estimation method can be performed on the whole network, including those known vertices and those unknown vertices, by using the equation (7.16). Once those parameters are obtained in this iteration, the posterior probabilities of each vertex  $p_i$  can then be updated according to equation (7.16) as well. Repeating this process for many times until both of them are stable. The step-by-step description of this Gibbs sampling procedure is given as follows.

### 1. Initialization:

Let  $t = 0$ . Initialize the prior probabilities for unknown vertices  $(p_1^{(0)}, p_2^{(0)}, \dots, p_n^{(0)})$  and known vertices  $p^{obs} = (p_{n+1}, p_{n+2}, \dots, p_{n+m})$ , respectively.

### 2. Parameter estimation:

Assign a configuration of  $x^{(t)} = (x_1^{(t)}, \dots, x_N^{(t)})$  and calculate the values of  $M_{i1}^{w'}$  and  $M_{i0}^{w'}$  according to the value of  $p^{(t)} = (p_1^{(t)}, \dots, p_n^{(t)}, p^{obs})$ . Estimate parameters  $\theta^{(t)}$  based on the equation (7.16).

### 3. Gibbs sampling:

$$\begin{aligned} p_1^{(t+1)} &\Leftarrow P(X_1 = 1 | p_2^{(t)}, \dots, p_n^{(t)}, p^{obs}, \theta^{(t)}) \\ p_2^{(t+1)} &\Leftarrow P(X_2 = 1 | p_1^{(t+1)}, p_3^{(t)}, \dots, p_n^{(t)}, p^{obs}, \theta^{(t)}) \\ &\vdots \\ p_n^{(t+1)} &\Leftarrow P(X_n = 1 | p_1^{(t+1)}, \dots, p_{n-1}^{(t+1)}, p^{obs}, \theta^{(t)}) \end{aligned}$$

4. Let  $t = t + 1$ , and go to 2, until  $t$  is larger than a predefined iteration step.

The details of how this predefined iteration step is set are given in Section 7.2.5. The improved Gibbs sampling procedure above is different from previous MRF methods [10, 13, 23] in two aspects. First, parameters of the improved method are estimated according to the configuration and the posterior probability of the whole network, while most previous MRF methods based on subnetworks that consist of only known vertices. Ignoring majority unknown vertices makes the value of  $M_{i1}$  and  $M_{i0}$  (or  $M_{i1}^{w'}$  and  $M_{i0}^{w'}$  in this study) inaccurate, and then cannot estimate parameters precisely. Predictions become unreliable if using those inaccurate parameters to identify human disease genes. Second, many previous MRF methods estimate parameters only once. Parameters are then fixed during the entire Gibbs sampling process [10, 13, 23]. This is very dangerous if the parameters are not estimated precisely. In our method, parameters are updated iteratively together with the change of all posterior probabilities. The Gibbs sampling process always takes the most updated parameters to estimate posterior probabilities for all unknown vertices, which is expected to generate more reliable predictions.

Finally, the proposed MRF method is extended for incorporating multiple types of biological networks. Suppose there are  $L$  networks  $H = (H^1, \dots, H^L)$ , where vertices represent genes and edges represent specific biological relationship between vertices. The equation (7.16) can be easily extended as

$$\log \frac{P(X_i = 1 | X_{[-i]}, \theta)}{1 - P(X_i = 1 | X_{[-i]}, \theta)} = \alpha + \sum_{l=1}^L [\beta_1^l M_{i1}^{w'l} + \beta_0^l M_{i0}^{w'l}], \quad (7.17)$$

by simply summing the effect of the weighted number of neighbors  $M_{i1}^{w'l}$  and  $M_{i0}^{w'l}$  for gene  $i$  from all  $L$  networks, where  $\theta = (\alpha, \beta_1^1, \beta_0^1, \dots, \beta_1^L, \beta_0^L)$  are parameters. The contribution of a network  $H^l$  can be adjusted through the value of  $\beta_1^l$  and  $\beta_0^l$  accordingly. The improved Gibbs sampling procedure can be easily performed by replacing (7.16) with equation (7.17), when estimating parameters and updating posterior probabilities during the iterations.

## 7.2.5 Experimental design

### Data sources

Known gene-disease associations are collected from the Morbid Map list of the Online Mendelian Inheritance in Man (OMIM) [26]. Goh et al. [27] classify all those diseases into 22 primary disease classes, including a ‘multiple’ class and an ‘unclassified’ class. The dataset consists of 1284 diseases and 1777 disease genes. In this study, we choose those disease classes that consist of at least 30 genes and exclude the ‘multiple’ class, the ‘unclassified’ class, the ‘cancer’ class and the ‘neurological’ class due to the lack of their class evidence and the class heterogeneity [27]. The final dataset consists of 815 genes in 12 disease classes.

Two sets of protein complexes are collected from the database of CORUM [28] and PCDq [29], which contain 1677 and 1103 protein complexes that consist of at least two proteins, respectively. All those protein complexes are employed to assign the prior probabilities for unknown vertices.

Three PPI networks are derived from the database of HPRD (Release 9) [30], BioGrid (Release 3.2.108) [31] and IntAct (downloaded on Jan 26, 2014) [32], respectively. Duplicated edges and loop edges are deleted. The HPRD PPI network consists of 9465 vertices and 37039 edges. The BioGrid PPI network consists of 15298 vertices and 127612 edges. The IntAct PPI network consists of 13449 vertices and 63825 edges. These PPI networks have been widely used to identify protein complexes [33–36] or essential proteins [37–39] and thus can be considered to be reliable data.

Pathway datasets are obtained from the database of KEGG [40], Reactome [41], PharmGKB [42] and PIN [43], which contain 280, 1469, 99 and 2679 pathways, respectively. A pathway co-existence network is constructed by taking individual proteins/genes as vertices. Edges are constructed between two vertices, if they co-exist in any pathway.

A gene co-expression network is constructed by using the dataset of BioGPS (GSE1133) [44, 45]. It contains 79 human tissues in duplicates which are measured by using the Affymetrix U133A array. Pair-wise Pearson correlation coefficients (PCC) are calculated and a pair of genes are linked by an edge if the PCC value is larger than 0.5, similar to the method used in [7, 27].

Overall, five biological networks are constructed and all protein (or gene) IDs are mapped onto the form of the gene symbol. In order to test the performance of multiple data integration of our method, we select those genes that appear at least in four networks. The final datasets consist of 7311 human genes, 815 out of which are known to be associated with 12 disease classes.

### **Estimating a prior probability**

To perform a Gibbs sampling procedure, a set of prior probabilities for all vertices is needed. Generally, the values of those prior probabilities do not have significant effect on the final stabled state of a Markov chain if enough iterations are performed. However, a good prior does help to reduce the time of iterations to achieve the stable state.

For those known disease genes, the prior probability of  $p^{obs} = (p_{n+1}, \dots, p_{n+m})$  can be assigned determinedly according to known gene-disease associations. The value of  $p_j, n+1 \leq j \leq n+m$  equals to 1 or 0, depending on the analyzed disease class and those known gene-disease associations.

For those unknown disease genes, since genes that encode proteins in a same complex tend to associate with similar diseases, we estimate their prior probabilities according to the protein complex information, similarly to the method used in Deng et al. [9, 10]. For a gene  $g_i$  that encodes protein in a complex, let

$$\hat{p}_i = A/B \tag{7.18}$$

be the prior probability, where  $A$  is the number of disease genes for a specific disease in the complex, and  $B$  is the number of all disease genes in the complex. If a gene appears in multiple protein complexes, we use the maximum value as the prior probability for the gene. For those genes that do not belong to any protein complex, let

$$\hat{p}_i = C/D \tag{7.19}$$

as the prior probability, where  $C$  is the number of all currently known disease genes for the specific disease, and  $D$  is the total number of genes in human genome.

### Specifying an iteration loop

During the Gibbs sampling procedure, a “burn-in period” and a “lag period” often need to be specified. The “burn-in period” is the period that a Markov chain takes to become stabilized. Simulation results in this period are discarded to reduce the effect of initial prior probabilities. The “lag period” is the period that needs to reduce the dependence of the Markov process. The posterior probabilities in this period are estimated by averaging simulation results during individual lag steps. In paper [20], we have shown that an additional “prediction period” is helpful to generate more stable and reliable predictions, which is the period used to generate final prediction by averaging all simulation results during this period.

In this study, the “burn-in period” takes 100 steps, the “lag period” takes 90 steps and the “prediction period” takes 100 steps. Simulation results are averaged every 10 steps in the “lag period”. There are 1100 steps in total for simulations.

### Validation method and evaluation criteria

The leave-one-out cross validation paradigm is employed to evaluate the proposed method. For each known disease gene with at least one annotated interaction partner in a biological network, we assume it is an unknown gene and predict its posterior probability by the proposed method. The receiver operating characteristic (ROC) curve [38, 39] is employed as one of the evaluation criteria, which shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) by varying the threshold for declaring positives. The area under the ROC curve (AUC) is also employed to show an overall performance of an

algorithm. The negative control set consists of known disease genes that do not belong to the current disease class, and they are also validated by using the leave-one-out cross validation paradigm. The application of the AUC score as the evaluation criterion is due to the fact that it is widely accepted by most researchers.

We compare the proposed method with four existing algorithms: (1) the random walk with restart (RWR) algorithm proposed by Köhler et al. [46]; (2) the data integration rank (DIR) algorithm proposed by Chen et al. [7]; (3) the original MRF method proposed by Deng et al. [10] (denoted as MRF-Deng hereafter) and (4) our previous improved MRF method for identifying human disease genes [20] (denoted as IMRF hereafter). The RWR algorithm [46] is a typical data integration method that uses a mixed network, where vertices and edges of several biological networks are simply merged together, while our proposed method integrates those networks separately. The comparison between those two algorithms can show which manner of multiple data integration is better. The DIR algorithm [7] has a very good performance in terms of multiple data integration. It also employs diffusion kernels to integrate different networks separately, which yields better performance than many other data integration methods [7]. The comparison with the other two existing MRF methods demonstrates how much improvement can be obtained from the proposed method as well.

### Decision score and declaration of positives

Different disease classes consist of different numbers of known disease genes, and thus the prediction results may not be good if a global threshold is used for all classes. Although one can directly use the posterior probabilities obtained from the Gibbs sampling to select candidate disease genes, we suggest to use a percentage as a decision score to generate the final predictions. Let  $p^{(T)} = (p_1^{(T)}, p_2^{(T)}, \dots, p_n^{(T)})$  be the set of final posterior probabilities for a specific disease class. The decision score  $q_i$  of vertex  $i$  is defined as

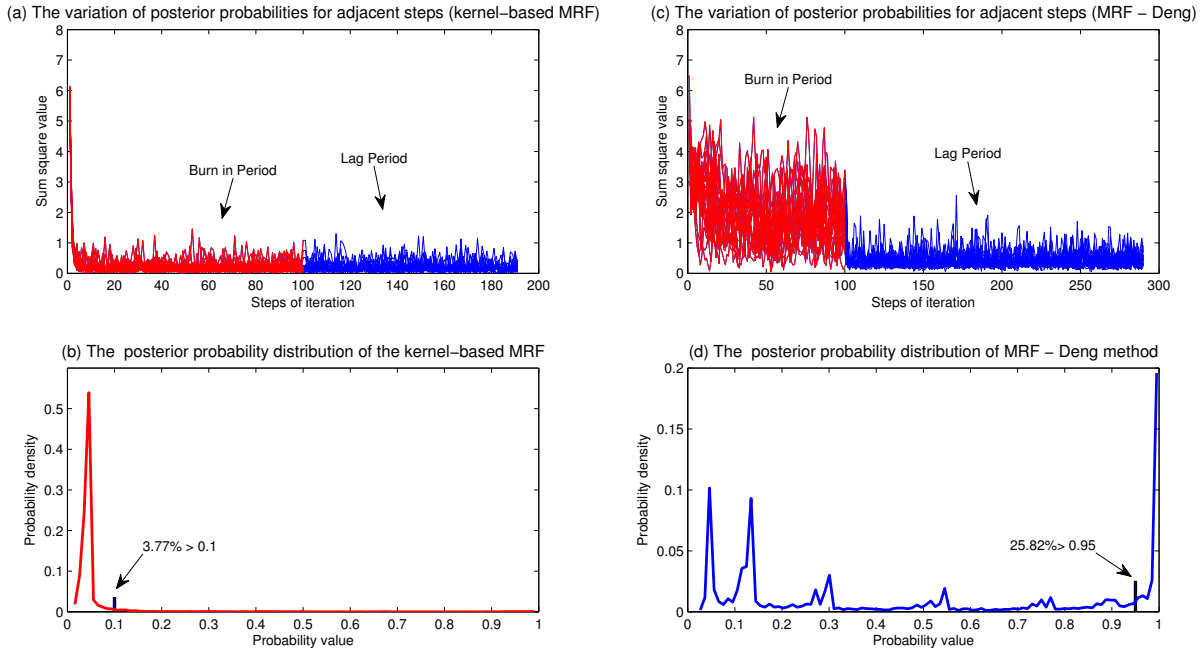
$$q_i = \frac{|\{s | p_i^{(T)} \geq p_s^{(T)}\}|}{n}.$$

The greater the decision score is for a gene, the more likely it is to associate with specific disease. All the ROC curves and the AUC scores of the proposed method are calculated according to the decision score hereafter.

## 7.3 Results

### 7.3.1 Stability and reliability of the kernel-based MRF method

We first investigate the stability and reliability of the kernel-based MRF method, by comparing the Markov processes of the proposed method and the MRF-Deng method. Fig 7.1 illustrates the variation of posterior



**Figure 7.1:** Analyses of stability and reliability of MRF methods (by using single HPRD PPI network for endocrine disease class). (a) The variation of posterior probabilities over iteration steps of the kernel-based MRF method. (b) The posterior probability distribution of the kernel-based MRF method. There are only 3.77% of unknown vertices which are predicted with probability larger than 0.1, which means only a small number significant vertices are predicted with higher probabilities. (c) The variation of posterior probabilities over iteration steps of the MRF-Deng method; (d) The posterior probability distribution of MRF-Deng method. There are almost 25.82% of unknown vertices that are predicted with probability larger than 0.95, which means too many vertices are predicted with very high probabilities.

probabilities over iteration steps and the final posterior probability distribution for the above two methods.

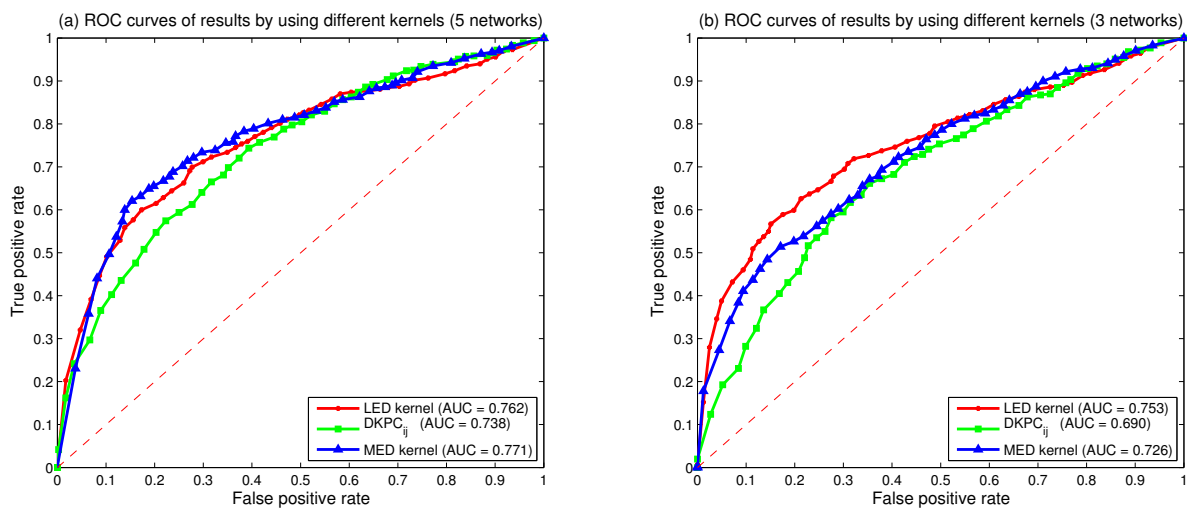
Firstly, by comparing Fig 7.1 (a) and Fig 7.1 (c), we can clearly find that the kernel-based MRF method is more stable than the MRF-Deng method. The change of posterior probability of the front method converges quickly and stays at a stable state.

Here, the variation of posterior probabilities for two consecutive steps is calculated from

$$Q(t) = \sum_{i=1}^n (p_i^{(t)} - p_i^{(t-1)})^2, \quad (7.20)$$

where  $P_i(t)$  is the posterior probability  $P(X_i = 1 | X_{[-i]}, \theta)$  of  $g_i$  obtained in the  $t^{th}$  iteration.

Secondly, predictions of the kernel-based MRF method are more reasonable compared with the MRF-Deng method. The parameters of the MRF-Deng method are estimated from subnetworks of known vertices, which may only be feasible when the subnetwork is enough large for estimating parameter precisely. When



**Figure 7.2:** Comparisons of different kernels by using the kernel-based MRF method. (a) Comparisons of ROC curves by integrated all five biological networks. (b) Comparisons of ROC curves by integrated only three PPI networks. The red lines are ROC curves by using the LED kernels. The green lines are ROC curves by using the  $\overline{DKPC}_{ij}$ . The blue lines are ROC curves by using the MED kernels. AUC values are listed in parentheses.

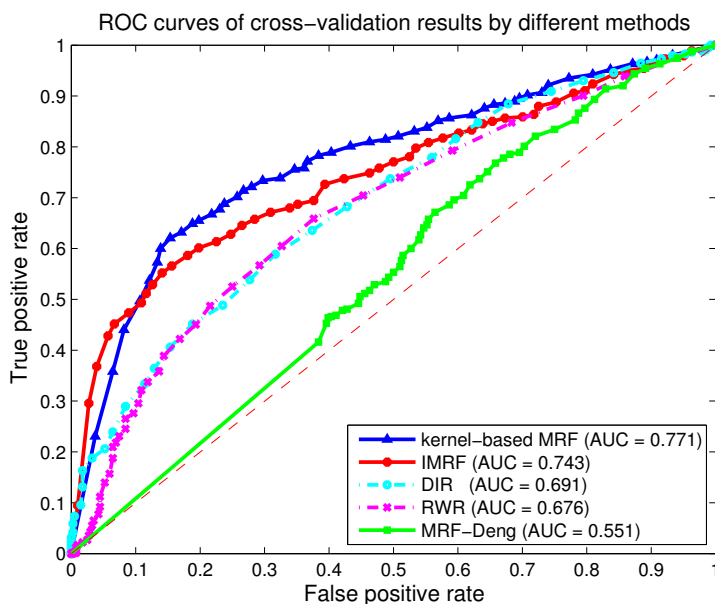
the MRF-Deng method is employed directly to identify human disease genes, there are approximately 25.82% unknown genes that are predicted as disease genes with a probability large than 0.95. This is unreasonably high in practice, since it contains too many false positive predictions. Fig 7.1 (d) shows the final posterior probability distribution of the MRF-Deng method as an example.

On the other hand, the kernel-based MRF method works very well. Taking the endocrine disease class for example, which is illustrated in Fig 7.1 (b), most genes are predicted with a probability small than 0.1. Only a few significant vertices are predicted with higher probabilities. Predictions of the kernel-based MRF method are more reliable than the MRF-Deng method.

### 7.3.2 Comparisons between different kernels

To test the contribution of graph kernels in the kernel-based MRF method, three types of kernels are employed in our experiments. Fig 7.2 illustrates the cross-validation results in terms of ROC curves and the AUC score by integrating only three PPI networks and all five biological networks, respectively. The LED kernel achieves the best performance (AUC = 0.753) when three PPI networks are integrated, while the MED kernel works best (AUC = 0.771) when all five PPI networks are integrated. The similar performance of those kernels also supports the stability of the kernel-based MRF method.





**Figure 7.3:** ROC curves of cross-validation results of different methods with integrating five biological networks. The blue solid line represents the ROC curve by using the kernel-based MRF method. The red solid line represents the ROC curve by using the IMRF method. The cyan dash-dot line represents the ROC curve by using the DIR method. The magenta dash-dot line represents the ROC curve by using the RWR method. The green solid line represents the ROC curve by using the MRF-Deng method. AUC values are listed in parentheses.

Generally, there is no such a kernel that works better than all other kernels in any situation. The LED kernel works better than the MED kernel when three networks are integrated. However, the difference of between those two AUC scores is not large in this situation. Besides, the MED kernel works much better than the LED kernel when five networks are integrated. Hence, the MED kernel is suggested to be used for multiple data integration if no particular information is obtained.

### 7.3.3 Comparisons with previous methods

The kernel-based MRF method is compared with the RWR algorithm, the DIR algorithm, the MRF-Deng algorithm and the IMRF method, respectively. Fig 7.3 illustrates ROC cross-validation results by integrating all five biological networks. It can be seen from the figure that the kernel-based MRF method performs best compared with the other four existing algorithms. The kernel-based MRF method achieves the highest AUC score at 0.771 by using the MED kernel, followed by the IMRF method (AUC = 0.743), the DIR algorithm (AUC = 0.691) and the RWR algorithm (AUC = 0.676). The MRF-Deng method achieves the AUC score only at 0.551.

## 7.4 Conclusions

In this paper, we have presented an improved kernel-based MRF method for identifying human disease genes by integrating five biological networks. The presented method is not only flexible in terms of integrating different types of biological data, but also reliable in terms of identifying human disease genes. Three kinds of graph kernels are employed to capture relationships of all vertices based on their global neighborhood characteristics. An improved Gibbs sampling procedure and a novel parameter estimation method are then developed for the presented MRF method. The use of different kernels brings great improvement for the previous MRF method. The proposed MED kernel works similar to the most commonly used LED kernel when three PPI networks are integrated, and it works best when five biological networks are integrated. Hence, the MED kernel is suggested to be used for the proposed algorithm when multiple data integration is involved to predict disease genes. Predictions by our presented method with integrating all five biological networks achieve the AUC score of 0.771 when the MED kernel is employed, which is very promising for identifying human disease genes.

## Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and it was also supported in part by the National Natural Science Foundation of China under Grant Nos. 61428209, 61232001.

## Author's contributions

FXW and BC initiated this study and designed algorithms and experiments. BC performed the experiments, analyzed the results, and drafted the manuscript. FXW, ML and JXW revised the manuscript. All authors have read and approved the final manuscript.

## Declarations

The authors declare that they have no competing interests.

## BIBLIOGRAPHY

- [1] Hwang T, Zhang W, Xie M, Liu J, Kuang R. Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics* 2011, **27**(19): 2692-2699.
- [2] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010, **6**(1): e1000641.
- [3] Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships *PLoS One* 2009, **4**(2): e4346.
- [4] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol* 2008, **4**:189.
- [5] Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007, **23**(2): 215-221.
- [6] Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007, **25**(3): 309-316.
- [7] Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. In silico gene prioritization by integrating multiple data sources. *PLoS One* 2011, **6**(6): e21137.
- [8] Strohman R. Maneuvering in the complex path from genotype to phenotype. *Science* 2002, **296**(5568): 701-703.
- [9] Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. *J Comput Biol* 2003, **10**(6): 947-960.
- [10] Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol* 2004, **11**(2-3): 463-475.
- [11] Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov random field analysis for protein function prediction based on network data. *PLoS One* 2010, **5**(2): e9293.
- [12] Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS* 2006, **10**(1): 40-55.

- [13] Deng M, Tu Z, Sun F, Chen T. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics* 2004, **20**(6): 895-902.
- [14] Letovsky S, Kasif S: Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 2003, **19**(Suppl. 1): i197-i204.
- [15] Wei Z, Li H. A Markov random field model for network-based analysis of genomic data. *Bioinformatics* 2007, **23**(12): 1537-1544.
- [16] kondor RI, Lafferty J. Diffusion kernels on graphs and other discrete input spaces *Machine Learning: Proceedings of the Nineteenth International Conference* 2002, In: Sammut C, Hoffmann AG (eds), San Mateo, CA: Morgan Kaufmann Publishers Inc,: 315-322.
- [17] Ma X, Chen T, Sun F. Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Brief Bioinform* 2014, **15**(5): 685-698.
- [18] Schölkopf B, Tsuda K, Vert JP. Kernel methods in computational biology. *The MIT press* 2004.
- [19] Chen B, Wang J, Wu FX. Prioritizing human disease genes by multiple data integration. *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on* 2013, 621.
- [20] Chen B, Wang J, Li M, Wu FX. Identifying disease genes by integrating multiple data sources. *BMC Genomics* 2014, **7**(Suppl 2): S2.
- [21] Li SZ. Markov random field modeling in image analysis. *third ed., Springer* 2009.
- [22] Besag J. Spatial interaction and the statistical analysis of lattice systems. *J Royal Statist Soc B* 1974, **36**(2): 192-236.
- [23] Kolaczyk ED. Statistical analysis of network data. *Springer* 2009.
- [24] Kamberova G. Markov random field models: a Bayesian approach to computer vision problems. *Department of Computer & Information Science Technical Reports, University of Pennsylvania* 1992.
- [25] Suess EA, Trumbo BE. Introduction to probability simulation and Gibbs sampling with R. *Springer New York* 2010.
- [26] McKusick VA. Mendelian Inheritance in man and its online version, OMIM. *Am J Hum Genet* 2007, **80**(4): 588-604.
- [27] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA* 2007, **104**(21): 8685-8690.
- [28] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501.

- [29] Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, Kanaya S, Imanishi T. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. *BMC Syst Biol* 2012, **6**(Suppl 2): S7.
- [30] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database - 2009 update *Nucleic Acids Res* 2009, **37**(Database issue): D767-D772.
- [31] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006, **34**(Database issue): D535-539.
- [32] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct - open source resource for molecular interaction data. *Nucleic Acids Res* 2007, **35**(Database issue): D561-565.
- [33] Zhao B, Wang J, Li M, Wu, FX, Pan, Y. Detecting protein complexes based on uncertain graph model. *IEEE/ACM Trans Comput Biol Bioinform* 2014, **11**: 486-497.
- [34] Wang J, Li M, Chen J, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**: 607-620.
- [35] Li M, Wu X, Wang J, Pan Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC Bioinformatics* 2012, **13**: 109.
- [36] Li M, Chen J, Wang J, Hu B, Chen G. Modifying the DPCLus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics* 2008, **9**: 398.
- [37] Wang J, Li M, Wang H, Pan, Y. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9**: 1070-1080.
- [38] Li M, Zheng R, Zhang H, Wang J, Pan Y. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. *Methods* 2014, **67**: 325-333.
- [39] Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans Comput Biol Bioinform* 2014, **11**: 407-418.
- [40] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, **28**(1): 27-30.

- [41] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007, **8**(3): R39.
- [42] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012, **92**(4): 414-417.
- [43] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. *Nucleic Acids Res* 2009, **37**(Database issue): D674-D679.
- [44] Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009, **10**(11): R130.
- [45] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004, **101**(16): 6062-6067.
- [46] Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008, **82**(4): 949-958.

## CHAPTER 8

# A FAST AND HIGH PERFORMANCE ALGORITHM FOR IDENTIFYING HUMAN DISEASE GENES

*Prepared as:* Chen B, Li M, Wang JX and Wu FX. A fast and high performance algorithm for identifying human disease genes (unpublished). The work is an extension of our conference paper: Chen B, Li M, Wang JX and Wu FX. A logistic-regression-based algorithm for identifying human disease genes. *Bioinformatics and Biomedicine (BIBM)*, 2014 *IEEE International Conference on:* 197-200.

In the previous chapter, a kernel-based MRF algorithm has been proposed to identify disease genes based on global topological characteristics of multiple biomolecular networks. Its prediction performance has been improved compared with MRF-based algorithms in terms of the AUC score. However, the MRF-based algorithms have to maintain Markov chains for unknown genes in biomolecular networks, which are very time-consuming.

In this chapter, to generalize the feature construction idea of the MRF-based algorithms, we directly formulate the disease gene identification issue as a binary logistic regression model and propose a fast and high performance logistic-regression-based algorithm. Numerical experiments show that the proposed algorithm not only generates predictions with high AUC score, but also runs very fast.

### Abstract

The identification of human disease genes is the primary step towards the understanding of genetic disease mechanisms. Although various algorithms have been proposed to identify disease genes, most of those algorithms either have poor prediction performance or are very time-consuming. In this study, we propose a fast and high performance disease gene identification algorithm based on logistic regression and multiple data integration. The issue of disease gene identification is first formulated as a two-class classification problem, where one class represents disease genes while the other class represents non-disease genes. A

logistic regression model is employed to calculate the posterior probability of each gene being a disease gene. Two kinds of prior probability estimation methods and three kinds of feature vector construction methods are developed to test the performance of the proposed algorithm. Numerical experiments show that the proposed algorithm outperforms existing methods such as MRF-based and RWR-based methods in terms of the AUC score and the running time.

## 8.1 Introduction

Many studies have shown that genes associated with the same or similar diseases tend to lie close to one another in various biomolecular networks [1–4]. Based on this principle, identifying human disease genes often requires integrating different kinds of biological data to capture complex relationships between disease genes and human genetic diseases. Oti et al. [3] use several sets of protein-protein interaction (PPI) data to predict disease genes. They argue that the use of PPI data can greatly increase the prediction performance for disease gene identifications. Fraser et al. [5] investigate both yeast and human functional genomic data and argue that protein complexes contain valuable information which is helpful for detecting disease genes. Li et al. [6] investigate genetic diseases from a pathway-based point of view. They find that individual pathways often enrich genes related to the same or similar diseases. Ma et al. [7] propose a method combining gene expression and protein interaction (CGI) information to prioritize genes associated with a specific phenotype or trait. Gene expression data are first employed to calculate association scores between each phenotype and individual genes. PPI data are then integrated to calibrate those association scores.

Besides various kinds of integrated datasets, disease gene identification algorithms have also been proposed by using different computational techniques. Lage et al. [8] build a phenome-interactome network by integrating PPI data, predicted protein complex data and phenotype similarity data. Wu et al. [4] use a linear regression method to calculate the concordance score between a PPI network and a phenotype network. A tool called CIPHER is developed to predict disease genes based on those concordance scores. Vanunu et al. [9] formulate a smoothness-related prioritization function in a PPI network that predicts not only disease genes but also disease-associated protein complexes. Zhang et al. [10] develop a Bayesian regression approach to explain similarities of disease phenotypes by using diffusion kernels of one or several PPI networks.

However, those previous algorithms often either have poor prediction performance or are very time-consuming. To improve the prediction performance, Köhler et al. [11] propose a random walk with restart (RWR) algorithm to prioritize disease genes by using a global network distance measure and random walk analysis. The RWR algorithm runs fast. Its prediction performance is much better than many previous methods which are based on local distance measures. However, when integrating multiple kinds of biological networks, the



RWR algorithm simply merges them into a mixed network. Although this strategy can integrate useful information from different data sources, it integrates noise from them as well. Predictions of the RWR algorithm from a mixed network do not always perform better than those from individual networks. To improve the data integration method, Chen et al. [12] define a data integration rank (DIR) score to select the most informative evidence among a set of data sources, which yields better performance than many data integration methods. However, the DIR algorithm is time-consuming due to the calculation of a normalized similarity measure for each gene pair that is calculated by comparing the weight of this gene pair with all other gene pairs'. We also propose an improved Markov random field (MRF) method [13, 14] to identifying human disease genes. The MRF method has better prediction performance than both the RWR algorithm and the DIR algorithm, but it spends more time than them under the same computing conditions. In our numerical experiments on the computer with specification in the subsection "Comparing with previous algorithms" of this paper, the MRF method takes 32.4 hours when one PPI network is used, and it increases to 92.7 hours when three biological networks are integrated.

In this paper, we propose a fast and high performance logistic-regression-based algorithm to further improving the MRF method. The MRF method [13, 14] formulates the problem of disease gene identification as a two-class classification problem, where one class represents disease genes and the other class represents non-disease genes. The idea of Bayesian inference is employed, where posterior probabilities of individual genes as disease genes are calculated according to a set of prior labels, feature vectors and logistic regressions. However, the Markovianity characteristic means that the MRF method can only consider direct neighbors to construct feature vectors, ignoring the contribution of indirect neighbors and other topological characteristics [15, 16]. In addition, the process of Gibbs sampling in the MRF method maintains Markov chains for every gene, which is very time-consuming. To overcome those drawbacks, we directly formulate the problem of disease gene identification as a binary logistic regression issue in this study. The proposed logistic-regression-based method is not only flexible in terms of feature vector constructions, but also very simple. Many topological attributes, such as direct neighbors, second-order neighbors, third- or higher-order neighbors, clustering coefficients, etc. can be used to construct feature vectors. It is a generalization of the previous MRF method. When integrating multiple kinds of biological datasets, the proposed method only needs to include features from different datasets into the feature vectors. The parameter estimation process of the proposed method tunes weights (parameters) of different data sources automatically according to a set of prior information of disease genes. By using Bayesian inference, each unknown gene is interpreted with a posterior probability that indicates its probability being a disease gene. The numerical experiments show that our proposed logistic-regression-based algorithm not only has high AUC score, but also runs very fast, and thus outperforms many previously published methods for identifying human disease genes.

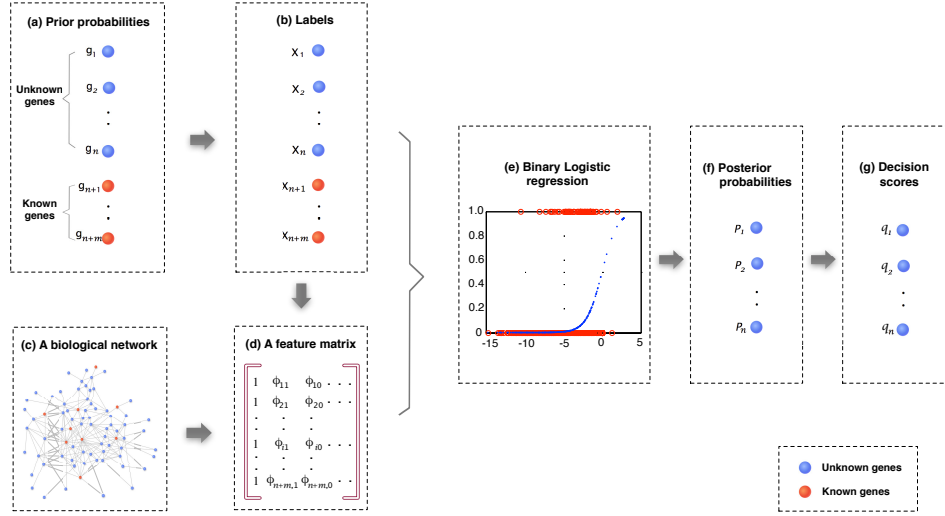
## 8.2 Methods and materials

### 8.2.1 Problem formulation

Suppose the human genome consists of a set of  $N$  genes  $\{g_1, g_2, \dots, g_N\}$ . Some of them are already known to be associated with  $r$  genetic diseases, while associations of most others are still not known and need to be determined. Without loss of generality, let  $g_{n+1}, g_{n+2}, \dots, g_{n+m}$  be those known disease genes, and  $g_1, g_2, \dots, g_n$  be all others, where  $N = n + m$ . Let  $\{D_1, D_2, \dots, D_r\}$  be the set of those  $r$  associated diseases, where each  $D_i$  consists of a set of known genes associated with the  $i^{th}$  disease. The number of all known disease genes equals to  $m = |D_1 \cup D_2 \cup \dots \cup D_r|$ , where  $|*|$  represents the cardinality of the set  $*$ .

For a specific disease, let  $x = (x_1, x_2, \dots, x_{n+m})$  be a vector of binary class labels (i.e. taking the value zero or one) defined on all genes, where  $x_i = 1$  represents gene  $g_i$  being a disease gene, and  $x_i = 0$  otherwise. Given known labels of  $x_{n+1}, x_{n+2}, \dots, x_{n+m}$ , the problem of disease gene identification is equivalent to finding labels of all those unknown genes  $x_1, x_2, \dots, x_n$ . The simplest method to achieve this is to find a *discriminant function* that directly assigns a specific label to each gene. However, it is not easy to find a good discriminant function that can classify individual genes with a high prediction performance. A more powerful method, alternatively, is to model a conditional probability  $p(x_i = 1 | \hat{x})$  for each gene first in an inference stage, and then use this probability to make an optimal decision subsequently in a decision stage [17]. Here  $\hat{x}$  is the vector of prior labels of  $x$ .

In this paper, we propose a logistic-regression-based algorithm to model the conditional probability of  $p(x_i = 1 | \hat{x})$  in the inference stage. A decision score is then calculated for each vertex based on the percentage of the inferred posterior probability and the number of disease genes in each disease class. The flow diagram of the proposed algorithm is depicted in Figure 8.1. To be more specific, a set of prior probabilities is first assigned to individual genes according to a predefined prior and known gene-disease associations. A vector of class labels is then assigned based on the prior probabilities. The connection relationship of those labelled vertices (genes) is reflected by the integrated biological network(s), and feature vectors of individual vertices can then be constructed according to the network(s). All feature vectors form a feature matrix. A binary logistic regression is conducted by taking class labels as categorical dependent variables and individual features as predictor variables. Then, a set of posterior probabilities can be obtained that represents the conditional probability of individual genes with label 1. Finally, in the decision stage, those posterior probabilities are transformed into the decision scores for generating final predictions.



**Figure 8.1:** The general idea of the proposed logistic-regression-based algorithm. (a) A prior probability of each gene is first predefined. (b) The class label of each gene is then assigned according to its prior probability. (c) A biological network shows how vertices connect with each other. (d) A feature matrix is constructed based on both labels and connections of individual vertices. (e) A binary logistic regression is conducted by using class labels as categorical dependent variables and individual features as predictor variables. (f) A posterior probability is obtained from the binary logistic regression for each unknown genes. (g) The posterior probability is transformed into a decision score for each unknown genes. (a) - (f) are the inference stage, while (g) is the decision stage.

## 8.2.2 Logistic regression

For a two-class classification problem, each gene is labelled with either 1 or 0. A vector of all binary values of  $x$  is called a *configuration* of  $x$ . In the previous MRF method [13, 14], the configuration of  $x$  is formulated as a Markov random field which follows a Gibbs distribution. However, the Markovianity characteristic of the MRF model means it only considers direct neighbors as feature vectors, which limits the capability of the method to use other topological attributes in a biological network. It is also very time-consuming to maintain Markov chains for every vector in  $x$ .

To generalize the formulation of feature vectors by using other topological attributes and reduce the computing time, we propose a novel logistic-regression-based algorithm in this study as follows. Let  $C_1$  be a set of genes with label 1 and  $C_0$  be a set of genes with label 0. Suppose the following four kinds of probabilities are given: the class-conditional densities  $p(x|C_1)$  and  $p(x|C_0)$ , which indicates the probability of the configuration  $x$  conditional on  $C_1$  and  $C_0$ , respectively, and the class prior densities  $p(C_1)$  and  $p(C_0)$ , which indicates the prior probability of genes in  $C_1$  and  $C_0$  being labelled with 1 and 0, respectively. The posterior probabilities of those genes in  $C_1$  that are labelled with 1 can be described as a logistic sigmoid function by using the

Bayes' rule as [17, 18]

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)} = \frac{e^t}{e^t + 1} \quad (8.1)$$

and the posterior probabilities of those genes in  $C_0$  that are labelled with 0 can be similarly written as

$$p(C_0|x) = \frac{p(x|C_0)p(C_0)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)} = \frac{1}{e^t + 1} \quad (8.2)$$

where the variable  $t$  is defined as

$$t = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_0)p(C_0)}, \quad (8.3)$$

which is related to the four kinds of probabilities.

However, the variable  $t$  is often unavailable for a real problem. In the previous MRF method [13, 14], numbers of direct neighbors that connects to disease genes and non-disease genes are employed to formulate the variable  $t$  for individual genes. In this study, we generalize the formulation of  $t = f(\cdot)$  as a function of different feature vectors. To be more specific, let  $\hat{x}$  be a prior configuration of all human genes and  $f$  be a function. For each gene  $g_i$ , let  $\phi_i$  be the feature vector of  $g_i$  that relates to the prior configuration  $\hat{x}$ . The posterior probability that the specific gene  $g_i$  has label 1 and 0 are

$$p(x_i = 1 | \phi_i, f) = \frac{\exp(f(\phi_i))}{\exp(f(\phi_i)) + 1}, \quad (8.4)$$

and

$$p(x_i = 0 | \phi_i, f) = \frac{1}{\exp(f(\phi_i)) + 1}. \quad (8.5)$$

respectively. Note that the sum of these two probabilities (8.4) and (8.5) must equal to 1 in the two-class classification problem.

In the simplest case, a linear function of  $\phi_i$  with coefficient (parameters)  $w$  is employed in this study. Hence, the posterior probabilities of (8.4) and (8.5) can be written as

$$p(x_i = 1 | \phi_i, w) = \frac{\exp(w^T \phi_i)}{\exp(w^T \phi_i) + 1}, \quad (8.6)$$

and

$$p(x_i = 0 | \phi_i, w) = \frac{1}{\exp(w^T \phi_i) + 1}, \quad (8.7)$$

respectively. The number of parameters equals to the dimension of individual feature vectors.

The parameter  $w$  can be estimated directly from a training dataset, where known disease genes are naturally available serving as the training data. However, as we previously discussed in [14], the number of known disease genes is far less than the number of all human genes. The exclusion of most unknown genes reduces the number of 0-labelled vertices significantly, thereby making the estimation of parameters inaccurate. Predictions from these inaccurate parameters are unreliable. Alternatively, we propose to estimate parameters

according to the prior configuration of all genes, where labels of those unknown genes are assigned based on the predefined prior probabilities.

Given a prior configuration  $\hat{x}$  of all vertices, the parameter  $w$  can be estimated by maximizing the following conditional likelihood, i.e.

$$\hat{w} = \arg \max_w \left( \prod_{i=1}^N p(x_i | \phi_i, w) \right), \quad (8.8)$$

where  $N$  is the number of all labelled genes,  $x_i$  is the label of  $g_i$ ,  $\phi_i$  is its feature vector that is calculated from  $\hat{x}$ , and  $p(x_i | \phi_i, w)$  is the posterior probability of  $x_i$  defined in (8.6) or (8.7), depending on which posterior probability is larger. If  $p(x_i = 1 | \phi_i, w) \geq p(x_i = 0 | \phi_i, w)$ , then  $x_i = 1$ . Otherwise,  $x_i = 0$ . Maximizing the conditional likelihood (8.8) is equivalent to maximize the log likelihood as follows

$$\hat{w} = \arg \max_w \mathcal{L}(w), \quad (8.9)$$

where the log likelihood  $\mathcal{L}(w)$ , after substitutions of (8.6) and (8.7) into (8.8) and some mathematical manipulations, is given as

$$\mathcal{L}(w) = \sum_{i=1}^N [x_i w^T \phi_i - \ln(1 + \exp(w^T \phi_i))]. \quad (8.10)$$

Although there is no analytic solution for the optimization problem (8.9), the log likelihood (8.10) is a convex function [19]. Therefore, the problem (8.9) is a convex optimization problem and thus has a global optimal solution. In this study, we use the standard MATLAB function *fminunc()* to find a numerical solution of (8.9) (by finding the minimum of  $-\mathcal{L}(w)$ ). The gradient of  $-\mathcal{L}(w)$  is provided to the *fminunc()* function, and the initial value of  $w$  is simply started at zero.

### 8.2.3 Feature vector constructions

The construction of feature vectors is the key step of the proposed logistic-regression-based algorithm. On the one hand, the employed feature vector directly affects the prediction performance of the algorithm. A better feature vector is helpful for identifying potential disease genes. On the other hand, the proposed algorithm is flexible in term of the feature selection. Any topological attributes related to vertex labels can be used to construct the feature vector. In this study, three kinds of feature vectors are proposed for both single biological network and multiple biological networks.

Firstly, similar to the MRF method employed, numbers of direct neighbors that are connected to 1- and 0-labelled vertices are employed to construct the feature vector for each gene. To be more specific, let  $\phi_{i1}$  and  $\phi_{i0}$  be the number of direct neighbors of  $g_i$  connected to 1- and 0-labelled vertices, respectively. By adding a dummy feature 1, the feature vector for  $g_i$  is then written as

$$\phi_i = (1, \phi_{i1}, \phi_{i0})^T. \quad (8.11)$$

It is a three dimensional vector, and is called the *basic feature vector* in a single network. All feature vectors of individual genes together form a feature matrix as

$$F_1 = \begin{bmatrix} 1 & \phi_{11} & \phi_{10} \\ 1 & \phi_{21} & \phi_{20} \\ \vdots & \vdots & \vdots \\ 1 & \phi_{N1} & \phi_{N0} \end{bmatrix}_{N \times 3}. \quad (8.12)$$

The corresponding parameter  $w = (w_0, w_1, w_2)^T$  is a three dimensional vector. Predictions generated by using (8.12) are denoted as  $F_1$  hereafter.

Secondly, the basic feature vector is extended within a single biological network by considering not only the number of direct neighbors of  $g_i$ , but also the number of its second-order neighbors as follows

$$\phi_i = (1, \phi_{i1}, \phi_{i0}, \phi'_{i1}, \phi'_{i0})^T \quad (8.13)$$

where  $\phi_{i1}$  and  $\phi_{i0}$  are the numbers of direct neighbors of  $g_i$  connected to 1- and 0- labelled vertices, respectively, while  $\phi'_{i1}$  and  $\phi'_{i0}$  are the numbers of the second-order neighbors of  $g_i$  connected to 1- and 0- labelled vertices, respectively. The contribution of those indirect neighbors has been investigated for predicting disease genes in [15, 16], etc. All those extended feature vectors together form a feature matrix as

$$F_2 = \begin{bmatrix} 1 & \phi_{11} & \phi_{10} & \phi'_{11} & \phi'_{10} \\ 1 & \phi_{21} & \phi_{20} & \phi'_{21} & \phi'_{20} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \phi_{N1} & \phi_{N0} & \phi'_{N1} & \phi'_{N0} \end{bmatrix}_{N \times 5}. \quad (8.14)$$

The corresponding parameter  $w = (w_0, w_1, w_2, w_3, w_4)^T$  is a five dimensional vector. Predictions generated by using (8.14) are denoted as  $F_2$  hereafter.

Thirdly, the basic feature vector  $F_1$  is extended to multiple biological network situation. Suppose there are  $K$  biological networks. Let  $\phi_{i1}^k, \phi_{i0}^k$  be the number of direct neighbors of  $g_i$  connected to 1- and 0- labelled vertices in the  $k^{th}$  network, respectively. By adding a dummy feature 1, the feature vector obtained from those  $K$  networks

$$\phi_i = (1, \phi_{i1}^1, \phi_{i0}^1, \dots, \phi_{i1}^K, \phi_{i0}^K)^T \quad (8.15)$$

is a  $2K + 1$  dimensional vector. All those feature vectors together form a feature matrix as

$$F_3 = \begin{bmatrix} 1 & \phi_{11}^1 & \phi_{10}^1 & \cdots & \phi_{11}^K & \phi_{10}^K \\ 1 & \phi_{21}^1 & \phi_{20}^1 & \cdots & \phi_{21}^K & \phi_{20}^K \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & \phi_{N1}^1 & \phi_{N0}^1 & \cdots & \phi_{N1}^K & \phi_{N0}^K \end{bmatrix}_{N \times (2K+1)}. \quad (8.16)$$

The corresponding parameter  $w = (w_0, w_1, w_2, \dots, w_{2K-1}, w_{2K})^T$  is a  $2K + 1$  dimensional vector as well. Predictions generated from (8.16) by integrating multiple networks is denoted as  $F_3$  hereafter.

## 8.2.4 Prior probability estimation

Estimating a prior probability for each unknown gene is the first step of the logistic-regression-based algorithm. In this study, two methods are developed to estimate the prior probability.

Firstly, when no additional prior information is available, it is reasonable to assign the prior probability as 0 for all unknown genes. The prediction results generated by using the zero prior is denoted as  $P_0$  hereafter.

Secondly, since genes that encode proteins in the same complexes tend to associate with similar diseases. Protein complex information can be employed to estimate a prior probability for each unknown gene as follows.

If a gene  $g_i$  encodes proteins in a complex, let

$$\hat{p}_i = \frac{A}{B} \quad (8.17)$$

be its prior probability, where  $A$  is the number of disease genes of the specific disease in the complex, and  $B$  is the number of all disease genes in the complex. If  $g_i$  appears in multiple protein complexes, we use the maximum value as its prior probability.

If  $g_i$  does not belong to any protein complex, let

$$\hat{p}_i = \frac{C}{D} \quad (8.18)$$

be its prior probability, where  $C$  is the number of all currently known disease genes of the specific disease, and  $D$  is the total number of genes in human genome. The prediction results generated by using protein complex prior is denoted as  $P_c$  hereafter.

## 8.2.5 Decision score

After the posterior probabilities are obtained at the inference stage, an optimal decision can be made according to those posterior probabilities in the decision stage. Although the value of the posterior probability can be employed directly as the decision score, by selecting those predictions with a posterior probability larger than a threshold, the posterior probability does not always work well. This is due to the fact that the number of disease genes largely varies among different classes and a global threshold cutoff of the posterior probability is an arbitrary decision for every disease class.

By considering the number of known disease genes in each class, a new decision score is designed to determine the positives from the predictions. The new decision score is defined as

$$q_i = \frac{|\{j|p_i \geq p_j\}|}{n}, \quad j = 1, 2, \dots, n \quad (8.19)$$

where  $(p_1, p_2, \dots, p_n)$  is the posterior probabilities of individual unknown genes for a specific disease class. Hence,  $q_i$  equals to the top percentage value of the posterior probability  $p_i$  among all unknown genes. The greater the decision score of a gene is, the more likely the gene is associated with a specific disease. All the evaluation criteria of the proposed logistic-regression-based algorithms are calculated according to the decision score hereafter.

### 8.2.6 Validation method and evaluation criteria

The leave-one-out cross validation paradigm is employed to evaluate the performance of the proposed logistic-regression-based algorithms. For each known disease gene with at least one annotated interaction partner in a biological network, we assume it is an unknown disease gene and predict its posterior probability by using the proposed algorithm. The receiver operating characteristic (ROC) curve is employed as one of the evaluation criteria, which shows the relation between the true positive rate (TPR) and the false positive rate (FPR) by varying the threshold for determining positives. The area under the ROC curve (AUC) is employed to show the overall performance of the algorithms.

The negative control genes are necessary to calculate false positives and true negatives. It is generally hard to report a true negative dataset [20]. In this study, those negative control genes are randomly selected from known disease genes that do not belong to the current disease class. If there are  $r$  known disease genes for a disease class, we randomly select  $\lfloor \frac{r}{2} \rfloor$  such genes as a negative control set. Since those genes have been widely studied as disease genes for other diseases, it is less likely for them being disease genes of a different specific disease. Each gene belongs to the negative control set is also validated by using the leave-one-out cross validation paradigm.

We compare the proposed logistic-regression-based algorithms with three previous algorithms: (1) the MRF method proposed by [14]; (2) the RWR algorithm proposed by [11]; and (3) the DIR algorithm proposed by [12]. The MRF method and the RWR algorithm work in either single or multiple network situation, while the DIR algorithm works in only multiple network situation. All three algorithms identify disease genes with high prediction performance and they work better than many previous methods [11, 12, 14].



## 8.2.7 Algorithm

The step-by-step description of the proposed logistic-regression-based algorithm is given as follows.

**Input:** The vector of all human genes  $(g_1, \dots, g_{n+m})$ , where  $(g_1, \dots, g_n)$  are unknown genes, and  $(g_{n+1}, \dots, g_{n+m})$  are known genes; a biological network  $G$ ; and a list of known disease gene for a specific disease.

**Output:** The vector of decision score for each unknown gene for the disease.

- 1: Calculating prior probabilities for all human genes, where the prior probability of unknown genes  $\hat{p}_1, \dots, \hat{p}_n$  are calculated according to (8.17) and (8.18). For each known gene  $g_{n+i}$ ,  $i = 1, \dots, m$ , if  $g_{n+i}$  is known to be associated with this disease class, let  $\hat{p}_{n+i} = 1$ . Otherwise, let  $\hat{p}_{n+i} = 0$ .
- 2: Assigning prior labels  $\hat{x} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n, \hat{x}_{n+1}, \dots, \hat{x}_{n+m})$  for all genes according to the prior probabilities  $(\hat{p}_1, \dots, \hat{p}_{n+m})$ , respectively.
- 3: Calculating the feature vector  $\phi_i$  for each  $g_i$  according to the biological network  $G$  and  $\hat{x}$ .
- 4: Estimating parameters  $\hat{w}$  by maximizing the  $\mathcal{L}(w)$  in (8.9) based on  $\hat{x}$  and  $\phi_i$ ,  $i = 1, \dots, n + m$ . A binary logistic regression is performed here by taking the vector  $\hat{x}$  as the categorical dependent variables and those label-related feature vectors  $\phi_i$ ,  $i = 1, \dots, n + m$  as predictor variables.
- 5: Calculating the posterior probability  $p_1, \dots, p_n$  for each unknown gene according to (8.6) by using  $\hat{w}$  and  $\phi_i$ .
- 6: Calculating the decision score  $q_1, \dots, q_n$  according to (8.19) by using  $p_1, \dots, p_n$ .

## 8.3 Results and discussions

### 8.3.1 Data sources

The data sources we used in this study are the same as those in our previous study for the MRF method [14]. To be more specific, known gene-disease associations are collected from the Morbid Map list of the Online Mendelian Inheritance in Man (OMIM) [21]. Goh et al. [2] manually classify all those diseases into 22 primary disease classes, including a ‘multiple class’ and an ‘unclassified’ class. The dataset contains 1284 diseases and 1777 disease genes. In order to make direct comparison with the method in [14], we choose the same disease classes as [14] that consist of at least 30 genes and exclude the ‘multiple class’ and the ‘unclassified class’ due to the lack of their class evidence, and the ‘cancer’ class and the ‘neurological’ class due to the lack of their class heterogeneity [2]. The final dataset consists of 815 genes that are classified into 12 disease classes.

The protein complex data are collected from the database of CORUM [22] and PCDq [23]. There are 1677 and 1103 protein complexes in datasets with at least two proteins, respectively. There are in total 3881 proteins in those protein complexes.

The PPI dataset is derived from the database of HPRD (Release 9) [24]. Duplicated edges between the same pair of vertices and self-loop edges are deleted. The final PPI network consists of 9465 vertices and 37039 edges. Another two PPI datasets are derived from the database of BioGrid (Release 3.2.108) [25] and the database of IntAct (downloaded on Jan 26, 2014) [26], respective, which are employed to select edges of biological networks.

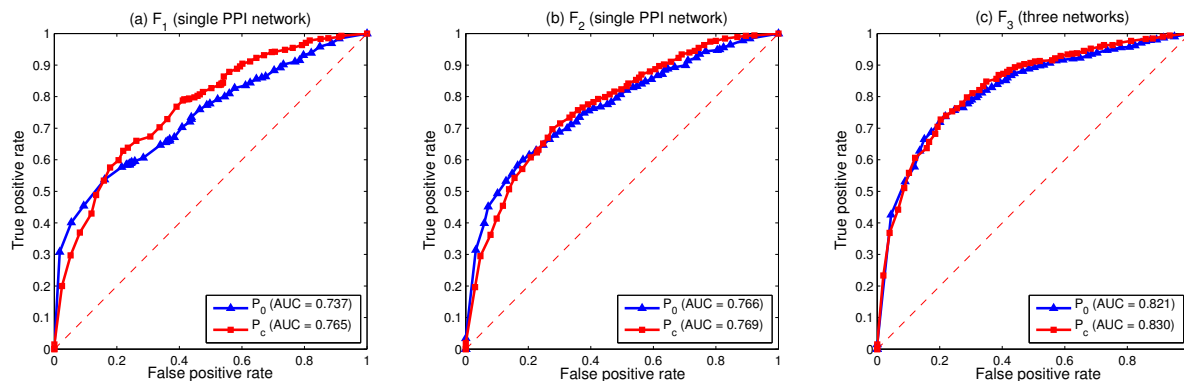
The pathway datasets are obtained from KEGG [27], Reactome [28], PharmGKB [29], and PIN [30]. There are 280, 1469, 99 and 2679 pathways in the datasets, respectively. There are in total 8614 proteins in those pathways. A pathway co-existence network is constructed by taking individual proteins/genes as vertices. Edges are constructed between two vertices if they co-existence in any pathway.

The human gene expression profiles are obtained from BioGPS (GSE1133) [31, 32], which contains 79 human tissues in duplicates, measured using the Affymetrix U133A array. Pairwise Pearson correlation coefficients (PCC) are calculated and a pair of genes are linked by an edge if the PCC value is large than 0.5, similar to the method used in [2, 12] to construct the gene co-expression network.

Overall, three kinds of biological networks are constructed and all protein (or gene) IDs are mapped onto the form of gene symbols. In order to test the performance of multiple data integration of our method, we selected those vertices that appear at least four times in all five biological networks (three PPI networks, a pathway co-existence network and a gene co-expression network). The final datasets consist of 7311 human genes, 815 out of which are known to be associated with 12 disease classes. The details of those datasets used in this study can be found in the “Availability of supporting data” section.

### 8.3.2 Comparisons between different priors

Figure 8.2 compares the logistic-regression-based algorithm by using either the zero prior  $P_0$  or the protein complex prior  $P_c$ . We can see from Figure 8.2 that protein complex prior works better than the zero prior with all three kinds of feature vectors in terms of the AUC score. The highest improvement from  $P_0$  to  $P_c$  is achieved when the basic feature vector  $F_1$  is employed, where the AUC score increases from 0.737 to 0.765, while there is only a slight improvement when  $F_3$  is employed in the multiple network situation, where the AUC score increases from 0.821 to 0.830. This may due to the fact the basic feature vector  $F_1$  using zero prior  $P_0$  achieves the lowest prediction AUC score for identifying disease genes. It has the highest potential to be improved. While the basic feature vector  $F_3$  using zero prior  $P_0$  in the multiple network situation



**Figure 8.2:** Comparisons between different priors of the logistic-regression-based algorithm by using three kinds of feature vectors. (a) The ROC curve of the proposed algorithm by using the basic features on the single HPRD PPI network. (b) The ROC curve of the proposed algorithm by using the extended features on the single HPRD PPI network. (c) The ROC curve of the proposed algorithm by using the basic features by integrating three biological networks: the HPRD PPI network, the pathway co-existence network and the gene co-expression network. AUC values are listed in parentheses.

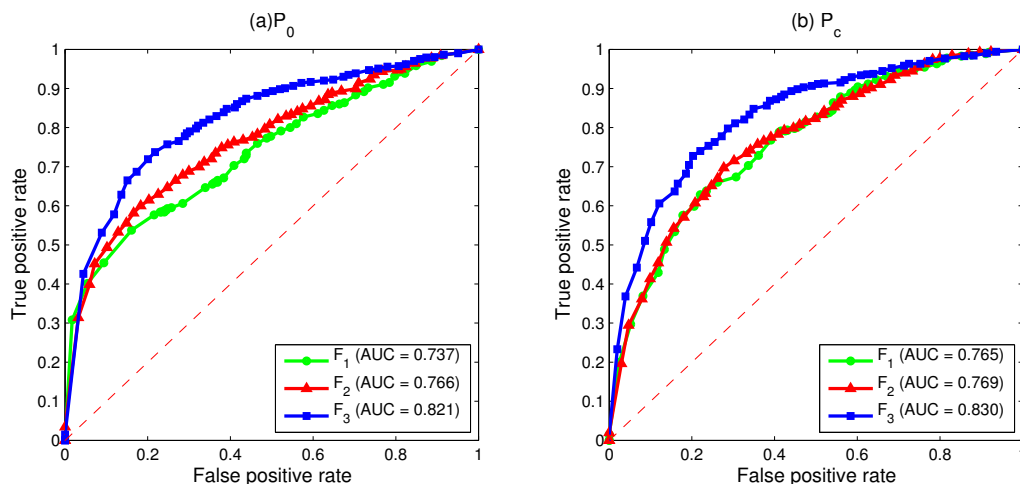
already achieves a very high AUC score, there is only a little room for it to be further improved by using prior information.

Although the improvement from the other two kinds of feature vectors are not so significant, the increased AUC score from the zero prior  $P_0$  to the protein complex prior  $P_c$  indicates that additional knowledge is helpful for improving the prediction performance. This makes the proposed logistic-regression-based algorithm more promising for identifying disease genes. The proposed algorithm is very flexible in terms of the usage of different prior information.

If there is no additional prior information available for the application of the proposed algorithm, zero prior  $P_0$  still works well in most situations. If there is general prior information available in practice (such as the protein complex information), the proposed algorithm works better than that using  $P_0$ . If there is some other specific prior information available, the proposed algorithm is expected to generate more reasonable prediction results.

### 8.3.3 Comparisons between different feature vectors

Figure 8.3 compares the logistic-regression-based algorithm by using different feature vectors. The basic feature vector  $F_1$  and the extended feature vector  $F_2$  are tested on the single HPRD PPI network, and the feature vector  $F_3$  is tested by integrating the following three biological networks: (1) the HPRD PPI network, (2) the pathway co-existence network and (3) the gene co-expression network. All three kinds of



**Figure 8.3:** Comparisons between different feature vectors of the logistic-regression-based algorithm. (a) The ROC curve of different feature vectors by using the zero prior. (b) The ROC curve of different feature vectors by using the protein complex prior. AUC values are listed in parentheses.

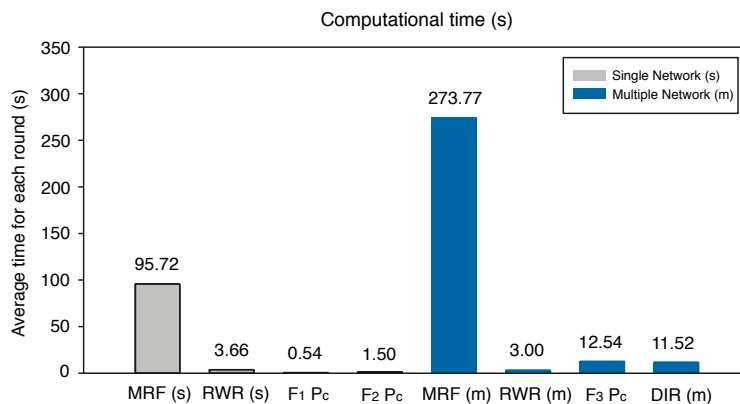
feature vector are tested by using both the zero prior  $P_0$  and the protein complex prior  $P_c$  in the numerical experiments. They are the same experimental results as Figure 8.2 shows, but from a different point of view.

We can see from Figure 8.3 that the feature vector  $F_3$  which integrates three biological networks achieves the highest AUC score in both the zero prior  $P_0$  situation and the protein complex prior  $P_c$  situation, while the basic feature vector  $F_1$  which uses only a single network obtains the lowest AUC score. In the zero prior situation, the basic feature vector  $F_1$  reaches the AUC score of only 0.737, the extended feature vector  $F_2$  on the same single PPI network reaches 0.766, while the feature vector  $F_3$  by integrating three networks achieves the AUC score of 0.821. In the protein complex prior situation, the AUC score of the basic feature vector  $F_1$  is 0.765. It increases to 0.769 by using the extended feature vector  $F_2$  on the same single PPI network, and it continues rising to 0.830 by integrating three networks of the feature vector  $F_3$ .

### 8.3.4 Comparing with previous algorithms

To test the efficiency of our proposed method, three previous algorithms are employed in both single network and multiple network situations. The RWR algorithm and the MRF algorithm work in both situations, while the DIR algorithm works only in the multiple network situation.

The comparison is first conducted in terms of the computational time. All those tests are conducted on a Windows 7 professional computer (Inter(R) Core(TM) i7 CPU, 3.07 GHz, 8.0 GB RAM, 64-bit OS). The MATLAB version is 7.10.0.499 (R2010a), 64-bit (win 64). Each algorithm is evaluated by using the leave-one-

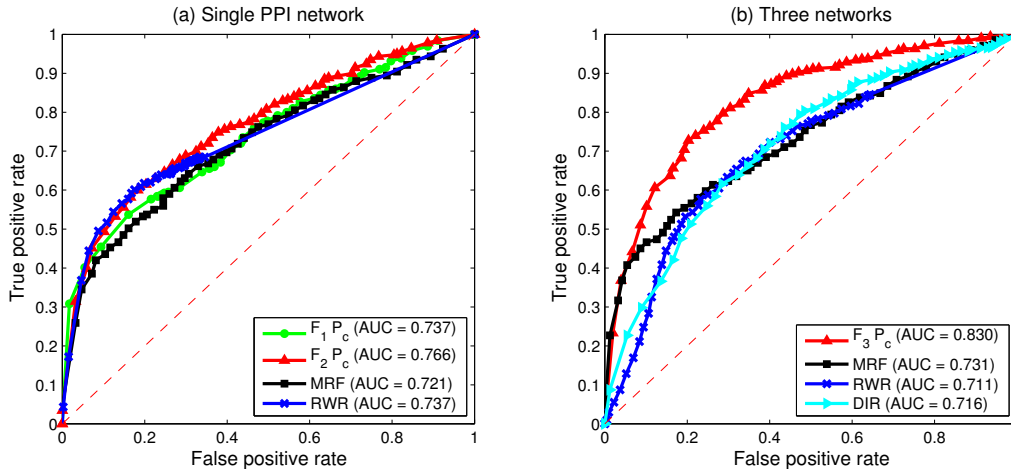


**Figure 8.4:** Comparison of the computational time among different algorithms. The grey bars illustrate the average time of different algorithms that work on the single HPRD PPI network. From left to right, they are the average computational time of the MRF method, the RWR algorithm, the proposed algorithm by using the  $F_1$  feature vector and the proposed algorithm by using the  $F_2$  feature vector, respectively. The blue bars illustrate the average time of different algorithms by integrating three biological networks. From left to right, they are the average computational time of the MRF method, the RWR algorithm, the proposed algorithm by using the  $F_3$  feature vector and the DIR algorithm, respectively. The number above each bar gives the average time (by second) for each leave-one-out experiment.

out cross validation paradigm, where one known gene is left out once, and the probability of each unknown gene (include the left out one) is calculated by each algorithm. The program takes around 2GB RAM if a single dataset is employed as input, and it takes around 4GB RAM if three datasets are employed as input. Figure 8.4 illustrates the average computational time for each leave-one-out experiment among the different algorithms.

We can see from Figure 8.4 that the MRF method is the slowest algorithm. A leave-one-out experiment spends around 95.72 seconds in the single network situation, and it increases to about 273.77 seconds when three biomolecular networks are integrated. The proposed logistic-regression-based algorithm runs very fast. It only spends approximate 0.54 seconds or 1.50 seconds in the single network situation when the basic feature vector  $F_1$  or the extend feature vector  $F_2$  is employed, respectively. Even when three biological networks are integrated, the computational time of the proposed algorithm increases to around 12.54 seconds, which is almost the same as the DIR algorithm (11.52 seconds). The computational time of the RWR algorithm does not vary too much. This is due to the fact that the RWR algorithm uses the mixed network as input. No matter how many networks are integrated, it combines them together as a single mixed network. Hence, the number of integrated networks does not affect the computational time significantly.

A comparison is then conducted in terms of the AUC score. When only the single HPRD PPI network is employed, as illustrated in Figure 8.5 (a), the proposed logistic-regression-based algorithm works better than



**Figure 8.5:** ROC curves of cross-validation results of the proposed logistic-regression-based algorithm and three previous methods. (a) The ROC curves of different algorithms conducted on the single HPRD PPI network. (a) The ROC curves of different algorithms conducted on the integrated three biological networks: the HPRD PPI network, the pathway co-existence network and the gene co-expression network. AUC values are listed in parentheses.

both the MRF method and the RWR algorithm. The AUC score is 0.766 when the extended feature vector  $F_2$  is used, which achieves 4.5% and 2.9% improvements compared with the the MRF method and the RWR algorithm, respectively. When three biological networks are employed, as illustrated in Figure 8.5 (b), the proposed logistic-regression-based algorithm achieves the highest AUC score among all these algorithms. The AUC score is 0.830 when protein complex prior  $P_c$  is used, which is 9.9%, 11.9% and 11.4% improvements compared with the MRF method, the RWR algorithm and the DIR algorithm under the same situation, respectively.

## 8.4 Conclusions

In this paper, we have proposed a logistic-regression-based algorithm to identify disease genes from both a single network and multiple networks. The posterior probability of each unknown gene being a disease gene is obtained by using a binary logistic regression model. The proposed algorithm is very flexible in terms of both the usage of different priors and the construction of different feature vectors. Much prior information about disease genes can be employed to estimate the prior probability, and many label-related topological attributes can be used to construct the feature vector.

Compared with previous methods, the proposed logistic-regression-based algorithm not only runs fast, but also generates predictions with very high AUC score. It takes only around 0.54 seconds or 1.50 seconds in

the single PPI network situation, and its AUC score is better than both the MRF method and RWR method. Although the running time in the multiple network situation is a little longer than the RWR algorithm and the DIR algorithm, it is still acceptable, and the AUC score of the proposed algorithm is much better than those two algorithms. Compared with the MRF method, the computational time has been significantly decreased while the AUC score has been significantly increased. The best AUC score is 0.766 in the single network situation, and it is 0.830 if three networks are integrated. The high prediction performance and the short computation time make the proposed algorithm very promising for identifying human disease genes.

## **Acknowledgement**

The publication costs for this article were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## **Authors' contributions**

FXW and BC initiated this study and designed algorithms and experiments. BC performed the experiments, analyzed the results, and drafted the manuscript. FXW, ML and JXW revised the manuscript. All authors have read and approved the final manuscript.

## **Declarations**

The authors declare that they have no competing interests.

## **Availability of the program package**

The Matlab code of the algorithm can be found in: <https://www.dropbox.com/s/xxxis233xh2n0gg/Package.zip>

## BIBLIOGRAPHY

- [1] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006, **43**(8): 691-698.
- [2] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA* 2007, **104**(21): 8685-8690.
- [3] Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007, **71**(1): 1-11.
- [4] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol* 2008, **4**: 189.
- [5] Fraser HB, Plotkin JB. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol* 2007, **8**(11): R252.
- [6] Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. *PLoS One* 2009, **4**(2): e4346.
- [7] Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007, **23**(2): 215-221.
- [8] Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007, **25**(3): 309-316.
- [9] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010, **6**(1): e1000641.
- [10] Zhang W, Sun F, Jiang R. Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach. *BMC Bioinformatics* 2011, **12**(Suppl 1): S11.
- [11] Köhler, S. Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008, **82**(4): 949-958.
- [12] Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. In silico gene prioritization by integrating multiple data sources. *PLoS One* 2011, **6**(6): e21137.



- [13] Chen B, Wang J, Wu FX. Prioritizing human disease genes by multiple data integration. *Bioinformatics and Biomedicine (BIBM)*, 2013 IEEE International Conference on 2013: 621.
- [14] Chen B, Wang J, Li M, Wu FX. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics* 2014, **7**(Suppl 2): S2.
- [15] Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel based logistic regression models for protein function prediction. *OMICS* 2006, **10**(1): 40-55.
- [16] Li SZ. Markov random field modeling in image analysis. London: Springer; 2009.
- [17] Bishop CM. Pattern recognition and machine learning. Singapore: Springer; 2006.
- [18] Shi J, Chen B, Wu FX. Unifying protein inference and peptide identification with feedback to update consistency between peptides. *Proteomics* 2013, **13**(2), 239-247.
- [19] Boyd SP, Vandenberghe L. Convex optimization. New York: Cambridge University Press; 2004.
- [20] Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM. Response to Chatr-aryamontri et al.: Protein interactions: to believe or not to believe? *Trends Biochem Sci* 2008, **33**(6): 242-243.
- [21] McKusick VA. Mendelian inheritance in man and its online version, OMIM. *Am J Hum Genet* 2007, **80**(4), 588-604.
- [22] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Res* 2010, **38**(Database issue): D497-D501.
- [23] Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, Kanaya S, Imanishi T. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. *BMC Syst Biol* 2012, **6**(Suppl 2): S7.
- [24] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database - 2009 update. *Nucleic Acids Res* 2009, **37**(Database issue): D767-D772.
- [25] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006, **34**(Database issue): D535-539.
- [26] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L,

- Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct - open source resource for molecular interaction data. *Nucleic Acids Res* 2007, **35**(Database issue): D561-D565.
- [27] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, **28**(1): 27-30.
- [28] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007, **8**(3): R39.
- [29] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012, **92**(4): 414-417.
- [30] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. *Nucleic Acids Res* 2009, **37**(Database issue): D674-D679.
- [31] Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009, **10**, R130.
- [32] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004, **101**(16): 6062-6067.

## CHAPTER 9

### CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK

#### 9.1 Conclusions

In this thesis, two topics related to the analyses of biomolecular networks have been studied: (1) the identification of protein complexes and (2) the identification of disease genes.

The identification of protein complexes provides insight information about how the ensemble of proteins is organized into functional units. Research studies related to this topic are presented from Chapter 2 to Chapter 5 in this thesis. Specifically, a comprehensive review about the identification of protein complexes and/or functional modules from static PPI networks to dynamic PPI networks is first presented. Then, an improved entropy-based algorithm is proposed to detect densely connected sub-graphs from PPI networks. It is a seed-growth-style algorithm, which starts from a set of seed vertices to search for local optimum clusters. The proposed algorithm uses cliques as the initial seeds, rather than using individual single vertices as in most other algorithms. The improved entropy-based algorithm can generate many meaningful predictions. However, it still cannot detect some known protein complexes in the “gold standard” dataset. Then, after investigating the topological characteristics of known protein complexes, we find that not all protein complexes exhibit dense structures in PPI networks. Many of them have a star-like structure, where only one hub protein connects with all other proteins within individual complexes. Therefore, a multiple-topological-structure-based algorithm to identify protein complexes in PPI networks is further proposed. Four single-topological-structure-based algorithms are employed to identify raw predictions with cliques, densely connected sub-graphs, core-attachment structures and star-like structures from PPI networks. Raw predictions are then be merged and/or trimmed based on their topological information or GO annotations. Numerical experiments have shown that our proposed algorithms not only generate more meaningful protein complexes, but also identify them with higher *f-score* compared with many existing algorithms.

The identification of disease genes helps us to understand the intricate association relationships between disease genes and genetic diseases. Research studies related to this topic are presented from Chapter 6

to Chapter 8 in this thesis. Specifically, the disease gene identification topic is formulated as a two-class classification problem, where one class represents disease genes while the other class represents non-disease genes. A MRF-based algorithm is proposed to calculate the post probabilities of individual genes to be disease genes by using a Bayesian analysis method. The proposed algorithm is not only flexible in easily incorporating different kinds of data, but also reliable in predicting candidate disease genes. Then, a kernel-based MRF algorithm is developed to combine the advantages of graph kernels and the MRF models. Kernels provide a general framework to represent data in the form of pair-wise similarities by considering the distant relationships among biomolecules globally. A kernel-based MRF algorithm yields better performance than the original MRF method which only considers direct neighbor information for feature constructions. Finally, to generalize the idea of feature constructions in the MRF-based algorithms, a fast and high performance logistic regression algorithm is proposed to identify disease genes. It directly formulates the issue of disease gene identification as a binary logistic regression problem, which not only is more flexible for feature constructions, but also runs very fast. Numerical experiments have shown that the proposed algorithms outperforms many previous algorithms in terms of the AUC score and the running time.

## 9.2 Contributions

Briefly, the thesis has provided a comprehensive review about the identification of protein complexes and has proposed several novel algorithms to identify protein complexes and disease genes from biomolecular networks. The major contributions of this study can be summarized as follows:

1. The review paper summarizes the current state of knowledge about algorithms for identifying protein complexes from PPI networks.
2. An improved graph-entropy-based algorithm is proposed to identify dense sub-graphs from PPI networks, which can generate many meaningful protein complexes.
3. Topological characteristics of known protein complexes are thoroughly studied in PPI networks. The results show that not all protein complexes have dense structures in PPI networks. Many of them have a star-like topological structures.
4. A multiple-topological-structure-based algorithm is proposed to identify protein complexes from PPI networks that detects protein complexes with cliques, dense sub-graphs, core-attachments structures and/or star-like structures.
5. A MRF-based algorithm is proposed to identify disease genes that considers not only edges of candidate genes to disease genes, but also edges of candidate genes to non-disease genes.

6. A kernel-based MRF algorithm is proposed to identify disease genes that combines the advantages of graph kernels and the MRF-based method. It extends the feature construction of the previous MRF method by considering global topological characteristics of biomolecular networks.
7. A logistic-regression-based algorithm is proposed to identify disease genes that generalizes the idea of feature constructions in MRF-based methods. It directly formulates the issue of disease gene identification into a binary logistic regression problem, which not only makes the algorithm more flexible and simple, but also has competing performance in terms of the AUC score.

### 9.3 Future work

Along with the research of this thesis, some directions of further work are listed as follows:

1. Predicting protein complexes and functional modules from dynamic PPI networks

Currently, most protein complex identification algorithms are applied on only static PPI networks. However, proteins and protein interactions are dynamic in real biological systems. Therefore, identifying protein complexes from dynamic PPI networks is more meaningful, but challenging. Time-course gene expression profiles and tissue-specific gene expression profiles provide information to construct dynamic PPI networks. It is also possible to distinguish protein complexes from functional modules in this situation, since the functional modules are groups of proteins that interact with each other at different times and places.

2. Identifying disease genes, drug targets and essential proteins

Although many algorithms have been proposed for identifying disease genes, the prediction accuracy is still limited and needs to be further improved. In this thesis, I have used MRFs, graph kernels and logistic regressions to develop identification algorithms. In the future, other advantageous computational approaches should be investigated, such as nonnegative matrix factorization (NMF), optimization, statistics, graph theory, and so on. The identification of drug targets and essential proteins are also suggested for future work, since these topics are strongly related to the topic of disease gene identifications.

3. Analyses of biomolecular networks through multiple data integration

Multiple data integration is necessary and indispensable when analyzing biomolecular networks. I have studied various kinds of biological data. In future work, other kinds of large-scale data can also be efficiently incorporated, such as protein domains, SNPs, protein microarrays, phenotype similarities, DNA methylation data, and so on.

# APPENDIX A

## LIST OF PUBLICATIONS

### Thesis-related Journal Papers:

- [J1] **Chen B**, Li M, Wang JX and Wu FX. A fast and high performance algorithm for identifying human disease genes. (unpublished).
- [J2] **Chen B**, Li M, Wang JX, Wu FX. Disease gene identification by using graph kernels and Markov random fields. *SCIENCE CHINA Life Sciences* 2014, **57**(11): 1054-1063.
- [J3] **Chen B**, Wang JX, Li M and Wu FX. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics* 2014, **7**(Suppl 2): S2.
- [J4] **Chen B**, Fan W, Liu J and Wu FX. Identifying protein complexes and functional modules - from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics* 2014, **15**(2): 177-194.
- [J5] **Chen B** and Wu FX. Identifying protein complexes based on multiple topological structures in PPI networks. *IEEE Transactions on Nanobioscience* 2013, **12**(3): 165-172.
- [J6] **Chen B**, Shi J, Zhang S and Wu FX. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics* 2013, **13**(2): 269-277.

### Thesis-related Conference Papers:

- [C1] **Chen B**, Li M, Wang JX and Wu FX. A logistic regression based algorithm for identifying human disease genes. *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 197-200.
- [C2] **Chen B**, Wang JX and Wu FX. Prioritizing human disease genes by multiple data integration. *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*: 621.
- [C3] **Chen B**, Shi J and Wu FX. Not all protein complexes exhibit dense structures in *S. cerevisiae* PPI network. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*: 470-473.
- [C4] **Chen B**, Yan Y, Shi J, Zhang S and Wu FX. An improved graph entropy-based method for identifying protein complexes. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*: 123-126.

### Thesis-unrelated Journal Papers:

- [J7] J Sun, **Chen B** and Wu FX. An improved peptide-spectral matching algorithm through distributed search over multiple cores and multiple CPUs. *Proteome Science* 2014 **12**: 18.
- [J8] Fan W, **Chen B**, Selvaraj G and Wu FX. Discovering biological patterns from short time-series gene expression profiles with integrating PPI data. *Neurocomputing* 2014, **145**: 3-13.
- [J9] Shi J, **Chen B** and Wu FX. Unifying protein inference and peptide identification with feedback to

updated consistency between peptides. *Proteomics* 2013, **13**(2): 237-247.

[J10] Yuan Z, Shi J, Lin WJ, **Chen B** and Wu FX. Features-based deisotoping method for tandem mass spectra. *Advances in Bioinformatics* 2011, Article ID 210805, 12 pages.

**Thesis-unrelated Conference Papers:**

[C5] Shi J, **Chen B** and Wu FX. Improving accuracy of peptide identification with consistency between peptides. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*: 191-196.

[C6] **Chen B**, Liu LZ and Wu FX. Inferring gene regulatory networks from multiple time course gene expression datasets. *Systems Biology (ISB), 2011 IEEE International Conference on*: 12-17.

## APPENDIX B

### COPYRIGHT PERMISSIONS

The copyright of the following papers:

Chen B, Fan W, Liu J and Wu FX. Identifying protein complexes and functional modules - from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics* 2014, **15**(2): 177-194.

Chen B, Shi J, Zhang S and Wu FX. Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy. *Proteomics* 2013, **13**(2): 269-277.

Chen B, Shi J and Wu FX. Not all protein complexes exhibit dense structures in *S. cerevisiae* PPI network. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on:* 470-473.

Chen B, and Wu FX. Identifying protein complexes based on multiple topological structures in PPI networks. *IEEE Transactions on Nanobioscience* 2013, **12**(3): 165-172.

Chen B, Wang JX, Li M and Wu FX. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics* 2014, **7**(Suppl 2): S2.

Chen B, Li M, Wang JX, Wu FX. Disease gene identification by using graph kernels and Markov random fields. *SCIENCE CHINA Life Sciences* 2014, **57**(11): 1054-1063.

are included in the following pages.



**OXFORD UNIVERSITY PRESS LICENSE  
TERMS AND CONDITIONS**

Oct 07, 2014

---

This is a License Agreement between Bolin Chen ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	3483820399487
License date	Oct 07, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Briefings in Bioinformatics
Licensed content title	Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks
Licensed content author	Bolin Chen, Weiwei Fan, Juan Liu, Fang-Xiang Wu
Licensed content date	March 1, 2014
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Identifying protein complexes and disease genes from biomolecular networks
Publisher of your work	n/a
Expected publication date	Jul 2015
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

[Terms and Conditions](#)

**STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL  
FROM AN OXFORD UNIVERSITY PRESS JOURNAL**

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following

territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from [www.oxfordjournals.org](http://www.oxfordjournals.org) Should there be a problem clearing these rights, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

**Questions? [customer care@copyright.com](mailto:customer care@copyright.com) or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

---

---

## JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS



Oct 07, 2014

This is a License Agreement between Bolin Chen ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	3483811316565
License date	Oct 07, 2014
Licensed content publisher	John Wiley and Sons
Licensed content publication	Proteomics
Licensed content title	Identifying protein complexes in protein-protein interaction networks by using clique seeds and graph entropy
Licensed copyright line	© 2012 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
Licensed content author	Bolin Chen, Jinhong Shi, Shenggui Zhang, Fang-Xiang Wu
Licensed content date	Nov 29, 2012
Start page	269
End page	277
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Identifying protein complexes and disease genes from biomolecular networks
Expected completion date	Jul 2015
Expected size (number of pages)	180
Total	0.00 USD
Terms and Conditions	

## TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking  accept  in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at <http://myaccount.copyright.com>).

## Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.
- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.
- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of

and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU
- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.
- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.
- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.
- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and

the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.
- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.
- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.
- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

## WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses:: Creative Commons Attribution (CC-BY) license [Creative Commons Attribution Non-Commercial \(CC-BY-NC\) license](#) and [Creative Commons Attribution Non-Commercial-NoDerivs \(CC-BY-NC-ND\) License](#). The license type is clearly identified on the article.

Copyright in any research article in a journal published as Open Access under a Creative Commons License is retained by the author(s). Authors grant Wiley a license to publish the article and identify itself as the original publisher. Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified as follows: [Title of Article/Author/Journal Title and Volume/Issue. Copyright (c) [year] [copyright owner as specified in the Journal]. Links to the final article on Wiley's website are encouraged where applicable.

### The Creative Commons Attribution License

The [Creative Commons Attribution License \(CC-BY\)](#) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-commercial re-use of an open access article, as long as the author is properly attributed.

The Creative Commons Attribution License does not affect the moral rights of authors, including without limitation the right not to have their work subjected to derogatory treatment. It also does not affect any other rights held by authors or third parties in the article, including without limitation the rights of privacy and publicity. Use of the article must not assert or imply, whether implicitly or explicitly, any connection with, endorsement or sponsorship of such use by the author, publisher or any other party associated with the article.

For any reuse or distribution, users must include the copyright notice and make clear to others that the article is made available under a Creative Commons Attribution license, linking to the relevant Creative Commons web page.

To the fullest extent permitted by applicable law, the article is made available as is and without representation or warranties of any kind whether express, implied, statutory or otherwise and including, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of defects, accuracy, or the presence or absence of errors.



## **Creative Commons Attribution Non-Commercial License**

The [Creative Commons Attribution Non-Commercial \(CC-BY-NC\) License](#) permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

## **Creative Commons Attribution-Non-Commercial-NoDerivs License**

The [Creative Commons Attribution Non-Commercial-NoDerivs License](#) (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

## **Use by non-commercial users**

For non-commercial and non-promotional purposes, individual users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles, as well as adapt, translate, text- and data-mine the content subject to the following conditions:

- The authors' moral rights are not compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be impugned).
- Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that content.
- If article content is copied, downloaded or otherwise reused for non-commercial research and education purposes, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on **Wiley Online Library**) should be maintained. Copyright notices and disclaimers must not be deleted.
- Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

## **Use by commercial "for-profit" organisations**

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

- Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;

- Copying, downloading or posting by a site or service that incorporates advertising with such content;
- The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)
- Use of article content (other than normal quotations with appropriate citation) by for-profit organisations for promotional purposes
- Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;
- Use for the purposes of monetary reward by means of sale, resale, licence, loan, transfer or other form of commercial exploitation such as marketing products
- Print reprints of Wiley Open Access articles can be purchased from:  
[corporatesales@wiley.com](mailto:corporatesales@wiley.com)

Further details can be found on Wiley Online Library <http://olabout.wiley.com/WileyCDA/Section/id-410895.html>

Other Terms and Conditions:

**v1.9**

**Questions? [customercare@copyright.com](mailto:customercare@copyright.com) or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

---

---



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Not AU protein complexes exhibit dense structures in *S. cerevisiae* PPI network

**Conference Proceedings:** Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on

**Author:** Bolin Chen; Jinhong Shi; Fang-Xiang Wu

**Publisher:** IEEE

**Date:** 4-7 Oct. 2012

Logged in as:

Bolin Chen

[LOGOUT](#)

Copyright © 2012, IEEE

## Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

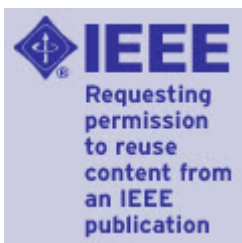
If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

[BACK](#)
[CLOSE WINDOW](#)

Copyright © 2014 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#)  
 Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Identifying Protein Complexes Based on Multiple Topological Structures in PPI Networks

Logged in as:  
Bolin Chen

[LOGOUT](#)

**Author:** Bolin Chen; Fang-Xiang Wu

**Publication:** NanoBioscience, IEEE Transactions on

**Publisher:** IEEE

**Date:** Sept. 2013

Copyright © 2013, IEEE

## Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

[BACK](#)
[CLOSE WINDOW](#)

Copyright © 2014 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#)  
Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)

## Copyright policy

### Research articles

Copyright on any research article in a journal published by BioMed Central is retained by the author(s).

Authors grant BioMed Central a license to publish the article and identify itself as the original publisher.

Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.

[Creative Commons Attribution License 4.0](#) formalizes these and other terms and conditions of publishing research articles.

In accordance with our [Open Data policy](#), the [Creative Commons CC 0 1.0 Public Domain Dedication waiver](#) applies to all published data in BioMed Central open access articles

Where an author is prevented from being the copyright holder (for instance in the case of US government employees or those of Commonwealth governments), minor variations may be required. In such cases the copyright line and license statement in individual articles will be adjusted, for example to state '© 2014 Crown copyright; licensee BioMed Central Ltd'.

### Other articles

In addition to fully open access research articles, seven of BioMed Central's journals publish commissioned content which is available by subscription for the first 6 or 12 months (depending on the journal) immediately following publication.

From January 2014, author(s) of such articles retain copyright but grant BioMed Central an exclusive license to publish and distribute the Article for the initial journal-dependent subscription period after online publication of the Article in the Journal. BioMed Central makes the Article available under the Creative Commons Attribution license after expiry of the initial subscription period, or earlier at BioMed Central's discretion.

### Authors' certification

In submitting a research article ('article') to any of the journals published by BioMed Central Ltd ('BioMed Central') authors are requested to certify that:

They are authorized by their co-authors to enter into these arrangements.

They warrant, on behalf of themselves and their co-authors, that:

the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;

they are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other obligation. If the law requires that the article be published in the public domain, they will notify BioMed Central at the time of submission;

the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy;

they have taken due care to ensure the integrity of the article. To their - and currently accepted scientific - knowledge all statements contained in it purporting to be facts are true and any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user.

they agree to all terms of the [Creative Commons Attribution License 4.0](#) and [Open Data policy](#).



# RightsLink®

[Home](#)[Account Info](#)[Help](#)

**Title:** Disease gene identification by using graph kernels and Markov random fields

Logged in as:  
Bolin Chen

**Author:** BoLin Chen

Account #:  
3000842825

**Publication:** SCIENCE CHINA Life Sciences

[LOGOUT](#)

**Publisher:** Springer

**Date:** Jan 1, 2014

Copyright © 2014, The Author(s)

## Permissions Request

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Springer and BioMed Central offer a reprint service for those who require professionally produced copies of articles published under Creative Commons Attribution (CC BY) licenses. To obtain a quotation, please email [reprints@springeropen.com](mailto:reprints@springeropen.com) with the article details, quantity(ies) and delivery destination. Minimum order 25 copies.

[CLOSE WINDOW](#)

Copyright © 2014 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement](#).  
Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)